SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos

Changan Chen[1] Kumar Ashutosh[1,2] Rohit Girdhar[2] David Harwath[1] Kristen Grauman[1,2]
[1]University of Texas at Austin [2]FAIR, Meta

Abstract

We propose a novel self-supervised embedding to learn how actions sound *from narrated in-the-wild egocentric videos. Whereas existing methods rely on curated data with known audio-visual correspondence, our multimodal contrastive-consensus coding (MC3) embedding reinforces the associations between audio, language, and vision when all modality pairs agree, while diminishing those associations when any one pair does not. We show our approach can successfully discover how the long tail of human actions sound from egocentric video, outperforming an array of recent multimodal embedding techniques on two datasets (Ego4D and EPIC-Sounds) and multiple cross-modal tasks.*

1. Introduction

Human activity often produces sounds. Closing a door, chopping vegetables, typing on a keyboard, talking with a friend—our interactions with the objects and people around us generate audio that reveals our physical behaviors. These sounds can be strongly associated with the subjects of our activity and how we perform it. For example, opening a water bottle sounds different than opening a cabinet; chopping sweet potatoes sounds different than chopping onions; chopping onions sounds different than mincing onions (the same object). Understanding the link between sounds and actions is valuable for a number of applications, such as multimodal activity recognition, cross-modal retrieval, content generation, or forecasting the physical effects of a person's actions.

How should AI learn about *sounding actions*? Existing work typically curates annotated datasets for supervised learning [9, 22, 28, 45], taking care to select events or actions that have associated sounds (e.g., lawnmowing, chopping), while others deliberately collect videos of object collisions (e.g., striking objects with a drumstick [43] or crashing into them with a robot [10, 17]), or develop physics-based simulations [16]. On the one hand, these approaches are appealing for their ability to focus on meaningful audio-visual correspondences. On the other hand, their curated nature risks limiting the scope of sounding actions that can be learned.

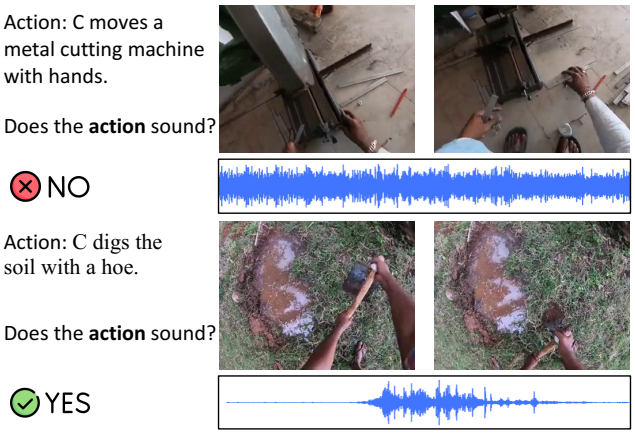

Figure 1. We aim to distinguish sounds that are directly caused by human actions (bottom) from those that are not (top). Given egocentric training videos with language descriptions of the camera wearer's ("C") current action, we learn an embedding where the audio and visual features of any given clip are best aligned only when both are also consistent with the language. This allows discerning clips where the audio and vision may be *correlated* (e.g., the cutting machine running making loud noise in top row) versus those where the sounds are *driven by human action* (digging in bottom row)—importantly, without language at inference time.

Instead, we aim to learn how human actions sound from narrated in-the-wild egocentric videos. See Figure 1. Given a pool of videos of everyday human activity, the goal is to learn a cross-modal representation where sounding actions cluster together based on how they look and sound. By sampling the videos freely, we can broaden the scope to discover the breadth of sounding actions without having to rely on a closed, pre-defined set of action categories. In particular, by focusing on *unscripted egocentric* video from wearable cameras in daily-life settings [11, 26], we aim to include subtle and long-tail scenarios unavailable in curated datasets, such as sounds of keys jangling when unlocking a door, scissors snipping when cutting the dog's fur, or fingernails scratching on one's own arm. Egocentric video is a particularly attractive source here because 1) human interaction sounds are more audible in near-field egocentric recordings and 2) passively captured long-form ego-video simply covers more

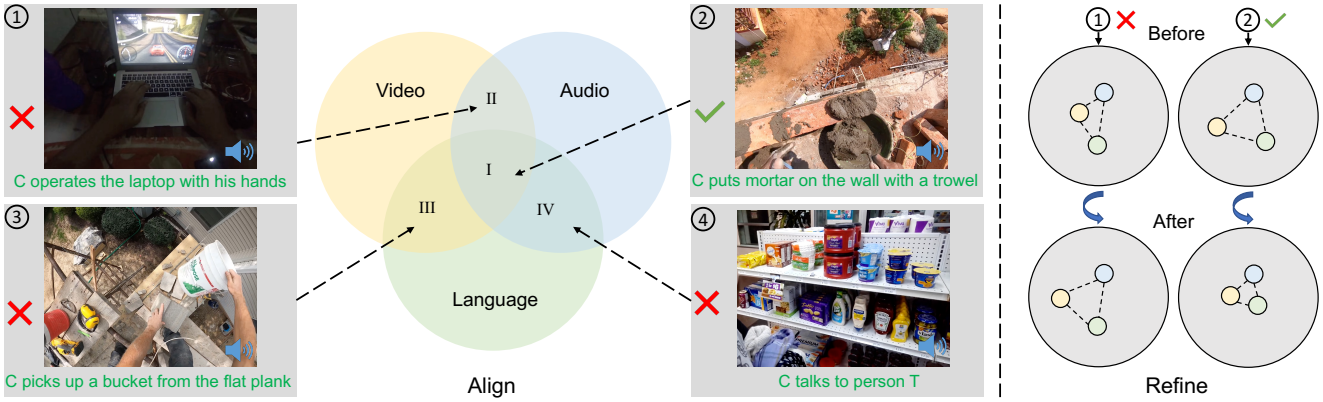

Figure 2. **Main idea**. On the left, the Venn diagram illustrates different ways audio (A), video (V) and language (L) modalities can overlap in the content they capture. C refers to the camera wearer. Regions II,III,IV are information that is only shared between two modalities but not the third, e.g., the racing game in ① where the game sounds correlate with the vision, yet are not about the camera wearer's described action (using hands on laptop), the lifting action in ③, where the visuals and language agree but the action is inaudible, and the off-screen talking action in ④, where talking is heard and described, but the camera wearer cannot be seen speaking. Region I is the information that corresponds to all modalities agreeing, e.g., the visible and audible plastering action in ②. Our model's "align" phase detects any such (dis)agreements via pairwise contrastive learning on the modalities. In the "refine" phase, we use the intersection of that agreement (region I) to refine the embedding. For example, on the right, we show what the three modality embeddings should look like after the "align" stage for examples 1 and 2. Embeddings of instances where all modalities agree will be closer in the embedding space and apart otherwise. In other words, for example 1, yellow (video) cannot be close to blue (audio) unless green is too (language).

everyday sounds, including the rare ones.

However, the learning task is challenging because some visible actions do not make any sound, and some sounds are the result of off-screen actions. Finally, other sounds may be *correlated* with on-screen objects (such as traffic noise and a city street), but are not directly related to the video by a salient and visible camera wearer action. For this reason, although existing self-supervised audio-visual methods [3, 5, 6, 20, 23, 33, 35, 42] are good at detecting audio-visual correspondences, they tend to capture general correlations rather than the action-specific correspondence.

To address this challenge, we propose a novel *multimodal consensus embedding* approach. Importantly, we suppose the in-the-wild egocentric training videos are accompanied by free-form natural language descriptions describing the actions of the camera wearer, as provided in the "narrations" of existing large-scale ego-video datasets [11, 26]. The main idea is to seek video samples where there is semantic agreement between all three modalities—the audio, visual, and language—while distancing those that do not. This *intersection* of the modalities with language assures that correspondences in the audio and visual streams stem from alignment on the sounding action.

To achieve this, the proposed model first aligns a preliminary embedding from contrastive losses imposed per instance on each pair of modalities. Next, we refine those embeddings with a consensus objective that targets a minimum (bottleneck) pairwise similarity. The latter pushes all pairs of inter-modality agreement towards this consensus—or lack thereof—while jointly continuing to optimize the paired-modalities' contrastive losses. In this way, we overcome the simplifying assumption made by existing multimodal embeddings that require all modalities to agree [1, 23, 55]. See Figure 2.

We demonstrate our approach by training with in-the-wild data from Ego4D [26] without audio labels and testing on both Ego4D and EPIC-Sounds [28]. To allow a formal large-scale evaluation of sounding actions, we introduce a dataset of professional annotations on 33K video clips spanning Ego4D. Our model successfully discovers sounding actions that agree with ground truth labels on both datasets. Compared to existing multimodal embedding paradigms [14, 23, 35, 55], our model not only better discovers sounding actions and learns embeddings for cross-modality retrieval, but also generalizes better to the audio classification benchmark on EPIC-Sounds. To our knowledge, this is the first result of its kind to show sounding actions discovered organically from narrated in-the-wild video.

2. Related Work

Action/interaction/impact sound. Some work [21, 30, 40] leverages audio to improve activity recognition on video datasets such as UCF101 [52] and ActivityNet [15], which have visual labels but no audio labels. Existing audio datasets such as AudioSet [22] and VGG-Sound [9] target general sound classes such as music, speech, and sports. EPIC-Sounds [28] provides an audio classification bench-

mark for actions in kitchen environments, but it has no labels for the correspondence between the visual action and the sound. The Greatest Hits dataset [43] contains videos where people hit and scratch object surfaces with a drumstick, which enables audio synthesis from videos. Interaction sound has also been studied in robotics, e.g., using a robotic platform to collect sounds and study the synergy between action and sounds [10, 17]. Impact sounds are modeled in a physics-based simulator [16]. Throughout, the existing work assumes a fixed, given taxonomy of action classes or audio labels of interest. In contrast, we learn how actions make sounds from in-the-wild narrated egocentric videos, without relying on a taxonomy of discrete labels for the audio events.

Audio-visual learning. As naturally co-occurring data, there are rich correspondences between video and audio, such as spatial correspondence [18, 29, 38], acoustic correspondence [8, 51], semantic correspondence [5, 6], and lip motion [19, 49]. Existing work typically either learns one type of correspondences by manually creating misalignment, e.g., shifting audio temporally to create temporal supervision [5, 33, 42] or down-mixing audio channels to create spatial supervision [18, 38], or it picks up any correspondence that emerges from the learning process [37, 39] (e.g., actions, objects, environments). In contrast, we aim to learn action-specific correspondence by leveraging the semantic grounding from human narrations.

Language+X learning. Language can provide semantic grounding for what we see or hear. Prior work exploring language with another modality includes image/video captioning [57], visual question answering [4], and audio captioning [36]. Recent results with large-scale image-text datasets show that language is excellent for guiding the learning of image features (such as CLIP [47]), video features [34, 59], or audio features [27]. However, there is limited work studying binding language to more than one modality, as we propose. Some recent work [1, 58] explores self-supervised learning with language, vision, and sound, where the language is typically the transcription of the audio. They construct self-supervised objectives under the assumption that modalities agree with each other. In contrast, we use language that is different from the speech transcription and we explore modality agreement in training.

Multi-modal/view representation learning. Recent work builds multimodal models using more than two modalities (e.g., audio, video, and language) for improving the representations of one modality [1, 2, 50], while others explore modality-invariant or view-invariant features [23, 53, 55]. For example, ImageBind [23] binds the visual features with other modalities in sequence, while CMC [55] proposes a contrastive multi-view loss that maximizes the mutual information between different views of the same scene. These methods assume there exists shared information among all modalities, which does not address how to find these examples. Our method implicitly discovers the shared information by analyzing the multimodal consensus, or lack thereof.

3. Task Formulation

We define a **sounding action** as a human-initiated action that produces sound during its execution due to interactions with the surrounding environment. We are particularly interested in learning how subtle and long-tail daily human actions sound. If hypothetically we were given a clip with audio a, video v, and label y indicating whether the clip contains a sounding action, our objective would be to minimize the distance between audio-visual embeddings if $y = 1$ and maximize the distance between them if $y = 0$, i.e., minimizing $(-1)^y \mathcal{D}(e_a, e_v)$, where $\mathcal{D}$ measures the distance and $e_{a,v}$ are their embeddings. However, we do not assume access to any such direct supervision; labeling sounding actions is expensive, both because many actions do not produce sounds, and because many clips do not contain actions. Instead, we aim to discover sounding actions in a weakly supervised fashion, while simultaneously learning multimodal embeddings that capture them well.

To this end, we leverage "narrations", a form of language description that is collected in recent egocentric video datasets such as Ego4D [26] and EPIC-Kitchens [11]. These narrations are timestamped free-form sentences describing the current action being performed by the camera-wearer. See Figure 2 for examples. Note that there may be other events in the video, too (e.g., a TV is playing), but these are *not* narrated. This is significant: the language specifically addresses near-field human interactions with objects, people, and the environment. The narrations offer two key benefits: 1) the timestamps provide *temporal* grounding of actions that occur in the video, indicating where potentially interesting clips are and 2) the language provides *semantic* grounding of actions—which our multimodal consensus idea will exploit to learn action-specific audio-visual correspondence.

Formally, given a video with frames $v \in \mathcal{R}^{T \times H \times W \times C}$, audio $a \in \mathcal{R}^S$, and language narration l, where T and S are the number of frames for video and audio respectively, the goal is to learn embeddings e_v and e_a that are close in the embedding space if both a and v capture the same human action described in l, and distant otherwise. If we plot how the three modalities overlap in a Venn diagram (Figure 2), we can see that what we are interested in learning is exactly region I, i.e., a camera-wearer action that sounds. From an information-theory perspective, this is equivalent to learning modality-invariant embeddings.

4. Multimodal Contrastive-Consensus Coding

Next we present our solution MC3 (Multimodal Contrastive-Consensus Coding) for learning modality-invariant embeddings, which consists of an *inter-sample contrastive loss* and

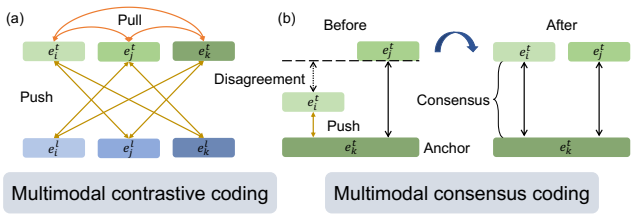

Figure 3. **Multimodal contrastive-consensus loss**. (a): Given three modality embeddings e_i^t, e_j^t, e_k^t, multimodal contrastive coding pulls each pair of modalities closer while pushing modality pairs from another sample further away. (b): However, not all modalities agree on how close they should be depending on the instance. Thus we set the furthest distance a feature has with respect to the anchor feature as the consensus and push the remaining embeddings away to meet this consensus.

an *intra-sample consensus loss*. See Fig. 3. We first present the two-stage training framework in Sec. 4.1 and then discuss the two losses in Sec. 4.2 and Sec. 4.3. For simplicity, we denote the n input modalities as $M_i, i \in [1, n]$.

4.1. Align-Refine Two-stage Training

We design a two-stage training paradigm. The high-level idea is to first optimize the pairwise agreement in an "align" stage, and then refine these embeddings with global consensus in the "refine" stage. See Fig. 2.

In the first stage, we train modality encoders with a contrastive loss $\mathcal{L}_{\text{contrastive}}$, which guides modality embeddings to have a good initial alignment that captures the pairwise similarity between modalities that capture the same underlying action, as opposed to random initialization.

In the second stage, we refine the pairwise-aligned embeddings with a globally established consensus. Specifically, we train the model with a consensus loss $\mathcal{L}_{\text{consensus}}$ that pushes all intra-sample modality agreement towards this consensus, while jointly optimizing the contrastive loss $\mathcal{L}_{\text{contrastive}}$, to maximally capture the shared information across modalities. The MC3 loss $\mathcal{L}_{\text{MC3}}$ combines the contrastive and consensus losses, and will be detailed below. We confirm experimentally that it is important to keep the contrastive loss in the second stage, although the main purpose of this stage is to refine embeddings with consensus.

4.2. Multimodal Contrastive Coding

Cross-modal contrastive learning has been shown to discover representations where modalities are informative of each other [39]. Prior work [46, 56] shows that minimizing the contrastive loss between M_i and M_j maximizes the lower bound on the mutual information $I(M_i; M_j)$. Inspired by this, we first use contrastive learning to optimize the pairwise similarities $\mathcal{S}(e_i, e_j) = e_i e_j$, where $e_{i,j}$ is the latent embedding normalized on the unit sphere for modality pair i, j. We use the InfoNCE [41] loss to optimize each individual $\mathcal{S}(e_i, e_j)$ as follows:

$$\mathcal{L}_{i,j} = -\frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}} \log \frac{\exp(e_i^t e_j^t/\tau)}{\sum_{l\in\mathcal{B}}\exp(e_i^t e_j^l/\tau)}, \tag{1}$$

where $\mathcal{B}$ is the batch and τ is the temperature. This loss treats modalities from the same sample as positive pairs and pulls them closer and it treats modalities from different samples as negative pairs and pushes them apart. See Fig. 3 (a). The total loss is the sum of losses enumerated over all pairs of modalities, i.e., $\mathcal{L}_{\text{contrastive}} = \sum_{i,j}\mathcal{L}_{i,j}$.

4.3. Multimodal Consensus Coding

The contrastive loss above attempts to bring all temporally co-occurring modalities closer assuming there are strong correspondences among them in the input space. However, naively doing so would be problematic for instances where not all modalities agree (cf. Figure 2). To tackle this issue, we propose a novel objective that leverages the consensus of inter-sample modalities discovered from the contrastive coding as additional supervision.

First of all, we choose an anchor modality M_a, which serves as the point of comparison for other modalities $M_i, i \in [1, n], i \neq a$. With the normalized embedding e_i^t of modality i and sample t, we then compute the cosine similarity score between each non-anchor modality and the anchor modality. Now, these similarity scores may or may not agree with each other. To only learn embeddings shared across all modalities, we set the consensus score as the minimum (bottleneck) score:

$$c^t = \mathcal{K}^{-1}(\min_{i,i\neq a}(\mathcal{K}_1(e_1^t e_a^t), ..., \mathcal{K}_n(e_n^t e_a^t))), \tag{2}$$

where $\mathcal{K}_i(x) = ((x+1)/2)^{\alpha_i}, x \in [-1, 1]$ is a modality-specific scaling function that first maps scores to [0, 1] and then adjusts the distribution with a tunable parameter α_i. $\mathcal{K}^{-1}$ is the inverse function that maps the scaled score back to the original space. The intuition behind $\mathcal{K}_i(x)$ is that different modalities carry different amounts of information and we want to normalize the score distributions among the different modality pairs, making them comparable.

The consensus score c^t is high if and only if all pairwise scores are high, and it is low if there exists at least one modality that does not agree with the anchor modality. After obtaining the consensus score, we design a loss that forces all modalities to follow this consensus, as follows:

$$\mathcal{L}_{\text{consensus}} = \frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}}\sum_{i,i\neq a} ||e_i^t e_a^t - c^t||_2 \tag{3}$$

The total loss $\mathcal{L}_{\text{MC3}}$ is the sum of the contrastive loss

(Eq. 1) and the consensus loss (Eq. 3):

$$\mathcal{L}_{\text{MC3}} = -\frac{1}{|\mathcal{B}|}\Big(\underbrace{\sum_{t\in\mathcal{B}}\sum_{i,j}\log\frac{\exp(e_i^t e_j^t/\tau)}{\sum_{l\in\mathcal{B}}\exp(e_i^t e_j^l/\tau)}}_{\text{Inter-sample}} - \underbrace{\sum_{i,i\neq a}||e_i^t e_a^t - c^t||_2}_{\text{Intra-sample}}\Big). \tag{4}$$

This loss pushes embeddings with a low consensus score apart while pulling together embeddings with a high consensus score, and thus aligns embeddings better in the joint embedding space. See Fig. 3.

Optimizing this loss is not trivial since it has both contrastive and reconstruction objectives. Indeed, directly optimizing the loss does not work well as shown in the ablation study (Sec. 6.1). The proposed two-stage training paradigm (Sec. 4.1) helps train the model stably.

4.4. Implementation Details

Our modalities of interest are $M_1 = A$ (audio), $M_2 = V$ (vision), and $M_3 = L$ (language). There are six pairwise contrastive losses for three modalities. When computing the modality consensus, we empirically find using audio as the anchor leads to the best results in our task (cf. Sec. 6). We set the scaling parameters α_l and α_v to 1 and 0.5 respectively, based on a hyperparameter search on the validation set. See ablations in Supp.

For extracting the feature representations, we use TimeSformer [7] as our video encoder, DistillBERT [48] as our text encoder, and AST [24] as our audio encoder. We initialize the video and language encoders with embeddings from [34], and the audio encoder with embeddings pretrained on ImageNet [13]. We train all encoders. We choose these initial encoders due to their good results in the literature; however, our MC3 loss is not specific to the choice of these encoders and others could be swapped in.

We train all models on 8 A40 GPUs with a learning rate of 3e$-$5 and batch size of 256 for 5 epochs for both stages, and take the final checkpoint for evaluation. We use the Adam optimizer [32]. Our implementation is based on the codebase from [34].

5. Training and Eval Data for Sounding Actions

Dataset. Ego4D [26] is a large-scale egocentric video dataset that has more than 3,600 hours of video recordings depicting hundreds of daily activities—and 2,113 of those hours have audio available. As discussed, it also has time-stamped narrations that are free-form sentences describing the current activity performed by the camera-wearer. However, Ego4D has no annotation of whether an action makes sounds, what sounds an action makes, or whether there exists other (non-action) sounds. It is thus non-trivial to detect if an action in the clip makes sound based on simple heuristics, e.g., the burst of sound energy, since many actions could produce continuous sounds with ambient-sound characteristics, e.g., wiping tables or sawing wood.

Wash	Close	Cut	Drop	Stir	Wipe	Rub	Touch	Lift	Hold
0.90	0.82	0.77	0.64	0.64	0.53	0.39	0.27	0.19	0.09

Table 1. Example verb groups and how frequently they sound

We construct the training dataset by extracting clips from each Ego4D video based on the narration timestamps. These clips cover a wide range of daily activities and environments, including construction sites, cooking, arts and crafts, shopping, farming, and many others. Since the timestamp is only an approximate point for where an action occurs, we sample the clip from 0.5 s before to 1 s after the timestamp (1.5 s duration) so that the clip is likely long enough to capture the action sound, if there is any, without introducing visuals that stray from the narrated action. See ablations on the duration in Supp. We sample a training set of 250K clips from 1,876 hours of video. From their narrations, we find there are 6,114 unique nouns (objects) and 2,819 unique verbs (actions).

Ground truth annotations for evaluation. Today's egocentric video datasets lack annotations for sounding actions. Thus, to determine how well our model learns long-tail sounding actions and facilitate future research, we collect a large ground truth evaluation set for Ego4D using professional annotators trained for the task. It consists of 33K clips manually labeled as to whether or not the camera wearer's action sounds, i.e., indicating whether the action described in the narration is both visible and audible in the clip.

To ensure annotation quality, in addition to providing concrete examples and annotation guidelines (see Supp.) and iterating with quality control feedback to the professional annotators, we assign three annotators per clip and take the majority vote as the correct answer. We split the 33K obtained annotations into 3K for validation and 30K for test. We stress that this is an eval set only; our training data (above) has no manual labels about sound, only free-form language narrations.

Action type analysis. In total, among the 33,000 resulting ground truth clips, 17,693 are positive and 15,307 are negative. The fact that only half of this in-the-wild clip distribution consists of sounding actions underscores the need for models that can tell the difference between audio that *co-occurs* with human action and *actions that sound*. To gain insight into the annotations, we group them by semantic similarity and analyze them at a group level. While narrations provide semantic descriptions of actions, using them for grouping would be too noisy since the same action could be described in different ways. To reduce the influence of narration variance, we utilize the taxonomy defined in Ego4D (for

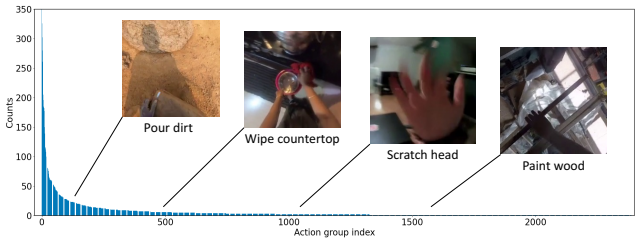

Figure 4. Long-tail distribution of sounding actions.

analysis only, not training). For example, “check", “examine", and “inspect" should belong to the same group (taxon). We first group these clips by verb alone, i.e., extracting verbs from narrations and then applying the taxonomy, which results in 106 unique groups. We then compute the percentage of clips in each group that make sounds. Tab. 1 shows 10 examples. We see that actions involving more significant human motions (wash, close, cut) are more often sounding, whereas more subtle movements (lift, hold) are often not. Importantly, there is not a one-to-one mapping between an action verb and its sounding label—how actions sound is scenario-dependent and hence must be mined from the data.

While grouping by verbs provides some insights, how actions make sounds also depends on the object that they interact with, e.g., cutting a carrot sounds different from cutting bread. To this end, we further group the 17K sounding clips by both verbs and nouns, which results in 2,388 unique action groups. We plot the long-tail distribution of them in Fig. 4 and show examples sampled from this distribution. This plot shows the diverse and long-tail nature of sounding actions and our test set annotations, which is not present in existing action datasets [10, 15, 17, 28, 43, 52].

6. Experiments

We compare our model with several baselines and ablations on three tasks: sounding action discovery (on Ego4D), sounding action retrieval (on Ego4d), and audio event classification (on EPIC-Sounds). We show our model outperforms an array of existing learning methods.

SotA Baselines. We consider two baselines that only use a contrastive loss for two modalities: CLAP [14] for audio-language and CM-ACC [35] for audio-video. For more than two modalities, we consider two more baselines: CMC [55] uses contrastive objectives between all pairs of viewpoints (modalities in our case), representing the joint training paradigm; ImageBind [23] learns the joint embedding by first performing vision-language pretraining and then freezing the vision encoder and training the vision-audio modality pair. This represents strategies that align modalities sequentially. For a fair comparison, we equip all baselines with the same encoder and the same initialization as ours (see Sec. 4.4) while keeping their original losses.

	🔊	📹	📄	*AV* ROC	*AV* PR	*AL* ROC	*AL* PR
Random	✗	✗	✗	0.500	0.559	0.500	0.559
CLAP [14]	✓	✗	✓	-	-	0.637	0.695
CM-ACC [35]	✓	✓	✗	0.540	0.590	-	-
CMC [55]	✓	✓	✓	0.550	0.601	0.635	0.693
ImageBind [23]	✓	✓	✓	0.554	0.605	0.642	0.685
w/o $\mathcal{L}_{\text{consensus}}$	✓	✓	✓	0.563	0.615	0.635	0.694
w/o $\mathcal{L}_{\text{contrastive}}$	✓	✓	✓	0.436	0.493	0.584	0.620
w/o align-stage	✓	✓	✓	0.448	0.507	0.464	0.521
MC3	✓	✓	✓	**0.598**	**0.666**	**0.658**	**0.715**

Table 2. Sounding action discovery. Area-under-curve (AUC) values are reported for both ROC and precision-recall (PR) curves, for audio-vision (AV) and audio-language (AL). Both are the higher the better. We train our model five times with different seeds; the standard deviation is always within 0.01.

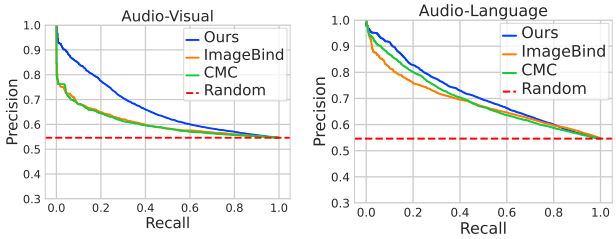

Figure 5. Sounding action discovery accuracy

6.1. Sounding Action Discovery

Human interactions with objects in our daily lives are complex and subtle. Due to many incidental background sounds, recognizing whether actions make sound is not trivial but can be useful for applications like multimodal video generation, e.g., verifying the generated action video and audio match. Towards this goal, we answer the question “what actions sound?" by performing sounding action discovery. In this experiment, we take the per-modality encoders learned on the narrated 250K Ego4D clips and apply them to the 30K test clips. Given a test clip, we feed the video and audio through their corresponding modality encoders, and compute the cosine similarity between the output embeddings. That score indicates how likely it is that the action in the video sounds. For completeness, instead of defining a hard threshold for positives, we plot the ROC and precision-recall (PR) curves by varying the positive threshold, and calculate the area-under-curve (AUC) values for them—common metrics for classification [12, 54] that are invariant to the absolute score values. For both metrics, higher values are better, indicating the model learns meaningful embeddings of sounding actions. Similarly, we can also evaluate discovery for audio-language, if narrations are available.

Results. Table 2 shows the results for sounding action discovery. We first look at discovery with audio-visual modalities alone at test time (“*AV*" columns). CM-ACC [35]

Figure 6. Example visual embedding cluster from our model

discovers sounding actions much better than random chance, showing that audio-visual contrastive learning captures both visual action embeddings and action sound embeddings. CMC [55] and ImageBind [23] do better—benefiting (like us) from the language modality at training time. However, neither the joint nor sequential training paradigm exploits modality agreement, resulting in weak cross-modal constraints, and thus only marginal performance improvement. In comparison, our model MC3 explicitly models the modality consensus and improves the discovery result substantially by learning embeddings most relevant to sounding actions.

We also report the discovery result from using audio-language modalities ("AL" columns). Since narrations provide action specifications, the discovery performance is better than AV, e.g., CLAP [14] vs CM-ACC [35]. While CMC's [55] and ImageBind's [23] joint training results are not much better than CLAP [14], our model improves the "AL" discovery by leveraging the video modality and imposing the trimodal consensus constraint.

Fig. 5 plots the precision-recall curves. For the audio-visual curve, our model always has higher precision compared to baselines, especially when recall is low. This is strong evidence of our model learning features of sounding actions, whereas baselines are limited to capturing general audio-visual correspondence—whether action-based or not. We observe a similar trend for audio-language discovery.

Ablations. To study the importance of each loss and the two-stage training, we first ablate the consensus loss in the second stage ("w/o $\mathcal{L}_{\text{consensus}}$" in Table 2), which trains the model contrastively for both stages. The model performance drops significantly, showing that exploring the modality consensus is key to learning how actions sound. We then ablate the contrastive loss in the second stage ("w/o $\mathcal{L}_{\text{contrastive}}$"), which harms performance even more. This suggests that $\mathcal{L}_{\text{consensus}}$ functions like a regularization term that forces the $\mathcal{L}_{\text{contrastive}}$ to learn sounding action embeddings. Lastly, we ablate the two-stage training strategy by removing the align stage ("w/o align-stage"), which optimizes $\mathcal{L}_{\text{MC3}}$ directly; this model fails badly. Aligning embeddings first is critical to making MC3's training stable.

Clustering. To visualize the learned embeddings, we group video embeddings in the test set with agglomerative clustering into 20 clusters. Fig. 6 shows the top 8 examples

	$V \rightarrow A$		$A \rightarrow V$		$L \rightarrow A$		$A \rightarrow L$	
	@5	@10	@5	@10	@5	@10	@5	@10
Random	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CLAP [14]	-	-	-	-	**49.8**	87.6	34.0	67.1
CM-ACC [35]	34.6	63.5	30.9	57.7	-	-	-	-
CMC [55]	36.5	67.9	33.8	63.7	44.1	81.8	32.8	64.3
ImageBind [23]	32.8	61.5	29.7	57.9	42.6	76.5	30.6	60.5
w/o $\mathcal{L}_{\text{consensus}}$	33.9	63.0	30.0	56.1	45.0	84.7	32.9	65.8
w/o $\mathcal{L}_{\text{contrastive}}$	3.3	3.7	6.4	12.5	3.1	4.7	3.3	8.0
w/o align-stage	10.0	19.4	5.9	11.8	11.6	20.9	6.5	12.6
MC3	**38.4**	**72.8**	**34.4**	**66.3**	46.2	**88.5**	**37.5**	**73.8**

Table 3. Sounding action retrieval. We report *Recall @5 and @10* for different query-retrieval modalities. See R@1 results in Supp.

of one cluster (more in Supp.). This cluster clearly captures the sound of water running, despite varying types of camera-wearer movement. Not only does it group videos with similar actions that make this sound, but also it shows the learned embeddings are agnostic of the background (the bathroom example).

6.2. Sounding Action Retrieval

Retrieving a different modality given a video, audio, or description is another useful application, such as adding sound effects to silent videos or retrieving captions for action sounds. To explore this setting, we answer the question "how do different actions sound?" by evaluating the cross-modal retrieval performance of long-tail sounding actions from the same category. Different from the binary classification task above, here we aim to retrieve other examples of the same action.

To do this, we utilize the action groups constructed in Sec. 5 based on verbs and nouns, and only keep groups that have more than two instances of sounding actions (such that there will be at least one true positive to retrieve for each query). We then divide each action group equally into a query pool of 7,559 examples and a retrieval pool of 7,032 examples. Given a query modality M_i of instance A, we compute its distance to other modalities M_j of all instances in the retrieval pool. A retrieval is correct if the retrieved instance B and A belong to the same action group.

Results. Table 3 shows the results for four different query-retrieval modality settings. For audio-visual retrieval, we observe that all models can retrieve video with audio (or audio with video) for similar actions with much higher recall than random chance. Our model strongly outperforms the baselines and ablations, benefiting from modeling the modality consensus explicitly. We also observe that retrieving audio with video is easier than the opposite, likely because audio can be vague sometimes, e.g., a collision sound might occur due to various actions while seeing a cutting action indicates the likely sound. For audio-language retrieval, our

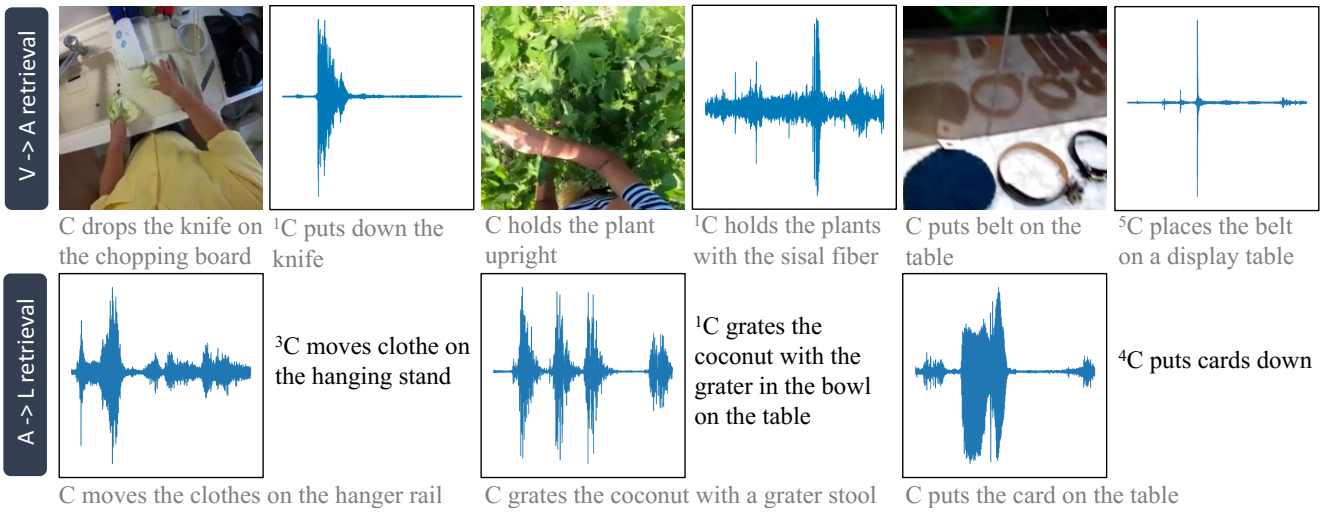

Figure 7. Qualitative examples for retrieval. The first row is video-to-audio retrieval, motivated by adding audio effects for silent videos. The second row is audio-to-text retrieval, motivated by audio captioning applications. For each row, we show three correct retrieval examples along with their text (gray indicates the text is not observed by the model). For the retrieved item, we show the ground truth rank as the superscript. All examples are long-tail sounding actions, showing how our model learns to capture the features of how actions sound.

model similarly outperforms the baselines by large margins.

Qualitative examples. In Fig 7, we show examples for video-to-audio and audio-to-language retrieval. Even though these actions are subtle, our model retrieves audio or captions that are very relevant. Listen to examples in Supp. video.

6.3. Audio Classification on EPIC-Sounds

Finally, we evaluate our learned representation on a standard audio benchmark. To assess the impact of our model's action sounds representation, we consider EPIC-Sounds [28], a challenging audio classification benchmark for sounds in kitchen environments. To our knowledge, EPIC-Sounds represents the only large-scale benchmark for audio in egocentric video. Note, this classification task is different from the sounding action discovery task in Sec. 6.1 in that here the model only takes an audio clip as input.

We consider both linear-probe and fine-tuning settings. In the linear-probe setup, we freeze the model weights and only train the last classification layer, which evaluates the quality of the pre-trained representations. In the fine-tuning setup, we fine-tune both the encoder and the last layer.

Table 4 shows the results. We compare with two SotA methods reported in EPIC-Sounds: SSAST [25] and ASF [31]. SSAST is pretrained on LibriSpeech [44] and shares the *same network architecture* as ours, while ASF is trained on VGG-Sound with supervised learning. With linear-probe, our model strongly outperforms SSAST [25], which, like us, is also pretrained in a self-supervised fashion with no audio labels. ASF [31] does better than both, likely due to its advantage of supervised audio classification pretraining. When fine-tuning, our model outperforms both prior methods in all but one metric when following the same fine-tuning and evaluation protocol. This shows our MC3 audio encoder—trained for sounding action discovery—learns generalizable action sound embeddings, improving the state of the art. The margins are naturally smaller in the fine-tuning regime, as is typical, since all models have time to adapt to the new domain.

		Top-1	Top-5	mCA	MAP	mAUC
Random	-	7.71	30.95	2.29	0.023	0.500
ASF [31]*	L	45.53	79.33	13.48	0.172	0.789
SSAST [25]	L	28.74	64.84	7.14	0.079	0.755
MC3	L	42.44	78.76	12.79	0.153	0.818
ASF [31]*	F	53.75	84.54	20.11	**0.254**	0.873
SSAST [25]	F	53.47	84.56	20.22	0.235	0.879
MC3	F	**55.97**	**85.86**	**21.65**	0.242	**0.885**

Table 4. Results of classification on EPIC-Sounds. L: Linear-Probe; F: Fine-tuning. * denotes pretraining with supervised audio classification while the rest are pretrained in a self-supervised fashion.

7. Conclusion

We explored learning how first-person actions sound from in-the-wild, narrated egocentric videos—without audio labels. Training with 250K clips from Ego4D, we show the promise of our novel multimodal consensus framework for accurately aligning representations to capture the long-tail of sounding actions in novel (unnarrated) videos, with clear impact on sounding action discovery, retrieval, and pre-training for audio classification. In the future, we plan to explore multimodal consensus from asynchronous multimodal streams.

Acknowledgements: UT Austin is supported in part by the IFML NSF AI Institute. KG is paid as a research scientist by Meta.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 2, 3

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020. 3

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[5] Relja Arandjelovi and Andrew Zisserman. Look, listen and learn. In *ECCV*, 2018. 2, 3

[6] Relja Arandjelovi and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2, 3

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 5

[8] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 3

[9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1, 2

[10] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *5th Annual Conference on Robot Learning*, 2021. 1, 3, 6

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3

[12] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006. 6

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[14] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022. 2, 6, 7

[15] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2, 6

[16] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel M. Bear, Dan Gutfreund, David D. Cox, James J. DiCarlo, Josh H. McDermott, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 1, 3

[17] Dhiraj Gandhi, Abhinav Gupta, and Lerrel Pinto. Swoosh! rattle! thump! - actions that sound. In *RSS*, 2022. 1, 3, 6

[18] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 3

[19] Ruohan Gao and Kristen Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 3

[20] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020. 2

[21] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 2

[22] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 1, 2

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 2, 3, 6, 7

[24] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 5

[25] 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*, pages 10699–10709, 2022. 8

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen,

Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 3, 5

[27] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 3

[28] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2023. 1, 2, 6, 8

[29] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *CVPR*, 2022. 3

[30] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[31] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859, 2021. 8

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[33] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 2, 3

[34] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 3, 5

[35] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *ICLR*, 2021. 2, 6, 7

[36] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Audio captioning transformer. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 211–215, Barcelona, Spain, 2021. 3

[37] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. In *NeurIPS*, 2022. 3

[38] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020. 3

[39] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Computer Vision and Pattern Recognition (CVPR), IEEE/CVF Conf. on*, 2021. 3, 4

[40] Arsha Nagrani, Shan Yang, Anurag Arnab, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion, 2021. 2

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[42] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2, 3

[43] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *CVPR*, 2016. 1, 3, 6

[44] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 8

[45] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. 1

[46] Ben Poole, Sherjil Ozair, Aaron van den Oord amd Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019. 4

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 5

[49] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 3

[50] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multimodal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 3

[51] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 3

[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 2012. 2, 6

[53] Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *ECCV*, 2020. 3

[54] A. Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2021. 6

[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2, 3, 6, 7

[56] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arxiv*, 2018. 4

[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society, 2015. 3

[58] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022. 3

[59] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 3