EquiAV: Leveraging Equivariance for Audio-Visual Contrastive Learning

Jongsuk Kim [*1] **Hyeongkeun Lee** [*1] **Kyeongha Rho** [*1] **Junmo Kim** [1] **Joon Son Chung** [1]

Abstract

Recent advancements in self-supervised audio-visual representation learning have demonstrated its potential to capture rich and comprehensive representations. However, despite the advantages of data augmentation verified in many learning methods, audio-visual learning has struggled to fully harness these benefits, as augmentations can easily disrupt the correspondence between input pairs. To address this limitation, we introduce EquiAV, a novel framework that leverages equivariance for audio-visual contrastive learning. Our approach begins with extending equivariance to audio-visual learning, facilitated by a shared attention-based transformation predictor. It enables the aggregation of features from diverse augmentations into a representative embedding, providing robust supervision. Notably, this is achieved with minimal computational overhead. Extensive ablation studies and qualitative results verify the effectiveness of our method. EquiAV outperforms previous works across various audio-visual benchmarks. The code is available on https://github.com/JongSuk1/EquiAV

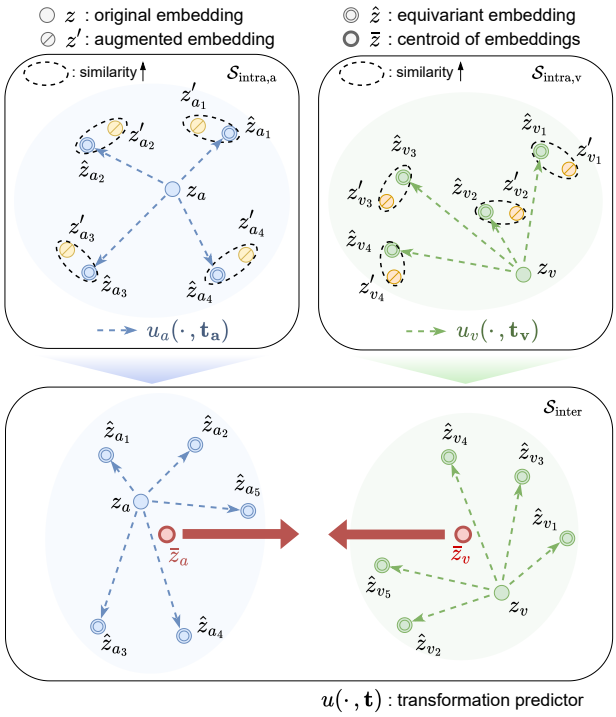


Figure 1. Conceptual illustration of EquiAV. Within the intra-modal latent space, the model learns augmentation-related information by leveraging equivariance. Extending equivariance to the inter-modal space provides robust cross-modal supervision.

1. Introduction

Audio and visual modalities play a pivotal role in how humans perceive their surroundings. Despite the differences in their characteristics, there exists an inherent correspondence between the two modalities. Learning such audio-visual correspondence from large-scale unlabeled video data in a self-supervised manner has recently become a major interest in the deep-learning research community. Among various approaches for audio-visual self-supervised learning, *Audio-Visual Contrastive Learning* has been popular due to its simplicity and effectiveness (Ma et al., 2021a;b; Morgado et al., 2021a;b; Patrick et al., 2021; Recasens et al., 2021; Wang et al., 2021; Sarkar & Etemad, 2023b). A key challenge in audio-visual contrastive learning is to enhance representational capability and diversity while maintaining the correspondence between two different modalities. The most straightforward and effective way to enrich the representation is to utilize data augmentation. However, data augmentation has not been extensively applied in the field of audio-visual contrastive learning. This is because correspondence between audio and visual modalities can be easily distorted due to data augmentations. For this reason, previous works utilize a very limited range of data augmentation and adopt alternatives to learn richer representations, such as incorporating masked data modeling (Gong et al., 2023; Huang et al., 2023) or utilizing temporal supervision (Sarkar & Etemad, 2023b; Jenni et al., 2023).

Meanwhile, there have been several recent studies (Dangovski et al., 2022; Devillers & Lefort, 2023; Garrido et al.,

[*]Equal contribution [1]Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. Correspondence to: Jongsuk Kim <jskpop@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

2023) on learning single-modal representations by employing equivariance, which can be complementary to learning representations that are invariant to data augmentations. Equivariant latent space learns to capture augmentation-related information, thereby enhancing the representational capability. In Particular, EquiMod (Devillers & Lefort, 2023) proposes the transformation predictor to estimate the displacement in the latent space that corresponds to the transformation applied to the input space.

In this paper, we analyze the impact of applying equivariance to the learning audio-visual correspondence and joint representations and then discuss the optimal way to incorporate it into the self-supervised audio-visual contrastive learning framework. We demonstrate that applying the equivariance in the intra-modal latent space is helpful for learning rich modality-unique information, thereby enhancing representational capability. However, it still cannot directly resolve the negative effects of distorted audio-visual correspondence caused by augmentations.

To address this issue, we propose EquiAV, a novel framework that efficiently leverages equivariance for audio-visual contrastive learning. Firstly, we discover that equivariant representation learned in the intra-modal latent space can be transferred to the inter-modal latent space via a shared transformation predictor. Taking advantage of this characteristic, we compute the centroid of a set of equivariant embeddings generated by the transformation predictor for each modality and use it for inter-modal contrastive learning. We verify that supervision from centroids improves representational capability while reducing the undesirable effect of augmentations. This approach requires minimal additional computational cost since the equivariant embeddings are produced from the single original inputs instead of encoding all augmented inputs. Besides, we devise the attention-based transformation predictor that accurately encodes the parameterized augmentation vector to the latent space. We demonstrate the effectiveness of each component of our framework through extensive ablation studies. EquiAV outperforms the existing state-of-the-art audio-visual self-supervised pretraining methods in diverse downstream tasks, including audio-visual event classification and zero-shot audio-visual retrieval tasks.

The contribution of our paper is summarized as follows:

- We propose EquiAV, a novel framework that incorporates equivariance to audio-visual contrastive learning.
- We effectively transfer equivariant representations from the intra-modal latent space to the inter-modal latent space with a shared transformation predictor. Using the centroid of equivariant embeddings enables the model to learn rich joint representations while avoiding the adverse effect of augmentations.
- We design the attention-based transformation predictor to encode the parameterized augmentation vector into the latent space.
- EquiAV outperforms existing audio-visual self-supervised learning methods in audio-visual event classification and zero-shot audio-visual retrieval tasks.

2. Related Works

Audio-Visual Representation Learning. Audio-visual contrastive learning has been one of the most popular approaches for learning the natural correspondence between audio and visual modalities, due to its simple intuition and powerful performance on the downstream tasks (Ma et al., 2021a;b; Morgado et al., 2021a;b; Patrick et al., 2021; Recasens et al., 2021; Wang et al., 2021; Owens & Efros, 2018; Sarkar & Etemad, 2023b). A common approach to audio-visual contrastive learning is to learn the context of the synchronous relationship between audio and visual inputs (Korbar et al., 2016; Alwassel et al., 2020a; Morgado et al., 2021b; Sarkar & Etemad, 2023b). On the other hand, several works (Georgescu et al., 2023; Haliassos et al., 2022) adopt masked modeling techniques that are designed to reconstruct the masked raw inputs or predict the masked context features. Recently, CAV-MAE (Gong et al., 2023) and MAViL (Huang et al., 2023) have combined contrastive learning and masked data modeling techniques to learn complementary representations.

Self-supervised Equivariant Representation Learning. Self-supervised learning (Chen et al., 2020b; He et al., 2020; Oord et al., 2018; He et al., 2022; Tong et al., 2022; Huang et al., 2022; Caron et al., 2021; Chen & He, 2021; Grill et al., 2020) with large-scale datasets has demonstrated promising performance across various domains. Many of these methods employ a joint-embedding framework, where the objective is to maximize the similarity between the embeddings of two augmented views derived from a single input. They focus on learning representations that are invariant to such augmentations. Recently, several studies (Lee et al., 2021; Dangovski et al., 2022; Devillers & Lefort, 2023; Garrido et al., 2023) have illustrated that models can achieve better representation learning by incorporating the principle of equivariance. Equivariance ensures that the semantic contents of representations adapt in response to the input data augmentation, thereby effectively capturing augmentation-related information within the representations. One approach to learning equivariance is to use the auxiliary task of predicting the specific augmentations applied to the input data (Dangovski et al., 2022; Lee et al., 2021). Another strategy is to model the mapping between the transformation in the latent space and the augmentations in the input space (Devillers & Lefort, 2023; Garrido et al., 2023).

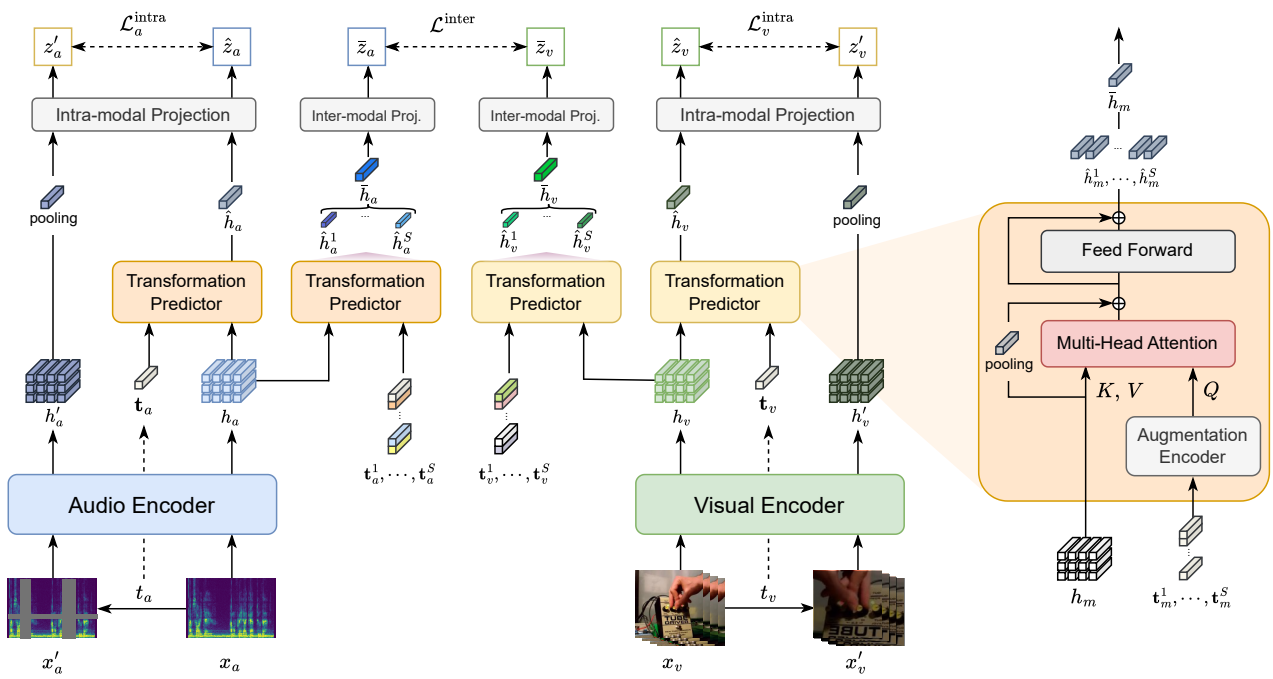


Figure 2. Overview of the proposed EquiAV framework. Given an audio-visual input pair and its augmented version, the audio encoder and the visual encoder encode them into representations. The transformation predictor takes the original representation h_m and the parameterized augmentation vector t_m as inputs and outputs the equivariant representation $\hat{h}_m$. In the intra-modal latent space, the model is learned to maximize the similarity of the equivariant embedding and the augmented embedding. In the inter-modal latent space, we sample multiple augmentation vectors $\{t_{m_i}\}_{i\in\{1,\ldots,S\}}$ and generate the corresponding equivariant representations $\{\hat{h}_{m_i}\}_{i\in\{1,\ldots,S\}}$. Then, the centroid $\bar{h}_m$ is computed and used for inter-modal contrastive learning.

3. Methods

3.1. Preliminary

One effective approach for providing self-supervision is to utilize the concept of *invariance*. It targets to align the representations of augmented inputs in the feature space. Given the encoder f and the augmentation distribution $\mathcal{T}$, the objective function can be described as

$$\forall t_1, t_2 \in \mathcal{T} \quad \min_f \mathcal{L}(f(t_1(x), f(t_2(x))), \tag{1}$$

where the function $\mathcal{L}(\cdot,\cdot)$ measures the dissimilarity between two inputs. On the other hand, some recent works (Dangovski et al., 2022; Devillers & Lefort, 2023; Garrido et al., 2023) have suggested that considering the discrepancy of input pairs in the feature space can lead to better representation learning. This concept, known as *equivariance*, aims to capture the variations between the original input and its augmented version in a way that preserves the underlying structure. To implement equivariance, a transformation predictor u along with augmentation t is employed to match the intra-modal pair in the feature space. The objective function of equivariant self-supervised learning can be represented as follows:

$$\forall t \in \mathcal{T}, \quad \min_{u,f} \mathcal{L}(u(f(x), t), f(t(x))). \tag{2}$$

Through this approach, the model can learn to encode not only the invariant features but also the specific transformations applied to input data.

3.2. EquiAV

Data Augmentations and Their Parameterization. In our framework, typical visual-based augmentations are employed for visual inputs. However, unlike the abundant augmentations available for an image or video, the options for the audio modality are relatively limited. To address this issue, we convert the audio signal into a spectrogram, followed by the application of visual-based augmentations in the same manner. The augmentation information is encoded into real vectors, denoted as $\mathbf{t}_a$ and $\mathbf{t}_v$. These augmentation vectors parameterize how much each augmentation is applied to the input data. A detailed explanation of data augmentation parameterization is provided in Appendix C.

Encoding Audio-Visual Representations. Given an audio-visual input pair (x_a, x_v) and its augmented version

(x'_a, x'_v), the audio encoder f_a and the visual encoder f_v encode them into representations. The encoding process can be represented as follows:

$$h_m = f_m(x_m), \quad h'_m = f_m(x'_m) = f_m(t_m(x_m)) \tag{3}$$

where $m \in \{a, v\}$ and t_m refers to the data augmentation applied to the input data. Note that a and v represent audio and visual modalities, respectively. We employ Vision Transformer (Dosovitskiy et al., 2021) for both encoders.

Equivariant Transformation. To predict the displacement in the latent space caused by the transformation in the input data space, a transformation predictor $u(\cdot)$ is required. In this paper, we design the architecture of $u(\cdot)$ based on the attention mechanism. First of all, we define an augmentation vector $\mathbf{t} \in \mathbb{R}^{S \times d_t}$ where S and d_t denote the number and the dimension of data augmentations t, respectively. The augmentation encoder f_t encodes the augmentation vector into the feature so that it can serve as the query of the attention layer. The representation h of the original input serves as the key and value of the attention layer to obtain the equivariant representation $\hat{h}$ as follows:

$$\begin{aligned} \hat{h} &= u(h, \mathbf{t}) \\ &= \mathrm{FFN}(\mathrm{MHA}(f_t(\mathbf{t}), h, h) + \mathrm{MeanPool}(h)), \end{aligned} \tag{4}$$

$$\begin{aligned} &\mathrm{MHA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathrm{Concat}\left(\mathbf{o}_1, \ldots, \mathbf{o}_H\right) w^O, \\ &\text{where } \mathbf{o}_j = \mathrm{Softmax}\left(\frac{\mathbf{q} w_j^Q \left(\mathbf{k} w_j^K\right)^\top}{\sqrt{d_H}}\right) \mathbf{v} w_j^V, \end{aligned} \tag{5}$$

where H is the number of heads, d_H is dimension of each attention head, $w_h^Q, w_h^K, w_h^V \in \mathbb{R}^{d \times d_H}$, $w^O \in \mathbb{R}^{H d_H \times d}$, and $j \in \{1, ..., H\}$. The Multi-Head Attention (MHA) layer calculates the score through query and key to identify the relevance of the augmentation feature in a patch-wise manner. These values are incorporated into the value to determine the displacement within the feature space. Then, the output of the multi-head attention layer is added to the representation of the original input image and passed through the Feed Forward Network (FFN).

Intra-modal Equivariant Learning. The goal of intra-modal representation learning is to learn the equivariant latent space so that augmentation-related information can be passed to inter-modal representation learning. To this end, the equivariant representation $\hat{h}_m$ and representation h'_m of augmented input are projected into the intra-modal latent space to get the equivariant embedding $\hat{z}_m$ and the augmented embedding z'_m as follows:

$$\hat{z}_m = g_m^{\text{intra}}(\hat{h}_m) = g_m^{\text{intra}}(u_m(h_m, \mathbf{t_m})) \tag{6}$$

$$z'_m = g_m^{\text{intra}}(\mathrm{MeanPool}(h'_m)), \tag{7}$$

where g_m^{intra} denotes the intra-modal projection head.

For intra-modal equivariant contrastive learning, the batch consisting of N equivariant embeddings $\{\hat{z}_m^i\}_{i \in \{1,\ldots,N\}}$ obtained by incorporating the augmentation information and original inputs, as well as N embeddings $\{z'^i_m\}_{i \in \{1,\ldots,N\}}$ of the augmented inputs is used. The pair $(\hat{z}_m^i, z'^i_m)$ generated from the same image forms a positive pair, while the remaining $2(N-1)$ embeddings within the batch serve as negative pairs. The set of embeddings for the negative pairs can be represented as follows:

$$Z_m^{i^-} = \left\{ \bigcup_{j=1}^{N} \{\hat{z}_m^j\} \setminus \{\hat{z}_m^i\} \right\} \cup \left\{ \bigcup_{j=1}^{N} \{z'^j_m\} \setminus \{z'^i_m\} \right\}. \tag{8}$$

The NT-Xent loss (Chen et al., 2020b) is employed to compute the contrastive loss. The loss for equivariant embeddings of the augmented input can be expressed as follows:

$$\ell^{\text{intra}}(\hat{z}_m, z'_m) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\hat{z}_m^i, z'^i_m)/\tau)}{\sum_{z \in Z_m^{i^-} \cup \{z'^i_m\}} \exp(\mathrm{sim}(\hat{z}_m^i, z/\tau)}, \tag{9}$$

where $\tau > 0$ is temperature. Then, the intra-modal loss is described by combining the losses of adopting the original and augmented inputs as anchors respectively,

$$\mathcal{L}^{\text{intra}} = \frac{1}{2}(\ell^{\text{intra}}(\hat{z}_m, z'_m) + \ell^{\text{intra}}(z'_m, \hat{z}_m)). \tag{10}$$

The process of aligning these two embeddings not only enhances the representation capability but also benefits in acquiring a representative vector for inter-modal correspondence learning.

Inter-modal Contrastive Learning. The most important factor in learning inter-modal correspondences is how to draw representative vectors from input data. A basic approach is to use the embedding of the original data as the representative value. It has the advantage of maximizing the correspondence information of the given data but suffers from the lack of input diversity. On the other hand, using the embedding of augmented input strengthens the representation capability, but may disrupt the correspondence of the given pair. Then, we can think of the expectation value of h' over the set D_h of representations of augmented inputs as follows:

$$h^{\text{rep}} = \mathbb{E}_{h' \sim D_h}\left[h'\right]. \tag{11}$$

The vector h^{rep} integrates information from multiple augmented data, securing more general features while also providing robustness against undesirable effects of augmentation. However, it is not feasible to generate all transformations and obtain the representative vector h^{rep}, we aim to approximate it with the centroid of sampled equivariant

representations.

$$h_m^{\text{rep}} \approx \bar{h}_m = \frac{1}{S}\sum_{i=1}^{S} \hat{h}_m^i = \frac{1}{S}\sum_{i=1}^{S} u_m(h_m^i, \mathbf{t}_m^i). \tag{12}$$

Note that the weight of u_m is shared between the intra-modal transformation predictor and the inter-modal transformation predictor. This allows the transfer of equivariance learned in the intra-modal space to the inter-modal space. Furthermore, since we use equivariant representations instead of augmented representations generated by passing through the encoder, there is minimal impact on computational cost. Lastly, the centroid is projected into the inter-modal latent space via the projection head g_a^{inter} and g_v^{inter} as follows:

$$\bar{z}_m = g_m^{\text{inter}}(\bar{h}_m). \tag{13}$$

In the training phase, the batch consisting of N paired embeddings $\{(\bar{z}_a^i, \bar{z}_v^i)\}_{i\in\{1,\dots,N\}}$, which are extracted without applying any augmentation to preserve the correspondence between input pairs, is used. Then, the contrastive loss for audio-visual learning can be expressed as follows:

$$\ell^{\text{inter}}(\bar{z}_a, \bar{z}_v) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\bar{z}_a^i, \bar{z}_v^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\bar{z}_a^i, \bar{z}_v^j)/\tau)}, \tag{14}$$

where $\tau > 0$ is temperature. Then, the inter-modal loss is represented by combining the losses of adopting the audio and visual embeddings as anchors respectively,

$$\mathcal{L}^{\text{inter}} = \frac{1}{2}(\ell^{\text{inter}}(\bar{z}_a, \bar{z}_v) + \ell^{\text{inter}}(\bar{z}_v, \bar{z}_a)). \tag{15}$$

Finally, the loss function of EquiAV can be expressed by incorporating scaling factors, it can be formulated as follows:

$$\mathcal{L}_{\text{EquiAV}} = \lambda^{\text{inter}}\mathcal{L}^{\text{inter}} + \lambda_a^{\text{intra}}\mathcal{L}_a^{\text{intra}} + \lambda_v^{\text{intra}}\mathcal{L}_v^{\text{intra}}, \tag{16}$$

where $\mathcal{L}_a^{\text{intra}}$ and $\mathcal{L}_v^{\text{intra}}$ denote intra-modal loss for audio and visual modalities, respectively.

Through these methods, EquiAV can maximize the benefits of equivariant representation learning while avoiding any detrimental impact on the learning of inter-modal pairs. In the next section, we will provide experimental evidence supporting our design choices along with benchmark performance. Algorithm 1 summarizes our method.

Algorithm 1 EquiAV

Input: backbone encoder f_a, f_v,
intra-modal projection head g_a^{intra}, g_v^{intra},
inter-modal projection head g_a^{inter}, g_v^{inter},
augmentation predictor u_a, u_v,
audio, visual augmentation distribution $\mathcal{T}_a$, $\mathcal{T}_v$
weight scaling factors λ^{inter}, λ_a^{inter}, λ_v^{inter}
batch size N, temperature τ
for sampled mini-batch $\{(x_a^k, x_v^k)\}_{k=1}^N$ **do**
  for $m \in \{a, v\}$ **do**
    for all $k \in \{1, \dots, N\}$ **do**
      sample augmentation vectors $\mathbf{t}_m^{\text{intra}} \sim \mathcal{T}_m$,
      $h_m^k = f_m(x_m^k), \quad h'^k_m = f_m(t_m^{\text{intra}}(x_m^k))$
      $\hat{h}_m^k = u_m(h_m^k, \mathbf{t}_m^{\text{intra}})$
      $z_m^{2k-1} = g_m^{\text{intra}}(\hat{h}_m^k)$
      $z_m^{2k} = g_m^{\text{intra}}(\text{MeanPool}(h'^k_m))$
      sample augmentation vectors $\{\mathbf{t}_{m_i}^{\text{inter}}\}_{i\in\{1,\dots,S\}} \sim \mathcal{T}_m$,
      $\mathbf{t}_m^{\text{inter}} = \left[t_{m_1}^\top, t_{m_2}^\top, \cdots, t_{m_S}^\top\right]^\top$
      $\bar{h}_m^k = \text{MeanPool}(u_m(h_m^k, \mathbf{t}_m^{\text{inter}}))$
      $\bar{z}_m^k = g_m^{\text{inter}}(\bar{h}_m^k)$
    end for
    for all $i \in \{1, \dots, 2N\}$ and $j \in \{1, \dots, 2N\}$ **do**
      $s_m^{i,j} = \exp\left(z_m^{i\top} z_m^j / (\tau\|z_m^i\|\|z_m^j\|)\right)$
    end for
    define $\ell_m^{\text{intra}}(i,j)$ **as**
    $\ell_m^{\text{intra}}(i,j) = -\log\left(\frac{s_m^{i,j}}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}s_m^{i,k}}\right)$
    $\mathcal{L}_m^{\text{intra}} = \frac{1}{2N}\sum_{k=1}^{N}\left[\ell_m^{\text{intra}}(2k-1, 2k) + \ell_m^{\text{intra}}(2k, 2k-1)\right]$
  end for
  for all $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, N\}$ **do**
    $s^{i,j} = \exp\left(\bar{z}_a^{i\top}\bar{z}_v^j / (\tau\|\bar{z}_a^i\|\|\bar{z}_v^j\|)\right)$
  end for
  define $\ell^{\text{inter}}(i)$ **as**
  $\ell^{\text{inter}}(i) = -\frac{1}{2}\left(\log\left(\frac{s^{i,i}}{\sum_{k=1}^{N}s^{i,k}}\right) + \log\left(\frac{s^{i,i}}{\sum_{k=1}^{N}s^{k,i}}\right)\right)$
  $\mathcal{L}^{\text{inter}} = \frac{1}{N}\sum_{i=1}^{N}\ell^{\text{inter}}(i)$
  $\mathcal{L} = \lambda^{\text{inter}}\mathcal{L}^{\text{inter}} + \lambda_a^{\text{intra}}\mathcal{L}_a^{\text{intra}} + \lambda_v^{\text{intra}}\mathcal{L}_v^{\text{intra}}$
  update networks to minimize $\mathcal{L}$
end for

4. Experiments

The model is pre-trained on the AudioSet-2M in a self-supervised manner, without the use of labels. The audio encoder f_a and visual encoder f_v are initialized with the self-supervised pre-trained ViT-B/16 model of MAE (He et al., 2022). We employ 3-layer MLP with layer normalization for intra-modal and inter-modal projection heads. Note that the multi-head attention layer and the feed-forward layer of the transformation predictor are similar to those used in Transformers (Vaswani et al., 2017; Chen et al., 2021), which include layer normalization layer and residual connection. We set λ^{inter}, λ_a^{inter} and λ_v^{inter} all the same as 1. More detailed experimental settings are explained in Appendix C and Appendix D.

4.1. Main Results

Audio-Visual Event Classification. We evaluate the representational capability of our model through fine-tuning on multi-modal benchmark datasets, AudioSet and VGGSound, in three settings: audio-only, visual-only, and audio-visual. During the fine-tuning process, intra-modal and inter-modal

Table 1. Audio-visual event classification performance on AudioSet and VGGSound. **A**: Audio-only, **V**: Visual-only, **A-V**: Audio-visual. IN SL: ImageNet supervised learning, SSL: Self-supervised learning, * Non-standard train/test split.

Method	Pretrain	AudioSet-20K (mAP)			AudioSet-2M (mAP)			VGGSound (acc.)		
		A	V	A-V	A	V	A-V	A	V	A-V
GBlend (Wang et al., 2020)	-	29.1	22.1	37.8	32.4	18.8	41.8	-	-	-
Perceiver (Jaegle et al., 2021)	-	-	-	-	38.4	25.8	44.2	-	-	-
Attn AV (Fayek & Kumar, 2021)	IN SL	-	-	-	38.4	25.7	46.2	-	-	-
MBT* (Nagrani et al., 2021)	IN21K SL	31.3	**27.7**	43.9	41.5	**31.3**	49.6	52.3	**51.2**	64.1
CAV-MAE (Gong et al., 2023)	SSL	37.7	19.8	42.0	46.6	26.2	51.2	59.5	47.0	65.5
AudiovisualMAE* (Georgescu et al., 2023)	SSL	-	-	-	46.6	31.1	51.8	57.2	50.3	65.0
MAViL (Huang et al., 2023)	SSL	41.8	24.8	44.9	48.7	30.3	53.3	60.8	50.9	**67.1**
EquiAV (ours)	SSL	**42.4**	25.7	**46.6**	**49.1**	30.1	**54.6**	**61.0**	50.7	**67.1**

Table 2. Audio classification accuracy of finetuning on ESC-50, SPC-v1, and FSD-50K. †Results reproduced on our environment.[2]

Method	ESC-50	SPC-v1	FSD-50K
CAV-MAE-Scale++† (Gong et al., 2023)	83.2	97.1	45.5
MAViL (Huang et al., 2023)	94.4	97.4	-
XKD (Sarkar & Etemad, 2024)	**96.5**	-	58.5
EquiAV (ours)	96.0	**97.8**	**62.6**

Table 3. Action recognition accuracy of transfer learning on UCF101, HMDB51, and Kinetics400. *High temporal resolution with industry-level computational resources.

Method	UCF101	HMDB51	Kinetics400
XDC (Alwassel et al., 2020b)	85.3	56.0	-
MMV (Alayrac et al., 2020)	83.9	60.0	-
BraVe* (Recasens et al., 2021)	**90.0**	63.6	-
AVID (Morgado et al., 2021c)	-	-	48.9
CrissCross (Sarkar & Etemad, 2023a)	87.7	56.2	50.1
XKD (Sarkar & Etemad, 2024)	88.4	62.2	56.5
EquiAV (ours)	89.7	**64.4**	**57.3**

Table 4. Zero-shot audio-visual retrieval results on the MSR-VTT, AudioSet, and VGGSound. †Results reproduced on our environment. *Models pre-trained on the much larger dataset.

Method	Video-to-Audio			Audio-to-Video		
	R@1	R@5	R@10	R@1	R@5	R@10
MSR-VTT						
Boggust*	9.3	20.7	28.8	7.6	21.1	28.3
Aranjelovic*	11.9	25.9	34.7	12.6	26.3	33.7
AVLnet*	**17.2**	26.6	**46.6**	**17.8**	**35.5**	**43.6**
CAV-MAE	13.3	29.0	40.5	7.6	19.8	30.2
EquiAV (ours)	13.8	**31.4**	43.0	14.4	33.3	43.0
AudioSet						
CAV-MAE-Scale++†	16.6	37.0	45.9	14.3	32.0	40.7
CAV-MAE	18.8	39.5	50.1	15.1	34.0	43.0
EquiAV (ours)	**30.1**	**53.3**	**62.9**	**29.6**	**53.7**	**63.1**
VGGSound						
CAV-MAE-Scale++†	15.5	35.3	45.1	16.4	35.0	44.7
CAV-MAE	14.8	34.2	44.0	12.8	30.4	40.3
EquiAV (ours)	**28.5**	**51.5**	**61.7**	**28.4**	**51.8**	**62.3**

projection heads are removed and a linear classifier is added on top of encoders. For audio-visual fine-tuning tasks, outputs of modality-specific encoders are concatenated to form the joint representation and used as the input for the linear classifier. As shown in Table 1, EquiAV outperforms previous methods by solely relying on contrastive learning.

Comparison on the Single-modal Benchmarks. To demonstrate the performance of our framework on the single-modal downstream tasks, we fine-tune the audio branch of EquiAV on audio classification datasets, including the ESC-50 (Piczak, 2015), SPC-v1 (Warden, 2018), and FSD-50K (Fonseca et al., 2021). According to the results reported in Table 2, our method shows superior performance compared to other self-supervised learning-based methods. We also evaluate the video branch of EquiAV on various action recognition datasets, including UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), and Kinetics400 (Kay et al., 2017). As demonstrated in Table 3, our method surpasses previous SOTA audio-visual SSL methods on uni-modal visual benchmark datasets. The comparison is based on results from other audio-visual SSL methods that also utilize the same pre-training dataset, AudioSet-2M.

Zero-Shot Audio-Visual Retrieval. Table 4 reports the performance of zero-shot retrieval on audio-visual datasets, based on the similarity between the audio-visual embeddings. The sample lists used in the experiments for AudioSet and VGGSound are the same as those used in CAV-MAE (Gong et al., 2023). Additionally, to demonstrate the robustness of our model across different datasets, we perform further zero-shot retrieval experiments on MSR-VTT. Our model consistently outperforms existing methods pre-trained on AudioSet and achieves comparable results to models pre-trained on HowTo100M, such as Boggust (Boggust et al., 2019), Aranjelovic (Arandjelovic & Zisserman, 2017), and AVLnet (Noroozi & Favaro, 2016).

[2]Weights from https://github.com/YuanGongND/cav-mae

Table 5. Zero-shot retrieval results on AudioSet and audio-visual event classification performance on AudioSet-20K with the variants of pre-training methods. **V2A**: Video-to-audio zero-shot retrieval R@1, **A2V**: Audio-to-video zero-shot retrieval R@1. **Inv.**: Invariant learning, **Equi.**: Equivariant learning, (z_a, z_v): Original embeddings, (z'_a, z'_v): Augmented embeddings, $(\hat{z}_a, \hat{z}_v)$: Equivariant embeddings, $(\bar{z}_a, \bar{z}_v)$: Centroids of equivariant representations.

		ZS Retrieval		Fine-Tuning (mAP)		
Intra.	Inter.	V2A	A2V	A	V	A-V
Inv.	(z_a, z_v)	25.2	24.3	40.1	22.9	42.3
Equi.	(z_a, z_v)	28.7	28.5	41.9	24.8	44.9
Equi.	(z'_a, z'_v)	28.1	27.5	41.5	24.2	43.9
Equi.	$(\hat{z}_a, \hat{z}_v)$	27.8	27.4	41.2	23.6	43.4
Equi.	$(\bar{z}_a, \bar{z}_v)$	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

4.2. Ablation Studies

This section presents ablation studies aimed at verifying the effectiveness of our framework. Firstly, we demonstrate the impact of employing equivariance and centroids of audio-visual embeddings produced by the transformation predictor in audio-visual contrastive learning. Next, we explore different architectures of the transformation predictor and training strategy. Lastly, we analyze the impact of the employed intra-modal equivariant loss function.

Impact of Single-modal Equivariance. As presented in Table 5, employing equivariance in the intra-modal latent space results in better zero-shot retrieval and fine-tuning performances compared to the invariance-based method. This underlines the role of intra-modal equivariance in capturing the modality-unique information, which in turn facilitates the learning of audio-visual correspondence and joint representation. When using the augmented inputs for inter-modal contrastive learning, there is a slight decrease in both zero-shot retrieval and fine-tuning performances. The result suggests the augmentation-related information learned through intra-modal equivariance cannot entirely compensate for the distortion of audio-visual correspondence caused by the data augmentations.

On the other hand, utilizing the centroids of the equivariant embeddings in the inter-modal latent space shows both enhanced zero-shot retrieval and fine-tuning performances. This provides two critical insights. Firstly, the equivariant representation is successfully transferred from the intra-modal latent space to the inter-modal latent space via the shared proposed transformation predictor. Moreover, although the individual augmented input pair may provide inaccurate supervision regarding the audio-visual correspondence, the centroid of these embeddings offers better supervision compared to the embeddings of the original input pair. Note that we use 16 equivariant embeddings for computing the centroid in this experiment.

Table 6. Zero-shot retrieval results on AudioSet and audio-visual event classification performance on AudioSet-20K with the number of equivariant representations used for computing the centroid in the inter-modal latent space.

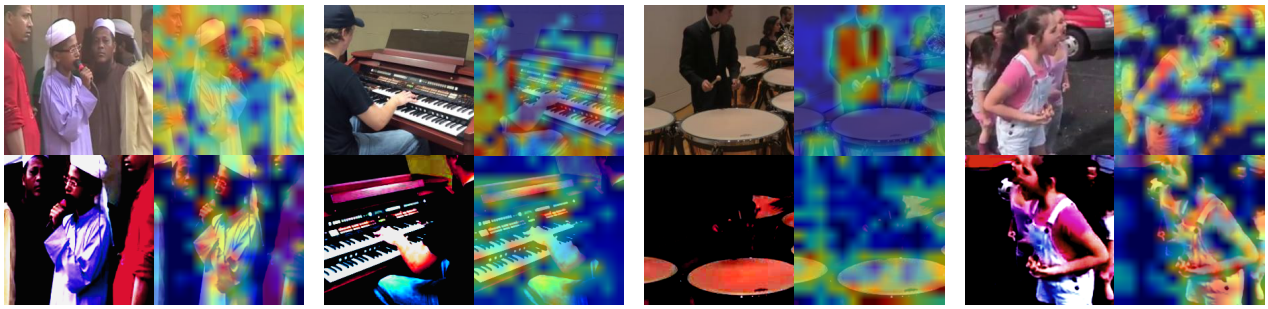
	ZS Retrieval		Fine-Tuning (mAP)		
# representations	V2A	A2V	A	V	A-V
1	27.8	27.4	41.2	23.6	43.4
4	28.3	29.0	41.4	24.4	44.9
8	29.4	29.2	41.9	24.9	45.5
16	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

Table 7. Averaged cosine similarity score and Event classification performance on AudioSet-20K based on varying architectures of transformation predictor.

	Similarity		ZS Retrieval		Fine-Tuning (mAP)		
Method	A	V	V2A	A2V	A	V	A-V
Linear	0.85	0.87	28.2	27.5	41.5	24.1	43.8
Hypernet.	0.87	0.90	29.6	29.6	42.0	25.2	44.5
MHA	**0.90**	**0.92**	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

Number of Representations for Computing Centroid. As explained in Section 3.2, the centroid gets closer to the true mean of the representations generated from the augmented inputs by increasing the number of equivariant representations. Table 6 shows that both zero-shot retrieval and fine-tuning performance consistently improve as the number of representations increases. However, since there still exists a variance in the centroids, our method can successfully overcome the trade-off between providing accurate audio-visual correspondence and increasing the diversity of training data. In addition, an advantage of our method is that the equivariant representations are generated from the representation of the original input using the transformation predictor, instead of passing the augmented inputs through the modality-specific encoders one by one. This can save the significant overhead of computation and training time. In our experiments, using 16 equivariant representations to compute the centroids increases the computational cost and training time by less than 1%, as illustrated in Appendix A.

Transformation Predictor Architectures. We explore different architectures of the transformation predictor that models the transformation in the latent space caused by input augmentations. Equimod (Devillers & Lefort, 2023) takes a concatenation of parameterized augmentation vector and embedding as the input to a linear layer. SIE (Garrido et al., 2023) points out that the predictor with such a simple architecture can collapse to the identity function and suggests a hypernetwork-based design that predicts the weights of the predictor. In Table 7, we compare the cosine similarity between the embedding of the augmented input z'_m and the equivariant embedding predicted by the predictor $\hat{z}_m$ for each architecture. The proposed attention-based architecture



(a) child singing (b) playing electronic organ (c) playing tympani (d) children shouting

Figure 3. Qualitative results of EquiAV using VGGSound. **Top Left**: Original input image, **Top Right**: Zero-shot sound source localization result for the original input image. **Bottom Left**: Augmented image whose embedding is the closest to the centroid of equivariant embeddings, **Botton Right**: Zero-shot sound source localization result for the augmented image.

Table 8. Zero-shot retrieval results on AudioSet and audio-visual event classification performance on AudioSet-20K with the different training strategies.

	ZS Retrieval		Fine-Tuning (mAP)		
Training Strategy	V2A	A2V	A	V	A-V
Two-Stage	27.0	25.6	41.0	23.1	43.8
Alternating iterative	28.5	29.0	41.6	24.8	44.7
Joint	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

Table 9. Zero-shot retrieval results on AudioSet and audio-visual event classification performance on AudioSet-20K with different loss functions.

	Zero-shot Retrieval		Fine-Tuning (mAP)		
Intra-modal Loss	V2A	A2V	A	V	A-V
without pos (Eq. A)	21.9	21.0	39.5	22.0	42.6
with pos (Eq. 9)	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

predicts displacement more precisely than others. Moreover, the well-trained equivariance leads to better performance in downstream tasks.

Training Strategies. We also conduct an ablation study on different training strategies for our framework. First, the two-stage training strategy divides the intra-modal equivariant representation learning phase and the inter-modal representation learning phase. In the first stage, the audio and visual encoders are trained using the intra-modal loss (Eq. 10), and in the second stage, both encoders undergo combined training through inter-modal loss (Eq. 15). Another training strategy employs an alternating iterative weight update of intra-modal loss and inter-modal loss. For a fair comparison, all strategies are trained with the same epoch. In Table 8, the two-stage and alternating iterative training strategies provide suboptimal results compared to the joint training. The result demonstrates joint training allows sufficient learning of intra-modal and inter-modal representations simultaneously without forgetting salient information.

Equivariant Loss Functions. Table 9 shows the results of using different equivariant loss functions for intra-modal pairs. The first row excludes the positive pair in the denominator according to Equation A, while the second row includes it as in Equation 9. When Equation 9 is used for the intra-modal contrastive loss, the weight updates due to hard positives are relatively larger than those due to easy positives. In the context of equivariant contrastive learning, learning from hard positive with the transformation predictor is more effective, leading to better representation quality. Consequently, the semantically rich intra-modal representation promotes effective learning of audio-visual correspondence and audio-visual joint representations. The experimental results support our hypothesis. For further analytical insights, refer to Appendix B.

4.3. Qualitative Results

The evaluation results and ablation studies indicate that the proposed representative embedding effectively captures the salient information of audio-visual inputs. We also explore how the relationship between the original embedding and the representative embedding is manifested in the input space. To this end, 1,000 augmentation vectors are randomly generated for each sample and then the equivariant embedding nearest to the centroid is selected as the substitute for the centroid. The corresponding transformation for this equivariant embedding is likely the one that best represents the image in the input space. The qualitative results in Figure 3 further validate our hypothesis, specifically by highlighting the parts of the image corresponding to the paired audio. In addition, the zero-shot sound source localization task is performed on the original images and the images selected by the above method. While the attention map on the original images is dispersed to parts beyond the main elements, the

attention for the augmented images, which correspond to the substitute embeddings of the centroids, is precisely and intensively concentrated on the areas that make the sound.

5. Conclusion

In this paper, we propose EquiAV, a novel self-supervised audio-visual contrastive learning framework that incorporates the principle of equivariance. EquiAV overcomes the limitations of applying augmentation in multi-modal representation learning by deriving a representative embedding with a proposed shared transformation predictor. Extensive quantitative and qualitative results support the validity of our approach. Furthermore, our approach can be adopted in other multi-modal domains, such as vision-language. In particular, a multi-modal dataset with paired data augmentations enables the diverse approaches of applying equivariance in multi-modal representation learning. We leave it as future work and we hope that EquiAV will help to push the boundaries in multi-modal representation learning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics), by the National Research Foundation of Korea grant funded by the Korean government (MSIT, RS-2023-00212845), and by the Information Technology Research Center (ITRC) support program (IITP-2024-RS-2023-00259991) supervised by Institute of Information & communications Technology Planning & Evaluation (IITP).

References

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. Self-Supervised MultiModal Versatile Networks. In *Advances in Neural Information Processing Systems*, 2020.

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., and Tran, D. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020a.

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., and Tran, D. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, 2020b.

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 609–617, 2017.

Boggust, A. W., Audhkhasi, K., Joshi, D., Harwath, D., Thomas, S., Feris, R. S., Gutfreund, D., Zhang, Y., Torralba, A., Picheny, M., et al. Grounding spoken words in unlabeled video. In *CVPR Workshops*, volume 2, 2019.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Chen, C.-F. R., Fan, Q., and Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *IEEE international conference on acoustics, speech and signal processing*, pp. 721–725. IEEE, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020b.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., and Soljacic, M. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022.

Devillers, A. and Lefort, M. Equimod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Fayek, H. M. and Kumar, A. Large scale audiovisual learning of sounds with weakly labeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions Audio, Speech and Language Processing*, 30:829–852, dec 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3133208. URL https://doi.org/10.1109/TASLP.2021.3133208.

Garrido, Q., Najman, L., and LeCun, Y. Self-supervised learning of split invariant equivariant representations. In *International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE international conference on acoustics, speech and signal processing*, pp. 776–780. IEEE, 2017.

Georgescu, M.-I., Fonseca, E., Ionescu, R. T., Lucic, M., Schmid, C., and Arnab, A. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16144–16154, October 2023.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. In *Interspeech*, pp. 571–575, 2021.

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., and Glass, J. R. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QPtMRyk5rb.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Haliassos, A., Ma, P., Mira, R., Petridis, S., and Pantic, M. Jointly learning visual and auditory speech representations from raw data. In *International Conference on Learning Representations*, 2022.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Huang, P.-Y., Xu, H., Li, J. B., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Huang, P.-Y., Sharma, V., Xu, H., Ryali, C., Fan, H., Li, Y., Li, S.-W., Ghosh, G., Malik, J., and Feichtenhofer, C. MAVil: Masked audio-video learners. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OmTMaTbjac.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Jenni, S., Black, A., and Collomosse, J. Audio-visual contrastive learning with temporal self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7996–8004, 2023.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

Korbar, B., Tran, D., and Torresani, L. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, volume 31, 2016.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. Hmdb: a large video database for human motion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2556–2563. IEEE, 2011.

Lee, H., Lee, K., Lee, K., Lee, H., and Shin, J. Improving transferability of representations via augmentation-aware self-supervision. *Advances in Neural Information Processing Systems*, 34:17710–17722, 2021.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ma, S., Zeng, Z., McDuff, D., and Song, Y. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021a.

Ma, S., Zeng, Z., McDuff, D., and Song, Y. Contrastive learning of global and local video representations. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021b.

Morgado, P., Misra, I., and Vasconcelos, N. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12945, 2021a.

Morgado, P., Vasconcelos, N., and Misra, I. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12475–12486, 2021b.

Morgado, P., Vasconcelos, N., and Misra, I. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021c.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213, 2021.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pp. 631–648, 2018.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pp. 2613–2617, 2019.

Patrick, M., Asano, Y. M., Kuznetsova, P., Fong, R., Henriques, J. a. F., Zweig, G., and Vedaldi, A. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9577–9587, October 2021.

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

Recasens, A., Luc, P., Alayrac, J.-B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1255–1265, 2021.

Sarkar, P. and Etemad, A. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Sarkar, P. and Etemad, A. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9723–9732, 2023b.

Sarkar, P. and Etemad, A. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, L., Luc, P., Recasens, A., Alayrac, J., and van den Oord, A. Multimodal self-supervised learning of general audio representations. *CoRR*, abs/2104.12807, 2021. URL https://arxiv.org/abs/2104.12807.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *ArXiv*, abs/1804.03209, 2018. URL https://api.semanticscholar.org/CorpusID:4719239.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

A. Appendix

A. Additional Ablation Studies

Weight-Scales of Loss Function. To understand how audio-visual contrastive learning is influenced by learning from each data modality, we explored the effects of setting $\lambda_{a,intra}$ and $\lambda_{v,intra}$ in Table A. When self-supervision is derived solely from the audio domain (i.e. $\lambda_v = 0$), it yielded the highest performance in audio-only tasks but showed suboptimal performance in visual-only and audio-visual tasks, and vice versa. Furthermore, adjusting the values of λ_a and λ_v to be either higher or lower relative to λ_{inter} also resulted in relatively lower performance. Consequently, we found that using equal λ values for both modalities and inter-modality learning produced balanced results.

Table A. Audio-visual event classification performance on AudioSet-20K with varying weights scales of loss function.

			Fine-Tuning		
λ_{inter}	$\lambda_{a;intra}$	$\lambda_{v;intra}$	A	V	A-V
1	1	0	**42.9**	24.0	43.4
1	0	1	41.2	**26.3**	43.2
1	2	2	42.2	25.5	45.5
1	0.5	0.5	41.3	25.1	46.0
1	1	1	42.4	25.7	**46.6**

Computational Cost & Scalability. Table B reports the computational costs required by generating equivariant representations and Table C shows how the size of the pre-training dataset affects performance on downstream tasks. The results demonstrate the computational efficiency and scalability of our framework.

Table B. Computation cost for generating equivariant representations in audio and visual domains.

# representations	Audio(GFLOPs)	Visual(GFLOPs)
Baseline	97.80	35.20
1	98.21 (+0.41)	35.36 (+0.16)
4	98.23 (+0.43)	35.38 (+0.18)
8	98.25 (+0.45)	35.40 (+0.20)
16	98.30 (+0.50)	35.44 (+0.24)

Table C. Zero-shot retrieval results on AudioSet and audio-visual event classification performance on AudioSet-20K with the variants of pre-training datasets.

	Zero-shot Retrieval		Fine-Tuning (mAP)		
Pre-training Dataset	V2A	A2V	A	V	A-V
AudioSet-20K	3.9	4.2	33.7	17.3	35.2
VGGSound	11.2	10.5	37.6	21.3	40.8
AudioSet-2M	**30.1**	**29.6**	**42.4**	**25.7**	**46.6**

B. Analysis on Equivariant Loss Functions

As mentioned in the ablation study on Table 9, our equivariant loss function differs from the equivariant contrastive loss function used in EquiMod (Devillers & Lefort, 2023), regarding whether a positive pair similarity is included in the denominator. The equivariant loss of EquiMod is the same as applying an indicator function to both summation terms in the

denominators of Equation 9, which can be represented as follows:

$$\ell_{\text{EquiMod}}(\hat{z}, z') = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\hat{z}_i, z'_i)/\tau)}{\sum_{j=1}^{N}\mathbb{1}_{[j\neq i]}\left[\exp(\text{sim}(\hat{z}_i, z'_j)/\tau) + \exp(\text{sim}(\hat{z}_i, \hat{z}_j)/\tau)\right]}, \tag{A}$$

$$\mathcal{L}_{\text{EquiMod}} = \frac{1}{2}(\ell_{\text{EquiMod}}(\hat{z}, z') + \ell_{\text{EquiMod}}(z', \hat{z})). \tag{B}$$

Consider intra-modal training batch embeddings as $\{z_i\} = \{\hat{z}_i\} \cup \{z'_i\}$. For each i-th embedding, let's denote the positive sample as p_i and the set of negative samples as N_i. Then, the equivariant loss terms of EquiMod and EquiAV can be simply rewritten as

$$\mathcal{L}_i^{\text{EquiMod}} = -\log\frac{s_{i,p_i}}{\sum_{n\in N_i} s_{i,n}}, \tag{C}$$

$$\mathcal{L}_i^{\text{EquiAV}} = -\log\frac{s_{i,p_i}}{s_{i,p_i} + \sum_{n\in N_i} s_{i,n}}, \tag{D}$$

where $s_{i,j} = \exp(\text{sim}(z_i, z_j)/\tau)$. Differentiating the above equations with respect to s_{i,p_i} yields the following expressions:

$$\frac{\partial\mathcal{L}_i^{\text{EquiMod}}}{\partial s_{i,p_i}} = -\frac{1}{s_{i,p_i}}, \tag{E}$$

$$\begin{aligned}\frac{\partial\mathcal{L}_i^{\text{EquiAV}}}{\partial s_{i,p_i}} &= -\frac{\sum_{n\in N_i} s_{i,n}}{s_{i,p_i}\left(s_{i,p_i} + \sum_{n\in N_i} s_{i,n}\right)} \\ &= \frac{\partial\mathcal{L}_i^{\text{EquiMod}}}{\partial s_{i,p_i}} \cdot \frac{\sum_{n\in N_i} s_{i,n}}{\left(s_{i,p_i} + \sum_{n\in N_i} s_{i,n}\right)}\end{aligned} \tag{F}$$

When we compare the EquiAV loss to the EquiMod loss through Equ.F and Equ. E, the EquiAV loss puts relatively more weight on hard positive compared to easy positive. This becomes particularly advantageous when stronger augmentations lead to an increased frequency of hard positives. Training with more hard positives and applying equivariance substantially enhances the model's capability to comprehend detailed features in single-modal contexts. As a result, models enhanced with this function show improved performance and effectively integrate different modalities.

C. Implementation Details

Datasets We utilize two prominent audio-visual datasets for our experiments: AudioSet (Gemmeke et al., 2017) and VGGSound (Chen et al., 2020a). AudioSet comprises 2 million 10-second YouTube clips, designed to classify events into 527 distinct classes, with each data having multiple labels. We download 1,893,278 clips for the AudioSet-2M, 21,074 clips for the AudioSet-20K, and 19,446 clips for evaluation. Particularly, AudioSet-20K is a subset of AudioSet-2M. VGGSound includes 200,000 10-second YouTube clips, encompassing 309 classes. The training and test splits of VGGSound consist of 183,730 and 15,446 downloaded clips, respectively. Unlike AudioSet, it has only one label for each clip. For the zero-shot retrieval evaluation, we collect 1,722 and 1,545 clips from AudioSet and VGGSound's evaluation set respectively. In addition, MSR-VTT (Xu et al., 2016) test set is also used in the zero-shot retrieval task.

Input Pre-processing. We follow AST (Gong et al., 2021) and ViT (Dosovitskiy et al., 2021) for pre-processing of audio and visual inputs, respectively. For audio, each 10-second audio waveform is transformed into a sequence of 128-dimensional log Mel filterbank features by using a 25-ms Hanning window and a 10-ms hop size, resulting in a 1024(time) $\times$ 128(frequency) spectrogram. For visual inputs, 10 frames are uniformly extracted from each 10-second video, and one frame is randomly selected as the input. Then the audio spectrogram and the video frame are tokenized to 16 $\times$ 16 patches and fed to the audio and visual encoders. In terms of input data augmentation for the visual modality, the typical visual-based augmentations are used; Random Resized Crop (RRC), Color Jitter (CJ), Gaussian Blur (GB), Horizontal Flip (HF), and

Gray Scale (GS). On the other hand, both audio- and visual-based augmentations are applied to the audio spectrogram. Specifically, SpecAugment (SA) (Park et al., 2019) and Time Shifting (TS) are utilized as audio-based augmentations. Meanwhile, the same augmentations pool used for the visual modality excluding GS is applied to audio spectrograms, as the spectrogram has only one channel.

Augmentation Parameterization. Audio and visual augmentations as well as their parameterization used in this work are listed as follows:

- *Random resized crop* (4 elements for both audio & visual): x and y coordinates of the top-left point, as well as the width and height of the crop (<0, 0, 0, 0>, is used as a default encoding).
- *Color jitter* (8 elements for visual & 4 elements for audio): the jitter factors for brightness, contrast, saturation, and hue of video frames, as well as the order of the application of transformation. We use the following mapping to encode the order of transformation: {0: brightness, 1: contrast, 2: saturation, 3: hue}. For instance, an encoding <2, 1, 3, 0> indicates that the saturation jitter is first applied, and then contrast, hue, and brightness (<0, 1, 2, 3> is used as default). On the other hand, we only use brightness and contrast jitters for audio spectrograms, as the audio spectrograms are originally grayscale. Then, we use the following mapping to encode the order of jitter transformation: 0: brightness, 1: contrast (<0, 1> is used as default).
- *Gaussian blur* (1 element for both audio & visual): the value of σ for Gaussian blurring kernel (0 as default).
- *Horizontal flip* (1 element for both audio & visual): 1 if an image or an audio spectrogram is horizontally flipped and 0 otherwise.
- *Grayscale* (1 element, for visual only): 1 if an image is converted to grayscale and 0 otherwise.
- *Random time shifting* (1 element, for audio only): the value of temporal shift of the audio spectrogram.
- *SpecAug* (Park et al., 2019) (4 elements, for audio only): starting and end points of the masking along the time and frequency axis of the audio spectrogram.

All augmentations except random resized crop (which is always applied) are applied with pre-defined probability. Therefore for each augmentation, we add an element, whose value is 1 when the augmentation is actually applied and 0 otherwise, to a parameterized vector. Consequently, the audio and visual augmentations are encoded into 24-dimensional and 18-dimensional vectors respectively.

D. Hyperparameter Details

The hyperparameter settings used in this paper are listed in Table D.

Table D. Hyperparameters used in pre-training and fine-tuning phase.

Stage	Pre-training	Fine-Tuning								
Dataset	AudioSet-2M	AudioSet-20K	AudioSet-2M	VGGSound	ESC-50	SPC-v1	FSD-50K	UCF101	HMDB51	K400
Optimizer	AdamW									
Optimizer momentum	β_1=0.9, β_2=0.95									
Weight decay	1e-5									
Learning rate scheduler	half-cycle cosine annealing (Loshchilov & Hutter, 2017)									
Initial learning rate	1e-6									
Peak learning rate	1e-4	1e-4	1e-4	1e-4	5e-4	1e-3	5e-4	1e-3	5e-4	5e-4
Warm-up epochs	2	1	1	1	4	1	4	0	0	4
Epochs	20	50	50	50	60	10	60	100	100	100
Batch size	256	256	512	256	64	256	64	32	32	512
Class Balancing Weight	No	No	Yes	Yes	No	No	No	No	No	No
Mixup	No	Yes	Yes	Yes	No	Yes	No	No	No	Yes
Loss Function	EquiAV Loss (Eq. 16)	BCE	BCE	CE	CE	BCE	BCE	CE	CE	CE
Temperature (τ)	0.07	-	-	-	-	-	-	-	-	-
Input Norm Mean	-4.346	-4.346	-4.346	-4.956	-6.627	-6.702	-6.627	-	-	-
Input Norm STD	4.332	4.332	4.332	4.486	5.359	5.448	5.359	-	-	-
GPUs	8 A6000	8 A5000	8 A6000	8 A5000	8 A5000	8 A5000	8 A5000	8 A5000	8 A5000	8 A5000