A. Datasets and Metrics

Audioset (AS) [18]. This dataset is used for both training and evaluation. It contains 10s videos from YouTube annotated into 527 classes. It consists of 3 pre-defined splits, the balanced split with about 20K videos, test split with 18K videos, and an unbalanced training split with about 2M videos. For **training**, we use the 2M unbalanced set without any labels, and only use it for audio-video matching. For **zero-shot evaluation** in Table 2, we use the test set and compute logits for each class using the textual class names along with the templates as described later in Appendix B.3. The metric used is top-1 accuracy.

ESC-50 (ESC) [58]. We use this dataset for evaluating the learned representations in a zero-shot manner. The task here is "Environmental Sound Classification" (ESC). It consists of 2000 5s audio clips classified into 50 classes. It has predefined 5 fold evaluation, each consisting of 400 test audio clips. In this work, we compute 0-shot predictions on the evaluation set for each fold and report the 5-fold average performance. For ablations we use only the first fold for computational ease. The metric used is top-1 accuracy.

Clotho (Clotho) [16]. This is a dataset of audio from the Freesound platform with textual descriptions. It consists of a dev and test set of 2893 and 1045 audio clips respectively, with each clip associated with 5 descriptions. We consider the text→audio retrieval task, and consider each of the 5 associated captions as a separate test query and retrieve from the set of audio clips. The metric used is recall@K, where a given test query is assumed to be correctly solved if the ground truth audio is retrieved within the top-K retrieved audio clips.

AudioCaps (AudioCaps) [36]. This is a dataset of audio-visual clips from YouTube accompanied by textual descriptions. It consists of clips from the Audioset dataset as described earlier. We use the splits as provided in [52],[1] which removes clips that overlap with the VGGSound dataset. We end up with 48198 training, 418 validation and 796 test clips. We only use the test set for zero-shot evaluation of our model. The task is text→audio retrieval, and evaluation is performed using recall@K.

VGGSound (VGGS) [8]. This dataset contains about 200K video clips of 10s length, annotated with 309 sound classes consisting of human actions, sound-emitting objects and human-object interactions. We only use the audio from the test set (with 14073 clips) for 0-shot classification. The evaluation is done using the top-1 accuracy metric.

SUN RGB-D (SUN). We use the registered RGB and Depth maps provided in the SUN RGB-D [67] dataset `train` set (∼5K pairs) for training our model. We follow [20] to post process the depth maps in two steps - 1) we use in-filled depth values and 2) convert them to disparity for scale normalization. This dataset is only used in training, so we do not use any metadata or class labels.

SUN Depth-only (SUN-D). We use only the ∼5K depth maps from the `val` split of the SUN RGB-D [67] dataset and denote them as SUN Depth-only. This dataset is only used for evaluation and we do not use the RGB images. We process the depth maps similar to SUN RGB-D (in-filled depth, converted to disparity). We use the 19 scene classes in the dataset and use their class names for constructing the zero-shot classification templates.

NYU-v2 Depth-only (NYU-D). We use the 794 `val` set depth maps from the NYU-v2 Depth-only [64] dataset for evaluation only. We post-process the depth similar to SUN Depth-only. We use the 10 scene class names in the dataset. The 10th scene class, called 'other', correspond to 18 different semantic classes – `['basement', 'cafe', 'computer lab', 'conference room', 'dinette', 'exercise room', 'foyer', 'furniture store', 'home storage', 'indoor balcony', 'laundry room', 'office kitchen', 'playroom', 'printer room', 'reception room', 'student lounge', 'study', 'study room']`. For zero-shot evaluation, we compute the cosine similarity of the 10th class as the maximum cosine similarity among these 18 classnames.

LLVIP (LLVIP). The LLVIP dataset [31] consists of RGB image and Thermal (infrared low-light) image pairs. The dataset was collected in an outdoor setting using fixed cameras observing street scenes and contains RGB images taken in a low-light paired with infrared images (8∼14um frequency). The RGB thermal pairs are registered in the dataset release. For training, we use the `train` set with 12025 RGB image and thermal pairs. For evaluation, we use the `val` set with 3463 pairs of RGB and thermal images. Since the original dataset is designed for detection, we post process it for a binary classification task. We crop out pedestrian bounding boxes and random bounding boxes (same aspect ratio and size as pedestrian) to create a balanced set of 15809 total boxes (7931 'person' boxes). For zero-shot classification, we use the following class names for the 'person' class - `['person', 'man', 'woman', 'people']`, and `['street', 'road', 'car', 'light', 'tree']` for the background class.

Ego4D (Ego4D) [22]. For the Ego4D dataset, we consider the task of scenario classification. There are 108 unique scenarios present in the 9,645 videos of the Ego4D dataset. We filter out all videos annotated with more than one scenario which yields 7,485 videos with a single scenario assigned. For each video, We select all time-stamps that contains a synchronized IMU signal as well as aligned narrations. We sample 5 second clips around each time-stamp. The dataset is split randomly such that we have 510,142 clips for train-

[1] https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/resources/benchmark-files/AudioCaps_retrieval_dataset.tar.gz

ing, and 68,865 clips for testing. During training we only use the video frames and their corresponding IMU signal. We use the test split to measure zero-shot scenario classification performance, where each clip of IMU signal is assigned the video-level scenario label as its ground-truth.

A.1. Data Representations

We use the standard RGB and RGBT representations for **images and videos**. For videos, we use 2-frame clips, inspired from recent work on ViT-style video architectures [15, 69], where a video patch is $2\times16\times16$ ($T\times H\times W$). We inflate the visual encoder's weights to work with spatiotemporal patches and and at inference time we aggregate features over multiple 2-frame clips. Hence, we can use models trained on image-text data directly on videos.

We used a single-channel image for the **thermal data** since it is the natural form in which current infrared thermal sensors return data [31]. For **single-view depth**, we experimented with different encodings – absolute depth [64] as returned by sensors like the Kinect, inverse depth [61], disparity [61], and HHA [24, 25]. Overall, we found that disparity representation (which is a single-channel image) worked the best. For **audio** we use the raw waveform processed into mel-spectrograms [21], as described in the main text. For **IMU** we use a $6\times T$ tensor to represent the sequence of IMU sensor readings over time.

B. Evaluation details

We now describe the evaluation setups used in this work.

B.1. Inference implementation details

Audio/Video: For both these temporal modalities (whether operated upon together during pre-training or separately during inference), we sample fixed length clips to operate on. During training, we randomly sample a clip, typically 2s in length. At inference time, we uniformly sample multiple clips to cover the full length of the input sample. For instance, for 5s ESC videos, we would sample $\lceil \frac{5}{2} \rceil = 3$ clips. For video clips, we sample a fixed number of frames from each clip. For audio, we process each raw audio waveform by sampling it at 16KHz followed by extracting a log mel spectrogram with 128 frequency bins using a 25ms Hamming window with hop length of 10ms. Hence, for a t second audio we get a $128\times100t$ dimensional input.
IMU: For IMU, we sample fixed length clips of 5 seconds, centered around time-stamps that are aligned with narrations. For each clip, we get a 6×2000 dimensional input and we measure the zero-shot performance for scenario classification using each clip as an independent testing sample.

B.2. Few-shot evaluation details

For the few-shot results in Figures 3 using the ESC and SUN datasets, we sampled k training samples per class, where $k \in \{1, 2, 4, 8\}$. We fix the k samples such that our model and the baselines use exactly the same samples during training. For all few-shot evaluations, including the baselines, we freeze the encoder parameters and only train a linear classifier.
Audio: For audio few-shot training with ESC, our model and the baselines are trained using AdamW with a learning rate of 1.6×10^{-3} and weight decay of 0.05 for 50 epochs.
Depth: For depth few-shot training with SUN, our model and the baselines are trained using AdamW with a learning rate of 10^{-2} and no weight decay for 60 epochs.

B.3. Zero-shot evaluation details

Query Templates. For all evaluations, we use the default set of templates from CLIP [59].[2] Note that we use the same templates for non visual modalities like audio and depth as well since we only use semantic/textual supervision associated with images.

B.4. Qualitative evaluation details

Cross-modal nearest neighbors. We perform the retrieval on the embedding feature after temperature scaling. The nearest neighbors are computed using cosine distance. In Figure 1, we show retrievals for audio from ESC, image retrievals from IN1K and COCO, depth from SUN-D, and text from AudioCaps.
Embedding arithmetic. For arithmetic, we again use the embedding features after temperature scaling. We ℓ_2 normalize the features and sum the embeddings after scaling them by 0.5. We use the combined feature to perform nearest neighbor retrieval using cosine distance, as described above. In Figure 1, we show combination of images and audio from IN1K and ESC, and show retrievals from IN1K.
Audio→Image Generation. For generating images form audio clips, we rely on an in-house reproduced implementation of DALLE-2 [60]. In DALLE-2, to produce images from text prompts, the image generation model relies on text embeddings produced by the pre-trained CLIP-L/14 text encoder. Since IMAGEBIND naturally aligns CLIP's-embedding space to that of other modalities proposed in the paper, we can upgrade the DALLE-2 model to generate images by prompting it with these new unseen modalities. We achieve zero-shot audio to image generation with DALLE-2 by simply using the temperature-scaled audio embeddings generated by IMAGEBIND's audio encoder as a proxy for the CLIP's text embeddings in the DALLE-2's image generation model.
Detecting objects using audio. We extract all audio descriptors from the validation set of ESC using an IMAGEBIND ViT-B/32 encoder, yielding 400 descriptors in total. We use an off-the-shelf CLIP-based Detic [86] model and

[2] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

use the audio descriptors as the classifier for Detic in place of CLIP text-based 'class' embeddings. We use a score threshold of 0.9 for the qualitative results in Figure 5.

C. Pretraining details

C.1. Best setup

In Table 9 we detail the hyperparameters used to pretrain each of the models reported in Table 4. Our experiments were done on 32GB V100 or 40GB A100 GPUs.

Config	AS	SUN	LLVIP	Ego4D
Vision encoder		ViT-Huge		
embedding dim.	768	384	768	512
number of heads	12	8	12	8
number of layers	12	12	12	6
Optimizer		AdamW		
Optimizer Momentum		$\beta_1 = 0.9, \beta_2 = 0.95$		
Peak learning rate	1.6e-3	1.6e-3	5e-4	5e-4
Weight decay	0.2	0.2	0.05	0.5
Batch size	2048	512	512	512
Gradient clipping	1.0	1.0	5.0	1.0
Warmup epochs		2		
Sample replication	1.25	50	25	1.0
Total epochs	64	64	64	8
Stoch. Depth [28]	0.1	0.0	0.0	0.7
Temperature	0.05	0.2	0.1	0.2
Augmentations:				
`RandomResizedCrop`				
`size`	-	224px		-
`interpolation`	-	Bilinear	Bilinear	-
`RandomHorizontalFlip`	-	$p = 0.5$	$p = 0.5$	-
`RandomErase`	-	$p = 0.25$	$p = 0.25$	-
`RandAugment`	-	9/0.5	9/0.5	-
`Color Jitter`	-	0.4	0.4	-
`Frequency masking`	12	-	-	-

Table 9. Pretraining hyperparameters

Contrastive loss batch size *vs*. modalities. While contrastive losses do require larger batch size, this requirement didn't increase with the number of modalities. As noted in Appendix B, our experiments (Table 2) sample a minibatch of one pair of modalities at a time: batch size of 2K for (video, audio), and 512 for (image, depth), (image, thermal), and (video, IMU). These batch sizes are smaller than the >32K batch sizes used in prior work [10, 59].

Combining modalities. In Table 4, we show results with combining the audio and video modalities. We combine them by extracting embeddings from both modalities per sample and computing a linear combinations of those embeddings. We used a weight of 0.95 for video and 0.05 for audio for this combination, which was found to perform the best.

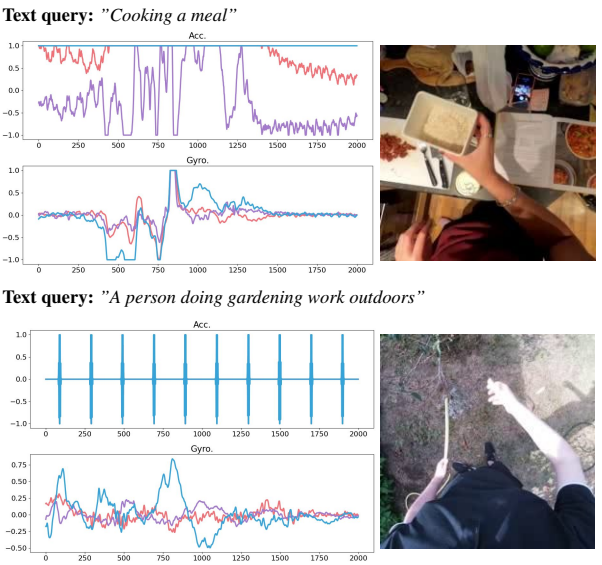

Figure 7. IMU retrievals. Given a text query, we show some IMU retrievals and corresponding video frames.

C.2. Ablation setup

The following setup was used for our evaluations in § 5. Different from the best setup, all ablation experiments uses ViT-Base both for the vision and the modality-specific encoders. The models are trained for 16 epochs, unless mentioned otherwise.

For Table 5b, the differences between the linear and MLP heads are detailed below: The MLP head did not improve performance in our experiments.

Linear	`Linear(in_dim, out_dim)`
MLP	`Linear(in_dim, in_dim), GELU, Linear(in_dim, out_dim)`

D. Additional Results

Qualitative results. We show additional results (along with audio) in the accompanying video.

Practical applications of disparate modalities. In general, a shared embedding space enables a variety of different cross-modal search and retrieval applications. *e.g.*, since IMU sensors are ubiquitous (in phones, AR/VR headsets, health trackers), IMAGEBIND can allow a user to search an IMU database using text queries (without training with IMU-text pairs). IMU-based text search has applications in healthcare/activity search. For instance, in Figure 7 we show examples of IMU (and accompanying video) retrieval given textual search query. The retrieved IMU sample, shown as 3-channel Accelerometer (Acc) and Gyroscope (Gyro) recording, matches the text query.

E. Additional Ablations

Design choices in losses. Since the modality-specific encoders are trained to align with a frozen image encoder, we tried using a ℓ_2 regression objective. For ZS SUN top-1 accuracy, we observed that regression led to good performance as the sole objective (25.17%) or jointly with contrastive (29.04%). However, it did not improve over using only the contrastive objective (31.74%).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020. 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 1, 2

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 1, 6

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 5

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 8

[7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 3

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 4, 12

[9] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP*, 2022. 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 7, 14

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020. 7

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2

[14] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2

[15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 13

[16] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *ACM international conference on Multimedia*, 2013. 4, 12

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2

[18] 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 *ICASSP*, 2017. 4, 12

[19] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 2, 3

[20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 2, 4, 5, 12

[21] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 3, 4, 13

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 4, 12

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[24] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013. 13

[25] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014. 13

[26] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 3, 4, 5

[27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[28] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 14

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 4, 5

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

[31] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, 2021. 4, 12, 13

[32] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 2

[33] Armand Joulin, Laurens van der Maaten, Allan Jabri, and

Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 2

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, AMustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP*, 2021. 5

[36] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 4, 12

[37] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[38] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022. 5

[39] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017. 2

[40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[41] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 2

[42] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, 2022. 2

[43] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 2

[44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. corr abs/2104.08860 (2021). *arXiv preprint arXiv:2104.08860*, 2021. 2

[45] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1

[46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2

[47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV*, 2019. 2

[48] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 5

[49] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 1, 2, 3, 5

[50] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 2, 4, 5

[51] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 5

[52] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021. 5, 12

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018. 3

[54] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1

[55] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *ICCV*, 2021. 1

[56] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 5

[57] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[58] Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015. 4, 12

[59] 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 *ICML*, 2021. 1, 2, 3, 4, 5, 7, 13, 14

[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 6, 13

[61] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 13

[62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 4

[63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[64] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4, 12, 13

[65] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022. 5

[66] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014. 2

[67] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 4, 12

[68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2, 3, 7

[69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 13

[70] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 8

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[72] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: Bert pretraining of video transformers. In *CVPR*, 2022. 2

[73] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2

[74] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[75] Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022. 6

[76] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 4

[77] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2, 5

[78] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 5

[79] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. 2

[80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2, 5

[81] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[82] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2

[83] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 2, 3

[84] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 7

[85] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014. 4

[86] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 6, 13