Looking Similar, Sounding Different: Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning

Nikhil Singh[1*] Chih-Wei Wu[2] Iroro Orife[2] Mahdi Kalayeh[2]
[1]Massachusetts Institute of Technology [2]Netflix
[1]nsingh1@mit.edu [2][chihweiw, iorife, mkalayeh]@netflix.com

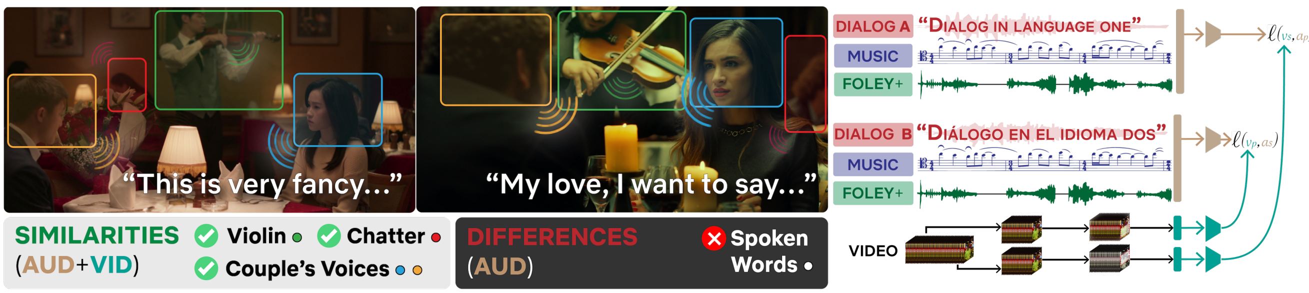

Figure 1. **(Left)** Audiovisual scenes can be perceptually similar even as the words spoken in them differ, which may be a challenge for self-supervised audiovisual representation learning. **(Right)** We propose to leverage movie dubs during training and show that it improves the quality of learned representations on a wide range of tasks.

Abstract

Audiovisual representation learning typically relies on the correspondence between sight and sound. However, there are often multiple audio tracks that can correspond with a visual scene. Consider, for example, different conversations on the same crowded street. The effect of such counterfactual pairs on audiovisual representation learning has not been previously explored. To investigate this, we use dubbed versions of movies and television shows to augment cross-modal contrastive learning. Our approach learns to represent alternate audio tracks, differing only in speech, similarly to the same video. Our results, from a comprehensive set of experiments investigating different training strategies, show this general approach improves performance on a range of downstream auditory and audiovisual tasks, without majorly affecting linguistic task performance overall. These findings highlight the importance of considering speech variation when learning scene-level audiovisual correspondences and suggest that dubbed audio can be a useful augmentation technique for training audiovisual models toward more robust performance on diverse downstream tasks.

1. Introduction

Can two videos look similar while sounding different? Consider the two scenes on the left in Fig. 1. These come from different sources, but share elements like a violinist in the background, other tables further away, and a couple's voices in an upscale restaurant environment; but what are they saying? This can vary considerably between the two scenes, even without changing other aspects. General-purpose self-supervised audiovisual representations are often focused on non-speech applications, evidenced by both existing training datasets and common downstream evaluation tasks. In audio alone, there is a myriad of applications beyond semantic speech processing, leading to recent benchmarks which evaluate generalization across and trade-offs between types of tasks [76, 81]. How then can we focus on learning robust representations from audiovisual content with speech mixed into it? Importantly, there are many non-semantic, or *paralinguistic*, speech processing tasks of interest, as speech is much more than audible text. These too require discovering other similarities beyond words.

Imagine a movie discussion scene, as in Fig. 2. Many audiovisual elements are present: background chatter, glasses clinking, music, footsteps, and characters' voices, but *a priori* this scene could contain many different dialogs without changing the fundamental scene attributes, beyond local

*Most of the work conducted during author's internship at Netflix.

features such as lip movements, and this indicates an explicitly counterfactual structure. Note that there are also other counterfactual cross-modal structures which relate to different problems, such as multiple videos of dancing to the same music. Differences in spoken words are one specific case of this which we explore.

In this work, we hypothesize that this *looking similar, while sounding different* problem, as it can occur in real-world audiovisual data distributions, may inhibit the performance of self-supervised audiovisual representation learners. Established approaches, such as cross-modal contrastive learning, where models learn to discriminate true audiovisual pairs from false ones, could be affected; linguistically different but otherwise similar audio-video pairs could act as confounders in this case. However, counterfactual versions of exactly the same scene with only different dialog are generally not available, even if the distribution of real-world audiovisual scenes exhibits this overall trend.

We propose to leverage a data source which naturally resembles this counterfactual-like structure as a proxy: *dubs*. Dubs are alternate versions of movie audio tracks where the speech is replaced with a second-language adaptation, and the rest of the sounds are generally unchanged. Recent works have shown how training on movie scenes can yield strong performance [13, 37], since they contain diverse audiovisual mixtures, compared with popular audiovisual datasets which are curated to focus on specific objects or actions. Although this distribution may help in learning representations focused on overall scene attributes rather than the dialog's semantics, which is our goal, contrastive training on aligned audio and video from movies does not explicitly account for scenes that look similar and sound different due to linguistic variation. We improve upon this strategy by leveraging multilingual dubbed versions of movies[1]. Specifically, we create a dataset of movies and television shows, each with up to seven audio tracks: English (**EN**), Spanish (**ES**), French (**FR**), Japanese (**JA**), German (**DE**), Italian (**IT**) and Korean (**KO**). We plug our training strategy into a well-established self-supervised contrastive learning formulation, *i.e.* SimCLR [14], and we show that this can improve performance in both multimodal and unimodal setups. Overall, this work contributes:

- An approach to improving self-supervised audiovisual representation learning using *dubs*, secondary audio language versions of movies.
- Extensive experiments showing that this approach not only improves performance on a range of auditory and audiovisual tasks but also yields new state-of-the-art on multiple benchmarks.
- Additional experiments to investigate potential trade-offs. These show that we can get an improvement without majorly affecting the performance on language identification, and semantic speech tasks.
- An example pipeline for producing counterfactual pairs in various languages; we apply the workflow to the LVU [83] dataset and demonstrate the possibility of creating alternate audio tracks that potentially empower the research community to further investigate the impact of spoken words in audiovisual representation learning.

[1]The pretraining data also includes episodes of television shows. To avoid clutter, we refer to all long-form content as movies unless it is necessary to specify.

Figure 2. Consider the pictured scene. Which of these dialog examples is more likely? Both are plausible within the scene, yet their phonetic-acoustic characteristics would create differences in the soundtrack.

2. Related Work

Self-supervised and Multimodal Learning Self-supervised learning relies on pretext tasks with engineered supervision based on data structure, rather than human labels, to learn useful representations [4, 20, 30, 32, 33, 47, 54, 88]. We focus on *contrastive learning*, which has shown strong performance by maximizing mutual information between views of the same instance [4, 14, 26, 33, 74, 74, 75]. These can then be adapted to novel tasks by fine-tuning, or by appending simple (often linear) models, both with smaller-scale task-specific supervision requirements. *Cross-modal* contrastive learning specifically leverages multimodal data like image and text [61], or, as in our case, video and audio [1–3, 23, 34, 41, 45, 51, 52, 56, 58, 59, 80, 85, 87].

Audiovisual Learning Audiovisual learning harnesses cross-modal correspondences for tasks like action [39, 41] and speaker [16, 50] recognition, source separation [9, 63, 78], media synthesis [25, 31, 57, 72], audio spatialization [27, 48, 86], acoustic simulation [12, 46, 69], and more. Much work takes a contrastive approach, recognizing that audio and video can be treated as two complementary sensory views of a single underlying phenomenon, and focuses on learning *coordinated* [5] representations. Prior work has found that cross-modal training can lead to better results than within-modal training [49], so we use this cross-modal setup as the basis for our framework. In this work, we rely on multilingual audio dubs and videos from long-form content, e.g. movies and television shows. Movies contain rich audiovisual correspondences mimicking real-world experiences, and are more diverse and novel than user-generated videos while being abundant and scalable [13, 35, 73].

General-purpose Audio Representation Learning and Evaluation Sound is heterogeneous, with speech, music, and environmental sounds having very different characteristics. Even within speech, for example, tasks like speech recognition [15, 44] and speech emotion recognition [68] differ dramatically. This has motivated developing general-purpose audio representations [53, 65] and benchmarks like HARES [81] and HEAR [76]. We focus our audio evaluation on HEAR [76] since it provides a consistent API. The central hypothesis is that if dub-augmented training in the cross-modal setting improves the generality of the representations, performance on various tasks should increase while avoiding a significant trade-off on language-related tasks.

Multilingual Audio Multilingual speech processing has enabled progress in areas like speech recognition [7] through pretraining on diverse data [17, 29]. Recently, speech-to-speech translation has been possible as well [43]. Speech translation in audiovisual media is often referred to as *dubbing*. This is a type of audiovisual translation [11] in which speech content from a media artifact (*e.g.* a movie) is re-recorded in another language. Dubs predominate over subtitles in many cultures [10]. This provides naturalistic multilingual data at scale, and offers a specific case for our hypothesis about audio-visual consistency: a dub's soundtrack differs from the original only in spoken language. We seek to leverage dubs' parallel primary and secondary audio, differing only in speech, to learn more robust audiovisual representations. We also produce a synthetic pipeline for creating counterfactual pairs, to demonstrate the concept of counterfactual cross-modal pairs, while enabling future exploration and validation from the research community.

3. Pretraining Dataset

Our dataset consists of ∼20K movies and ∼33K television episodes, which constitutes ∼59K video-hours in total. We have paid extra attention to the diversity of titles used in our pretraining dataset in order to minimize the potential implicit biases in our learned representations, and limited ourselves to only a small part of the catalog to investigate this question. Fig. 3 provides details on the distribution of genre, and original language of the titles included in our dataset[2]. Each title contains a video track, as well as up to seven audio tracks: English (**EN**), Spanish (**ES**), French (**FR**), Japanese (**JA**), German (**DE**), Italian (**IT**) and Korean (**KO**). Most titles have only a single audio track, which is almost always their original language while about a quarter of the dataset is multilingual where on average 2.8 audio tracks are available for each title. Such a dataset allows us to explore the impact of spoken words in audio for self-supervised audiovisual representation learning. Having multiple dub options enables us to investigate trade-offs between secondary languages, and whether "multilingual" models might further strengthen downstream performance.

[2]Further details are given in the supplementary material.

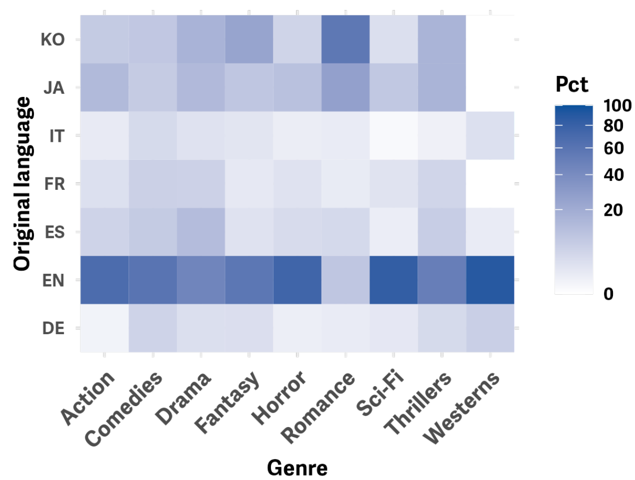

Figure 3. Movies and television episodes included in our pretraining dataset are chosen from a diverse set of original languages and genres. Our goal is to minimize potential content and story biases that could potentially impact our self-supervised models. Note that beyond curating the dataset, we do not use this metadata for representation learning.

We recognize that this kind of data has the potential to significantly benefit research. We are actively investigating the necessary legal steps to potentially release a variant of it for non-commercial use. Fig. 4 illustrates a few samples but readers are encouraged to check out our supplementary material for more examples[3].

4. Methodology

4.1. Approach

Our pretraining dataset is denoted by $\mathcal{X} = \{\mathcal{X}_n | n \in [1 \cdots N]\}$, where $\mathcal{X}_n = \{x_{n,m} | m \in [1 \cdots M_n]\}$ contains M_n non-overlapping snippets which are temporally segmented from the duration of the n^{th} title in the dataset. $\mathcal{Q}$ is a function class which we use to create quadruplet training instances $(v_p, a_p, v_s, a_s) \sim \mathcal{Q}(x_{n,m})$[4] where v_p and v_s are obtained through spatio-temporal augmentation of video modality in $x_{n,m}$. Similarly are a_p and a_s for the audio modality, yet, unlike video, we do have the opportunity to further add dub-augmentation to audio instances. When more than one language is available this would ensure that a_p and a_s are similar except in their spoken language.

[3]nikhilsinghmus.github.io/lssd

[4]subscripts stand for primary and secondary

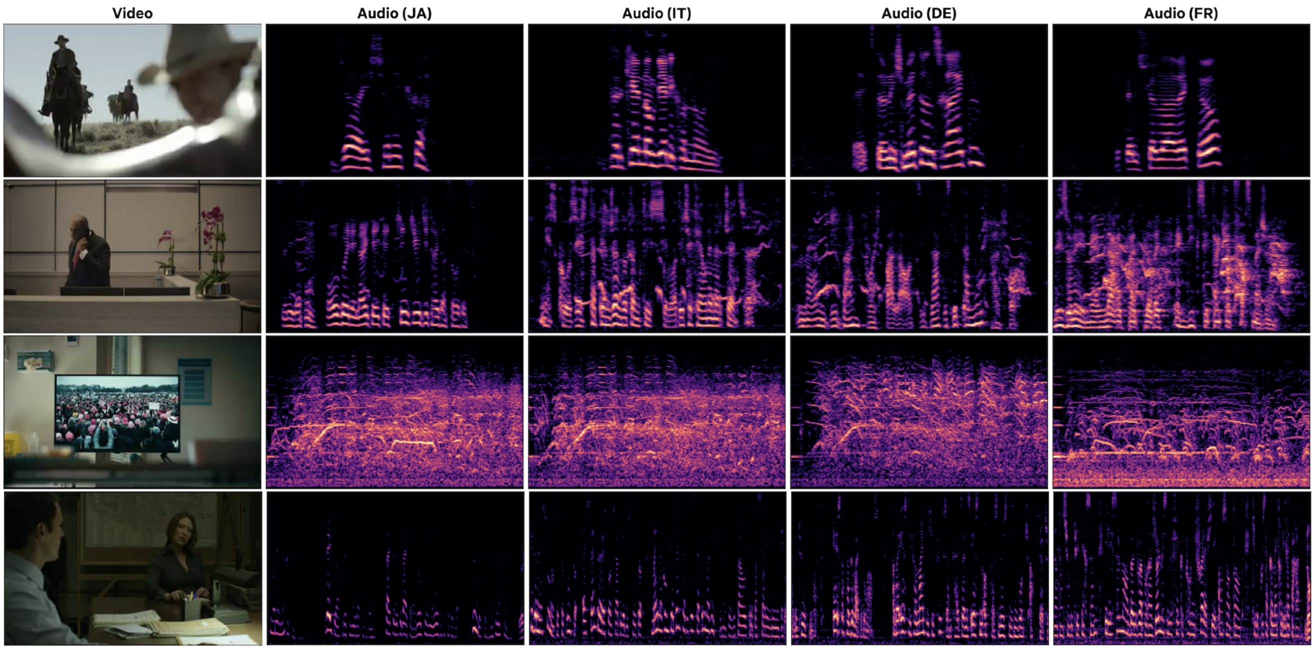

Figure 4. Example clips from our pretraining dataset, showing video stills and mel spectrograms for each of the audio tracks.

Randomly sampling negatives, the traditional approach in metric and contrastive learning, has been observed to be suboptimal [45, 67]. A number of recent works develop methods for so-called *hard negative mining*, where the goal is to populate the negative set with challenging examples [55, 64]. In our case, the data is hierarchical; snippets are naturally nested within source long-form titles, and those from the same title share several common attributes including characters, places, objects, voices, and aesthetics. Hence, following prior work [37], to create a mini-batch $\mathcal{B} = \{x_i | i \in [1 \cdots B]\}$, we first uniformly sample a title, $n \sim \mathbb{U}(1, N)$, and then draw multiple distinct snippets from $\mathcal{X}_n$. This ensures that for each instance in $\mathcal{B}$, there are always a sufficient number of samples from the same title to act as hard negatives. This is important since $B \ll N$, hence for $n \sim \mathbb{U}(1, N)$ and $m \neq m'$, $\mathsf{P}(x_{n,m} \in \mathcal{B} \wedge x_{n,m'} \in \mathcal{B}) \rightarrow 0$. In other words, the naive random sampling policy of $x_i \sim \bigcup_{n=1}^{N} \mathcal{X}_n$ would mainly lead to easy cross-title negatives.

We can now formulate the training objective. Considering a cross-modal setup, $\mathcal{B} = \{(v_i, a_i) | i \in [1 \cdots B]\}$ represents a minibatch of size B, where video and audio modalities of the i^{th} instance are denoted by v_i and a_i. We use z_v^i and z_a^i to represent their respective embeddings. For the i^{th} element in the minibatch, (z_v^i, z_a^i) serves as the positive pair, while assuming negative pairs for both modalities, $\mathcal{N}_i = \{(z_v^i, z_a^j), (z_v^j, z_a^i) | j \in [1 \cdots B], i \neq j\}$ constitutes the set of negative pairs. With that, Equation 1 shows the cross-modal normalized temperature-scaled cross-entropy objective [14] associated with the i^{th} instance. Since $(v, a) \in \{(v_p, a_s), (v_s, a_p)\}$, in practice we optimize Equation 2 which aggregates over all available instances.

$$\ell_i(v, a) = -\log \left(\frac{e^{((z_v^i)^\intercal (z_a^i))/\tau}}{e^{((z_v^i)^\intercal (z_a^i))/\tau} + \sum_{(z'_v, z'_a) \in \mathcal{N}_i} e^{((z'_v)^\intercal (z'_a))/\tau}} \right) \tag{1}$$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{B} \Big(\ell_i(v_p, a_s) + \ell_i(v_s, a_p) \Big) \tag{2}$$

$$\mathcal{L}_v = \sum_{i=1}^{B} \ell_i(v_p, v_s), \quad \mathcal{L}_a = \sum_{i=1}^{B} \ell_i(a_p, a_s) \tag{3}$$

Equation 3 shows the within-modal variants of the loss function for video and audio modalities. Unless explicitly mentioned otherwise, we train our models from scratch and *cross-modally*, *i.e.* we compute the contrastive loss between modalities as shown in Eq. 2. We do this based on the observation in our early experiments that, when training from scratch without tuning additional scaling parameters, the within-modal contrastive task is too easy comparatively and results in early convergence on the corresponding terms. This approach is also supported by prior literature [49]. Despite not directly optimizing for within-modal terms, we track $\mathcal{L}_v$ and $\mathcal{L}_a$ during self-supervised pretraining and observe that they diminish as a byproduct of minimizing $\mathcal{L}$. There are variants in our modeling where $\mathcal{L}_v$ and

$\mathcal{L}_a$ are included in total loss function (*e.g* $\mathcal{L}+\lambda_v\mathcal{L}_v+\lambda_a\mathcal{L}_a$) which we'll discuss later in Sec. 5.2.

4.2. Architecture

As we seek to validate the effect of our data and training approach, we rely on standard backbone architectures. Our video model is a multi-scale vision transformer [22], specifically MViT-S, and our audio model follows a similar architecture except a slight modification to allow processing audio spectrograms as input. Note that we train all our models from scratch on our pretraining dataset detailed in Sec. 3. We use a single (weight sharing) audio backbone which processes all audio spectrograms, regardless of language. As is common in contrastive learning, we use multi-layer perceptron (MLP) projection heads, one for each modality, to further reduce the dimensionality of representations during training, prior to computing the contrastive loss. These additional layers are discarded after pretraining.

5. Experiments

5.1. Downstream Tasks

Audio Tasks and Benchmarks We evaluate on a diverse set of auditory tasks to probe the quality of our learned representations, taken from the HEAR [76] challenge benchmark. We subselect tasks relevant to our hypotheses, and focus on those which use pooled (rather than temporally dense) representations.

Sound and Scene Classification: These tasks are firmly non-linguistic, and we hypothesize performance on them should benefit from de-emphasizing language in training. We include ESC-50 [60], FSD50K [24], and Vocal Imitations (VI) [40]. VI is a query-by-vocalization (QBV) task, however since it is based on AudioSet [28] ontology sound events, we place it in this category. *Non-Semantic Speech*: Many non-semantic or *paralinguistic* attributes of speech or vocal signals may be shared between languages, and such signals are important for a range of tasks. We include here CREMA-D [8] for emotion recognition, GTZAN [77] for music/speech discrimination, and LibriCount [71] for speaker count estimation. We hypothesize performance should improve, if our scheme increases focus on non-linguistic speech attributes. *Semantic Speech*: To probe a potential trade-off, we evaluate on semantic speech tasks. We consider keyword understanding as a proxy for speech recognition that uses pooled representations. To do so, we employ the *full* version of Speech Commands [82] implemented in HEAR [76]. *Language*: Another way to measure a possible trade-off is by evaluating how models perform on an audio-based language identification task, to see if features useful for this are preserved in learned representations. We include VoxLingua107 Top10 [79] for this reason.

	# data	init.	(λ_v, λ_a)	original language	avg. # dubs	dub augment
A.1	4.6M	rand.	(0,0)	**ESF**	2.8	✗
A.2	4.6M	rand.	(0,0)	**ESF**	2.8	✓
A.3	4.6M	A.1	(0,0.2)	**ESF**	2.8	✗
A.4	4.6M	A.2	(0,0.2)	**ESF**	2.8	✓
B.1	11.8M	rand.	(0,0)	**EN**	1.0	✗
B.2	9.8M	rand.	(0,0)	$\mathbb{U}\backslash$**EN**	0.2	✗
B.3	19.4M	rand.	(0,0)	$\mathbb{U}$	0.6	✗
B.4	5.1M	B.3	(0,0)	$\mathbb{U}$	2.8	✓
B.5	5.1M	B.3	(0.2,0.2)	$\mathbb{U}$	2.8	✓
C.1	19.4M	rand.	(0,0)	$\mathbb{U}$	0.6	✓
C.2	5.1M	C.1	(0.1,0.1)	$\mathbb{U}$	2.8	✓

Table 1. Details of different pretraining model variants. Here, **ESF** := {**EN**, **ES**, **FR**} is denoting the union of three languages. $\mathbb{U}$ represents the universal set including all the seven languages.

Visual and Audiovisual Tasks We also evaluate the visual representations independently, and coordinated with, the auditory representations. Following recent work on representation learning from long-form content [13], we include the LVU [83] benchmark covering various aspects of long-form video understanding to our evaluation suite. LVU [83] contains small-scale tasks covering a wide range of aspects of long-form videos, including content understanding (*relationship*, *speaking style*, *scene/place*), and movie metadata prediction (*director*, *genre*, *writer*, *movie release year*). Among the LVU tasks, we explore benefits and potential trade-offs using both visual and auditory representations. In general, we expect improvement except for *speaking style*, where it is not *a priori* clear whether de-emphasizing spoken words during pretraining is harmful for such a downstream task.

Evaluation Once the self-supervised pretraining is over, we discard the projection heads and use the backbone architectures to extract features from audio and video assets. Unless mentioned otherwise, we do spatio-temporal mean pooling on the output tensors in order to obtain a d-dimensional vector embedding for each data instance in the downstream tasks. We then train either an MLP or linear probe on these representations following the prescribed approaches in the relevant benchmarks. More implementation details can be found in the supplementary material.

5.2. Models

In total, we train **11** model variants, detailed in Table 1, and evaluate them on **15** different tasks across audio and video modalities.

First (A) group of model variants demonstrates a small-scale multilingual pretraining regime, as a first study of the impact of dub-augmentation. We sample English (**EN**),

	HEAR							LVU						
	ESC	LibCnt	CREMA	VI	FSD	Speech	VoxLng	Director	Genre	Relation	Scene	Speak	Writer	Year
A.1	77.20	67.29	59.52	10.37	44.52	74.83	27.16	44.86	54.42	36.59	**45.12**	42.86	**38.10**	41.84
A.2	75.95	67.94	59.76	11.14	44.23	73.80	23.87	47.66	56.63	36.59	41.46	40.74	33.33	41.84
A.3	82.00	67.87	**62.69**	11.39	48.90	**79.47**	**28.70**	**49.53**	57.65	43.90	39.02	43.92	33.93	46.10
A.4	**83.05**	**68.65**	61.95	**12.57**	**49.42**	74.38	26.55	44.86	**59.01**	**46.34**	**45.12**	**48.15**	29.17	**47.52**
B.1	84.15	67.12	61.00	**13.05**	50.29	82.31	24.69	47.66	57.14	51.22	41.46	42.33	32.14	**45.39**
B.2	82.00	67.10	61.98	11.86	49.07	82.90	28.09	42.99	55.95	48.78	42.68	47.62	30.36	44.68
B.3	**85.60**	66.31	62.79	11.55	**53.69**	**83.82**	**30.35**	50.47	**60.20**	46.34	42.68	48.68	37.50	**45.39**
B.4	83.75	68.88	63.18	10.82	51.61	77.12	28.19	**51.40**	59.69	**56.10**	46.34	**49.21**	**38.10**	44.68
B.5	85.25	**69.16**	**63.27**	11.38	52.48	76.99	27.98	**51.40**	58.33	51.22	**52.44**	48.68	36.31	**45.39**
C.1	84.10	67.57	63.70	**12.12**	51.96	**81.88**	29.42	42.99	**58.84**	48.78	46.34	41.27	38.69	41.13
C.2	**85.50**	**68.90**	**64.28**	11.90	**52.55**	77.14	**29.94**	**48.60**	57.65	48.78	**51.22**	**50.79**	**39.88**	**49.65**

Table 2. **Ablation results with audio.** All metrics are top-1 accuracy, except for FSD50K [24] and Vocal Imitation [40] (Mean Average Precision). We have followed the prescribed evaluation strategy from HEAR [76] benchmark; training an MLP on frozen embeddings of the downstream tasks. For LVU [83], we use the official data splits and train a linear probe. Results are shown on the test split where the best epoch to report is chosen based on the same metric on the validation set. All model variants obtained 100.0 top-1 accuracy on GTZAN, hence we did not include that task here. We denote the top performance(s) within each ablation group with **bold**. The HEAR [76] tasks from left to right are ESC-50, LibriCount, CREMA-D, Vocal Imitation, FSD-50k, SpeechCommands (Full), and VoxLingua107 Top10.

	HEAR									LVU						
	ESC	LibCnt	CREMA	VI	FSD	Speech	VoxLng	GTZAN		Director	Genre	Relation	Scene	Speak	Writer	Year
Bench [76]	96.65	78.53	75.21	22.69	65.48	97.79	72.02	99.23	Obj Tr [83]	58.90	56.10	54.70	60.00	40.30	35.10	40.60
Bench (SSL)	80.50	78.53	75.21	18.48	50.88	96.87	71.40	96.86	M2S [13]	70.90	55.90	71.20	68.20	42.20	53.70	57.80
GURA [84]	74.35	68.34	75.21	18.48	41.32	94.68	71.40	93.59	ViS4mer [36]	62.61	54.71	57.14	67.44	40.79	48.80	44.75
PaSST [42]	94.75	66.01	61.04	18.20	64.09	63.87	25.93	97.69	SCALE [66]	49.09	58.97	76.47	74.02	42.27	62.76	39.23
CLAP [21]	96.70	77.83	64.36	–	58.59	96.83	–	100.0	STCA [19]	66.70	56.62	59.25	69.15	41.62	52.93	53.30
Ours																
B.3 (A)	85.60	66.31	62.79	11.55	53.69	83.82	30.35	100.0	B.3 (V)	69.16	60.88	60.98	63.41	46.03	48.81	52.48
B.4 (A)	83.75	68.88	63.18	10.82	51.61	77.12	28.19	100.0	B.4 (V)	67.29	61.73	60.98	65.85	47.62	41.67	55.32
B.5 (A)	85.25	69.16	63.27	11.38	52.48	76.99	27.98	100.0	B.5 (V)	69.16	64.29	58.54	64.63	46.03	41.07	52.48

Table 3. State-of-the-art results across HEAR [76] (adding GTZAN Music/Speech) and LVU [83] tasks we evaluate on. On HEAR, we compare to (1) the best result on each task, on the HEAR leaderboard, (2) same as (1) but considering only self-supervised models, (3) GURA Fuse HuBERT [84], the best performer on average, (4) CP-JKU PaSST 2lvl+mel [42], the strongest average performer after the GURA models, (5) the recent CLAP model [21]. On LVU, we compare to the Object Transformer from the original LVU paper [83], along with recent advances: ViS4mer [36], the SVT SCALE model [66], STCA [19], and Movies2Scenes [13]. Movies2Scenes uses movie metadata, which introduces task-specific supervision. When reporting our results, (A) indicates audio representations only, and (V) means video representations only.

Spanish (**ES**), or French (**FR**) titles which have at least one dub available, so we can systematically study the effect of dub-augmentation. For each title, we sample dubs from *all* seven total languages. **A.3** and **A.4** variants incorporate an explicit within-modal term, *i.e* $\mathcal{L}_a$. We hypothesize that, with dub-augmentation, $\lambda_a > 0$ may yield a broader gap on linguistic and language identification tasks. This is because the optimization explicitly maximizes the similarity of audio embeddings that are only different in their spoken language, rather than just implicitly through $\mathcal{L}$. Importantly, the total number of pretraining steps is the same for **A.3** and **A.4** , similarly when one compares **A.1** and **A.2** .

Second (B) group of model variants aims at understanding the impact of data scale and language diversity. We approximately double the number of pretraining instances compared to experiments in group **A** and study whether this leads to higher quality representations. This is important since self-supervised pretraining is computationally expensive and it is not clear *a priori* if bigger and more diverse pretraining data necessarily leads to better models. **B.3** is trained on all pretraining instances including all languages to test the limit of multilingual pretraining *without* dub-augmentation. By comparing **B.4** and **B.5** , we hope to shed light on the behavior of the within-modal objective function which the latter uses.

Third (C) group of experiments explore the impact of deeper architectures, namely MViT-B [22] (vs MViT-S [22] as our default). We keep the data scale and diversity the

	Director	Genre	Relation	Scene	Speak	Writer	Year
A.1	53.27	54.59	43.90	52.44	34.39	36.90	42.55
A.2	53.27	55.44	41.46	50.00	**41.27**	35.12	42.55
A.3	57.01	**57.48**	**46.34**	**57.32**	39.68	**38.69**	46.10
A.4	**63.55**	**57.48**	36.59	53.66	36.51	33.93	**47.52**
B.1	60.75	55.78	**48.78**	53.66	38.10	35.71	42.55
B.2	54.21	57.65	46.34	51.22	37.04	**38.69**	44.68
B.3	**65.42**	57.48	41.46	53.66	39.68	38.10	45.39
B.4	62.62	**58.50**	36.59	**59.76**	**43.39**	35.12	46.81
B.5	62.62	58.16	43.90	**59.76**	39.15	37.50	**49.65**
C.1	**63.55**	55.10	43.90	57.32	**40.74**	**39.29**	**45.39**
C.2	61.68	**56.63**	**46.34**	**60.98**	40.21	36.90	43.97

Table 4. **Ablation results with video**. All metrics are top-1 accuracy. We have followed prescribed data split from LVU benchmark and trained a linear probe on frozen **video** embeddings of the downstream tasks. Results are shown on the test split where the best epoch to report is chosen based on the validation set. We denote the top performance within each ablation group with **bold**.

same as in the **B.3** , **B.4** and **B.5** variants. Similarly to these, here we initially train on the entire data, then fine-tune from the final checkpoint of **C.1** only on a subset of titles which have more than one audio tracks. This ensures that dub-augmentation is present in every optimization step of **C.2** .

We are now set to comprehensively study various aspects of multilingual and multimodal representation learning, thanks to a wide variety of pretrained models and downstream tasks across audio and video modalities.

5.3. Ablation Study

Does dub-augmented pretraining help? To address this, we start by looking at the **first (A)** group of model variants in Table 2. We've hypothesized that dub-augmentation should improve the performance on sound/scene classification and non-semantic speech tasks. On the HEAR [76] benchmark, with the exception of CREMA-D [8], our quantitative results confirm this. LVU [83] tasks are also considered non-linguistic and Table 2 shows that, in most of them, dub-augmented variants lead to large performance gains over their baseline counterparts. Our second hypothesis was that dub-augmentation should impact linguistic and language identification tasks as it aims at diminishing the influence of spoken words in audio representations. Indeed, we can see **A.4** which utilizes dub-augmentation is underperforming **A.3** on Speech Commands and VoxLingua. Table 2 also suggests that dub-augmentation benefits from within-modal objective *i.e.* $\mathcal{L}_a$, and for this approach to be effective, we actually need as expected, sufficient number of instances with alternative audio tracks during pretraining.

Can dub-augmented models still recognize language and conduct linguistic tasks? Results shown in Table 2 on VoxLingua demonstrate that enforcing dub-augmentation in both small (**A** variants) and large-scale (**B** variants) regimes clearly affects language identification performance. We measure this by comparing **A.2** vs. **A.1** , or **B.4** vs. **B.3** . We observe similar behavior for Speech Commands, our proxy for linguistic performance implemented as keyword spotting. However, in both cases, the degradation is not large enough to prevent dub-augmented models from recognizing language or conducting linguistic tasks. We hypothesized this modeling trade-off, *i.e.* that while performance might reduce, the significance of this would be limited.

Is the quality of video representations impacted? To answer this, we look at Table 4 where LVU tasks are evaluated via a linear probe on frozen video embeddings. In the small-scale pretraining regime, we observe a mixed pattern where dub-augmented variants, *i.e.* **A.2** and **A.4** , outperform their counterparts in 3 tasks ("Director", "Speaking Way", and "Year") while being either worse or on par on the rest. In the large-scale pretraining regime, we see a more clear trend where **B.4** and **B.5** show improvements over **B.3** in 5 out of 7 LVU tasks demonstrating that on a diverse evaluation set, dub-augmented pretraining is overall helpful to even video-only tasks.

How does language diversity influence pretraining? Properly addressing this research question demands a closer look at **B.1** , **B.2** , and **B.3** . It is worth reiterating that despite a different number of pretraining instances (see Table 1), we have trained all three of these model variants with approximately the same number of gradient optimization steps to establish a fair comparison. In general, across both audio (ref. Table 2) and video (ref. Table 4) we observe performance gains when we maximize language variation (ref. **B.3**). However, the inclusion of English (**EN**) language titles, as our most dominant original language (see Fig 3), during pretraining seems to be crucial. Table 2 illustrates a clear pattern for VoxLingua and Speech Commands, where greater language diversity during pretraining leads to significant gains *e.g.* absolute 5.6% on VoxLingua.

Is a deeper architecture better? For each task in Tables 2 and 4, we can compare the strongest **B** model variants against **C** variants. With a few exceptions, our quantitative results do not indicate that using MViT-B [22] with $\sim$40% more parameters provides a meaningful boost over the smaller MViT-S [22] to justify the significant additional computation during pretraining. We acknowledge that this conclusion might not have held if downstream tasks where evaluated by fine-tuning (instead of probing), especially for large-scale tasks in HEAR [76].

Additional Experiments In the supplementary material, we provide additional results on a small dubbed audiovisual dataset with matched smaller backbone architectures, where we have exact parity between four languages (over 700 **EN** titles with all of **ES**, **FR**, and **JA** available). We also

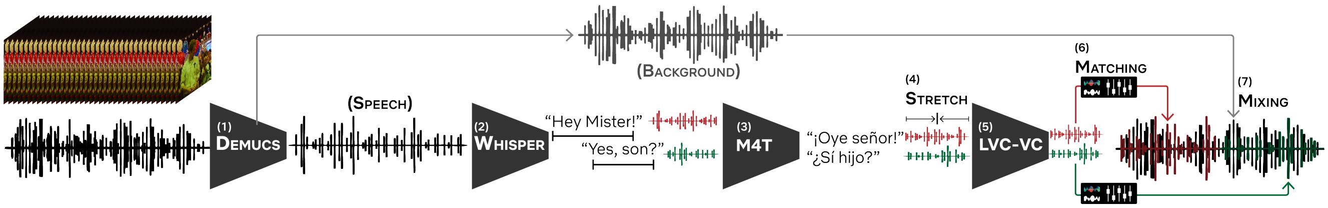

Figure 5. Pipeline to produce the synthetic counterfactual pairs.

compare to a speech-removal strategy, where we source-separate the full dataset and remove the speech part as an alternate strategy for de-emphasizing the speech. Since we have language parity, we also evaluate "bilingual" models with specific dub-augmentation pairs (e.g. **EN+ES**). These results show systematically that dub-augmented training is beneficial even in this smaller-scale setup, that it outperforms the speech removal strategy, and that multilingual models (with multiple dubs, randomly sampled as in our main results here) can add further robustness.

5.4. Comparison with State-of-the-Art

HEAR Table 3 compares our results to several strong results on HEAR [76] tasks. On ESC-50, FSD50K, and GTZAN Music/Speech, our results beat the top self-supervised result on the HEAR Leaderboard and at least one more result. On most tasks (except Vocal Imitation), we beat at least one of the models, showing robustness across these different tasks.

LVU Also in Table 3, we compare our strongest models with state-of-the-art results on 7 LVU [83] tasks. Our models achieve new state-of-the-art performance on the *Genre* and *Speak* tasks, showing substantial improvements over prior results. Without considering Movies2Scenes [13], which uses movie metadata, we also get state-of-the-art results on *Director* and *Year* (4/7 total). On the remaining tasks, our results are highly competitive. This demonstrates that models pretrained on our dataset with dub-augmentation can match or exceed the performance of the best available models on a diverse range of video understanding benchmarks. Overall, these results highlight the effectiveness of our approach.

6. Synthetic Counterfactual Pairs

To encourage the study of counterfactual pairs in audiovisual representation learning, we propose a modular pipeline, shown in Fig. 5, for simulating dub-like counterfactual pairs that are similar to the one-to-many audio-visual distribution from our pretraining data on arbitrary target clips. The proposed pipeline, while being limited in terms of the synthetic quality, serves as a simple tool to alleviate the data constraint for the research community when conducting a similar study.

The steps are (1) Isolate speech from background sounds using Demucs [18], (2) Transcribe and segment the speech using Whisper [62], producing timestamped segments (3) Translate speech (or, optionally, text) into the target language(s) with SeamlessM4T [6] (4) Align translations to original segments using stretching (5) Convert voices to match original actors' using LVC-VC [38] (6) Loudness-normalize and EQ-match the output with the original using Pyloudnorm [70] and matching[5] (7) re-place segments into their original locations, remix with background audio, and mux with original videos. The pipeline also implements other intermediate steps, such as resampling, to bridge between the main steps.

As a proof-of-concept resource for the community, we use this pipeline to produce a multilingual version of LVU [83]. LVU-M demonstrates the feasibility of generating counterfactual data at scale (examples included in the supplementary material). We open-source the pipeline to enable creating such *"looking similar, sounding different"* datasets. We also hope that future advancements can improve the quality and enable deeper research of such data structure.

7. Conclusion

In this work, we introduced the *looking similar, while sounding different* problem, wherein perceptually similar scenes can have different speech content. We showed we can leverage a similarly structured counterfactual data source, dubbed movies, to improve audiovisual representation learning in a well-established cross-modal contrastive learning scheme. Our experiments with a large pretraining dataset of movies and television shows demonstrated that this improves performance across a range of auditory and audiovisual tasks. Dub-augmented training is, as such, a scalable and effective approach for learning more robust audiovisual representations without supervision.

References

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

[5] https://github.com/sergree/matchering

in Neural Information Processing Systems, pages 9758–9770, 2020. 2
[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.
[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 2
[4] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019. 2
[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2
[6] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023. 8
[7] William Byrne, Peter Beyerlein, Juan M Huerta, Sanjeev Khudanpur, Bhaskara Marthi, John Morgan, Nino Peterek, Joe Picone, Dimitra Vergyri, and T Wang. Towards language independent acoustic modeling. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, pages II1029–II1032. IEEE, 2000. 3
[8] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 5, 7
[9] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021. 2
[10] Frederic Chaume. The turn of audiovisual translation: New audiences and new technologies. *Translation spaces*, 2(1): 105–123, 2013. 3
[11] Frederic Chaume. *Audiovisual translation: dubbing*. Routledge, 2020. 3
[12] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2
[13] Shixing Chen, Chun-Hao Liu, Xiang Hao, Xiaohan Nie, Maxim Arap, and Raffay Hamid. Movies2scenes: Using movie metadata to learn scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6535–6544, 2023. 2, 3, 5, 6, 8
[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 4
[15] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE, 2021. 3
[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2
[17] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020. 3
[18] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019. 8
[19] Ali Diba, Vivek Sharma, Mohammad Arzani, Luc Van Gool, et al. Spatio-temporal convolution-attention video network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869, 2023. 6
[20] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2
[21] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6
[22] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 5, 6, 7
[23] Haytham M Fayek and Anurag Kumar. Large scale audio-visual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020. 2
[24] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021. 5, 6
[25] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, pages 758–775. Springer, 2020. 2
[26] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP (1)*, 2021. 2
[27] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021. 2
[28] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[29] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. Multilingual training of deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7319–7323. IEEE, 2013. 3

[30] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[31] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2

[32] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 2

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[34] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[35] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pages 709–727. Springer, 2020. 3

[36] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 6

[37] Mahdi M Kalayeh, Shervin Ardeshir, Lingyi Liu, Nagendra Kamath, and Ashok Chandrashekar. On negative sampling for audio-visual contrastive learning from movies. *arXiv preprint arXiv:2205.00073*, 2022. 2, 4

[38] Wonjune Kang, Mark Hasegawa-Johnson, and Deb Roy. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions. In *Proc. INTERSPEECH 2023*, pages 2303–2307, 2023. 8

[39] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[40] Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology. In *DCASE*, pages 148–152, 2018. 5, 6

[41] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2

[42] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021. 6

[43] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021. 3

[44] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE, 2020. 3

[45] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020. 2, 4

[46] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *arXiv preprint arXiv:2206.04006*, 2022. 2

[47] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[48] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 2

[49] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2, 4

[50] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 2

[51] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020. 2

[52] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 2

[53] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 3

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[55] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 4004–4012, 2016. 4

[56] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2

[57] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 2

[58] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[59] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021. 2

[60] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015. 5

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[62] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 8

[63] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, 2014. 2

[64] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 4

[65] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021. 3

[66] Sepehr Sameni, Simon Jenni, and Paolo Favaro. Spatio-temporal crop aggregation for video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5674, 2023. 6

[67] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019. 4

[68] Ben Shneiderman. The limits of speech recognition. *Communications of the ACM*, 43(9):63–65, 2000. 3

[69] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295, 2021. 2

[70] Christian J Steinmetz and Joshua Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021. 8

[71] Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. Libricount, a dataset for speaker count estimation, 2018. 5

[72] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33:3325–3337, 2020. 2

[73] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 3

[74] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2

[75] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019. 2

[76] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. Hear 2021: Holistic evaluation of audio representations. *arXiv preprint arXiv:2203.03022*, 2022. 1, 3, 5, 6, 7, 8

[77] George Tzanetakis. Gtzan music/speech collection, 1999. 5

[78] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 2

[79] Jörgen Valk and Tanel Alumäe. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658, 2021. 5

[80] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021. 2

[81] Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. Towards learning universal audio representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4597. IEEE, 2022. 1, 3

[82] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. 5

[83] Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *CVPR*, 2021. 2, 5, 6, 7, 8

[84] Tung-Yu Wu, Tsu-Yuan Hsu, Chen-An Li, Tzu-Han Lin, and Hung-yi Lee. The efficacy of self-supervised speech models for audio representations. In *HEAR: Holistic Evaluation of Audio Representations*, pages 90–110. PMLR, 2022. 6

[85] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[86] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 2

[87] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2

[88] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2