LEARNING COMPRESSED AUDIO-VISUAL REPRESENTATIONS VIA CONDITIONAL CAUSAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual distillation compresses large multimodal teachers by transferring predictions or representations to compact students. Although existing objectives employ different knowledge carriers, they generally assume that teacher targets provide reliable supervision. However, teachers learned from observational data may entangle task semantics with latent multimodal priors arising from scene context, modality quality, and cross-modal correspondence. Direct imitation can thus preserve both useful knowledge and context-dependent response patterns, especially under limited student capacity. We present Causal Audio-Visual Distillation (CAVD), a teacher-student framework that standardizes supervision before compression. CAVD captures recurrent multimodal context patterns with a compact latent state set and models teacher responses conditioned on each state. It then standardizes these responses over a shared reference distribution, producing supervision that is less dependent on individual sample context composition. The student learns both standardized predictions and state-conditioned response variations without auxiliary annotations or modality-specific bias modules. Operating in latent representation space, CAVD provides a unified formulation for mutually exclusive, multi-label, and dense outputs. Experiments on temporal localization, video parsing, and sounding-object segmentation demonstrate consistent improvements over distillation baselines, with compact students approaching teacher performance while using approximately 80–95% fewer parameters and FLOPs.

1 INTRODUCTION

Audio-visual learning combines visual appearance with acoustic dynamics, but high-capacity multimodal models remain costly for resource-constrained deployment. Knowledge distillation (KD) reduces this cost by transferring supervision from a large teacher to a compact student. Existing objectives mainly use softened predictions Hinton et al. (2015), intermediate features Chen et al. (2021), attention responsesZagoruyko & Komodakis (2016), or cross-modal relations Bang et al. (2025); Park et al. (2026) as carriers of teacher knowledge. These designs differ in the form and granularity of transferred information, but share a common assumption that reproducing teacher outputs or representations provides reliable supervision for compact models.

This assumption can become unreliable when the teacher is learned from observational audio-visual data. As shown in Figure 1(a), teacher responses can entangle task semantics with latent multimodal priors, including scene co-occurrence and visual dominance. Scene co-occurrence captures recurring associations between acoustic events and surrounding environments, while visual dominance reflects the tendency of visual representations to exploit salient appearance cues even when they are not causally responsible for the prediction. Some of these patterns represent stable task structure, whereas others act as shortcuts specific to the training distribution. Changing the distillation carrier does not resolve this issue: logits may preserve context-dependent decisions, features may preserve the supporting representations, and relational objectives may retain the same dependencies through cross-modal interactions, as summarized in Figure 1(b). Under limited student capacity, these shortcut correlations can be preferentially retained during compression. This issue differs from ordinary modality uncertainty. Uninformative modalities usually produce weak or uncertain responses, whereas prior-induced shortcuts can remain confident and misleading. Confidence weighting adjusts supervision strength but does not alter the contextual dependency embedded in the target.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

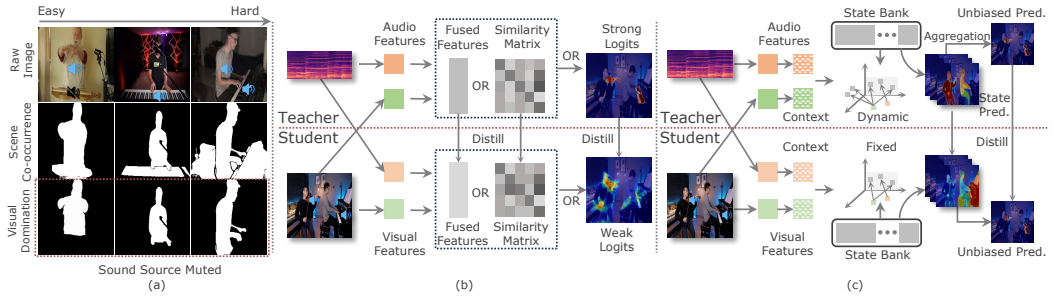


Figure 1: Motivation for causal audio-visual distillation. Conventional objectives transfer predictions or representations produced under the teacher's observational training distribution, where task semantics and latent multimodal priors can be entangled. CAVD marginalizes teacher supervision under a sample-independent reference distribution before transferring it to the compact student.

Reducing contextual influence through student regularization provides a possible solution, but the supervision target remains generated under the original observational distribution. Weak regularization may leave shortcuts intact, whereas overly strong regularization can suppress useful teacher knowledge. Removing context-related information entirely is also undesirable, since contextual conditions can indicate when audio-visual cues are informative. The desired supervision should therefore reduce dependence on sample-specific context composition while preserving prediction behavior across represented conditions. We formulate this requirement through a compact structural causal model of teacher supervision. A latent joint prior captures recurring multimodal configurations and influences observed modalities and the fused prediction pathway, allowing contextual frequencies to affect teacher responses through latent backdoor paths. The goal is not to remove contextual variation, but to separate stable task behavior from context-dependent response patterns. CAVD evaluates the teacher under represented latent states while keeping the current audio-visual input fixed, and marginalizes the resulting predictions over a sample-independent reference distribution. This process standardizes the contribution of represented conditions while preserving input-specific task information.

Technically, we instantiate this formulation as CAVD with three coordinated components. Cross-Fitted Joint-Prior Discovery (JPD) constructs a compact representation of recurring joint prior variation while reducing the influence of dominant task semantics. Prior-Conditional Teacher Response (PTR) models teacher responses under different latent states and standardizes them through a sample-independent reference distribution. Prior-Conditional Student Distillation (PSD) transfers both the standardized prediction and the centered state-conditional variation. As illustrated in Figure 1(c), CAVD provides a unified distillation framework without auxiliary annotations or modality-specific bias branches. The formulation operates across different prediction interfaces rather than requiring separate designs for individual modalities or tasks. To our knowledge, CAVD is the first framework to formulate audio-visual KD as causal adjustment of the teacher prediction transferred during distillation. Our contributions are summarized as follows:

- We identify prior-induced bias in observational supervision as a carrier-independent limitation of audio-visual distillation and distinguish systematic prior dependence from ordinary modality uncertainty.
- We propose CAVD, a unified teacher-student framework that discovers latent multimodal prior structure, standardizes the supervision source, and transfers task-relevant conditional behavior without auxiliary annotations or modality-specific bias branches.
- Extensive evaluations on multiple benchmarks and perturbation settings achieve a promising balance between compression and performance.

2 Related Work

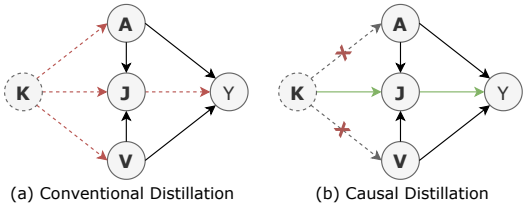
Multimodal Knowledge Distillation. Closely related to audio-visual compression, EM-KTD Park et al. (2026) incorporates cross-stream token relations with reliability weighting. Together with

classical KD Hinton et al. (2015) and SPKD Tung & Mori (2019), these methods shift from output/feature transfer to structural token-level knowledge transfer.

3 METHODOLOGY

3.1 PROBLEM SETUP AND CAUSAL ANALYSIS

Let $\mathbf{X} = (\mathbf{V}, \mathbf{A})$ denote synchronized visual and audio random variables and let Y denote the task outcome, with $\mathbf{x}$ and y denoting their realizations. For task τ, the output encoding lies in $\mathbb{R}^{C_\tau}$. The Structural Causal Model (SCM) Pearl et al. (2003) introduces a latent joint prior state $\mathbf{K}$ that affects the observed modalities and the prior-conditioned fused representation $\mathbf{J}$. A fixed reference distribution $p^{\mathtt{ref}}(\mathbf{K})$ provides the weights used to marginalize responses across states.



(a) Conventional Distillation (b) Causal Distillation

Figure 2: SCMs for audio-visual distillation.

Observational Distillation. Figure 2(a) describes conventional distillation. The joint prior state $\mathbf{K}$ affects the observed audio $\mathbf{A}$, visual input $\mathbf{V}$, and prior-conditioned representation $\mathbf{J}$, while Y depends on the modality inputs and the mediated representation. Consequently, the backdoor paths $\mathbf{A} \dashleftarrow \mathbf{K} \dashrightarrow \mathbf{J} \dashrightarrow Y$ and $\mathbf{V} \dashleftarrow \mathbf{K} \dashrightarrow \mathbf{J} \dashrightarrow Y$ remain open, whereas the ordinary prediction paths are written with solid arrows, such as $\mathbf{A} \to \mathbf{J} \dashrightarrow Y$. For model index $\nu \in \{T, S\}$, where T and S denote the teacher and student, respectively, the observational response is:

$$p^{\nu,\mathtt{obs}}(Y = y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}_{\mathbf{K}\sim p(\mathbf{K}\mid\mathbf{x})}[p^{\nu}(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{K})]\,, \tag{1}$$

where $p(\mathbf{K} \mid \mathbf{x})$ assigns different state weights to each input. Directly matching this response transfers both the teacher's state-conditional predictions and these input-dependent weights.

Fixed-Reference Adjustment. Figure 2(b) removes the backdoor edges $\mathbf{K} \dashrightarrow \mathbf{A}$ and $\mathbf{K} \dashrightarrow \mathbf{V}$. CAVD keeps the modality pair fixed, evaluates the response under every represented state, and replaces the input-dependent weights with a fixed reference distribution,

$$p^{\nu,\mathtt{adj}}(Y = y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}_{\mathbf{K}\sim p^{\mathtt{ref}}(\mathbf{K})}[p^{\nu}(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{K})]\,, \tag{2}$$

where $p^{\mathtt{ref}}(\mathbf{K})$ is the same for every sample. If $\mathbf{K}$ is an admissible adjustment set and $p^{\mathtt{ref}}(\mathbf{K}) = p(\mathbf{K})$, Eq. 2 gives the backdoor-adjusted response $p^{\nu}(Y \mid do(\mathbf{X} = \mathbf{x}))$. Other reference distributions simply specify different fixed weights over the represented states.

3.2 CROSS-FITTED JOINT-PRIOR DISCOVERY

The objective is to represent the latent joint prior $\mathbf{K}$ with a finite state bank for the sum in Eq. 2. For task τ, site n of sample i is indexed by $s = (i, n)$, with input $\mathbf{x}_s$, random outcome Y_s, and observed label y_s. A site may be a global token, temporal segment, or spatial query. Directly clustering teacher features often groups task semantics, since these features are strongly associated with Y_s. Using outcome classes as states creates the same problem by treating the prediction target as a cause that precedes the input. We therefore subtract a class-conditioned mean before state discovery. Let $\mathbf{h}^T_{m,s}$ denote the teacher representation at site s for stream $m \in \{a, v, f\}$, corresponding to the audio, visual, and fused streams. The decomposition is:

$$\mathbf{h}^T_{m,s} = \underbrace{\mathbb{E}[\mathbf{h}^T_{m,s} \mid Y_s]}_{\text{class-semantic component}} + \underbrace{\big(\mathbf{h}^T_{m,s} - \mathbb{E}[\mathbf{h}^T_{m,s} \mid Y_s]\big)}_{\text{class-conditioned residual}}, \tag{3}$$

and the state bank is learned from the residuals together with audio-visual relations. Subtracting the class mean reduces dominant task semantics, although the residual can retain task information and the class mean can contain prior variation. Let $\omega_{s,c}$ be the ground-truth weight of class c at site s, or the corresponding stopped teacher probability when labels are unavailable. Define $\bar{\omega}_{s,c} = \omega_{s,c}/(\sum_{c'=1}^{C_\tau} \omega_{s,c'} + \varepsilon)$ with $\varepsilon > 0$. The training samples are partitioned into F disjoint folds, and

$\mathcal{F}_b$ contains all sites from samples in fold b. For $s \in \mathcal{F}_b$, the leave-fold-out class center is:

$$\boldsymbol{\mu}_{m,c}^{T,(-b)} = \frac{\sum_{s' \notin \mathcal{F}_b} \omega_{s',c} \mathbf{h}_{m,s'}^T}{\sum_{s' \notin \mathcal{F}_b} \omega_{s',c} + \varepsilon}, \quad m \in \{a, v, f\}, \quad c \in \{1, \ldots, C_\tau\}, \tag{4}$$

where the sites in fold b do not contribute to the center used for that fold. For a multi-label site, the normalized weights $\bar{\omega}_{s,c}$ combine the relevant leave-fold-out centers. We then remove this class-associated component from each stream,

$$\boldsymbol{\delta}_{m,s}^T = \mathbf{h}_{m,s}^T - \sum_{c=1}^{C_\tau} \bar{\omega}_{s,c} \boldsymbol{\mu}_{m,c}^{T,(-b)}, \quad m \in \{a, v, f\}. \tag{5}$$

The resulting $\boldsymbol{\delta}_{m,s}^T$ retains variation not explained by the estimated class centers. Cross-fitting ensures that its class centers are estimated from other folds, so the held-out site does not affect the statistics used to compute its own residual. When the leave-fold-out centers approximate the class means, $\mathbb{E}[\boldsymbol{\delta}_{m,s}^T \mid Y_s = c] \approx 0$, which reduces first-order class separation in the state-discovery space. Pairwise distances then depend less on class-specific mean offsets and more on variation among sites with comparable task semantics. Centering does not erase the full task signal, since within-class structure and audio-visual dependence can remain in $\boldsymbol{\delta}_{m,s}^T$. This motivates coupling the residuals with relation statistics before assigning latent states.

Joint Audio-Visual Descriptor. Residuals from separate streams do not capture how audio and visual features correspond. We append compact relation statistics, such as pooled cross-attention or temporal similarity (task-dependent), and use a learned projection to form a joint descriptor $\mathbf{u}_s^T \in \mathbb{R}^{d_K}$. Let $\mathcal{E}^T = \{\mathbf{e}_k^T \in \mathbb{R}^{d_K}\}_{k=1}^M$ be a bank of M latent states. Distances to the state vectors define the assignment weights,

$$q_{s,k} = \frac{\exp\left(-\|\mathbf{u}_s^T - \mathbf{e}_k^T\|_2^2 / \tau_K\right)}{\sum_{j=1}^M \exp\left(-\|\mathbf{u}_s^T - \mathbf{e}_j^T\|_2^2 / \tau_K\right)}, \tag{6}$$

where $\tau_K > 0$ controls assignment sharpness and $\mathbf{q}_s = (q_{s,1}, \ldots, q_{s,M})$ is the assignment vector for site s. Soft assignment lets nearby states share sites during training. The same $\mathbf{q}_s$ later forms the observed mixture in the teacher anchoring loss.

State-Bank Optimization. The state bank is learned by reconstructing the joint descriptor while promoting sharp site-level assignments and balanced state usage across the training set. Here, $\mathbb{E}_s$ denotes the empirical average over training sites, $\mathcal{H}$ the categorical entropy, and $\mathcal{D}_{\mathrm{KL}}$ the categorical KL divergence. We further define $\bar{\mathbf{q}} = \mathbb{E}_s[\mathbf{q}_s]$ as the mean assignment and π_0 as a fixed target distribution over the M states. The resulting objective is:

$$\mathcal{L}_{\texttt{prior}} = \mathbb{E}_s \left[\sum_{k=1}^M q_{s,k} \|\mathbf{u}_s^T - \mathbf{e}_k^T\|_2^2 \right] + \mathbb{E}_s[\mathcal{H}(\mathbf{q}_s)] + \mathcal{D}_{\mathrm{KL}}(\bar{\mathbf{q}} \| \pi_0), \tag{7}$$

where the reconstruction term ties states to recurring patterns, the entropy term separates assignments at each site, and the aggregate KL term prevents the bank from concentrating on only a few states. The target π_0 regularizes only the average assignment during discovery. After the bank is frozen, the reference weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ are set to the empirical average $\pi_k = \mathbb{E}_s[q_{s,k}]$. Thus, $\boldsymbol{\pi}$ and π_0 serve different roles and coincide only when the same weights are chosen for both. We further use $\mathrm{Cat}(\boldsymbol{\pi})$ to denote the label distribution with $p^{\texttt{ref}}(\mathbf{K} = k) = \pi_k$.

3.3 Prior-Conditional Teacher Response

Backdoor adjustment requires a teacher response for each state before applying the fixed weights $\boldsymbol{\pi}$, but a conventional teacher exposes only $p^{T,\texttt{obs}}(Y \mid \mathbf{X})$. For the finite state bank, the gap between the available teacher prediction and the desired reference-adjusted prediction is:

$$p^{T,\texttt{obs}}(Y \mid \mathbf{x}) - p^{T,\texttt{adj}}(Y \mid \mathbf{x}) = \sum_{k=1}^M [p(\mathbf{K} = k \mid \mathbf{x}) - \pi_k] \, p^T(Y \mid \mathbf{x}, \mathbf{K} = k). \tag{8}$$

The gap depends on how the input-dependent weights differ from $\boldsymbol{\pi}$ and on how the teacher prediction varies across states. The adjusted prediction therefore cannot generally be recovered from the observational prediction alone. We first model the state-conditional responses and then marginalize them with $\boldsymbol{\pi}$. At site s, let $p_s^{T,\texttt{obs}}$ be the original teacher-head prediction. The map Ψ_τ converts

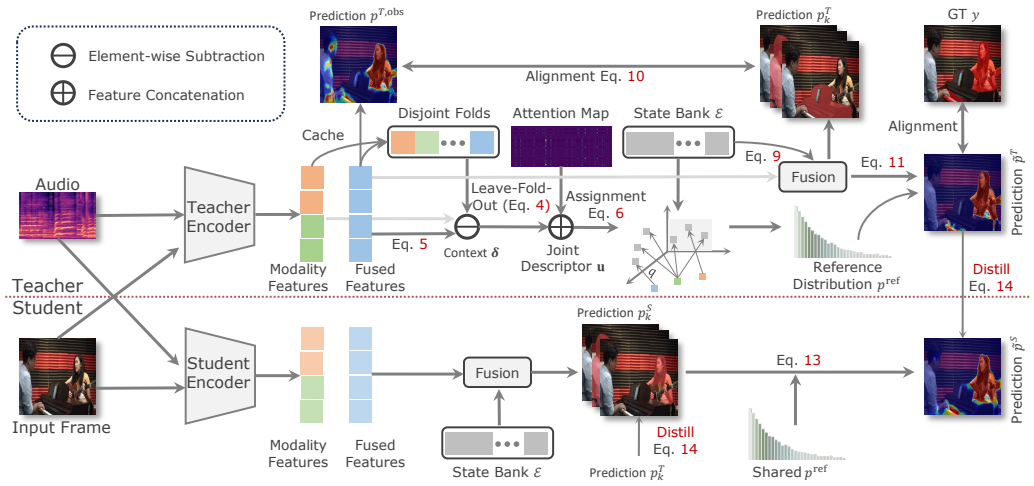


Figure 3: Overview of the CAVD pipeline. Given audio-visual inputs, CAVD discovers a joint latent prior state bank and models the teacher response conditioned on each state. It further transfers the state-conditional responses and reference-marginalized prediction to the student.

scores to probabilities using softmax for mutually exclusive outputs and sigmoid for multi-label or dense outputs. We use $\mathcal{D}_\tau$ for categorical KL or element-wise Bernoulli divergence and $\mathcal{L}_{\texttt{task}}$ for the per-site task loss. Each site distributes its assignment mass unevenly across the state bank, so an independent head for every state receives uneven supervision and can learn unrelated state-specific offsets. The SCM instead places state influence along the shared $\mathbf{K} \to \mathbf{J} \to Y$ pathway, motivating structured variation around a common response. We therefore use the fused teacher representation $\mathbf{h}_{f,s}^T$ and factorize the state-dependent change,

$$\boldsymbol{\eta}_{s,k}^T = \mathcal{G}^T(\mathbf{h}_{f,s}^T) + \mathcal{U}^T(\mathbf{h}_{f,s}^T)\mathbf{e}_k^T, \tag{9}$$

where $\mathcal{G}^T$ captures the response shared across states and $\mathcal{U}^T$ maps each state coordinate to a site-dependent modification, with $p_{s,k}^T = \Psi_\tau(\boldsymbol{\eta}_{s,k}^T)$. After removing the shared term, the state-varying scores factorize through the common coordinates $\mathbf{e}_k^T$, so their rank is bounded by d_K. This coupling shares information across unevenly observed states, limits arbitrary extrapolation, and enables parallel evaluation after one teacher forward pass.

Observed-State Anchoring. Low-rank modulation restricts how state responses can vary but does not determine which responses agree with the teacher on observed inputs. In particular, several collections of state responses can yield the same reference-marginalized prediction. The assignment $\mathbf{q}_s$ provides the training-time link between the current observation and the state bank, so the observational teacher prediction supervises their assignment-weighted aggregate,

$$\mathcal{L}_{\text{anchor}}^T = \mathbb{E}_s\left[\mathcal{D}_\tau\left(\texttt{sg}(p_s^{T,\text{obs}}) \middle\| \sum_{k=1}^M \texttt{sg}(q_{s,k}) p_{s,k}^T\right)\right], \tag{10}$$

where $\texttt{sg}(\cdot)$ denotes stop-gradient. The loss reconstructs the observed teacher response through the assignment-weighted conditional responses, following the aggregation in Eq. 1. It therefore anchors the observed mixture without forcing every compatible state to reproduce the same output. Stopping $q_{s,k}$ prevents response learning from changing the state assignments, while variation in $\mathbf{q}_s$ across sites and the shared functions $\mathcal{G}^T$ and $\mathcal{U}^T$ constrain the individual responses.

Sample-Independent Reference Marginalization. Anchoring uses $\mathbf{q}_s$ to identify states compatible with the observation, but using the same input-dependent weights for the final prediction would retain the observational aggregation in Eq. 1. The final teacher target therefore separates local anchoring from reference marginalization. For input $\mathbf{x}_s$, it evaluates $p_{s,k}^T$ for every state and combines them with the sample-independent $\boldsymbol{\pi}$,

$$\widetilde{p}_s^T = \mathbb{E}_{k\sim\text{Cat}(\boldsymbol{\pi})}\left[p^T(Y \mid \mathbf{x}_s, \mathbf{K} = k)\right] = \sum_{k=1}^M \pi_k p_{s,k}^T. \tag{11}$$

The replacement of $\mathbf{q}_s$ by $\boldsymbol{\pi}$ removes site-specific state weighting from the target while retaining input information inside every $p^T_{s,k}$. The order also matters since $\mathbb{E}_{k\sim\mathrm{Cat}(\boldsymbol{\pi})}[\Psi_\tau(\boldsymbol{\eta}^T_{s,k})] \neq \Psi_\tau(\mathbb{E}_{k\sim\mathrm{Cat}(\boldsymbol{\pi})}[\boldsymbol{\eta}^T_{s,k}])$ in general. Marginalizing in probability space preserves the competition or confidence produced within each state, and the finite sum evaluates the expectation exactly. We jointly constrain the individual responses and calibrate their reference-marginalized prediction,

$$\mathcal{L}^T_{\texttt{total}} = \mathcal{L}^T_{\texttt{anchor}} + \mathbb{E}_s\big[\mathcal{L}_{\texttt{task}}(y_s, \widetilde{p}^T_s)\big]\,. \tag{12}$$

Applying labels only to $\widetilde{p}^T_s$ avoids imposing the same target on every state. The teacher backbone, original head, assignment model, and state bank remain fixed, leaving only the low-rank response module trainable.

3.4 Prior-Conditional Student Distillation

The student is deliberately designed differently from the teacher response module. It neither discovers states nor uses $\mathbf{q}_s$ to form a prediction. Allowing the compact model to estimate its own input-dependent state weights would reproduce the observational aggregation in Eq. 1, even if its conditional responses matched the teacher. Instead, the student inherits the state identities and $\boldsymbol{\pi}$, while the corresponding teacher responses serve only as targets in the transfer objective. Let $\mathbf{h}^S_s$ denote its task-site representation. A learned projection $\mathcal{P}_K$ maps each frozen teacher state to $\mathbf{e}^S_k = \mathcal{P}_K(\texttt{sg}(\mathbf{e}^T_k))$. Using student response maps $\mathcal{G}^S$ and $\mathcal{U}^S$, it forms $p^S_{s,k} = \Psi_\tau\big(\mathcal{G}^S(\mathbf{h}^S_s) + \mathcal{U}^S(\mathbf{h}^S_s)\mathbf{e}^S_k\big)$ for state k. Its prediction rule is:

$$\widetilde{p}^S_s = \mathbb{E}_{k\sim\mathrm{Cat}(\boldsymbol{\pi})}\big[p^S(Y \mid \mathbf{x}_s, \mathbf{K} = k)\big] = \sum_{k=1}^{M} \pi_k p^S_{s,k}. \tag{13}$$

The stop-gradient keeps the teacher state bank frozen while $\mathcal{P}_K$ adapts its coordinates to the compact representation. Across sites, neither the inherited states nor $\boldsymbol{\pi}$ changes, so $\mathbf{x}_s$ enters the prediction only through $p^S_{s,k}$. Although the conditional parameterization follows Eq. 9, its role differs: the teacher completes responses from observational anchors, whereas the student approximates the frozen teacher responses within a compact backbone. When $\mathbf{K}$ is an admissible adjustment set and $p^{\texttt{ref}}(\mathbf{K}) = p(\mathbf{K})$, Eq. 13 has the backdoor-adjustment interpretation $p^S(Y \mid do(\mathbf{X} = \mathbf{x}_s))$. Otherwise, it is the fixed-weight response defined by the chosen $\boldsymbol{\pi}$.

The reference prediction alone does not determine how each conditional response departs from it, while matching every absolute state response overlaps with the reference-prediction constraint. We therefore transfer the response changes centered at each model's reference prediction together with the reference prediction itself,

$$\mathcal{L}_{\texttt{transfer}} = \mathbb{E}_s\left[\sum_{k=1}^{M} \pi_k \left\|\texttt{sg}(p^T_{s,k} - \widetilde{p}^T_s) - (p^S_{s,k} - \widetilde{p}^S_s)\right\|_2^2\right] + \mathbb{E}_s\big[\mathcal{D}_\tau\big(\texttt{sg}(\widetilde{p}^T_s)\|\widetilde{p}^S_s\big)\big]\,. \tag{14}$$

The centered changes have zero $\boldsymbol{\pi}$-weighted mean for both models. The first term transfers how the conditional response varies around the corresponding reference prediction, while the second preserves the prediction used at inference. This separation avoids reproducing in the first term the common response already aligned by the second. The complete student objective $\mathcal{L}^S_{\texttt{total}}$ is,

$$\mathcal{L}^S_{\texttt{total}} = \mathbb{E}_s\big[\mathcal{L}_{\texttt{task}}(y_s, \widetilde{p}^S_s)\big] + \mathcal{L}_{\texttt{transfer}}, \tag{15}$$

where the task loss fits downstream labels and $\mathcal{L}_{\texttt{transfer}}$ transfers both the centered response variation and the reference prediction. Raw observational logits and direct feature regression are excluded to avoid restoring the input-dependent weighting removed above.

4 Experiments

4.1 Downstream Tasks and Datasets

We evaluate CAVD on three audio-visual downstream tasks with distinct prediction interfaces. Audio-visual event localization (AVEL) uses the AVE dataset Tian et al. (2018) to localize events that are jointly visible and audible over temporal segments. Audio-visual video parsing (AVVP) uses the LLP dataset Tian et al. (2020) for fine-grained temporal parsing of audible, visible, and

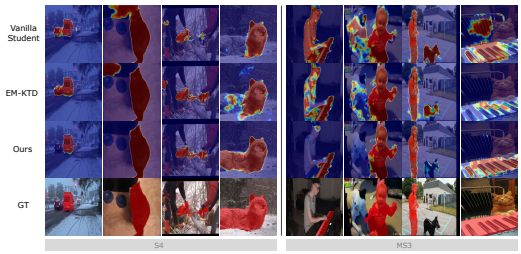


Figure 4: Qualitative comparison of sounding-object segmentation on the S4 and MS3 subsets.

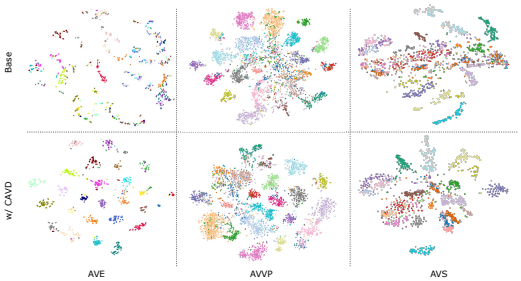


Figure 5: t-SNE visualization of compact student features with and without CAVD.

Table 1: Comparison with state-of-the-art methods on the LLP dataset.

Method	Audio: HTS-AT												Audio: SSAST-Tiny											
	Segment-level						Event-level						Segment-level						Event-level					
	A	V	AV	Type	Event	Avg.	A	V	AV	Type	Event	Avg.	A	V	AV	Type	Event	Avg.	A	V	AV	Type	Event	Avg.
Baselines & Teacher Models (Visual: Swin-V2-L)																								
AVMoE	62.1	60.0	54.4	58.8	59.0	58.9	51.8	55.7	47.6	51.7	50.2	51.4	-	-	-	-	-	-	-	-	-	-	-	-
Mettle	64.3	55.9	50.9	57.0	61.2	57.9	56.1	52.2	45.1	51.1	53.1	51.5	-	-	-	-	-	-	-	-	-	-	-	-
Student Models (Visual: Swin-V2-T)																								
Mettle	57.4	53.4	47.1	52.6	56.0	53.3	48.8	48.1	40.9	45.9	47.3	46.2	53.8	52.6	46.1	50.8	52.2	51.1	44.5	47.2	39.8	43.8	41.7	43.4
+ Vanilla KD	57.6	53.0	46.5	52.4	55.0	52.9	48.9	48.1	40.4	45.8	46.3	45.9	53.1	52.3	45.8	50.4	52.4	50.8	44.2	47.0	39.4	43.5	41.5	43.1
+ AT	57.6	52.8	46.3	52.2	57.1	53.2	48.5	47.0	40.2	45.2	45.1	45.2	54.6	53.0	47.2	51.6	52.1	51.7	44.1	46.6	39.8	43.5	42.0	43.2
+ SPKD	58.4	53.8	47.6	53.3	57.4	54.1	49.5	48.2	41.2	46.3	48.3	46.7	55.5	53.7	47.2	52.1	52.5	52.2	45.8	48.3	40.7	44.9	44.3	44.8
+ MTST	59.1	54.7	48.3	54.0	57.9	54.8	50.1	49.1	41.5	46.9	47.9	47.1	56.5	54.4	48.2	53.0	53.4	53.1	45.8	48.1	41.1	45.0	43.5	44.7
+ EM-KTD	59.7	54.5	48.5	54.2	58.1	55.0	50.7	49.3	41.9	47.3	48.8	47.6	56.1	53.9	47.9	52.6	52.5	52.6	46.6	48.4	41.3	45.4	44.8	45.3
+ CAVD (Ours)	**62.0**	**55.4**	**49.8**	**55.7**	**60.1**	**56.6**	**53.5**	**51.1**	**43.6**	**49.4**	**51.4**	**49.8**	**58.4**	**55.0**	**49.2**	**54.2**	**54.7**	**54.3**	**48.7**	**49.8**	**43.0**	**47.2**	**47.3**	**47.2**

joint audio-visual events. Audio-visual segmentation (AVS) uses the S4 and MS3 subsets of AVS-Bench Zhou et al. (2022) to predict pixel-level masks of sounding objects.

4.2 Main Results

Quantitative Comparison. CAVD achieves the best aggregate performance across the evaluated tasks and backbones. AVE requires temporal alignment between audible and visible events. As shown in Table 3, consistent gains across AVMoE Cheng et al. (2024), DG-SCT Duan et al. (2023), HTS-AT, and SSAST-Tiny show that CAVD is not tied to a particular fusion design or audio encoder. LLP further evaluates audio, visual, and joint events at both segment and event levels. The results in Table 1 indicate that CAVD preserves modality-specific information while maintaining cross-modal consistency across temporal granularities. AVSBench extends the evaluation to pixel-level source delineation. In Table 2, the improvements hold for both the single-source S4 and multi-source MS3 subsets, as well as the VGGish and EfficientAT settings. Consistent performance across these prediction interfaces indicates that reference-standardized transfer preserves task-relevant audio-visual correspondence during compression.

Table 3: Comparison with state-of-the-art methods on the AVE benchmark.

Method	Audio: HTS-AT		Audio: SSAST-Tiny	
	AVMoE	DG-SCT	AVMoE	DG-SCT
Teacher Model (Visual: Swin-V2-L)				
Teacher Model	82.60	82.20	-	-
Student Models (Visual: Swin-V2-T)				
Student Baseline	77.14	76.46	70.42	68.18
+ Vanilla KD	77.92	76.22	71.31	69.09
+ AT	77.72	77.31	71.08	69.35
+ SPKD	78.79	77.87	71.67	68.02
+ MTST	78.06	78.83	71.46	69.81
+ EM-KTD	78.75	78.64	72.42	70.13
+ CAVD (Ours)	**80.86**	**80.52**	**75.10**	**73.25**

Qualitative Comparison. Figure 4 shows that CAVD produces masks that more closely follow sounding objects in both single-source and multi-source scenes. This advantage becomes more apparent when visual saliency alone cannot determine which object is acoustically active, where standardizing teacher responses reduces the influence of scene-specific co-occurrence. Figure 5 provides a complementary representation-level comparison across the three tasks. Relative to the conventional student, CAVD yields more compact feature distributions with reduced overlap among

Table 2: Comparison with state-of-the-art methods on the AVS benchmarks.

Method	Audio: VGGish						Audio: EfficientAT					
	AVS-Bench-S4			AVS-Bench-MS3			AVS-Bench-S4			AVS-Bench-MS3		
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	Avg.	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	Avg.	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	Avg.	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	Avg.
Baselines & Teacher Models (Visual: PVTv2-b5)												
TPAVI	78.74	87.90	83.32	54.00	64.50	59.25	-	-	-	-	-	-
DiffusionAVS	81.38	90.20	85.79	58.18	70.90	64.54	-	-	-	-	-	-
AVSegFormer	82.06	89.90	85.98	58.36	69.30	63.83	-	-	-	-	-	-
UFE-AVS (Teacher)	83.15	90.40	86.78	61.95	70.90	66.43	-	-	-	-	-	-
Student Models (Visual: PVTv2-b0)												
AVSegFormer	77.20	85.81	81.51	55.93	66.48	61.21	74.32	82.28	78.30	54.52	62.97	58.75
+ Vanilla KD	77.05	86.02	81.54	56.08	66.21	61.15	74.50	84.11	79.31	54.38	63.32	58.85
+ AT	77.92	86.51	82.22	55.81	66.23	61.02	75.19	85.11	80.15	55.13	64.41	59.77
+ SPKD	78.37	86.17	82.27	56.63	67.26	61.95	75.64	84.97	80.31	55.26	63.77	59.52
+ MTST	78.69	87.15	82.92	56.52	67.04	61.78	75.92	85.19	80.56	55.08	64.93	60.01
+ EM-KTD	78.65	86.25	82.45	57.15	67.26	62.21	76.15	85.55	80.85	55.90	65.30	60.60
+ CAVD (Ours)	**80.58**	**87.84**	**84.21**	**59.82**	**68.52**	**64.17**	**78.96**	**87.93**	**83.45**	**57.82**	**67.93**	**62.88**

Table 4: Pillar-level ablation of CAVD.

ID	Causal Pillar			Task Average				
	JPD	PTR	PSD	AVEL	AVVP		AVS	
					Segment Avg.	Event Avg.	S4 Avg.	MS3 Avg.
I	✗	✗	✗	68.18	51.1	43.4	78.30	58.75
II	✗	✓	✓	70.50	52.8	45.8	80.60	60.80
III	✓	✗	✓	69.63	52.2	44.7	79.55	59.95
IV	✓	✓	✗	71.04	53.5	46.5	81.30	61.75
V	✓	✓	✓	**73.25**	**54.3**	**47.2**	**83.45**	**62.88**

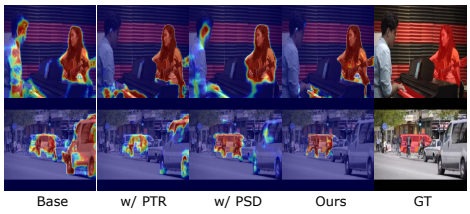


Figure 6: Progressive attention changes. PTR/PSD implies JPD by default.

task-relevant groups. The consistent pattern across different prediction interfaces suggests that standardized teacher supervision promotes more structured compact representations.

4.3 Analysis and Discussion

We report the average results across three tasks.

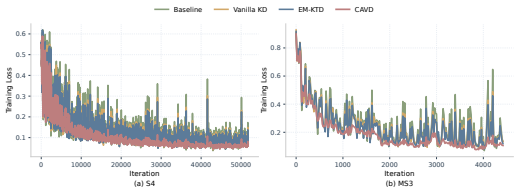


Figure 7: Training loss convergence of distillation methods on AVSBench S4 and MS3.

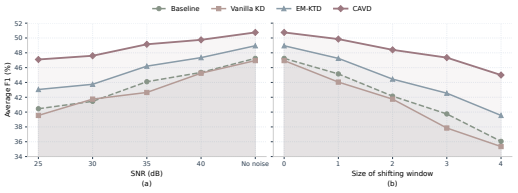


Figure 8: Robustness on LLP to audio noise and shifts of the visual window toward later frames.

Convergence and Robustness. Figure 7 shows more stable late-stage convergence for CAVD on AVSBench S4 and MS3, consistent with the use of shared reference weights across samples. In Figure 8, we further evaluate LLP parsing under Gaussian audio noise at controlled signal-to-noise (SNR) ratios and visual temporal shifts, where the visual sampling window moves toward later frames while the audio window remains fixed. Audio noise weakens audible-event information, whereas temporal shifts misalign visual events with the sound and may alter joint-event boundaries. Under both perturbations, the baseline and other KD methods degrade more rapidly as severity increases, while CAVD maintains a more gradual decline. The shared aggregation rule prevents perturbed inputs from directly changing state weights, while conditional responses remain sensitive to event content. Under severe noise or large temporal shifts, however, degraded acoustic information and cross-modal misalignment still limit parsing performance.

Core Causal Pillars. In Table 4, removing PTR causes the largest performance drop, highlighting the reference-standardized teacher target as the central link between state discovery and student learning. Without JPD, raw teacher descriptors tend to organize states around dominant task seman-

Table 5: Representations for latent-state construction.

Setting	AVEL	AVVP	AVS
Raw Feature	70.50	49.3	70.70
Class Index	70.68	49.4	70.94
Residual Only	71.69	50.1	71.58
Joint Residual	**72.94**	**50.6**	**72.71**

Table 6: Teacher weighting and anchoring.

Setting	AVEL	AVVP	AVS
Observed	70.20	48.9	70.05
+ Anchoring	71.68	50.0	71.38
+ Reference	71.40	49.6	71.02
+ Both	**72.76**	**50.5**	**72.58**

Table 7: Student transfer loss settings.

Setting	AVEL	AVVP	AVS
Task only	68.18	47.3	68.53
+ Prediction	71.55	50.2	71.67
+ Variation	71.09	49.8	71.22
+ Both	**72.82**	**50.4**	**72.76**

tics rather than recurrent contextual conditions. Removing PSD transfers the standardized prediction without the centered variations needed to preserve state-conditioned response differences. Figure 6 provides a complementary qualitative view. As the three components are progressively integrated, attention shifts from the visually co-occurring man toward the sounding piano in one case, and from a visually salient car toward the actual sounding car in another.

Latent-State Construction. In Table 5, state construction improves as the representation shifts from semantic or label-defined descriptions toward cross-fitted joint residuals. `Raw Feature` is dominated by task semantics, while `Class Index` suppresses the contextual diversity within each category. `Residual Only` removes first-order category content, but in-sample centers still introduce self-influence and the descriptor lacks explicit audio-visual correspondence cues. `Joint Residual` instead combines centers estimated from held-out folds with cross-modal relations, capturing recurring conditions beyond category identity while preserving multimodal structure.

Teacher Response Formation. In Table 6, `Anchoring` and `Reference` address complementary aspects of teacher response formation. Observed retains the assignment weights of the current input, allowing sample-specific state composition to directly influence the target. `Anchoring` ties the conditional response family to the original teacher output, reducing drift in weakly represented states. `Reference` instead uses shared weights to standardize state contributions across samples while preserving each state-conditional prediction. Together, they first ground the conditional responses in observed teacher behavior and then construct a sample-independent target.

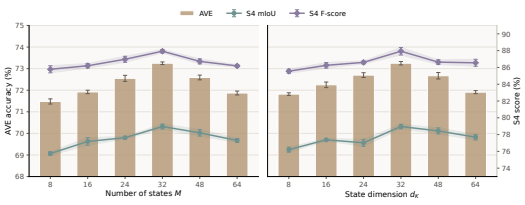


Figure 9: Sensitivity analysis of latent-state number M and dimension d_K.

Student Transfer Objectives. In Table 7, `Prediction` and `Variation` provide complementary supervision beyond the Task-only objective. `Prediction` aligns the reference-weighted output used at inference, but leaves the organization of the underlying conditional responses unconstrained. `Variation` transfers deviations from the shared reference, preserving state-dependent changes without reintroducing input-specific state weights. Together, they align both the inference-time prediction and the conditional variations centered around it.

Hyperparameter Sensitivity. In Figure 9, the number of states M controls the granularity of the proxy adjustment set. A small M merges distinct prior conditions, whereas an overly large M leaves insufficient support for reliable state-conditional responses. The state dimension d_K governs representation capacity and modulation rank. Small values may conflate different prior mechanisms, while large values can retain outcome-related or instance-specific variation.

5 Conclusion

Drawing on causal theory, we propose CAVD, a causal framework for compressed audio-visual representation learning. CAVD decouples contextual condition identification from contextual effects on prediction, and constructs supervision under a sample-independent reference distribution. This design reduces dependence on training composition while preserving conditional task behavior, reframing audio-visual distillation from reproducing observational teacher outputs to transferring standardized supervision. CAVD requires neither additional annotations nor modality-specific bias branches and generalizes across prediction interfaces. Experiments on diverse audio-visual tasks and settings demonstrate consistent effectiveness and robustness.

AI USE STATEMENT

In this work, we employed generative AI tools to improve the clarity, fluency, and grammatical quality of the English writing, including the abstract, introduction, methodology, and experimental sections. They were not involved in aspects such as literature search or data analysis.

REFERENCES

Geonho Bang, Minjae Seong, Jisong Kim, Geunju Baek, Daye Oh, Junhyung Kim, Junho Koh, and Jun Won Choi. Rctdistill: Cross-modal knowledge distillation framework for radar-camera 3d object detection with temporal fusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25315–25324. IEEE, 2025.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19846–19856. IEEE, 2025.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650. IEEE, 2022.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10797–10806. IEEE, 2020.

Mengxi Chen, Linyu Xing, Yu Wang, and Ya Zhang. Enhanced multimodal representation learning with cross-modal kd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11766–11775, 2023.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. *Advances in Neural Information Processing Systems*, 37:36218–36242, 2024.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5006–5015. IEEE, 2021.

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24087–24096, 2025.

Tongtong Cheng, Rongzhen Li, Yixin Xiong, Tao Zhang, Jing Wang, and Kai Liu. Mcam: Multimodal causal analysis model for ego-vehicle-level driving video understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5479–5489. IEEE, 2025.

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

Euntae Choi, Youshin Lim, Byeong-Yeol Kim, Hyung Yong Kim, Hanbin Lee, Yunkyu Lim, Seung Woo Yu, and Sungjoo Yoo. Masked token similarity transfer for compressing transformer-based asr models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22899–22910. IEEE, 2025.

Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27435–27445. IEEE, 2024.

Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36:56075–56094, 2023.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4178–4188. IEEE, 2025.

Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Em-kd: distilling efficient multimodal large language model with unbalanced vision tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21111–21119, 2026.

Mengyu Gao and Qiulei Dong. Causality-guided prompt learning for vision-language models via visual granulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1141–1151. IEEE, 2025.

Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12155–12163, 2024.

Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision*, pp. 106–121. Springer, 2018.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C 2 kd: Bridging the modality gap for cross-modal knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015. IEEE, 2024.

Dino Ienco and Cassio Fraga Dantas. Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. *arXiv preprint arXiv:2408.07080*, 2024.

Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum. Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Sungnyun Kim, Sungwoo Cho, Sangmin Bae, Kangwook Jang, and Se-Young Yun. Multi-task corrupted prediction for learning robust audio-visual speech representation. In *International Conference on Learning Representations*, volume 2025, pp. 65598–65619, 2025.

Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X 3 kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13343–13353. IEEE, 2023.

Chi-Hsi Kung, Frangil Ramirez, Juhyung Ha, Yi-Ting Chen, David Crandall, and Yi-Hsuan Tsai. What changed and what could have changed? state-change counterfactuals for procedure-aware video representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12294–12306. IEEE, 2025.

Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. Coca: Collaborative causal regularization for audio-visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 12995–13003, 2023.

Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2373–2383. IEEE, 2023.

Chaoyang Li, Jianyang Qin, Jinhao Cui, Zeyu Liu, Ning Hu, and Qing Liao. Joint scheduling of causal prompts and tasks for multi-task learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25124–25134. IEEE, 2025.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12458–12468. IEEE, 2024.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7450–7460, 2026.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640. IEEE, 2023.

Xusheng Liang, Lihua Zhou, Nianxin Li, Miao Xu, Ziyang Song, Dong Yi, Jinlin Wu, Jiawei Ma, Hongbin Liu, Zhen Lei, et al. Multimodal causality-driven representation learning for generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13670–13679, 2026.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18020–18029. IEEE, 2022.

Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, et al. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22266–22275. IEEE, 2025.

Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26318–26329. IEEE, 2024a.

Ye Liu, Zihan Ji, and Hongmin Cai. Plug-and-play feature causality decomposition for multimodal representation learning. *Advances in Neural Information Processing Systems*, 38:58912–58938, 2026.

Ying Liu, Ge Bai, Lu Chenji, Shilong Li, Zhang Zhang, Ruifang Liu, and Wenbin Guo. Eliminating the language bias for visual question answering with fine-grained causal intervention. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024b.

Dwarikanath Mahapatra, Abhijit Das, Manish Pandey, Joy Dhar, Sudipta Roy, Zongyuan Ge, Behzad Bozorgtabar, and Imran Razzak. Proximity-constrained counterfactual decoding for hallucination-robust medical vqa. In *European Conference on Computer Vision*, pp. 421–437. Springer, 2026.

Sijie Mai and Shiqin Han. Learning invariant modality representation for robust multimodal learning from a causal inference perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45698–45721, 2026.

Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Dong Li, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *IEEE Transactions on Image Processing*, 34:4108–4119, 2025.

Roy Miles, Ismail Elezi, and Jiankang Deng. $v_\{k\}d$: Improving knowledge distillation using orthogonal projections. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15720–15730. IEEE, 2024.

Hiroshi Morioka and Aapo Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International conference on artificial intelligence and statistics*, pp. 3399–3426. PMLR, 2023.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705. IEEE, 2021.

Hyoungseob Park, Lipeng Ke, Pritish Mohapatra, Huajun Ying, Alex Wong, et al. Entropy-monitored kernelized token distillation for audio-visual compression. In *International Conference on Learning Representations*, volume 2026, pp. 132845–132868, 2026.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Judea Pearl et al. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685): 46, 2003.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5190–5201. IEEE, 2023.

Mohammed Rakib and Arunkumar Bagavathi. G 2 d: Boosting multimodal learning with gradient-guided distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4059–4068. IEEE, 2025.

Florian Schmid, Khaled Koutini, and Gerhard Widmer. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *ICASSP 2023-2023 IEEE international Conference on acoustics, Speech and signal processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. Looking similar, sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26897–26908. IEEE, 2024.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740. IEEE, 2024.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3713–3722. IEEE, 2020.

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

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

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Haoyu Wang, Le Wang, Sanping Zhou, Jingyi Tian, Zheng Qin, Yabing Wang, Gang Hua, and Wei Tang. Towards precise embodied dialogue localization via causality guided diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13350–13360. IEEE, 2025.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International conference on medical image computing and computer-assisted intervention*, pp. 216–226. Springer, 2023.

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13139–13150. IEEE, 2024.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022.

Jie Wen, Yicheng Liu, Chao Huang, Chengliang Liu, Yong Xu, and Xiaochun Cao. Causal interventional prompt tuning for few-shot out-of-distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15669–15680, 2024.

Ehsan Yaghoubi, Andre Peter Kelm, Timo Gerkmann, and Simone Frintrop. Acoustic and visual knowledge distillation for contrastive audio-visual localization. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 15–23, 2023.

Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12996–13010, 2021a.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9842–9852. IEEE, 2021b.

Chufan Yi, Hongyu Qu, Shiyu Xuan, Rui Yan, Xiangbo Shu, Fang Zhao, and Guo-Sen Xie. Test-time counterfactual calibration for hallucination-resistant temporal grounding. In *European Conference on Computer Vision*, pp. 191–209. Springer, 2026.

Heeseung Yun, Joonil Na, and Gunhee Kim. Dense 2d-3d indoor prediction with sound via aligned cross-modal distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7829–7838. IEEE, 2023.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19027–19036. IEEE, 2023.

Mingfeng Zha, Tianyu Li, Guoqing Wang, Peng Wang, Yangyang Wu, Yang Yang, and Heng Tao Shen. Implicit counterfactual learning for audio-visual segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22349–22360. IEEE, 2025.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in neural information processing systems*, 33:655–666, 2020.

Jing-Xuan Zhang, Genshun Wan, Jianqing Gao, and Zhen-Hua Ling. Audio-visual representation learning via knowledge distillation from speech foundation models. *Pattern Recognition*, 162: 111432, 2025.

Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5404–5413. IEEE, 2023.

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 *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 133(4):1644–1664, 2025.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5116–5125. IEEE, 2023.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Cmad: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636. IEEE, 2025.

A More Related Work **16**

B Pseudocode **17**

C More Experiments **17**

C.1 Dataset Details and Evaluation Metrics 17

C.2 More Experimental Setup 19

C.3 Implementation Details 19

C.4 Task-Wise Efficiency Comparison 20

C.5 Fine-Grained Ablation Study 20

C.6 Qualitative Analyses 22

D Derivations and Theoretical Properties **23**

D.1 Observational and Reference Mixtures 23

D.2 Class-Conditional Centering and Cross-Fitting 23

D.3 Entropy-Regularized State Assignment 24

D.4 Probability-Space Reference Marginalization 24

D.5 Low-Rank Structure of the Conditional Response 24

D.6 Observed-Mixture Anchoring 24

D.7 Decomposition of Reference-Centered Transfer 25

D.8 Sensitivity to the Reference Distribution 25

E Limitations and Broader Impact **25**

A More Related Work

Knowledge Distillation. KD transfers supervision from a high-capacity teacher to a compact student, progressing from softened predictions to intermediate features, attention maps, and relational geometry Buciluǎ et al. (2006); Hinton et al. (2015); Zagoruyko & Komodakis (2016); Tung & Mori (2019); Park et al. (2019); Tian et al. (2019). Transformer compression further shifts transfer toward token-level relations. MTST Choi et al. (2023) preserves token-similarity distributions across unequal representation dimensions, while logit and projection constraints improve transfer under response-scale and architectural mismatch Sun et al. (2024); Miles et al. (2024). Cross-modal KD introduces an additional modality gap when teacher and student observe different streams. Early privileged-modality learning assumes stronger modalities are available only during training Gupta et al. (2016); Garcia et al. (2018). Subsequent methods address this gap through spatial or geometric alignment Yun et al. (2023); Lee et al. (2023); Klingner et al. (2023); Zhou et al. (2023); Zhang et al. (2023), information decomposition Chen et al. (2023); Li et al. (2023), and adaptive transfer Wang et al. (2023); Radevski et al. (2023). C2KD Huo et al. (2024) further links modality imbalance with soft-label misalignment through customized, selective, and progressive transfer. Missing-modality learning extends this direction by balancing robustness, recovering cross-sample relations, modeling incomplete observations, disentangling transferable components, or reweighting modalities Dai et al. (2024); Li et al. (2024); Chen et al. (2024); Ienco & Dantas (2024); Zhuang et al. (2025); Rakib & Bagavathi (2025). Related developments span audio-visual localization and speech Yaghoubi et al. (2023); Kim et al. (2025); Zhang et al. (2025), multi-sensor 3D perception Kim et al. (2024); Liu et al. (2025); Bang et al. (2025), and compact multimodal models Feng et al. (2025); Cao et al. (2025); Chowdhury et al. (2025); Feng et al. (2026).

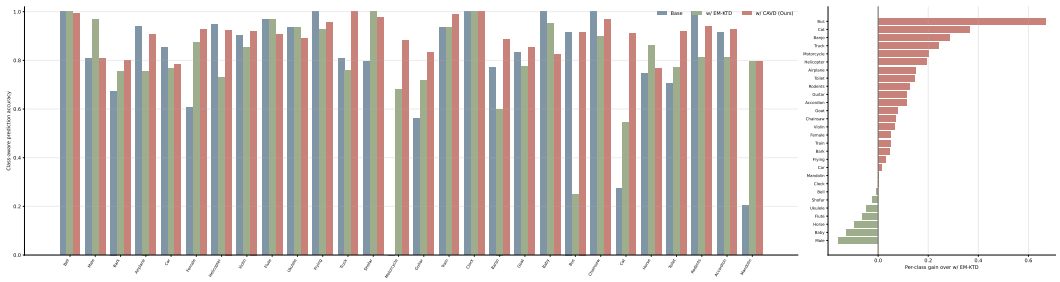


Figure 10: Class-aware AVEL prediction accuracy and per-class gains over the comparison method.

Causal Inference. Pearl's causal ladder organizes learning into association, intervention, and counterfactual reasoning Pearl et al. (2003). At the association level, multimodal methods recover latent causal components or richer video–language dependencies, although recurring context and modality imbalance can still mimic causal structure Morioka & Hyvarinen (2023); Zang et al. (2023). Intervention addresses this ambiguity by controlling selected pathways. Back-door formulations adjust observed visual, contextual, or linguistic confounders Zhang et al. (2020); Tang et al. (2020); Liu et al. (2022), while front-door formulations use attention or intermediate representations as mediators Yang et al. (2021b;a). Recent works extend intervention to modality-specific and cross-modal pathways, including audio-visual question answering Lao et al. (2023), navigation Wang et al. (2024), and generalizable recognition Liang et al. (2026). Other formulations localize bias within vision–language interactions, intervene on fine-grained VQA signals, assign branch-specific adjustments, or estimate causal contributions across tasks and sensors Weng et al. (2024); Liu et al. (2024b); Chen et al. (2025); Li et al. (2025); Wang et al. (2025); Cheng et al. (2025); Wen et al. (2025); Mai & Han (2026); Liu et al. (2026). Counterfactual reasoning further compares observations with alternative states, from removing critical visual or linguistic clues Chen et al. (2020); Niu et al. (2021); Li et al. (2026); Mahapatra et al. (2026) to controlled audio-visual pairs, latent counterfactual variation, and alternative temporal outcomes Singh et al. (2024); Zha et al. (2025); Kung et al. (2025); Gao & Dong (2025); Yi et al. (2026).

B Pseudocode

We summarize the complete three-stage training procedure in Algorithm 1.

C More Experiments

C.1 Dataset Details and Evaluation Metrics

AVE. The AVE dataset Tian et al. (2018) contains 4,143 ten-second videos from 28 event categories, following the standard split of 3,339 training, 402 validation, and 402 testing videos. The task aims to localize temporal segments in which audible and visible events are semantically aligned. We report localization accuracy across different fusion models and audio encoders. In addition to the fully supervised setting used in the main comparison, the appendix includes a weakly supervised setting trained with video-level labels, providing a complementary evaluation when segment-level annotations are unavailable.

LLP. The LLP dataset Tian et al. (2020) contains 11,849 videos from 25 event categories. Following the standard AVVP protocol, 10,000 videos with video-level annotations are used for training, 649 for validation, and 1,200 videos with segment-level annotations for testing. Evaluation covers audio, visual, and audio-visual predictions at both segment and event levels. The main table reports A, V, AV, Type, Event, and their arithmetic mean.

AVSBench. AVSBench Zhou et al. (2022) evaluates pixel-level sounding-object segmentation. The S4 subset contains single-source scenes with 3,452 training, 740 validation, and 740 testing videos, while MS3 contains multi-source scenes with 296 training, 64 validation, and 64 testing videos. We report region similarity $\mathcal{M}_{\mathcal{J}}$ and contour accuracy $\mathcal{M}_{\mathcal{F}}$.

Algorithm 1: Complete CAVD Training Pipeline

Input: Training samples, frozen teacher T, student S, state count M, folds F, occupancy target π_0, and stage-specific epoch counts
Output: Reference-standardized student S

```
// Stage 1: Cross-fitted joint-prior discovery
```

Partition the training samples into F folds, let $\mathcal{F}_b$ contain all sites of samples in fold b, and initialize the joint descriptor projection and $\mathcal{E}^T$
foreach *fold* $b \in \{1, \dots, F\}$ **do**
  Set class weights from labels or stopped teacher probabilities, then compute $\{\boldsymbol{\mu}_{m,c}^{T,(-b)}\}_{m,c}$
for *each state-discovery epoch* **do**
  foreach *mini-batch of training samples* **do**
    Form cross-fitted residuals, relation descriptors, and $\mathbf{u}_s^T$ for every site s in the mini-batch
    for *state* $k = 1, \dots, M$ **do**
      compute $q_{s,k}$ from $\|\mathbf{u}_s^T - \mathbf{e}_k^T\|_2^2$ for those sites
    Update the joint descriptor projection and $\mathcal{E}^T$ with $\mathcal{L}_{\texttt{prior}}$
Recompute $q_{s,k}$ over all training sites, set $\pi_k \leftarrow \mathbb{E}_s[q_{s,k}]$, and freeze $\boldsymbol{\pi}$ and the state bank

```
// Stage 2: Reference-marginalized teacher response
```

Initialize the teacher response mappings $\mathcal{G}^T$ and $\mathcal{U}^T$
for *each teacher-response epoch* **do**
  foreach *mini-batch of training samples and each site* s *in that mini-batch* **do**
    for *state* $k = 1, \dots, M$ **do**
      Compute $\boldsymbol{\eta}_{s,k}^T = \mathcal{G}^T(\mathbf{h}_{f,s}^T) + \mathcal{U}^T(\mathbf{h}_{f,s}^T)\mathbf{e}_k^T$
      if *the task has mutually exclusive outputs* **then**
        set $p_{s,k}^T \leftarrow \mathrm{softmax}(\boldsymbol{\eta}_{s,k}^T)$
      else
        set $p_{s,k}^T \leftarrow \mathrm{sigmoid}(\boldsymbol{\eta}_{s,k}^T)$
    Form $\widetilde{p}_s^T = \sum_k \pi_k p_{s,k}^T$ and update $\mathcal{G}^T, \mathcal{U}^T$ with $\mathcal{L}_{\texttt{total}}^T$

```
// Stage 3: Transfer the reference prediction and centered variation
```

Initialize $\mathcal{P}_K$, $\mathcal{G}^S$, and $\mathcal{U}^S$
for *each student-training epoch* **do**
  foreach *mini-batch of training samples and each site* s *in that mini-batch* **do**
    for *state* $k = 1, \dots, M$ **do**
      set $\mathbf{e}_k^S \leftarrow \mathcal{P}_K(\mathtt{sg}(\mathbf{e}_k^T))$ and compute $p_{s,k}^S$
    Form $\widetilde{p}_s^S$, compute $\mathcal{L}_{\texttt{total}}^S$, and update S, $\mathcal{P}_K$, $\mathcal{G}^S$, $\mathcal{U}^S$

Table 8: Task-wise teacher-student efficiency comparison.

Task	Role	Visual Backbone		Audio Backbone		Overall Efficiency	
		Params (M)	FLOPs (G)	Params (M)	FLOPs (G)	Params (M)	FLOPs (G)
AVEL	Teacher	228.78	69.05	31.33	14.14	260.11	83.19
	Student	28.35 (-87.6%)	9.01 (-87.0%)	5.87 (-81.3%)	7.94 (-43.8%)	34.22 (-86.8%)	16.95 (-79.6%)
AVVP	Teacher	228.78	69.05	31.33	14.14	260.11	83.19
	Student	28.35 (-87.6%)	9.01 (-87.0%)	5.87 (-81.3%)	7.94 (-43.8%)	34.22 (-86.8%)	16.95 (-79.6%)
AVS	Teacher	81.44	23.43	72.14	1.73	153.58	25.16
	Student	3.41 (-95.8%)	1.13 (-95.2%)	0.98 (-98.6%)	0.04 (-97.7%)	4.39 (-97.1%)	1.17 (-95.3%)

Table 9: Pillar-level ablation under lightweight visual and audio backbones.

Causal Pillar			AVEL		AVVP					AVS	
JPD	PTR	PSD	Fully	Weakly	Segment-level		Event-level		Average	S4	MS3
			Accuracy	Accuracy	Type	Event	Type	Event		$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{J}}$
✗	✗	✗	68.18	63.02	50.8	52.2	43.8	41.7	47.1	74.32	54.52
✗	✓	✓	70.50	65.31	52.6	53.0	46.0	45.6	49.3	76.25	55.85
✓	✗	✓	69.63	64.44	52.0	52.4	45.0	44.4	48.5	75.30	55.05
✓	✓	✗	71.04	65.92	53.3	53.7	46.8	46.2	50.0	76.86	56.75
✓	✓	✓	**73.25**	**68.12**	**54.2**	**54.7**	**47.2**	**47.3**	**50.9**	**78.96**	**57.82**

Table 10: Ablation of latent-state construction.

State Construction					AVEL		AVVP					AVS	
Raw Feature	Class Index	Semantic Residual	Cross-Fitting	Audio-Visual Relation	Fully	Weakly	Segment-level		Event-level		Average	S4	MS3
					Accuracy	Accuracy	Type	Event	Type	Event		$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{J}}$
✓	✗	✗	✗	✗	70.50	65.26	52.6	53.0	46.0	45.6	49.3	76.25	55.85
✗	✓	✗	✗	✗	70.68	65.48	52.8	53.0	46.2	45.6	49.4	76.48	56.06
✗	✗	✓	✗	✗	71.69	66.55	53.4	53.8	46.8	46.2	50.1	77.02	56.70
✗	✗	✓	✓	✗	72.45	67.30	53.7	53.9	47.0	46.6	50.3	77.34	57.03
✗	✗	✓	✓	✓	**72.94**	**67.81**	**54.0**	**54.2**	**47.3**	**46.9**	**50.6**	**78.40**	**57.50**

C.2 More Experimental Setup

For the visual backbones, we adopt PVT-v2-B5 and PVT-v2-B0 from PVT-v2 Wang et al. (2022) as the large and lightweight variants, respectively. For audio, we consider the Transformer-based HTS-AT Chen et al. (2022) and SSAST-Tiny Gong et al. (2022), together with the CNN-based VGGish Hershey et al. (2017) and EfficientAT Schmid et al. (2023), representing large and lightweight models within each architectural family. For AVE, we use AVMoE Cheng et al. (2024) and DG-SCT Duan et al. (2023) as teacher models, while Mettle and UFE-AVS Liu et al. (2024a) serve as the teachers for AVVP and AVS, respectively. For reference, we also include representative AVS methods, including TPAVI Zhou et al. (2025), DiffusionAVS Mao et al. (2025), and AVSegFormer Gao et al. (2024). For distillation strategies, we compare CAVD with Vanilla KD Hinton et al. (2015), AT Zagoruyko & Komodakis (2016), SPKD Tung & Mori (2019), MTST Choi et al. (2023), and EM-KTD Park et al. (2026).

C.3 Implementation Details

Model Setup. Each experiment follows the teacher-student backbone pair and task head of the corresponding downstream baseline. Visual resolution, audio sampling, clip length, normalization, and patch configuration are kept consistent with each baseline, while VGGish, EfficientAT, HTS-AT, and SSAST use their standard log-mel inputs. By default, CAVD uses $M = 32$ latent states with a 32-dimensional representation for each state. Within each benchmark, all distillation methods share the same input processing, student architecture, task loss, and evaluation protocol to ensure fair comparison.

Training and Inference Details. The pretrained teacher backbone and original prediction head remain frozen throughout training. Training proceeds in three stages. First, state discovery computes cross-fitted class centers, constructs joint residual descriptors, and learns the joint descriptor projection and state bank. Second, the discovered states and empirical reference weights are frozen while training the conditional teacher module. Finally, the entire teacher side is fixed, and only the student and its conditional response module are optimized. All experiments use PyTorch with AdamW, cosine learning-rate decay, linear warm-up, and mixed precision when supported by the backbone. Base learning rate, weight decay, batch size, training epochs, assignment temperature, and fold count follow fixed benchmark-specific settings. Training is conducted on NVIDIA A100 and RTX 4090 GPUs. Frozen teacher features can be cached, while joint descriptors are recomputed as their projection is optimized. At inference, the student backbone is evaluated once, with

Table 11: Ablation of assignment regularization and stage separation with descriptor reconstruction retained.

State Bank			AVEL		AVVP					AVS	
Sharp Assignment	Balanced Occupancy	Stage Separation	Fully Accuracy	Weakly Accuracy	Segment-level Type	Segment-level Event	Event-level Type	Event-level Event	Average	S4 $\mathcal{M}_{\mathcal{J}}$	MS3 $\mathcal{M}_{\mathcal{J}}$
✗	✗	✓	70.22	65.06	52.2	52.6	44.9	44.3	48.5	75.70	54.85
✓	✗	✓	70.80	65.62	52.8	53.2	45.6	45.2	49.2	76.31	55.63
✗	✓	✓	70.62	65.45	52.7	52.9	45.4	44.8	49.0	76.10	55.39
✓	✓	✗	71.46	66.33	53.5	53.9	46.5	46.1	50.0	76.86	56.64
✓	✓	✓	**72.71**	**67.58**	**53.9**	**54.1**	**47.0**	**46.6**	**50.4**	**77.93**	**57.30**

Table 12: Ablation of teacher anchoring and reference marginalization.

Teacher Configuration				AVEL		AVVP					AVS	
Observed Anchor	Reference Weights	Probability Space	Task Calibration	Fully Accuracy	Weakly Accuracy	Segment-level Type	Segment-level Event	Event-level Type	Event-level Event	Average	S4 $\mathcal{M}_{\mathcal{J}}$	MS3 $\mathcal{M}_{\mathcal{J}}$
✓	✗	✓	✓	71.68	66.54	53.4	53.8	46.6	46.2	50.0	76.85	56.50
✗	✓	✓	✓	71.40	66.27	52.9	53.3	46.2	45.8	49.6	76.55	56.12
✓	✓	✗	✓	71.80	66.66	53.6	54.0	46.8	46.4	50.2	77.18	56.79
✓	✓	✓	✗	71.54	66.40	53.2	53.6	46.4	46.0	49.8	76.82	56.35
✓	✓	✓	✓	**72.76**	**67.62**	**53.8**	**54.2**	**47.2**	**46.8**	**50.5**	**78.35**	**57.35**

all 32 state modifications processed in parallel. The teacher, joint descriptor projection, cross-fitting statistics, and site-specific assignments are excluded from deployment.

C.4 Task-Wise Efficiency Comparison

In Table 8, we report the visual and audio encoder costs separately, together with the total parameter count and its reduction relative to the teacher. Visual FLOPs are measured with a 224×224 input, while audio FLOPs follow the standard spectrogram configuration of each backbone. At inference, the student backbone runs only once, with all M state modifications evaluated in parallel. The teacher, joint descriptor projection, fold statistics, and site-specific assignments are excluded from deployment.

C.5 Fine-Grained Ablation Study

Task-Wise Pillars. In Table 9, removing PTR produces the largest decline since the student again receives an observation-dependent teacher target. Without JPD, the response conditions lack a structured description of recurring multimodal variation. Without PSD, the student loses the transfer of conditional response differences. Combining JPD, PTR, and PSD connects state construction, target standardization, and compact-model transfer.

Joint-Prior State Construction. In Table 10, `Raw Feature` and `Class Index` introduce different limitations: the former follows dominant feature geometry, whereas the latter suppresses variation within each category. `Semantic Residual` shifts the comparison from overall semantic similarity to deviations around comparable task content, allowing recurring contextual conditions to form distinct states. `Cross-Fitting` provides a complementary correction, since residualization determines what information is removed, while cross-fitting controls whether the estimate is influenced by the sample being represented. `Audio-Visual Relation` further distinguishes cases with similar unimodal residuals but different cross-modal correspondence, a distinction unavailable from either modality alone.

State-Bank Optimization. In Table 11, descriptor reconstruction is retained across all settings and provides the geometry on which the state bank is organized. Reconstruction alone, however, can yield overlapping assignments or concentrate many sites in a small subset of states. `Sharp Assignment` and `Balanced Occupancy` address these issues at complementary scales: the former promotes locally distinct memberships, while the latter encourages globally balanced us-

Table 13: Ablation of student weighting and transfer losses.

Student Configuration					AVEL		AVVP					AVS	
Direct KD	Input-Dependent Weights	Reference Weights	Prediction	Variation	Fully Accuracy	Weakly Accuracy	Segment-level Type	Segment-level Event	Event-level Type	Event-level Event	Average	S4 $\mathcal{M}_{\mathcal{J}}$	MS3 $\mathcal{M}_{\mathcal{J}}$
✓	✗	✗	✗	✗	69.09	63.91	50.6	51.0	43.3	42.9	47.0	75.20	54.90
✗	✓	✗	✓	✓	70.86	65.72	52.9	53.3	46.1	45.7	49.5	76.58	56.20
✗	✗	✓	✓	✗	71.55	66.41	53.6	54.0	46.7	46.3	50.2	77.08	56.83
✗	✗	✓	✗	✓	71.09	65.96	53.3	53.7	46.3	45.9	49.8	76.65	56.35
✗	✗	✓	✓	✓	**72.82**	**67.68**	**53.8**	**54.2**	**47.0**	**46.6**	**50.4**	**78.52**	**57.48**



Figure 11: Audio-visual video parsing cases on the LLP dataset.

age. The two are not interchangeable, as confident assignments can still collapse onto a few states, whereas balanced usage can still arise from diffuse memberships. `Stage Separation` further limits reorganization of the bank to accommodate the later transfer objective, preserving a consistent state basis for teacher target formation and student learning.

Teacher Response and Reference Marginalization. In Table 12, `Observed Anchor` and `Reference Weights` constrain complementary aspects of target formation. The anchor ties the conditional response family to an accessible teacher prediction, but does not specify how those responses are aggregated at inference. Reference weighting fixes this aggregation rule, yet does not by itself determine the compatible state-conditional responses. `Probability Space` controls whether averaging preserves the class competition expressed by each conditional prediction and therefore complements, rather than replaces, these two components. `Task Calibration` operates on the combined target to correct aggregate task behavior without forcing all states toward the same output, preserving the state-resolving role of the anchor.

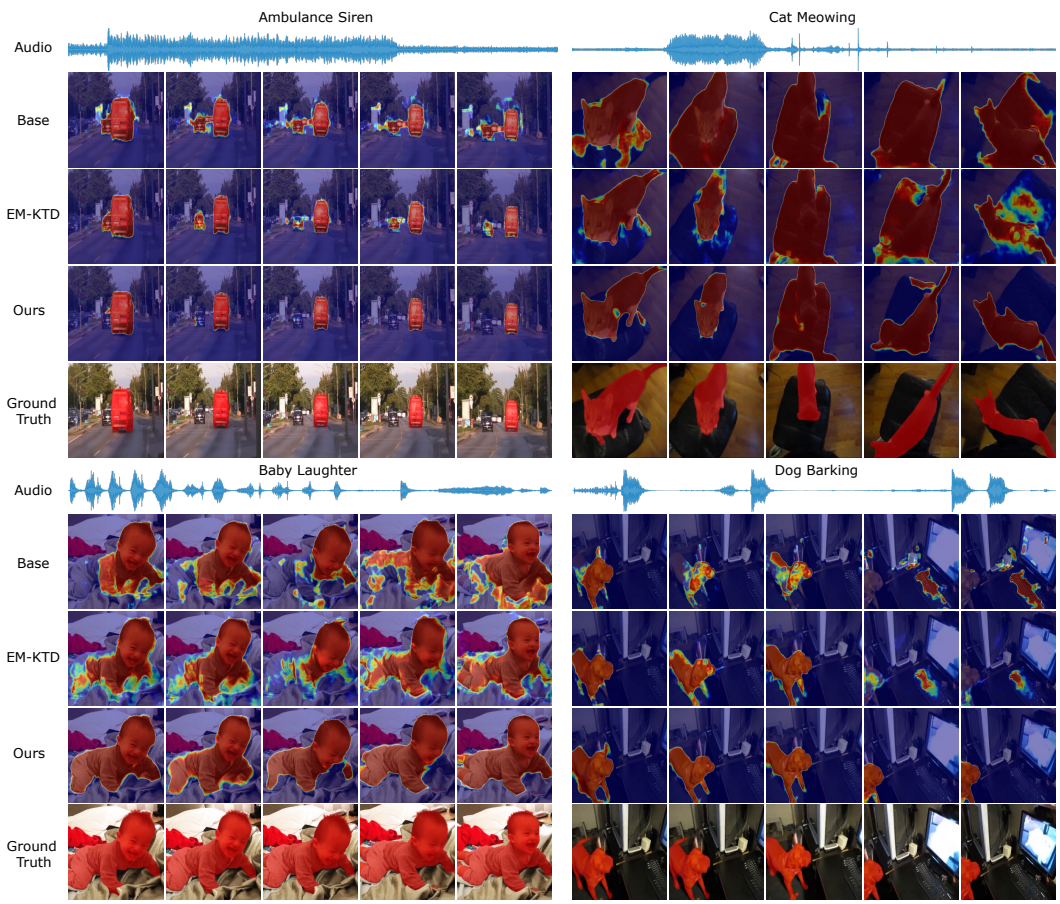


Figure 12: More Sounding-object segmentation cases on the AVSBench.

Student Transfer. In Table 13, `Direct KD` and `Input-Dependent Weights` retain different forms of dependence on the observational teacher, either through its final target or through sample-specific aggregation. `Reference Weights` removes the latter dependence, but matching the weighted prediction alone still leaves the organization of conditional responses underconstrained. `Prediction` fixes their shared weighted center, whereas `Variation` preserves how individual responses deviate from that center. These objectives are complementary: variation alone leaves the common response undetermined, while prediction alone does not retain the teacher's state-dependent structure. Their combination therefore transfers a directly anchored output together with the relative changes that characterize its behavior across represented conditions.

C.6 Qualitative Analyses

In Figure 10, we present the aggregate class-wise distribution of temporal localization results. Improvements span a broad range of event classes rather than being concentrated on a few dominant categories, indicating broader coverage under the standardized target. The nonuniform gains further reflect variations in visual ambiguity, acoustic distinctiveness, and recurring scene associations across classes. Figure 11 complements this distribution-level analysis with two temporal parsing examples. The first involves singing, speech, and guitar events, where accurate localization requires preserving distinct event intervals instead of extending a dominant label across neighboring segments. The second contains Chicken Rooster and Speech with significant object motion and frame-to-frame appearance changes. CAVD maintains more consistent temporal predictions as visual content and position vary, associating decisions with persistent audio-visual correspondence rather than isolated frame similarity. Figure 12 further provides sounding-object segmentation examples under diverse visual conditions. For instance, when a sounding car is surrounded by visually

similar vehicles, CAVD maintains localization on the active source while suppressing nearby distractors across consecutive frames.

D Derivations and Theoretical Properties

D.1 Observational and Reference Mixtures

For model index $\nu \in \{T, S\}$ and a finite latent state space, the law of total expectation expands the observational response as:

$$p^{\nu,\mathtt{obs}}(Y = y \mid \mathbf{X} = \mathbf{x}) = \sum_{k=1}^{M} p(\mathbf{K} = k \mid \mathbf{x}) p^{\nu}(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{K} = k). \tag{16}$$

Reference marginalization replaces only the state weights,

$$p^{\nu,\mathtt{adj}}(Y = y \mid \mathbf{X} = \mathbf{x}) = \sum_{k=1}^{M} \pi_k p^{\nu}(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{K} = k). \tag{17}$$

Subtracting the two responses gives:

$$p^{\nu,\mathtt{obs}}(Y = y \mid \mathbf{x}) - p^{\nu,\mathtt{adj}}(Y = y \mid \mathbf{x}) = \sum_{k=1}^{M} \left[p(\mathbf{K} = k \mid \mathbf{x}) - \pi_k\right] p^{\nu}(Y = y \mid \mathbf{x}, \mathbf{K} = k). \tag{18}$$

The gap requires both input-dependent state weights and state-varying conditional responses. It vanishes when $p(\mathbf{K} \mid \mathbf{x}) = \boldsymbol{\pi}$ or when the conditional response is identical across all supported states. For vector-valued categorical predictions, the triangle inequality and $\|p^{\nu}(\cdot \mid \mathbf{x}, k)\|_1 = 1$ yield,

$$\left\|p^{\nu,\mathtt{obs}}(\cdot \mid \mathbf{x}) - p^{\nu,\mathtt{adj}}(\cdot \mid \mathbf{x})\right\|_1 \leq \left\|p(\mathbf{K} \mid \mathbf{x}) - \boldsymbol{\pi}\right\|_1. \tag{19}$$

For sigmoid outputs, each of the C_τ coordinates lies in $[0, 1]$, giving the same bound with the right side multiplied by C_τ. These bounds describe the effect of replacing one mixture with another. If $\mathbf{K}$ blocks the backdoor paths from $\mathbf{X}$ to Y and positivity holds, the formula gives:

$$p^{\nu}(Y = y \mid do(\mathbf{X} = \mathbf{x})) = \sum_{k=1}^{M} p^{\nu}(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{K} = k) p(\mathbf{K} = k). \tag{20}$$

Eq. 17 equals this adjusted prediction when $\pi_k = p(\mathbf{K} = k)$. When the assignments estimate state membership, the empirical $\boldsymbol{\pi}$ used by CAVD estimates these shared state weights. This interpretation is conditional on the learned states forming an admissible adjustment set, positivity over their support, and invariance of the state-conditional prediction mechanism under the intervention.

D.2 Class-Conditional Centering and Cross-Fitting

The ideal class-semantic center is $\mathbb{E}[\mathbf{h}^T_{m,s} \mid Y_s]$. Subtracting it gives a conditionally centered residual:

$$\mathbb{E}\big[\mathbf{h}^T_{m,s} - \mathbb{E}[\mathbf{h}^T_{m,s} \mid Y_s] \mid Y_s\big] = \mathbf{0}. \tag{21}$$

Thus ideal centering removes the class-conditional first moment rather than every semantic dependency. Higher-order class information can remain. The implemented residual $\boldsymbol{\delta}^T_{m,s}$ in Eq. 5 instead uses estimated centers. Since $\mathcal{F}_b$ contains every site of each held-out sample, $\boldsymbol{\mu}^{T,(-b)}_{m,c}$ excludes the current site and the other sites of its sample. This removes direct self-influence, whereas an in-sample center contains the feature whose residual it estimates. The soft class weights extend centering to multi-label targets.

The separation does not imply lower residual variance or exact conditional centering. If $\boldsymbol{\mu}^{T,(-b)}_{m,c}$ is unbiased for the class mean and its estimation error is conditionally independent of $\mathbf{h}^T_{m,s}$ given $Y_s = c$, then

$$\mathbb{E}[\mathbf{h}^T_{m,s} - \boldsymbol{\mu}^{T,(-b)}_{m,c} \mid Y_s = c] = \mathbf{0}. \tag{22}$$

Under the same conditions, the covariance is

$$\mathrm{Cov}(\mathbf{h}^T_{m,s} - \boldsymbol{\mu}^{T,(-b)}_{m,c} \mid Y_s = c) = \mathrm{Cov}(\mathbf{h}^T_{m,s} \mid Y_s = c) + \mathrm{Cov}(\boldsymbol{\mu}^{T,(-b)}_{m,c} \mid Y_s = c). \tag{23}$$

The second term records center-estimation uncertainty. Soft weights and finite-sample smoothing do not in general satisfy the unbiasedness condition, so Eq. 22 describes an idealized property rather than an exact identity for the implemented residual.

D.3 Entropy-Regularized State Assignment

The assignment in Eq. 6 is the exact solution of a local entropy-regularized problem. For a fixed site, suppress s and define $d_k = \|\mathbf{u}^T - \mathbf{e}_k^T\|_2^2$. Over the probability simplex Δ^{M-1}, consider

$$\mathbf{q}^\star = \arg\min_{\mathbf{q}\in\Delta^{M-1}} \left\{ \sum_{k=1}^{M} q_k d_k + \tau_K \sum_{k=1}^{M} q_k \log q_k \right\}. \tag{24}$$

The stationarity condition under $\sum_k q_k = 1$ gives $d_k + \tau_K(1 + \log q_k) + \xi = 0$, where ξ is the Lagrange multiplier. Normalization then yields $q_k^\star \propto \exp(-d_k/\tau_K)$, which recovers Eq. 6. Thus τ_K controls the smoothness of the distance-induced assignment, while the separate sharpness penalty in Eq. 7 discourages persistently diffuse assignments during state-bank learning. The two regularizers in Eq. 7 constrain different degeneracies. Let $\boldsymbol{\epsilon}_j$ denote vertex j of the probability simplex. If every site selects $\mathbf{q}_s = \boldsymbol{\epsilon}_j$, then $\mathbb{E}_s[\mathcal{H}(\mathbf{q}_s)] = 0$, but $\mathcal{D}_{\mathrm{KL}}(\bar{\mathbf{q}}\|\pi_0) = \mathcal{D}_{\mathrm{KL}}(\boldsymbol{\epsilon}_j\|\pi_0) > 0$ for a nondegenerate occupancy target. Conversely, setting $\mathbf{q}_s = \pi_0$ for every site makes the aggregate divergence zero but retains entropy $\mathcal{H}(\pi_0) > 0$. The sharpness and occupancy terms therefore distinguish global state collapse from uniformly diffuse local assignments. Semantic validity instead depends on the residual and relation construction examined in Table 10.

D.4 Probability-Space Reference Marginalization

For a fixed site, we suppress s and write $p_k^T = \Psi_\tau(\boldsymbol{\eta}_k^T)$ and $\widetilde{p}^T = \sum_k \pi_k p_k^T$. Averaging the scores first would instead produce $\Psi_\tau(\sum_k \pi_k \boldsymbol{\eta}_k^T)$. The two operations coincide for an affine Ψ_τ but not generally for softmax or sigmoid,

$$\sum_{k=1}^{M} \pi_k \Psi_\tau(\boldsymbol{\eta}_k^T) \neq \Psi_\tau\left(\sum_{k=1}^{M} \pi_k \boldsymbol{\eta}_k^T\right). \tag{25}$$

The left side is the mixture distribution specified by the adjustment formula. The right side is the prediction obtained from averaged scores and can change within-state class competition or confidence. Enumerating the finite state bank therefore gives the exact reference expectation required by the model rather than an approximation in score space.

D.5 Low-Rank Structure of the Conditional Response

The teacher response module does not fit an independent task head for every state. For one task site, suppress s and collect the state coordinates as $\mathbf{E}^T = [\mathbf{e}_1^T, \ldots, \mathbf{e}_M^T]$ and the corresponding score vectors as $\mathbf{L}^T = [\boldsymbol{\eta}_1^T, \ldots, \boldsymbol{\eta}_M^T]$. Let $\mathbf{1}_M$ be the all-ones vector. Eq. 9 gives:

$$\mathbf{L}^T = \mathcal{G}^T(\mathbf{h}_f^T)\mathbf{1}_M^\top + \mathcal{U}^T(\mathbf{h}_f^T)\mathbf{E}^T. \tag{26}$$

Subtracting the $\boldsymbol{\pi}$-weighted score center from each column cancels the shared response, giving

$$\mathrm{rank}\big(\mathbf{L}^T - (\mathbf{L}^T\boldsymbol{\pi})\mathbf{1}_M^\top\big) = \mathrm{rank}\big(\mathcal{U}^T(\mathbf{h}_f^T)\left[\mathbf{E}^T - (\mathbf{E}^T\boldsymbol{\pi})\mathbf{1}_M^\top\right]\big) \leq d_K. \tag{27}$$

The same bound applies to the student scores, with the dimension of $\mathbf{e}_k^S$ bounding their state-dependent variation. The state dimension therefore controls both the learned state geometry and the maximum rank of reference-centered score variation. This coupling prevents the conditional modules from becoming collections of unrelated per-state classifiers.

D.6 Observed-Mixture Anchoring

For one site, the observational response is reconstructed under its assignment weights. Categorical KL is jointly convex, and summing the coordinate-wise Bernoulli divergences preserves joint convexity. Hence, either form of $\mathcal{D}_\tau$ used in the main text satisfies,

$$\mathcal{D}_\tau\left(p^{T,\mathrm{obs}}\middle\|\sum_{k=1}^{M} q_k p_k^T\right) \leq \sum_{k=1}^{M} q_k \mathcal{D}_\tau\big(p^{T,\mathrm{obs}}\|p_k^T\big). \tag{28}$$

Thus, the mixture-level objective in Eq. 10 directly fits the response exposed by the observational teacher while imposing a weaker constraint than matching every compatible state to that response. The reverse implication does not hold, leaving the conditional responses free to differ when their assignment-weighted aggregate remains accurate. Assignments that vary across sites and the shared low-rank functions $\mathcal{G}^T$ and $\mathcal{U}^T$ provide additional constraints. States with little aggregate assignment mass remain weakly anchored and receive correspondingly small empirical reference weights.

D.7 Decomposition of Reference-Centered Transfer

For one site, let $\Delta_k = p_k^T - p_k^S$ and $\overline{\Delta} = \sum_k \pi_k \Delta_k = \widetilde{p}^T - \widetilde{p}^S$. The weighted bias-variance identity gives:

$$\sum_{k=1}^{M} \pi_k \|p_k^T - p_k^S\|_2^2 = \|\widetilde{p}^T - \widetilde{p}^S\|_2^2 + \sum_{k=1}^{M} \pi_k \left\|(p_k^T - \widetilde{p}^T) - (p_k^S - \widetilde{p}^S)\right\|_2^2. \tag{29}$$

To verify the identity, write $\Delta_k = (\Delta_k - \overline{\Delta}) + \overline{\Delta}$ and expand the squared norm. The cross term vanishes since $\sum_k \pi_k(\Delta_k - \overline{\Delta}) = \mathbf{0}$. The first term on the right measures disagreement in the reference prediction, while the second measures disagreement in conditional changes around that prediction. Hence the centered component excludes the common response already represented by the reference component. Eq. 14 uses this centered squared discrepancy for response variation and $\mathcal{D}_\tau$ for direct distributional alignment of the inference output.

D.8 Sensitivity to the Reference Distribution

For either the teacher or student, let p_k denote its state-k conditional prediction and $\widetilde{p}(\boldsymbol{\pi}) = \sum_k \pi_k p_k$. Consider another supported reference vector $\boldsymbol{\pi}'$. For categorical predictions, every p_k has unit ℓ_1 norm, therefore we have:

$$\|\widetilde{p}(\boldsymbol{\pi}') - \widetilde{p}(\boldsymbol{\pi})\|_1 \le \sum_{k=1}^{M} |\pi_k' - \pi_k| \|p_k\|_1 = \|\boldsymbol{\pi}' - \boldsymbol{\pi}\|_1. \tag{30}$$

For sigmoid predictions, $\|p_k\|_1 \le C_\tau$ yields an upper bound of $C_\tau \|\boldsymbol{\pi}' - \boldsymbol{\pi}\|_1$. The bounds separate sensitivity to the chosen distribution from the support question. A state-conditional prediction can still be unreliable when that state has little training support. In CAVD, $\boldsymbol{\pi}$ is the empirical average of the assignments in Eq. 6, as specified after Eq. 7. The perturbation bounds above apply directly to any estimation error in this vector without treating correlated temporal or spatial sites as independent observations.

E Limitations and Broader Impact

Limitations. CAVD models recurring multimodal conditions through latent states, so its effectiveness depends on the coverage of the training data. Rare conditions or substantial distribution shifts may therefore be insufficiently represented.

Broader Impact. Efficient causal audio-visual distillation can further reduce the computational and memory demands of multimodal large language models (MLLMs), particularly audio-visual MLLMs, supporting deployment on resource-constrained devices and broader multimodal applications. At the same time, compact students may inherit systematic dependencies from the teacher or training distribution. Practical deployment therefore benefits from evaluation across diverse data distributions, acoustic conditions, visual environments, and cross-modal configurations, especially under domain shift and modality corruption