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

Supplementary Material

Abstract

In this supplementary material, we present information about our pretraining procedures and results from additional experiments. We also showcase examples of both our pretraining data and synthetic data. This document is laid out as shown below.

Contents

A. Pretraining Details

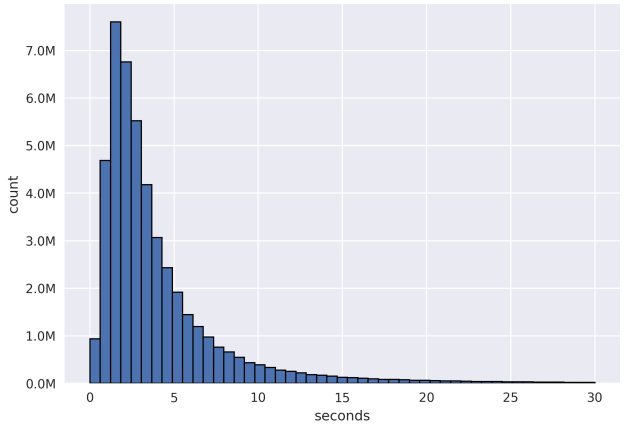

Figure A.1. Distribution of shot lengths observed in our dataset.

Data Preprocessing We temporally segment long-form content into shots (camera changes). Fig. A.1 shows the distribution of shot lengths. We ignore shots that are shorter than 3 and longer than 12 seconds. The former constraint is to make sure the snippet is long enough for our models, while the latter is to improve training throughput. The total number of shots in each pretraining setting is shown in Table 1 of the main submission under the column #data. When creating a minibatch during pretraining, we ensure that $\frac{1}{8}$ of each batch comes from the same long-form content source (e.g. the same movie) to create hard negatives. The process of generating quadruple training instances (v_p, a_p, v_s, a_s) is as follows:

1. Randomly pick a shot.
2. Temporal jitter: randomly select two temporal windows of $\max(T, 3)$ seconds where T is the length of the shot. These two snippets, derived from the same shot, are our primary and secondary instances. For the secondary instance, the language of the audio is different from the one in the primary instance, if an alternate audio track (i.e. dub) is available.
3. For each pair of audio and video:
   Video: Resample to 25 fps, uniformly sample 16 frames, randomly scale the shorter side of video within the range of 256-320, then perform a random crop of 224x224.
   Audio: Resample to 48kHz, convert audio to mel-spectrogram (n_fft=1024, hop_length=501, num_mels=96), convert to the decibel scale, and apply time and frequency masking with maximum value of 50 percent of the corresponding axis.

Model and Pretraining Hyperparameters The MLP projection heads have an output dimensionality of 512. The latent embeddings (z) are L2 normalized prior to computing the loss. The temperature factor τ in the objective function is set to 0.07. We use the AdamW optimizer [10] with a learning rate of 3e-4, and weigh decay of 5e-2. We train for 12 epochs on 32 NVIDIA A100 GPUs, with a batch size of 64 per GPU, using a half-cosine learning rate annealing which kicks off after 2 warm-up epochs.

B. Additional Experiments

B.1. Video-Only Results

We report results on **UCF101** [15] and **HMDB51** [9], well-known benchmarks, to assess the video-only performance of our models, shown in Table B.1. Performance between our model variants is comparable, showing that the dub-augmented training does not necessarily decrease video-only performance. Additionally, we compare to recent

Model	UCF101 [15]	HMDB51 [9]
B.3	88.90	69.35
B.4	88.20	68.91
B.5	87.99	69.43
FIMA [26]	76.40	47.30
FAME [4]	72.20	42.20

Table B.1. Performance of video models on UCF101 [15] and HMDB51 [9] datasets, comparing with recent results that *do not* involve fine-tuning.

Model	VGGSound
B.3	43.49
B.4	41.95
B.5	42.96
LAION-CLAP [23]	46.20
BLAT [25]	42.90

Table B.2. Performance of audio models on the VGGSound [2] dataset, comparing with recent results that *do not* involve fine-tuning on the downstream dataset. The LAION-CLAP result reported uses keyword-to-caption augmentation.

state-of-the-art results which, like us, do not use fine-tuning. Note that these results use linear probes, vs. our MLP probes which were derived from a grid search over probing strategies. Nevertheless, the fact that we significantly beat these results without fine-tuning (>12% absolute) demonstrates the value of our learned representations.

B.2. VGGSound Results

We report results on **VGGSound** [2], an audiovisual benchmark on which we focus on audio results, shown in Table B.2. Once again, performance between our model variants is comparable, and our results are competitive with recent state-of-the-art results which don't use fine-tuning.

B.3. Controlled Dataset and Models

In this section, we discuss the methods and results from a smaller-scale, more controlled set of experiments. The pretraining dataset consists of 748 movies, about 1300 video-hours of content. Each movie contains a video track, as well as four audio tracks: English (**EN**) as the primary language, and three dubbed versions, Spanish (**ES**), French (**FR**), and Japanese (**JA**), all languages for which we find dubs are relatively commonly available. Having multiple dub options allows us to investigate trade-offs between secondary languages, and whether "multilingual" models might further strengthen performance.

The video model is a medium X3D [5], which is an efficient ResNet-based model. Our audio model is an Acoustic ResNet50 [24], which takes audio spectrograms as input. Both models output 1024-dimensional representations per clip. We share backbone weights (i.e. Acoustic ResNet50) across audio variants with primary and secondary (dubbed) languages. We do not share weights for primary vs. secondary audio, to allow for more flexibility. As in our primary experiments, we mainly train these models *cross-modally*, i.e. we compute the contrastive cost between modalities.

We train these models on 4 A100 GPUs for 10 epochs with a batch size of 26 per GPU. We use a negative sampling parameter k (samples drawn from the same movie as the positive clip), which we set to 12 per GPU. We use the AdamW optimizer [10] with β=(0.9, 0.999), a learning rate of 0.001, weight decay of 0.05, and a cosine learning rate schedule with a half-epoch warmup.

In all, we compare the following model variants in these smaller-scale, more controlled, experiments:

1. ***Monolingual*** (**EN**): In this baseline, we consider models trained with two differently-augmented primary (English) audio treated as "primary" and "secondary" (a_p=**EN**; a_s=**EN**) audio respectively. This is to account for any possible effect of two augmentations per seen sample, as occurs for the dub-augmented cases, although it does not modify the data distribution. This is a SimCLR-based setup, with two audio paths each contrasted with video.
2. ***Bilingual*** (**ES**, **FR**, **JA**): We introduce one secondary audio at a time to explore the dub-augmented training hypothesis (a_p=**EN**; a_s= **ES** OR **FR** OR **JA**).
3. ***Multilingual*** (+**EFJ**): Here, we effectively randomly select a secondary audio from the given list (Spanish, French, and Japanese) per batch (a_p=**EN**; $a_s \in_R$ {**ES**, **FR**, **JA**}). The order of samples is randomized, so in practice we simply circle through the list round-robin. We aim to explore whether there are additional benefits or drawbacks to having more than one secondary audio.
4. ***No-Speech*** (**SEP**): We establish another baseline where the speech is separated and we only train on video + non-speech audio. This allows us to examine whether simply removing the speech is enough for a performance gain on non-speech-focused tasks. We use the pretrained Hybrid Demucs v3 model [3] to separate the vocal from the rest, mixing the other stems back together. There is no secondary audio here (a_p=**EN** SEP). Note that this variant is trained with 44.1kHz audio, as this is the input and output sample rate for the Demucs models. Although Demucs is trained for music separation, we find that it works well on speech in practice on our dataset. We use the default (`mdx_extra_q`) pretrained model.
5. ***Audio-Only*** (Monolingual: **AUD**, and Multilingual: **AUD**$_{+\text{EFJ}}$): Finally, we examine two audio-only models. The data is similar to the ***monolingual*** and ***multilin-***

gual setups, except without video. The objective function is now *within-modal*, between the two audio clips. The monolingual version represents standard audio contrastive training with two augmented copies. These models cannot work on visual or audiovisual tasks, but here we seek to evaluate whether and how much dub-augmented training contributes improvements in the absence of video.

B.3.1 Evaluation

Evaluation Tasks Beyond the HEAR [17] tasks used in our main experiments, we include results from additional audio tasks to this controlled setup to gain a more complete picture in the controlled setup. First, we add audio tasks from HARES [20]; specifically, TUT18 [13] for acoustic scene recognition, Fluent Speech Commands [11] for speech command recognition, and VoxForge [12] for language identification, complementing existing HEAR tasks. As in the supplementary material for our main results, we include the video-only action recognition tasks HMDB51 [9] and UCF101 [15]. Finally, we add an *audiovisual* task (VGGSound [2]) to facilitate a better comparison with **SEP**, since this baseline sees no speech altogether. We hypothesize that **SEP** will be a strong performer in some cases, but that dub-augmented models will be stronger in general as they preserve the audiovisual relationship between speech actions visually occurring and sounding.

For the visual and audiovisual tasks, we train the probes for 200 epochs using Stochastic Gradient Descent and a learning rate of 0.2 following a cosine schedule. We train on 2 A10 GPUs with a total batch size of 1024. For HEAR tasks, we use the provided API's strategy and the 48kHz data. For HARES tasks, we follow the authors' specifications [20]: in general, with 400K training steps and a learning rate schedule consisting of 5K linear warmup steps and a cosine decay for the rest (max. learning rate of 0.0002, with the Adam [8] optimizer). We train on 2 GPUs with a total batch size of 64. In all relevant cases, we duplicate mono audio to the second channel to form a pseudo-stereo input to match our model's architecture.

B.3.2 Results

In total, we trained 8 different model variants and evaluated them on 15 different tasks. Table B.3 shows our main tasks on which we hypothesized improvement (N=8), grouped by modality and task type.

Does dub-augmented pretraining help? For all tasks in Table B.3, one or more dub-augmented models outperform the monolingual **EN** model. In 6/8 tasks, *all* dub-augmented variants outperform **EN**, except for the two easiest tasks (TUT18 and GTZAN). We hypothesized this outcome for the sound and scene classification tasks, where we consistently observe substantial gains, as well as the non-semantic speech tasks. This supports the results from the main paper.

Is the improvement due only to de-emphasizing speech? We examine the source-separated version to address this question, since it offers the extreme case where the speech is removed altogether (as much as possible). The source-separated variant presents a strong baseline on the sound/scene classification tasks, despite mostly being outperformed by one or more dub-augmented models. We expect this is due to re-focusing on non-speech elements. However, despite strong performance in these cases, this variant has drawbacks. First, it results in lower performance than all other models on VGGSound (audiovisual classification) and both visual tasks (shown in the trade-off results in Table B.4). We suspect this is because there is a clear discrepancy between the auditory and visual channels in the source-separated version, i.e. speech. When a person is speaking, and there is little or no speech content in the auditory stream accompanying the visual, this may act as a confounder for coordinating the two representations. Note that *People* is a large category in VGGSound[1].

Second, **SEP** significantly underperforms on non-semantic speech tasks and (in Table B.4) language identification, with the exception of GTZAN which we find is an easier task in general. This intuitively makes sense: this variant does not see speech, effectively, and performs lower than the monolingual variant as well. These results illustrate a trade-off: source-separation as a preprocessing method, in addition to being very computationally expensive and weakening the self-supervision assumption (by dependence on a third-party supervised model), results in poor performance on paralinguistic tasks, which require attention to aspects of speech beyond language.

Are more languages better? Given the strength of dub-augmented training, we ask whether introducing more languages into the mix improves performance further. Our results don't indicate this to be the case, but note that in Table B.3, the **EFJ** model is least commonly the lowest-performing dub-augmented variant (1/8 tasks). Additionally, the multilingual variant performs well on 2/3 non-semantic speech tasks. Even though paralinguistic features can vary by language, commonalities exist that may be useful and many practical scenarios could benefit from diverse examples. The robustness of the multilingual model suggests that it could be a reasonable default choice assuming little knowledge about the downstream tasks, and we use

[1] www.robots.ox.ac.uk/ vgg/data/vggsound

	Task	M	Baselines (SimCLR) AV	SEP	A	Dub-Augmented ES	FR	JA	EFJ	A_{+EFJ}
Snd/Scn	ESC-50 [14]	A	.527±.012	.570±.028	.220±.027	.580±.019	.575±.031	**.590±.036**	.587±.009	.550±.026
	FSD50K [6]	A	.296	.307	.109	**.317**	.313	.311	.313	.277
	TUT18 [13]	A	.853	.857	.682	**.884**	.881	.849	.867	.801
	VocalImitation [7]	A	.042	.051	.022	.045	.047	.045	.050±.006	**.055**
	VGGSound [2]	AV	.303	.287	—	**.323**	.314	.314	.311	—
NonSem	CREMA-D [1]	A	.514±.012	.489±.009	.354±.022	.528±.009	.540	.520±.010	**.548±.012**	.530±.011
	GTZAN Mus/Sp [18]	A	.954±.054	.931±.099	.866±.119	.946±.082	.891±.142	.931±.092	**.969±.054**	.954±.054
	LibriCount [16]	A	.654±.026	.608±.016	.505±.014	.671±.025	**.706±.021**	.681±.016	.676±.013	.678±.022

Table B.3. **Controlled experiments evaluation results.** All metrics are top-1 accuracy, except FSD50K and VocalImitation (Mean Average Precision). Results in **bold** indicate the highest score, and in gray indicate the lowest. The task types are **Snd/Scn** = Sound/Scene Classification and **NonSem** = Non-Semantic Speech.

a similar multilingual approach in our larger scale experiments in the main paper.

Is dub-augmentation beneficial even without video? The A_{+EFJ} variant always outperforms the **AUD** model (including on all audio tasks we examine later for trade-offs, shown in Table B.4). **AUD** is the weakest performer on all relevant tasks, indicating the benefits of cross-modal training. Additionally, on some tasks, the multilingual variant comes close to or even outperforms (as in on VocalImitation) the cross-modal variants. Of course, this variant cannot work on visual or multimodal tasks, and still largely underperforms the multimodal dub-augmented models, but it demonstrates the significant value of even unimodal dub-augmented training.

B.3.3 Exploring Trade-Offs

Results on the 7 tasks in Table B.4 help us evaluate possible trade-offs in the smaller-scale and controlled setup, to complement the previous results.

Can dub-augmented models still recognize language? The dub-augmented variants generally perform similarly or slightly worse on VoxLingua but appear to do better on VoxForge, both language identification tasks. The latter is a large-scale user-submitted dataset, which may have different auditory characteristics from the former as a result. Taking these results together, we expect that the dub-augmented models are able to retain information useful for language identification in their pre-MLP features. It is possible that more general auditory features, which do not encode speech semantics, are still discriminative in these tasks.

Are they discriminative between spoken words? As in our results from the main paper, we do not observe major degradations on linguistic tasks. This suggests that the features learned by our dub-augmented models preserve speech-related information that can be used to, for instance, recognize words or commands. However, the source-separated models' features appear useful for these tasks, which suggests that non-speech features and more general representations of the sound signals may be helpful. We further investigate this below, where our results show that the background noise in one of these datasets (Fluent Speech Commands) may provide useful signal for performance.

Is performance on video-only tasks impacted? On the visual action recognition tasks, the results from the dub-augmented variants appear similar to the baseline. The baseline performs slightly better on HMDB51 and slightly worse on UCF101. This suggests that the overall video-only performance of the model may not be significantly affected by dub-augmented pretraining, similar to what is shown in Table B.1 for our main model variants.

C. Examples of Synthetic Counterfactual Pairs

Fig. C.2 highlights clips from a synthetically generated version of the LVU dataset [22], which we refer to as **LVU-M**, as noted in the main paper. Similar to Fig. **??**, the spectrograms show variation and commonalities between alternate audio tracks of the same clip. The examples, arbitrarily selected, show both consistency with the visual (e.g. voices, general timing, etc.) and divergence from it due to artifacts, lack of full acoustic context (e.g. reverberation), and other current limitations of the proposed pipeline. We only show the middle 10 seconds of these clips, to allow easy inspection.

D. Code for Modular Pipeline

Finally, we include the codebase for our modular pipeline for simulating counterfactual pairs. The `README.md` file lists the main dependencies and components, and provides

	Task	M	Baselines (SimCLR) AV	SEP	A	Dub-Augmented ES	FR	JA	EFJ	A_{+EFJ}
SemSp	FlSpComm [11]	A	.379	.400	.263	.391	**.410**	.402	.373	.368
	SpComm5h [21]	A	.298	**.372**	.144	.362	.344	.325	.300	.231
	SpCommFull [21]	A	.471	.489	.162	.477	**.537**	.530	.491	.298
Lang	VoxForge [12]	A	.546	.516	.504	.580	.584	**.592**	.571	.543
	VoxLingua10 [19]	A	**.251**±.045	.226±.033	.111±.012	.229±.016	.237±.050	.246±.032	.227±.043	.201±.009
Act	HMDB51 [9]	V	**.341**	.319	–	.330	.324	.322	.333	–
	UCF101 [15]	V	.531	.496	–	.540	.523	.538	**.542**	–

Table B.4. **Controlled experiments potential trade-offs: Does dub-augmentation negatively impact performance on linguistic or vision-only tasks?** The tasks in this table include **Semantic Speech** (FlSpComm [11], SpComm5h [21], and SpCommFull [21]) and **Language ID** (VoxForge [12] and VoxLingua10 [19]), and 2 **Action Recognition** video-only tasks (HMDB51 [9] and UCF101 [15]). The results vary and often reflect relatively small differences in either direction, suggesting overall that performance is not majorly affected on language-focused and vision-only tasks.

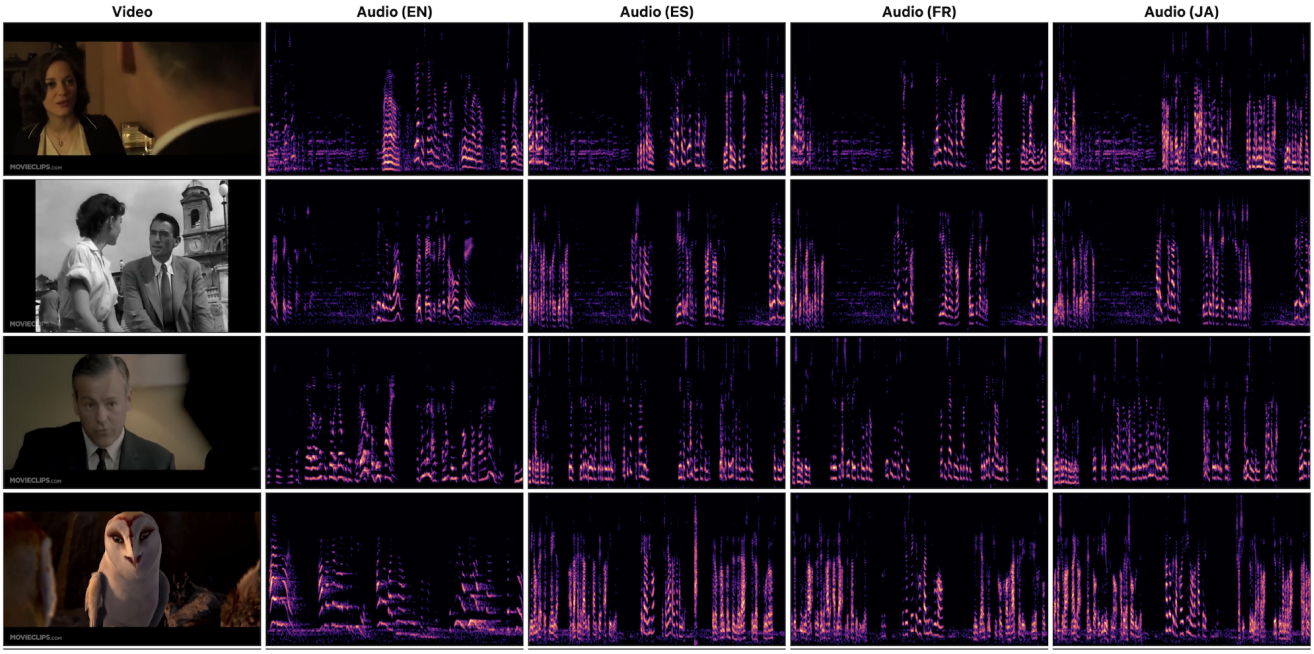

Figure C.2. Examples of clips from **LVU-M**.

instructions for configuring and running the pipeline on video datasets.

References

[1] 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. 4

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 3, 4

[3] 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. 2

[4] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022. 2

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 203–213, 2020. 2

[6] 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. 4

[7] 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. 4

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[9] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 2, 3, 5

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2

[11] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*, 2019. 3, 5

[12] Ken MacLean. Voxforge. *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*, 2018. 3, 5

[13] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. *arXiv preprint arXiv:1807.09840*, 2018. 3, 4

[14] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 4

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 3, 5

[16] Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. Libricount, a dataset for speaker count estimation, 2018. 4

[17] 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. 3

[18] George Tzanetakis. Gtzan music/speech collection, 1999. 4

[19] 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

[20] 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. 3

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

[22] Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *CVPR*, 2021. 4

[23] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[24] 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

[25] Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data. *arXiv preprint arXiv:2303.07902*, 2023. 2

[26] Minghao Zhu, Xiao Lin, Ronghao Dang, Chengju Liu, and Qijun Chen. Fine-grained spatiotemporal motion alignment for contrastive video representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4725–4736, 2023. 2