Attribution-Guided Shared-Private Decoupling for Noise-Reduced Audio-Visual Representation Learning

Linge Wang[1,2] **Yingying Chen**[1,2,†] **Bingke Zhu**[1,2] **Lu Zhou**[1,2]
Haixin Wang[1,2] **Zhipiao Liu**[4] **Jinqiao Wang**[1,2,3,†]

[1] Foundation Model Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing, China
[3] Objecteye Inc., Beijing, China [4] Huawei Technologies Ltd., Beijing, China
wanglinge2024@ia.ac.cn
{yingying.chen,bingke.zhu,lu.zhou,jqwang}@nlpr.ia.ac.cn
[†]Corresponding authors

Abstract

Audio-visual representation learning commonly aligns paired audio and visual signals with sample-level contrastive objectives. However, real-world audio-visual pairs usually contain heterogeneous patch tokens: only a subset carries reliable cross-modal shared semantics, while many others mainly encode modality-private residuals. Aggregating all tokens into a single global representation can therefore introduce semantic noise into cross-modal alignment. Motivated by the minimal-sufficiency principle, we propose an attribution-guided shared-private decoupling framework to mitigate this issue. Our key idea is to identify shared-cover subviews that preserve dominant cross-modal information while suppressing modality-private residuals, and use them to supervise decoupled shared and private representations. To identify such subviews efficiently, we draw on transformer attribution method and introduce Layer-Truncated Attribution (LTA), which provides a lightweight estimate of token-level relevance to cross-modal shared information. We further adopt a teacher-student framework, where the teacher provides attribution-induced supervision and the student internalizes shared-private decoupling into its representations, without introducing extra inference cost. Experiments on retrieval, classification, and sound-prompted semantic segmentation show consistent gains over strong baselines. Compared with baseline, our method improves the average zero-shot R@1 by 4.0 across AudioSet and VGGSound retrieval, and improves AS20K classification mAP by 3.1, demonstrating the effectiveness of fine grained attribution guidance for reducing semantic noise. Code will be released at `https://github.com/wanglg20/LTA`.

1 Introduction

Multimodal representation learning aims to learn transferable representations by exploiting the complementary information across different sensory modalities. Early studies have shown that jointly modeling multiple modalities can improve feature learning, cross-modal retrieval, and semantic grounding [1, 2, 3, 4, 5, 6]. Among various multimodal settings, audio-visual representation learning is particularly important because real world videos naturally provide massive paired audio and visual signals, where sounds are often correlated with visible objects, events, and temporal dynamics. This has motivated a series of self-supervised audio-visual learning methods that exploit audio-video

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

correspondence from unlabeled videos [7, 8, 9, 10, 11]. More recently, representative frameworks [12, 13, 14] combine masked autoencoding [15] with contrastive audio-visual alignment based on InfoNCE [16], leading to strong transferable representations for downstream tasks such as cross-modal retrieval, event recognition, and sound source localization.

Despite these advances, real-world audio-visual pairs are rarely perfectly aligned. From the perspective of partial information decomposition, paired modalities generally contain both shared information that is redundant across modalities and unique information that is specific to each modality [17, 18, 19]. In audio-visual data, the shared component is often reflected in acoustically grounded visual events, such as a speaking person, a musical instrument being played, or an object action synchronized with sound. In contrast, modality-private components may correspond to off-screen sounds, silent visual regions, scene context, or appearance details that have no direct counterpart in the other modality. Most existing audio-visual contrastive frameworks, however, encode each modality into a single global token and align audio and visual tokens with a sample-level InfoNCE objective. This design has two inherent limitations. First, sample-level contrastive learning mainly supervises the cross-modal shared component, since the correspondence between paired audio and visual signals is induced by their shared semantics. As a result, modality-private cues, although potentially useful for downstream recognition, are not explicitly captured and may instead be absorbed into the same global representation. Second, the global token is obtained by aggregating all patch tokens, although these tokens are semantically heterogeneous, only a subset carries reliable cross-modal correspondence, while many others are dominated by modality-private residuals. As a result, low-relevance patches can be inadvertently involved in cross-modal alignment, introducing semantic noise into the learned representation. Recent studies have explored decoupled or disentangled multimodal representations [19, 20, 21, 22], but they mainly focus on modality-level or representation-level decomposition and do not explicitly address how shared and private information should be separated at the fine-grained patch token level.

To address these issues, we take inspiration from the minimal-sufficiency principle of the information bottleneck [23, 24] and seek a representation that preserves the shared information while suppressing modality-private residuals. Our key idea is to construct a shared-cover subview, a subset of patch tokens that retains sufficient cross-modal shared information but filters out tokens dominated by modality-private information. Based on this subview, we decouple each modality representation into two learnable tokens, a shared token z_s^m for cross-modal alignment and a private token z_p^m for modality-private residual information. The main challenge is how to identify such shared-relevant tokens without manual annotations. To this end, we build on Transformer attribution methods[25, 26, 27, 28] and introduce LTA, which estimates token-level relevance from the gradient of the cross-modal matching score with respect to the last layer attention map. LTA provides an efficient and objective-conditioned ranking signal for selecting shared-cover subviews. Finally, we instantiate the framework with a teacher-student training scheme. The teacher produces attribution-induced high-relevance and low-relevance subview anchors, and the student learns to associate its shared and private tokens with the corresponding anchors. After training, the student internalizes the attribution-guided decoupling behavior and requires no extra computation during inference.

In summary, our main contributions are as follows:

- We develop an attribution-guided teacher-student framework for shared-private decoupling, where two learnable tokens are trained to capture cross-modal shared information and modality-private residuals through high-relevance and low-relevance subview supervision.
- We introduce Layer-Truncated Attribution (LTA), an efficient token-level relevance estimator that uses objective conditioned gradient-attention signals from the last Transformer block to identify patch tokens carrying cross-modal shared information without full relevance propagation.
- Extensive experiments on large scale real world audio-visual data show state-of-the-art performance on retrieval, classification, and sound-prompted semantic segmentation, demonstrating the effectiveness of fine-grained attribution guidance for reducing semantic noise.

2 Related Works

Audio-Visual Representation Learning. Audio-visual representation learning has long explored the natural correspondence between sound and vision, from early statistical modeling of audio-visual association [29, 30] to deep self-supervised learning from unlabeled videos [7, 31, 32]. A

major line of work learns cross-modal correspondence by discriminating synchronized or paired audio-visual samples [33, 34, 35], which also enables weakly supervised localization of sounding objects and events [36, 37]. More recently, masked modeling has become a dominant paradigm for audio-visual pretraining, reconstructing masked audio and visual tokens or exploiting cross-modal prediction [8, 10, 11]. Representative methods further combine masked autoencoding with contrastive alignment, achieving strong transferable representations for retrieval, classification, and localization [12, 13, 14]. Despite these advances, most existing methods rely on sample-level objectives and aggregate heterogeneous patch tokens into a single global representation, leaving fine-grained token-level information decoupling underexplored. Our work takes a step toward this direction by studying how audio-visual representations can be separated at the token level to reduce the influence of semantically noisy regions.

Transformer Attribution. Model attribution aims to explain neural predictions by identifying input regions or features that contribute most to a target output [38, 39]. Early studies developed feature visualization and gradient- or relevance-based explanations for deep networks [40, 41, 42, 43, 44]. For Transformers, raw attention weights are often insufficient as faithful explanations [45, 46], motivating attention propagation methods such as attention rollout and attention flow [47]. Subsequent works further combine attention with gradients or relevance propagation to obtain class- or objective-conditioned explanations for vision and multimodal Transformers [27, 28, 48]. Recent methods improve this line by enhancing computational efficiency or providing more principled relevance propagation for Transformer architectures [49, 26, 25].

3 Theoretical Motivation

3.1 Problem Setup

Considering a paired audio-visual sample $X = (A, V)$, existing audio-visual representation learning frameworks typically process the two modalities with modality-specific encoders and then employ a unified block to aggregate token-level representations before applying a cross-modal alignment objective [13, 12]. For each modality, the input can be viewed as a sequence of patch tokens. We write the audio and visual inputs as

$$A = \{A_i\}_{i=1}^{N_A}, \qquad V = \{V_j\}_{j=1}^{N_V}, \tag{1}$$

where A_i denotes the i-th audio patch token and V_j denotes the j-th visual patch token.

Let Y denote the downstream task variable. For an audio-visual pair (A, V), the task relevant information may arise from both cross-modal shared semantics and modality-specific cues.

Conceptually, the information contained in each modality can be attributed to two latent sources: shared information and modality-private residual information. We write this shared-private abstraction as

$$A \leftarrow (S, P_A), \qquad V \leftarrow (S, P_V), \qquad Y \leftarrow (S, P_A, P_V), \tag{2}$$

where S denotes the information shared across modalities, and P_A and P_V denote audio-private and visual-private residual information, respectively.

Representing each modality with a single global token tends to mix cross-modal shared information with modality-private residuals. Following the shared-private abstraction in Eq. (2), we instead represent each modality with two learnable tokens: a shared token z_s^m for S, and a private token z_p^m for P_m. The key challenge is that shared and modality-private components are entangled at the patch-token level. We formalize this challenge next from a minimal-sufficiency perspective.

3.2 Minimal-Sufficient Subviews for Shared Information Representation

We first characterize the desired shared representation from the information bottleneck[24] perspective. For modality $m \in \{A, V\}$, the shared token z_s^m should preserve information about the cross-modal shared information S while suppressing modality-private residual information P_m. We therefore use the following IB style criterion to describe an ideal shared representation:

$$\max_{z_s^m} \; I(z_s^m; S) - \beta I(z_s^m; P_m), \tag{3}$$

where $\beta > 0$ balances shared sufficiency and private minimality. To instantiate this criterion, we consider the audio modality as an example. Let $A = \{A_i\}_{i=1}^{N_A}$ denote the audio patch tokens. Since shared and modality-private information is entangled at the token level, the full token sequence may contain redundant private residuals under the minimal-sufficiency criterion. We therefore seek a shared-cover subview that preserves the dominant information about S while filtering out tokens with low shared relevance. A natural construction is to rank tokens according to their contribution to the shared information S. We introduce an ideal shared-relevance score:

$$r_i^S = \mathcal{R}(A_i; S), \tag{4}$$

where r_i^S measures the contribution of token A_i to the shared information S.

Sorting tokens by r_i^S in descending order gives a permutation π such that

$$r_{\pi_1}^S \geq r_{\pi_2}^S \geq \cdots \geq r_{\pi_{N_A}}^S. \tag{5}$$

Given a retention ratio $\rho \in (0, 1]$, we keep the top $K = \lfloor \rho N_A \rfloor$ tokens as the selected subview:

$$A_{\mathcal{T}} = \{A_{\pi_i}\}_{i=1}^{K}, \qquad A_{\mathcal{R}} = A \setminus A_{\mathcal{T}}, \tag{6}$$

where $A_{\mathcal{R}}$ denotes the discarded complement. For this score-induced subview to improve the minimal-sufficiency objective, we characterize its expected behavior through two sufficient conditions.

Condition 1 (Shared Cover)**.** *Let $A_{\mathcal{T}}$ be the selected shared-cover subview and $A_{\mathcal{R}} = A \setminus A_{\mathcal{T}}$ be its discarded complement. There exists $\epsilon_S \geq 0$ such that*

$$I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}) \leq \epsilon_S. \tag{7}$$

This condition captures the desired effect of shared-relevance selection: the retained subview should preserve most information about the shared information S, leaving only limited additional shared information in the discarded complement.

Condition 2 (Private Tail)**.** *For the decomposition $A = A_{\mathcal{T}} \cup A_{\mathcal{R}}$, there exists $\epsilon_P > 0$ such that*

$$I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}) \geq \epsilon_P. \tag{8}$$

where $\epsilon_P > 0$. This condition reflects that low-shared-relevance tokens are not necessarily information-free. Although they contribute little additional information about the shared information S, they may still encode audio-private residual information.

For any token subset $B \subseteq A$, we define its input-level IB criterion as

$$F(B) = I(B; S) - \beta I(B; P_A), \tag{9}$$

Proposition 1 (Shared-Cover Subview Improves the IB Objective)**.** *Under the Shared Cover and Private Tail conditions, the selected subview $A_{\mathcal{T}}$ satisfies*

$$F(A_{\mathcal{T}}) - F(A) \geq \beta\epsilon_P - \epsilon_S. \tag{10}$$

In particular, if $\beta\epsilon_P > \epsilon_S$, then $F(A_{\mathcal{T}}) > F(A)$.

This proposition shows that selecting a shared-cover subview can be more suitable than using the full input under the minimal-sufficiency objective. The proof is provided in Appendix A.

4 Proposed Method

4.1 Efficient Token-Relevance Estimation via Layer-Truncated Attribution

The theoretical motivation above suggests that constructing a shared-cover subview relies on a reliable token-level scoring function: tokens with high scores should preserve the dominant shared factor, while tokens with low scores are more likely to contain modality-private residual information. Therefore, the key question is how to estimate the shared relevance of each token without introducing additional supervision or task-specific annotations.

A straightforward choice is to use raw attention weights as token importance scores. However, attention weights only describe the forward aggregation pattern and are not necessarily faithful to the

cross-modal matching objective [46, 45]. Attribution methods provide a more suitable alternative by measuring token contributions with respect to a specified output or objective, which matches our need to identify tokens that drive cross-modal matching. Inspired by CheferCAM [27], which combines attention maps with their objective gradients, we use gradient-attention relevance as an objective-conditioned token scoring signal:

$$\mathrm{CAM}^{(l)} = I + \mathbb{E}_h \left[\left(\nabla A_h^{(l)} \odot A_h^{(l)} \right)^+ \right], \tag{11}$$

where $A_h^{(l)}$ denotes the attention map of head h at layer l, $\nabla A_h^{(l)}$ is the gradient of the cross-modal matching score with respect to $A_h^{(l)}$, $\mathbb{E}_h[\cdot]$ denotes averaging over attention heads, $\odot$ denotes element-wise multiplication, and $(\cdot)^+$ keeps positive relevance contributions.

Standard Transformer attribution methods usually propagate relevance through the full network to obtain a complete input-level explanation. In contrast, our goal is to construct shared-cover subviews, which only requires a relative ranking of token contributions to the cross-modal matching objective. We therefore propose Layer-Truncated Attribution (LTA), which computes Eq. (11) only at the last Transformer block, where the CLS token directly aggregates patch tokens for cross-modal matching. This preserves the objective conditioned nature of attribution while avoiding full network relevance propagation. Empirically, LTA scores provide a strong ranking signal for subview construction: the selected subviews substantially improve cross-modal retrieval over full-view inputs and other token scoring baselines in Sec. 5.3, while LTA also produces CheferCAM-like localization patterns with only 6.54% extra computation over a forward pass, as analyzed in Appendix C.

Concretely, we instantiate LTA by measuring how each patch token contributes to the cross-modal matching score through the last layer CLS attention. Fig. 1(b) provides an overview of the LTA computation. For an audio-visual pair, let $s_{av} = \langle z_{\mathrm{cls}}^a, z_{\mathrm{cls}}^v \rangle$ be the cross-modal matching score. For modality $m \in \{a, v\}$, we define the relevance score of token i as

$$r_i^m = \mathbb{E}_h \left[\left(\frac{\partial s_{av}}{\partial A_{h,\mathrm{cls},i}^{m,L}} A_{h,\mathrm{cls},i}^{m,L} \right)^+ \right], \tag{12}$$

where $A_{h,\mathrm{cls},i}^{m,L}$ denotes the attention weight from the CLS token to token i at head h of the last Transformer block, and $\mathbb{E}_h[\cdot]$ denotes averaging over attention heads.

4.2 Training Framework

The preceding sections establish two key ingredients: a shared-cover subview can reduce modality-private residuals at the input level, and LTA provides an efficient way to estimate such subviews. However, applying LTA at inference would require paired audio-visual inputs, gradient-based scoring, and explicit subview construction. We therefore adopt a teacher-student training framework, as shown in Fig. 1(a), to distill the attribution-induced decoupling behavior into the student. The teacher generates high-relevance shared-cover targets and low-relevance residual complement targets, while the student learns to align its shared and private tokens with the corresponding responses. Once trained, the student internalizes this behavior and introduces no additional parameters or computational overhead during inference.

4.2.1 Pretraining an Early Converged Relevance Teacher

The teacher is pretrained from scratch following the baseline framework of [14], which separates masked autoencoding (MAE) and contrastive learning (CL) into two forward passes and uses a lower masking ratio for the CL branch. At this stage, the teacher is trained with the original random masking strategy only, without attribution-guided subview selection. Rather than serving as a fully optimized representation model, its role is to provide token-relevance guidance for student training. As analyzed in Appendix B.3, LTA-induced subviews become stable before the full training schedule is completed. Therefore, we stop teacher pretraining at epoch 18 instead of following the original 35-epoch TG-DP schedule.

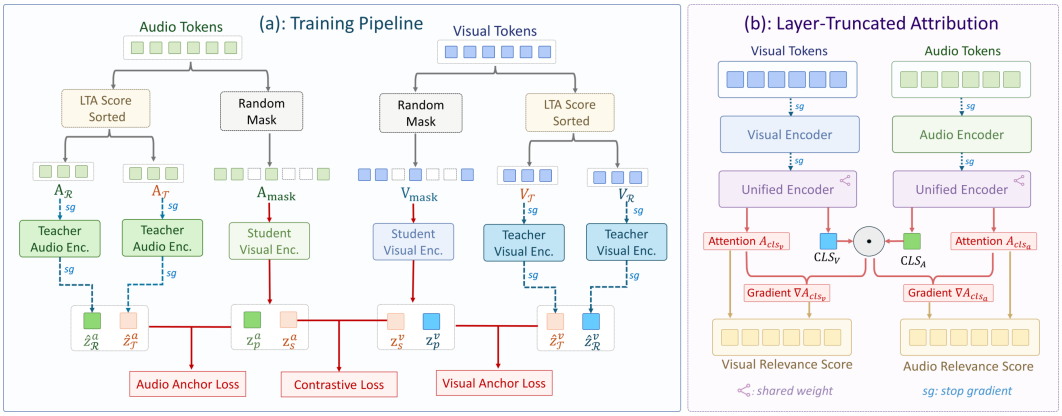


Figure 1: Overall pipeline of our proposed framework. (a) Attribution-guided teacher-student training. A pretrained teacher generates high-relevance shared-cover subviews and low-relevance residual complements, which serve as stop-gradient anchors for training the student's shared and private tokens. For clarity, we omit the MAE branch inherited from the baseline [14]. (b) Layer-Truncated Attribution (LTA). Given the cross-modal matching score, LTA computes CLS-to-token gradient-attention relevance from the last Transformer block.

4.2.2 Student Training

After teacher pretraining, we freeze the teacher and use it to generate attribution-induced subview targets for each audio-visual pair. As shown in Fig. 1(a), for modality $m \in \{a, v\}$, the teacher computes token relevance scores following Eq. (12), and separates the input tokens into a high-relevance subview $X_{\mathcal{T}}^m$ and a low-relevance complement $X_{\mathcal{R}}^m$. The frozen teacher then processes these two subviews and produces stop-gradient anchor targets:

$$\hat{z}_{\mathcal{T}}^m = \mathrm{sg}\left(f_{\theta_T}(X_{\mathcal{T}}^m)\right), \qquad \hat{z}_{\mathcal{R}}^m = \mathrm{sg}\left(f_{\theta_T}(X_{\mathcal{R}}^m)\right), \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient.

The student follows the same backbone architecture, but replaces the original single learnable global token with two learnable tokens for each modality: a shared token and a private token. Their output representations are denoted as z_s^m and z_p^m, respectively. The shared token is encouraged to match the teacher response on the shared-cover subview, while the private token is encouraged to absorb the residual information from the low-relevance complement:

$$\mathcal{L}_{s\text{-anc}} = \sum_{m\in\{a,v\}} \ell_{\cos}\left(z_s^m, \hat{z}_{\mathcal{T}}^m\right), \qquad \mathcal{L}_{p\text{-anc}} = \sum_{m\in\{a,v\}} \ell_{\cos}\left(z_p^m, \hat{z}_{\mathcal{R}}^m\right), \tag{14}$$

where $\ell_{\cos}(x, y) = 1 - \cos(x, y)$ denotes the cosine similarity loss.

To further stabilize representation learning under different masking ratios, we introduce a consistency regularizer on the contrastive token. Specifically, the CL branch with mask ratio 0.5 produces the standard contrastive representation $z_{\mathrm{cl},0.5}^m$, which is detached and used as the target. We then reuse the intermediate features from the MAE branch and obtain the corresponding CL representation under the higher mask ratio 0.75, denoted as $z_{\mathrm{cl},0.75}^m$. The consistency loss is defined as

$$\mathcal{L}_{\mathrm{cons}} = \sum_{m\in\{a,v\}} \left\| z_{\mathrm{cl},0.75}^m - \mathrm{sg}\left(z_{\mathrm{cl},0.5}^m\right) \right\|_2^2. \tag{15}$$

This regularizer encourages the contrastive representation to remain stable across different visible token ratios, while avoiding an additional full forward pass by reusing the MAE branch computation.

The final student objective is

$$\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{base}} + \lambda_a(\mathcal{L}_{s\text{-anc}} + \mathcal{L}_{p\text{-anc}}) + \lambda_c \mathcal{L}_{\mathrm{cons}}, \tag{16}$$

where $\mathcal{L}_{\text{base}}$ denotes the original decoupled MAE and contrastive training objective of [14]. In our framework, the contrastive objective is applied only to the shared tokens. Through the attribution-induced anchor losses and the consistency regularizer, the student learns to associate the shared token with shared-cover subviews and the private token with residual complements, while maintaining robust contrastive representations under different masking ratios. The resulting model requires no additional computation during inference.

5 Experiments

5.1 Experimental Setup

Following standard practice in audio-visual representation learning [12, 13, 14], we pretrain our model on AudioSet-2M [50]. We evaluate the pretrained representations on three downstream tasks: cross-modal retrieval, audio-visual classification, and sound-prompted semantic segmentation. Detailed dataset descriptions and training recipes are provided in Appendix B. For VGGSound retrieval, AVSBench localization, and AS20K classification, we report means over repeated runs, with the repetition protocols and standard deviations detailed in Appendix B.4.

Computational cost. Under the same hardware setting of 8 NVIDIA A100 GPUs, the overall pretraining time increases from 19.0 hours for the adopted backbone to 27.1 hours for our method. A stage-wise breakdown of the training cost is provided in Appendix C.1. Since the teacher and attribution computation are used only during training, our method introduces no extra computational overhead during inference.

5.2 Main Results

Cross-modal retrieval. We evaluate zero-shot audio-visual retrieval on the VGGSound and AudioSet-20K evaluation sets provided by [13]. Following prior work, we compute cosine similarity between frozen audio and visual representations and report Recall@1, Recall@5, and Recall@10 in both audio-to-visual and visual-to-audio directions. For our model, we use the shared token as the global representation for retrieval, since this task primarily evaluates cross-modal shared correspondence.

As shown in Table 1, our method achieves state-of-the-art zero-shot retrieval performance on both AudioSet-20K and VGGSound[51]. Compared with the adopted backbone model TG-DP [14], our method improves R@1 by 3.8 and 4.0 on AudioSet-20K for V→A and A→V retrieval, respectively. On VGGSound, it further improves R@1 by 4.0 and 3.7 in the two directions. These consistent gains demonstrate the effectiveness of our attribution guided shared-private decoupling framework for cross-modal alignment.

	Vision → Audio						**Audio → Vision**					
	AudioSet Eval Subset			VGGSound Eval Subset			AudioSet Eval Subset			VGGSound Eval Subset		
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CAV-MAE [12]	16.1	38.6	49.3	14.7	35.3	45.9	9.5	22.6	32.4	8.3	23.8	32.4
CAV-MAE$^{\text{Scale+++}}$ [13]	18.8	39.5	50.1	14.8	34.2	44.0	15.1	34.0	43.0	12.8	30.4	40.3
LanguageBind [52]	6.4	20.2	28.3	10.3	30.1	39.7	4.4	15.0	22.5	6.5	22.7	33.5
AVSiam [53]	19.7	–	–	19.0	–	–	17.6	–	–	20.4	–	–
ImageBind [9]	22.1	43.2	52.6	21.6	43.4	52.9	20.8	42.6	51.6	20.7	43.2	53.4
DenseAV [54]	26.3	49.9	58.5	24.2	48.2	58.3	20.8	42.6	51.6	25.1	50.1	59.1
CAV-MAE Sync [13]	35.2	58.3	67.6	27.9	51.7	61.8	27.9	52.4	62.2	23.2	46.2	58.1
TG-DP[14]	37.4	59.2	68.1	31.3	54.4	63.9	37.1	59.7	68.2	30.3	53.1	64.2
Ours	**41.2**	**62.8**	**70.8**	**35.3**	**59.8**	**69.5**	**41.1**	**63.3**	**71.3**	**34.0**	**58.9**	**69.5**

Table 1: Zero-shot cross-modal retrieval on the AudioSet-20K and VGGSound evaluation subsets. We report Recall@1, Recall@5, and Recall@10 for both Vision→Audio (V→A) and Audio→Vision (A→V) retrieval using frozen representations.

Audio-visual classification. We evaluate the representation quality of the pretrained models through *attention probing*, following the evaluation protocol of [13]. Specifically, we freeze the pretrained

Method	AS20K mAP	VGGSound Acc.
CAV-MAE$^{\text{Scale+++}}$ [13]	25.3	51.6
MaViL [8]	30.0	–
CAV-MAE Sync [13]	30.5	52.7
TG-DP [14]	32.0	52.7
Ours	**35.2**	**54.4**

Table 2: Audio-visual classification results under attention probing [13].

	AVSBench		ADE20K-Sound	
Method	**mIoU**	**mAP**	**mIoU**	**mAP**
DAVENet [55]	–	–	16.8	17.0
CAV-MAE [12]	28.0	24.5	21.2	20.9
CAV-MAE Sync [13]	22.3	25.2	22.6	22.7
TG-DP [14]	21.8	24.3	21.5	22.3
Ours	**29.9**	**37.3**	**26.1**	**24.0**

Table 3: Sound-prompted semantic segmentation results on AVSBench and ADE20K-Sound.

audio and visual encoders and train a lightweight attention-based classification head on AudioSet-20K and VGGSound. For our model, we average the shared and private tokens as the final global token for classification, ensuring that all compared downstream models have the same computational cost. We report mAP on AudioSet-20K and top-1 accuracy on VGGSound.

As shown in Table 2, our method achieves state-of-the-art audio-visual classification performance on both AudioSet-20K and VGGSound. Compared with TG-DP [14], the adopted backbone model of our framework, our method improves AS20K mAP from 32.0 to 35.2. On VGGSound, our method improves top-1 accuracy from 52.7 to 54.4. These results demonstrate that attribution guided shared-private decoupling improves not only cross-modal retrieval but also the transferability of the learned representations to downstream recognition tasks.

Sound-prompted semantic segmentation. We further evaluate whether the learned representations capture fine-grained audio-visual correspondence through sound-prompted semantic segmentation. Following the protocol of DenseAV [54], we use the similarity heatmap between visual patch tokens and the audio shared CLS token as the segmentation prediction. We evaluate on AVSBench [56] and ADE20K-Sound [57, 54], and report mean intersection over union (mIoU) and mean average precision (mAP). As shown in Table 3, our method achieves the best sound-prompted semantic segmentation performance on both benchmarks. On AVSBench, it improves the adopted backbone TG-DP from 21.8 to 29.9 in mIoU and from 24.3 to 37.3 in mAP, showing a substantial gain in fine-grained audio-visual localization. On ADE20K-Sound, our method improves mIoU from 21.5 to 26.1 and mAP from 22.3 to 24.0. These results indicate that attribution-guided shared-private decoupling leads to more discriminative patch-level representations.

5.3 Empirical Validation of LTA-Induced Subviews

To verify whether LTA can construct more reliable shared-cover subviews, we evaluate the induced subviews directly with the pretrained teacher model. We freeze the teacher, select visible tokens according to different scoring functions under the same retention ratio of 0.50, and feed the resulting subviews back into the teacher encoder. The output representations are then evaluated by zero-shot cross-modal retrieval.

	V→A			A→V		
Subview Score	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Full view	31.7	56.0	65.9	30.5	54.0	65.3
Similarity	21.6	46.2	56.4	23.2	44.8	54.0
Attention	28.5	54.9	64.2	28.7	52.7	63.7
CheferCAM [27]	37.2	61.6	70.6	37.5	60.9	68.3
LTA	**44.1**	**68.5**	**77.3**	**41.2**	**67.2**	**75.8**

Table 4: Empirical validation of LTA-induced subviews on VGGSound. We compare teacher retrieval performance using subviews constructed by different token scoring functions.

As shown in Table 4, LTA-induced subviews substantially improve teacher retrieval performance over the original full view. On VGGSound, LTA improves R@1 from 31.7 to 44.1 for V→A retrieval and from 30.5 to 41.2 for A→V retrieval. In contrast, subviews constructed by token-level similarity or raw attention perform worse than the full view, indicating that these scores do not reliably capture cross-modal shared relevance. We also compare with the original CheferCAM, which performs full relevance propagation through the Transformer. Although CheferCAM improves over the full view, it remains clearly behind LTA. This result suggests that full relevance propagation is not necessarily optimal for subview selection.

To further examine the separation induced by LTA, we perform the same retrieval diagnostic using only the low-relevance residual tokens. Their R@1 drops to 8.9 for V→A and 8.4 for A→V, compared with 44.1 and 41.2 for the high-relevance subviews. This contrast shows that LTA concentrates cross-modal retrieval cues in the selected subviews, while leaving substantially weaker shared cues in the residual complements.

	Similarity		Attention		LTA	
Pair condition	mIoU	mAP	mIoU	mAP	mIoU	mAP
Matched	22.3	25.2	34.2	42.8	44.2	60.1
Mismatched	22.2	25.4	34.4	43.2	34.7	45.7

Table 5: Sound-source localization under matched and mismatched audio-video pairing on AVSBench. For mismatched pairs, the audio is replaced while the visual input and its original sound-source annotation are retained. Both mIoU and mAP are reported in percentage.

We further examine whether LTA responds to audio-video correspondence through a mismatched-pair diagnostic on AVSBench. For each visual sample, we replace its matched audio with a mismatched audio clip and evaluate the resulting localization map against the original visual sound-source annotation. As shown in Table 5, mismatched pairing reduces LTA localization performance by 9.5 mIoU and 14.4 mAP, whereas similarity and raw attention change by at most 0.4 points. These results indicate that LTA is sensitive to the correspondence between the paired modalities, supporting its use as a cross-modal relevance signal for subview selection.

5.4 Verification of Shared-Private Token Decoupling

We further examine whether the learned shared and private tokens are functionally decoupled in the final representation. Since cross-modal retrieval directly depends on shared audio-visual information, we use retrieval performance to measure how much shared information each token captures. As shown in Table 6, the shared token consistently outperforms the private token for zero-shot retrieval on VGGSound. Specifically, the shared token achieves 35.4 R@1 for V→A and 34.3 R@1 for A→V retrieval, while the private token obtains 29.5 and 28.2 R@1, respectively. This gap shows that the shared token carries stronger cross-modal correspondence than the private token, which is consistent with our decoupled design. At the same time, the private token still achieves non-trivial retrieval performance, indicating that the two tokens are not completely separated in terms of shared information. This is also consistent with our formulation, since the residual complement is selected by relatively lower shared relevance rather than by the absence of shared semantics.

Token	Direction	R@1	R@5	R@10
Private	V→A	29.5	53.9	63.9
Private	A→V	28.2	52.9	63.8
Shared	V→A	35.4	59.6	70.5
Shared	A→V	34.3	57.9	69.8

Table 6: VGGSound zero-shot retrieval using different tokens as the global representation.

5.5 Ablation Studies

Setting	**Anchor**	**Cons.**	**Dec.**	**Avg. R@1**	**AS20K mAP**	**Seg. mAP**
Backbone				30.8	32.0	24.3
+ Anchor	✓			33.0	32.2	28.3
+ Consistency	✓	✓		34.3	32.0	26.3
+ Decoupled tokens	✓	✓	✓	**34.9**	**35.1**	**38.2**

Table 7: Incremental ablation of the main training components. Anchor, Cons., and Dec. denote attribution-induced anchor supervision, consistency regularization, and shared-private token decoupling, respectively. Avg. R@1 is averaged over the two retrieval directions on VGGSound. Seg. mAP denotes sound-prompted semantic segmentation mAP on AVSBench.

Effect of training components. We conduct an incremental ablation to examine the contribution of the main components in our training framework. Starting from the adopted backbone, we progressively add attribution-induced anchor supervision, consistency regularization, and shared-private token decoupling. As shown in Table 7, anchor supervision improves the average R@1 from 30.8 to 33.0, showing that attribution-induced subview anchors provide useful guidance for cross-modal alignment. Adding consistency regularization further improves retrieval performance, suggesting that the regularized objective stabilizes global representation learning. Finally, introducing decoupled shared and private tokens leads to the best overall performance across all metrics. In particular, it improves AS20K classification mAP from 32.0 to 35.1 and localization mAP from 26.3 to 38.2. These gains indicate that explicit shared-private decoupling not only transfers subview supervision to fine-grained patch-level correspondence, but also preserves modality-private residual information that benefits recognition tasks requiring complementary audio-visual cues.

Effect of retention ratio. We further study the retention ratio used for constructing attribution-induced subviews. As shown in Table 8, increasing the ratio from 0.25 to 0.50 consistently improves retrieval, suggesting that overly compact subviews may discard useful shared information. Further increasing the ratio to 0.75 brings little retrieval gain and noticeably degrades localization, indicating that excessive token retention can reintroduce modality-private or weakly related regions. Classification remains relatively stable across ratios, showing that global recognition is less sensitive to this hyperparameter. We therefore use 0.50 as the default retention ratio, which offers the best retrieval performance while maintaining strong classification and localization results with moderate teacher side computation.

Retention	Avg. R@1	AS20K mAP	Seg. mAP
0.25	32.6	35.3	**38.3**
0.35	33.9	35.2	37.3
0.50	**34.9**	35.1	38.2
0.75	34.8	**35.5**	31.5

Table 8: Effect of the retention ratio.

6 Conclusion

We presented an attribution guided shared-private decoupling framework for noise reduced audio-visual representation learning. Our key idea is to identify shared-cover subviews with layer-truncated attribution, and to use the resulting high-relevance and low-relevance teacher responses to supervise decoupled shared and private student tokens. This allows the final model to internalize fine-grained attribution guidance without introducing extra inference cost. Experiments on cross-modal retrieval, audio-visual classification, and sound-prompted semantic segmentation demonstrate that our method improves both global audio-visual alignment and patch-level semantic correspondence. We hope that this work encourages further exploration of token-level supervision and shared-private decomposition in multimodal representation learning.

Acknowledgments and Disclosure of Funding

This work was supported by National Key R&D Program of China under Grant No. 2023ZD0120400, in part by National Natural Science Foundation of China (No. 62472423, 62276260).

References

[1] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress.

[2] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[3] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[4] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[6] 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*, pages 8748–8763. PmLR, 2021.

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[8] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36:20371–20393, 2023.

[9] 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*, pages 15180–15190, 2023.

[10] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023.

[11] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731, 2024.

[12] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] 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*, pages 18794–18803, 2025.

[14] Linge Wang, Yingying Chen, Bingke Zhu, Lu Zhou, and Jinqiao Wang. Semantic noise reduction via teacher-guided dual-path audio-visual representation learning, 2026.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[17] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[18] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[19] Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 5408–5432, 2025.

[20] Lu Gao, Wenlan Chen, Daoyuan Wang, Fei Guo, and Cheng Liang. Disentangled cross-modal representation learning with enhanced mutual supervision. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Ye Liu, Zihan Ji, and Hongmin Cai. Plug-and-play feature causality decomposition for multimodal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36:63529–63541, 2023.

[23] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000.

[24] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

[25] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR, 21–27 Jul 2024.

[26] Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. *ArXiv*, abs/2404.03214, 2024.

[27] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[28] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, 2021.

[29] John W. Fisher, Trevor Darrell, William T. Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of the 14th International Conference on Neural Information Processing Systems*, NIPS'00, page 742–748, Cambridge, MA, USA, 2000. MIT Press.

[30] John W. Fisher and Trevor Darrell. Probabalistic models and informative subspaces for audiovisual correspondence. In *Proceedings of the 7th European Conference on Computer Vision-Part III*, ECCV '02, page 592–603, Berlin, Heidelberg, 2002. Springer-Verlag.

[31] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning, 2016.

[32] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9758–9770. Curran Associates, Inc., 2020.

[33] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[34] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7774–7785, Red Hook, NY, USA, 2018. Curran Associates Inc.

[35] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12486, 2021.

[36] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

[37] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International Conference on Learning Representations*, 2021.

[38] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[40] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[41] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[42] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recogn.*, 65(C):211–222, May 2017.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[45] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[46] Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[47] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.

[48] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971, May 2021.

[49] Jiamin Chen, Xuhong Li, Lei Yu, Dejing Dou, and Haoyi Xiong. Beyond intuition: Rethinking token attributions inside transformers. *Transactions on Machine Learning Research*, 2023.

[50] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[51] 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)*, pages 721–725. IEEE, 2020.

[52] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

[53] Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *European Conference on Computer Vision*, pages 303–321. Springer, 2024.

[54] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13117–13127, 2024.

[55] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input, 2018.

[56] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 386–403, Berlin, Heidelberg, 2022. Springer-Verlag.

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

A Proof of Proposition 1

We provide the proof for the audio modality, while the same argument applies to the visual modality. Recall that the full audio input is decomposed into the selected shared-cover subview and its discarded complement:

$$A = A_{\mathcal{T}} \cup A_{\mathcal{R}}, \qquad A_{\mathcal{T}} \cap A_{\mathcal{R}} = \emptyset. \tag{17}$$

For any token subset $B \subseteq A$, the input-level IB criterion is defined as

$$F(B) = I(B; S) - \beta I(B; P_A), \tag{18}$$

where S denotes the cross-modal shared information and P_A denotes the audio-private residual information.

Using the chain rule of mutual information, we have

$$I(A; S) = I(A_{\mathcal{T}}; S) + I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}), \tag{19}$$

and similarly,

$$I(A; P_A) = I(A_{\mathcal{T}}; P_A) + I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}). \tag{20}$$

Substituting the two decompositions into $F(A)$ gives

$$\begin{aligned} F(A) &= I(A; S) - \beta I(A; P_A) \\ &= I(A_{\mathcal{T}}; S) + I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}) - \beta I(A_{\mathcal{T}}; P_A) - \beta I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}) \\ &= F(A_{\mathcal{T}}) + I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}) - \beta I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}). \end{aligned} \tag{21}$$

Therefore,

$$F(A_{\mathcal{T}}) - F(A) = \beta I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}) - I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}). \tag{22}$$

By the Shared Cover Condition,

$$I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}) \leq \epsilon_S, \tag{23}$$

and by the Private Tail Condition,

$$I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}) \geq \epsilon_P. \tag{24}$$

It follows that

$$F(A_{\mathcal{T}}) - F(A) \geq \beta\epsilon_P - \epsilon_S. \tag{25}$$

Consequently, if $\beta\epsilon_P > \epsilon_S$, then

$$F(A_{\mathcal{T}}) > F(A). \tag{26}$$

This proves Proposition 1.

The equality above also reveals the exact condition under which a selected subview improves the input-level IB objective:

$$F(A_{\mathcal{T}}) > F(A) \quad \Longleftrightarrow \quad \beta I(A_{\mathcal{R}}; P_A \mid A_{\mathcal{T}}) > I(A_{\mathcal{R}}; S \mid A_{\mathcal{T}}). \tag{27}$$

In other words, the selected subview is more suitable than the full input for learning the shared representation when the discarded tokens contain sufficiently more conditional private residual information than conditional shared information, after weighting by β.

B Implementation Details

B.1 Datasets Details

AudioSet. AudioSet-2M [50] is a large-scale audio-visual dataset originally curated for audio event classification. It contains approximately two million 10-second video clips collected from YouTube, with audio event annotations over 527 classes. Due to its large scale and diverse acoustic events, AudioSet-2M has been widely used for audio-visual self-supervised pretraining. However, the dataset was not explicitly designed for audio-visual correspondence learning. Therefore, some clips contain weakly correlated or mismatched audio-visual content, such as off-screen sound sources, silent visual regions, or background audio unrelated to the visible scene. This property makes AudioSet-2M a realistic but noisy pretraining source for cross-modal representation learning.

In practice, not all AudioSet videos are accessible because the original data are hosted on YouTube and may become unavailable due to removal, privacy settings, or regional restrictions. In our implementation, approximately 350K samples could not be downloaded. Among the downloaded videos, we further remove clips with missing audio tracks, corrupted frames, or invalid decoding results. After filtering, we obtain 1,390,395 valid audio-visual pairs for pretraining. For downstream audio-visual classification, we use AudioSet-20K following the standard probing protocol. For zero-shot cross-modal retrieval, we follow the AudioSet evaluation split provided by CAV-MAE [12].

VGGSound. VGGSound [51] is a large-scale audio-visual dataset collected from YouTube, containing around 200K 10-second video clips across 309 sound event categories. Unlike AudioSet, VGGSound was curated with explicit audio-visual correspondence in mind: each clip is expected to contain a visible object or event that is semantically related to the accompanying sound. This property makes VGGSound a relatively clean benchmark for evaluating audio-visual alignment, since it contains fewer off-screen or weakly matched audio-visual pairs. In our experiments, we use VGGSound for both downstream classification and zero-shot cross-modal retrieval evaluation. Specifically, we train the probing classifier on the VGGSound training split, and follow the retrieval evaluation split provided by CAV-MAE Sync [13] for zero-shot audio-visual retrieval.

ADE20k-Sound Prompted ADE20K-Sound is a sound-prompted semantic segmentation benchmark constructed from ADE20K [57] and VGGSound [51]. It contains 106 images covering 20 ADE20K object categories, where each image is paired with a semantically corresponding sound clip sampled from VGGSound. The benchmark is designed to evaluate whether an audio-visual model can localize image regions that are relevant to the given sound prompt.

AVSBench AVSBench [56] is a pixel-level audio-visual segmentation benchmark designed to evaluate whether a model can localize sounding objects in videos. Different from ADE20K-Sound, which is constructed from static images paired with sound prompts, AVSBench provides video-based audio-visual samples with manually annotated segmentation masks for sound-producing objects. The original AVSBench-object benchmark contains two settings: single sound source segmentation (S4) and multiple sound source segmentation (MS3), covering 5,356 videos and 23 object categories in total. This makes it a larger and more dedicated benchmark for evaluating fine-grained audio-visual localization.

B.2 Training Pipeline and Implementation Details

Preprocessing. We follow the CAV-MAE Sync preprocessing pipeline for both pretraining and downstream evaluation. Given a video, we uniformly sample $T = 16$ frames. Each frame is resized and center-cropped to 224×224, and then patchified into 16×16 visual patches, producing 196 visual patch tokens per frame. The audio waveform is converted into 128-bin log-Mel filterbanks using a 25 ms Hanning window and a 10 ms hop size. For each sampled frame, we extract a temporally aligned 4-second spectrogram segment of size 128×416 by mapping the frame timestamp to the corresponding spectrogram coordinate. Each spectrogram segment is also patchified with a 16×16 patch size, resulting in 208 audio patch tokens. Standard normalization is applied to the audio spectrograms with mean -5.081 and standard deviation 4.485.

During pretraining, we randomly sample one frame and its temporally aligned spectrogram segment from the $T = 16$ candidates as the input pair. During downstream classification, all $T = 16$ frame-spectrogram pairs are used to construct a temporal sequence for attention probing.

Dual-path pretraining backbone. Our training framework inherits the dual-path design of TG-DP [14], which separates masked reconstruction and contrastive alignment into two objective-specific forward passes. For each audio-visual pair, the model processes the input twice with different masking configurations. The reconstruction branch adopts a high masking ratio of 0.75 and optimizes masked autoencoding, while the contrastive branch adopts a lower masking ratio of 0.50 and optimizes cross-modal alignment. The two branches share the same modality-specific encoders and unified block, but their losses are computed from different masked views.

In the reconstruction branch, audio and visual patch tokens are randomly masked and then encoded by the modality-specific encoders. The visible audio and visual tokens are fed into the unified block

and decoded to reconstruct the masked patches. The reconstruction loss is defined as

$$\mathcal{L}_{\mathrm{rec}} = \sum_{m\in\{a,v\}} \frac{1}{|\mathcal{M}^m|} \sum_{i\in\mathcal{M}^m} \|\hat{x}_i^m - x_i^m\|_2^2 , \tag{28}$$

where $\mathcal{M}^m$ denotes the masked token set of modality m, and $\hat{x}_i^m$ is the reconstructed patch token.

In the contrastive branch, the student receives a lower-mask-ratio view and produces modality-level representations for cross-modal alignment. Different from the original single-global-token design, our student uses two learnable tokens for each modality: a shared token z_s^m and a private token z_p^m. The shared tokens are used for cross-modal contrastive learning:

$$\mathcal{L}_{\mathrm{cl}} = \frac{1}{2}\left(\mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}\right), \tag{29}$$

where

$$\mathcal{L}_{a\to v} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\langle z_{s,i}^a, z_{s,i}^v\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle z_{s,i}^a, z_{s,j}^v\rangle/\tau)}, \tag{30}$$

and $\mathcal{L}_{v\to a}$ is defined symmetrically. Here, N is the batch size and τ is the temperature parameter. The base dual-path objective is therefore written as

$$\mathcal{L}_{\mathrm{base}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{cl}}\mathcal{L}_{\mathrm{cl}}. \tag{31}$$

Early-converged relevance teacher. Before student training, we pretrain a relevance teacher using the same dual-path backbone as TG-DP [14]. The teacher is trained from scratch with the original random masking strategy, without LTA-based subview construction or shared-private token decoupling. The full TG-DP training schedule contains 35 epochs, while we stop the teacher at epoch 18. This early stopping is motivated by the empirical observation that token-relevance rankings become sufficiently stable in the middle stage of training, so a fully converged teacher is not necessary for generating subview supervision. After pretraining, the teacher is frozen and used only to compute LTA scores and produce stop-gradient anchor targets during student training.

Attribution-guided student training. The student is initialized from scratch and trained for 25 epochs using the same dual-path backbone. For each audio-visual pair, the frozen teacher computes LTA scores from the last unified block and ranks patch tokens within each modality. We use a retention ratio of $\rho = 0.50$ to split tokens into a high-relevance shared-cover subview and a low-relevance residual complement. The teacher responses on these two subviews serve as stop-gradient anchors for the student's shared and private tokens, respectively.

Downstream evaluation. For zero-shot cross-modal retrieval, we use the shared token z_s^m as the global representation of each modality, since retrieval mainly evaluates cross-modal shared correspondence. For audio-visual classification, we follow the attention probing protocol and freeze the pretrained encoders. At each of the $T = 16$ temporal steps, we average the shared and private tokens to form a single modality-level token, concatenate the audio and visual tokens over time, and train a lightweight attention-based classification head. For sound-prompted semantic segmentation, we use the similarity between the sound representation and visual patch tokens to produce the localization heatmap, following the zero-shot evaluation protocol used in prior audio-visual grounding benchmarks.

Table 9 summarizes the main hyperparameters used for teacher pretraining, student pretraining, and downstream probing.

B.3 Teacher Training Dynamics

To justify the use of an early-converged relevance teacher, we evaluate the quality of LTA-induced subviews at different teacher checkpoints. For each checkpoint, we freeze the teacher, compute LTA scores, select the top-ρ tokens with $\rho = 0.50$, and feed the resulting subviews back into the teacher encoder. We then measure zero-shot cross-modal retrieval performance using the subview representations, and report the average Recall@1 and Recall@5 over the two retrieval directions.

As shown in Fig. 2, the quality of teacher-induced subviews improves rapidly during the early training stage. The average R@1 increases from 20.0 at epoch 3 to 42.0 at epoch 18, while the average R@5

Setting	Teacher Pretrain	Student Pretrain	AS20K Attn. Probe	VGGSound Attn. Probe
Dataset	AS-2M	AS-2M	AS-20K	VGGSound
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	2×10^{-4}	2×10^{-4}	5×10^{-3}	1×10^{-3}
LR Scheduler	Cosine	Cosine	Cosine	Cosine
Epochs	18	25	10	10
Warmup Epochs	3.5	2.5	1.5	1
Batch Size	8×64	8×64	36	96
GPUs	8×A100	8×A100	3×A100	8×A100
Audio Input Size	128×416	128×416	$16 \times 128 \times 416$	$16 \times 128 \times 416$
Visual Input Size	224×224	224×224	$16 \times 224 \times 224$	$16 \times 224 \times 224$
Patch Size (Audio/Visual)	16×16	16×16	16×16	16×16
Frames per Clip Used	1	1	16	16
MAE Masking Ratio	0.75	0.75	–	–
CL Masking Ratio	0.50	0.50	–	–
LTA Retention Ratio ρ	–	0.50	–	–
LTA Layer	–	Last unified block	–	–
Encoder Frozen	No	No	Yes	Yes
Class-Balanced	No	No	No	Yes
Mixup	No	No	Yes	Yes
Random Shift	Yes	Yes	Yes	Yes
Loss Function	$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{student}}$	BCE	CE
Reconstruction Weight λ_{rec}	1.0	1.0	–	–
Contrastive Weight λ_{cl}	0.1	0.1	–	–
Contrastive Temperature τ	0.05	0.05	–	–
Anchor Weight λ_a	–	0.5	–	–
Consistency Weight λ_c	–	0.5	–	–
Weight Averaging	No	No	Yes	Yes
Norm Mean	-5.081	-5.081	-5.081	-5.081
Norm STD	4.485	4.485	4.485	4.485

Table 9: Hyperparameters for pretraining and attention probing. The MAE and CL masking ratios apply to the reconstruction and contrastive branches, respectively, in both modalities. During student pretraining, LTA uses the frozen teacher's last unified block with retention ratio $\rho = 0.50$ for both modalities. The encoder-freezing setting refers to the encoder being optimized or probed in each column; the supervising teacher remains frozen during student training. A dash denotes not applicable.

increases from 42.6 to 67.2. After epoch 18, the retrieval performance becomes largely saturated: later checkpoints only bring marginal fluctuations, with average R@1 staying around 42–43 and average R@5 staying around 67–68. This indicates that the token-relevance ranking used for subview construction has already become sufficiently stable by the middle stage of training.

The stopping epoch is selected using this teacher-side subview retrieval diagnostic. We therefore stop teacher pretraining at epoch 18 instead of following the full 35-epoch TG-DP schedule. This early stopping reduces the teacher-side training cost while preserving reliable subview supervision for student training.

Sensitivity to the teacher stopping epoch. We further evaluate how the teacher stopping epoch affects final student performance. Using teachers stopped at epochs 15, 18, and 21, we train the corresponding students following the recipe in Appendix B.2 and evaluate cross-modal retrieval on VGGSound and localization on AVSBench. As shown in Table 10, average retrieval R@1 remains within 34.50–34.86, and the variation in either retrieval direction is at most 0.65 R@1. Localization performance ranges from 29.5 to 30.6 mIoU and from 36.5 to 38.2 mAP. These results indicate that final student performance is relatively stable across the tested stopping epochs near the saturation stage, supporting the use of an early-stopped teacher.

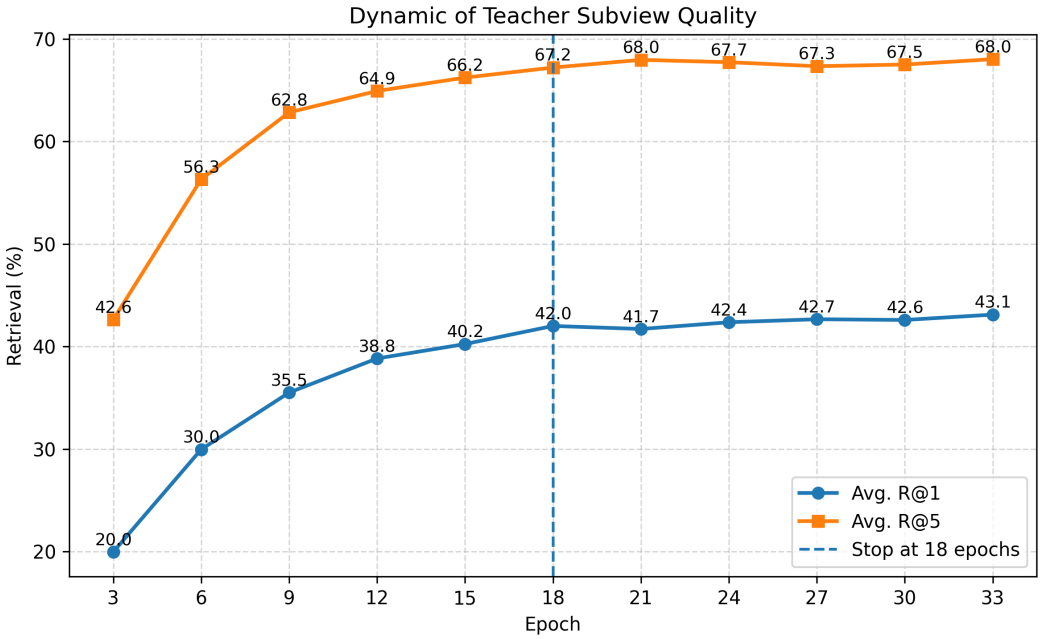


Figure 2: Dynamics of teacher-induced subview quality during teacher pretraining. For each teacher checkpoint, we construct LTA-induced subviews with retention ratio $\rho = 0.50$ and evaluate zero-shot cross-modal retrieval using the resulting subview representations. Avg. R@1 and Avg. R@5 denote the mean retrieval performance over the two retrieval directions. The subview quality improves rapidly in the early stage and becomes saturated around epoch 18, after which later checkpoints provide only marginal gains. We therefore stop teacher pretraining at epoch 18 rather than using the full 35-epoch training schedule.

Teacher stop	**Retrieval**			**Localization**	
epoch	V→A R@1	A→V R@1	Avg. R@1	mIoU	mAP
15	35.38	33.62	34.50	29.6	37.1
18	35.44	34.27	34.86	30.6	38.2
21	35.19	33.81	34.50	29.5	36.5

Table 10: Sensitivity of final student performance to the teacher stopping epoch. The default stopping epoch is shaded. Avg. R@1 is averaged over both retrieval directions.

B.4 Repeated Runs

To assess the variability of the reported results, we conduct five independent pretraining runs with different random seeds and evaluate the resulting models on VGGSound retrieval and AVSBench localization. For AS20K classification, we repeat attention probing five times using the same pretrained checkpoint with the encoders frozen. The classification results therefore measure variability in downstream probing, while the retrieval and localization results reflect variability across independent pretraining runs. Table 11 reports the mean and standard deviation for each metric.

The standard deviations of VGGSound retrieval R@1 are 0.35 and 0.70 for V→A and A→V, respectively. On AVSBench, the standard deviations are 0.44 mIoU and 0.57 mAP, while repeated AS20K probing yields a standard deviation of 0.13 mAP. These results support the stability of the reported performance under the tested training and evaluation settings.

Method	AVSBench localization		AS20K classification
	mIoU	mAP	mAP
Ours	29.88 ± 0.44	37.26 ± 0.57	35.20 ± 0.13

(a) Localization and classification

Method	Vision → Audio			Audio → Vision		
	R@1	R@5	R@10	R@1	R@5	R@10
Ours	35.32 ± 0.35	59.76 ± 0.64	69.54 ± 0.63	33.95 ± 0.70	58.93 ± 0.82	69.51 ± 0.30

(b) VGGSound retrieval

Table 11: Repeated-run results, reported as mean ± standard deviation. Retrieval and localization results use five independent pretraining runs with different random seeds. Classification results use five attention-probing runs with the same pretrained checkpoint. The corresponding means in Tables 1–3 are rounded to one decimal place.

C Additional Analysis of Layer-Truncated Attribution

C.1 Computation Cost

We report the overall pretraining time and the per-pair computation cost of LTA scoring separately. The teacher, attribution computation, and subview construction are used only during pretraining and introduce no additional inference-time overhead.

Overall pretraining cost. Under the same hardware setting of 8 NVIDIA A100 GPUs, the adopted TG-DP backbone requires 19.0 hours for 35 epochs of pretraining. Our method first trains a teacher for 18 epochs and then trains the student for 25 epochs, following Appendix B.2. Table 12 provides the corresponding breakdown in wall-clock hours. Teacher pretraining takes 9.77 hours, while the student stage consists of 13.57 hours of base computation, 2.40 hours of additional consistency regularization, and 1.39 hours of teacher-anchor construction and LTA scoring. These components sum to 27.13 hours, reported as 27.1 hours in the main text. The main source of additional training cost is the one-time pretraining of the early-stopped teacher.

Training component	Time (hours)
TG-DP baseline (35 epochs)	19.00
Teacher pretraining (18 epochs)	9.77
Student base computation (25 epochs)	13.57
Additional consistency regularization	2.40
Teacher-anchor construction and LTA scoring	1.39
Our total pretraining time	**27.13**

Table 12: Breakdown of pretraining time under the same hardware setting of 8 NVIDIA A100 GPUs. Times are reported in wall-clock hours. The four components below the baseline sum to our total pretraining time; the last three comprise the student stage.

Per-pair cost of LTA scoring. To isolate the model computation involved in LTA scoring, we exclude data loading and audio/image preprocessing and report the cost per audio-visual pair. As shown in Table 13, a standard teacher forward pass costs 71.78 GFLOPs, while the teacher forward together with LTA scoring costs 76.47 GFLOPs. The additional attribution computation therefore requires 4.69 GFLOPs, or 6.54% of the teacher forward cost. Memory usage increases from 746.83 MB to 779.10 MB, an increment of 32.27 MB. The 6.54% figure describes the incremental FLOPs of attribution relative to a teacher forward pass; it does not represent the increase in total pretraining time. Likewise, the 1.39-hour component in Table 12 includes both teacher-anchor construction and LTA scoring.

Component	GFLOPs	Memory (MB)	Relative FLOPs
Teacher forward	71.78	746.83	1.00×
Teacher forward + LTA	76.47	779.10	1.07×
LTA overhead	+4.69	+32.27	+6.54%

Table 13: Computation cost of LTA scoring per audio-visual pair, excluding data loading and audio/image preprocessing. The full scoring process includes one teacher forward pass and the additional attribution computation. Relative FLOPs are computed with respect to the teacher forward.

C.2 Segmentation Ability of LTA

We further evaluate whether LTA captures fine-grained audio-visual correspondence beyond global retrieval. To this end, we use sound-prompted semantic segmentation as an external localization test. Given an audio prompt and a visual image or video frame, the model produces a spatial heatmap indicating regions that are relevant to the input sound. We compare three types of heatmaps: the original audio-visual similarity heatmap, the full CheferCAM attribution heatmap, and our LTA heatmap.

As shown in Table 14, attribution-based heatmaps substantially improve over the original similarity heatmap on both ADE20K-Sound and AVSBench. On ADE20K-Sound, LTA improves mIoU from 21.9 to 31.1 and mAP from 25.4 to 39.6. On AVSBench, LTA improves mIoU from 22.3 to 44.2 and mAP from 25.2 to 60.1. These results indicate that gradient-attention attribution provides a much stronger localization signal than raw cross-modal similarity.

Compared with full CheferCAM, LTA achieves very similar segmentation performance. On AVS-Bench, the gap is only 0.3 in mIoU and 1.1 in mAP, showing that last-layer gradient-attention relevance already preserves most of the localization ability needed for identifying sound-related regions. Although CheferCAM performs full relevance propagation through the Transformer, its marginal gain is small for this subview-selection purpose. Therefore, LTA offers a favorable trade-off: it retains strong fine-grained localization ability while avoiding the computational cost of full-network attribution propagation. This supports our choice of LTA as an efficient token-relevance estimator for constructing shared-cover subviews.

C.3 Visualization of LTA

We provide qualitative visualizations to further examine whether LTA captures sound-relevant regions at the token level. Fig. 3 compares heatmaps generated by different scoring functions, including LTA, CheferCAM, raw attention, and token-level similarity. For each example, warmer regions indicate tokens assigned with higher relevance scores.

The visualization shows that LTA produces compact and semantically meaningful heatmaps around sound-related objects or regions. For example, it highlights the musical instrument in the first row, the sounding object region in the second row, the object being interacted with in the third row, and the mouth/face region of the crying infant in the fourth row. These regions are consistent with the expected audio-visual correspondence in each sample. CheferCAM produces similar localization patterns, but it requires full relevance propagation through the Transformer, which introduces substantial computational overhead. The close visual agreement between LTA and CheferCAM suggests that last-layer gradient-attention relevance preserves the core attribution behavior needed for subview selection, while avoiding expensive full-network propagation.

In contrast, raw attention and similarity-based scores are less reliable. Attention maps often spread over visually salient but acoustically irrelevant areas, since attention weights only reflect forward token aggregation and are not explicitly conditioned on the cross-modal matching objective. Similarity maps can also be noisy or biased toward broad appearance-level correlations, leading to less precise localization. These qualitative results are consistent with the quantitative segmentation results in Table 14, and further support the use of LTA as an efficient token-level scoring function for constructing shared-cover subviews.

Method	AVSBench		ADE20K-Sound	
	mIoU	mAP	mIoU	mAP
Original heatmap	22.3	25.2	21.9	25.4
CheferCAM	**44.5**	**61.2**	**32.3**	**41.9**
LTA	44.2	60.1	31.1	39.6

Table 14: Sound-prompted semantic segmentation results of different heatmaps on AVSBench and ADE20K-Sound. We report mean IoU (mIoU) and mean average precision (mAP) in percentage.

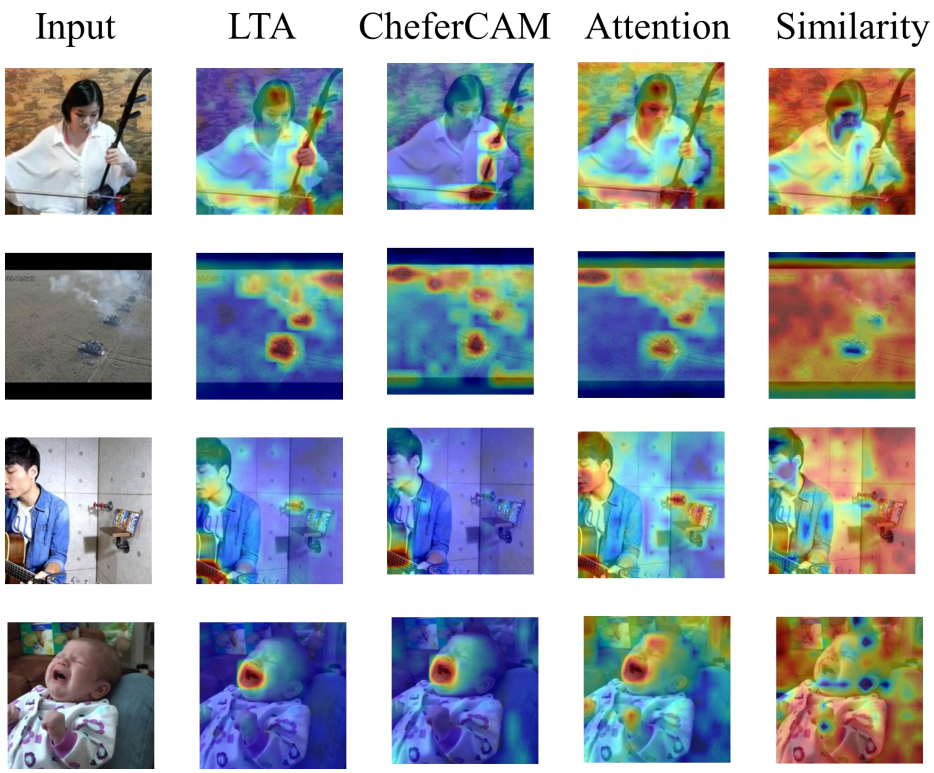


Figure 3: Visualization of different score type.

D Limitations

Our work has several limitations. First, although the motivation of shared-private decoupling is generally applicable to multimodal representation learning, our experiments are focused on the audio-visual setting. Due to the high cost of large-scale multimodal pretraining, we have not yet validated the proposed framework on other modality pairs, such as image-text or audio-language data. Extending attribution-guided shared-private decoupling to broader multimodal settings is an important direction for future work.

Second, the decoupling ability internalized by the student is still limited. As shown in Table 6, the private token still retains a non-negligible amount of cross-modal shared information, indicating that the learned shared and private representations are not completely separated. Moreover, the retrieval performance of the student shared token does not fully match the teacher-induced shared-cover subview in Table 4. This suggests that the current teacher-student training transfers the attribution-induced decoupling behavior only partially. More effective distillation objectives or architectural designs may further improve the separation between shared and private representations.

Third, LTA is a heuristic approximation to the ideal shared-relevance score used in our theoretical analysis. The theory assumes a ranking function that reflects each token's contribution to the cross-

modal shared factor, while LTA estimates this relevance using the last-layer gradient-attention signal of the matching objective. Although our empirical results show that LTA provides an effective ranking signal for subview construction, it does not provide a strict guarantee of recovering the theoretically ideal score. Developing more principled and theoretically grounded token-level relevance estimators remains an interesting future direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are aligned with the theoretical analysis and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: we discuss the limitations in Appendix D

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical result is stated under explicit conditions, and the complete proof is provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Appendix B.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets used in this work are publicly available. Code will be made available along with publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide implementation details in Appendix B.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We do not report error bars because repeated large-scale pretraining runs are computationally expensive. Instead, we evaluate under standard protocols and report consistent results across multiple benchmarks and ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As described in Appendix B.2, all experiments were conducted on 8 NVIDIA A100 GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available datasets and does not involve human subjects or private data collection. We disclose experimental settings, compute resources, assumptions, and limitations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We briefly discuss potential positive applications in audio-visual understanding and localization.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer:[N/A]

    Justification: We do not release any new datasets or high-risk foundation models. Our work uses standard, publicly available audio-visual datasets and provides training and evaluation code for reproducibility, which does not introduce additional misuse risks beyond the existing public availability of these resources.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available audio-visual datasets and existing code/model resources, all of which are properly cited in the paper. We follow the licenses and terms of use specified by the original creators.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[N/A]

    Justification:[N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: LLM is used only for writing the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.