Supplementary Material for "*MMG-Ego4D*: Multi-Modal Generalization in Egocentric Action Recognition"

Xinyu Gong[2*†], Sreyas Mohan[1*], Naina Dhingra[1], Jean-Charles Bazin[1], Yilei Li[1], Zhangyang Wang[2], Rakesh Ranjan[1]
[1]Meta Reality Labs, [2]The University of Texas at Austin

1. Details of MMG-Ego4D Dataset

Dataset Statistic. *MMG-Ego4D* comprises 79 classes, composed of 65 base classes and 14 novel classes. Fig. 1 shows the final distribution of data points per label in base classes and novel classes. The base and novel classes' names are listed in class_name.json.

Why Ego4D? We illustrate why we choose Ego4D [5] other than other datasets as our source dataset. We focus on egocentric activities captured by head-mounted devices (applications in AR/VR). In this context, IMU, audio, and video are relevant modalities. The MMAct dataset provides IMU data from legs and hands but not from head movements, and does not include audio. The EPIC-KITCHENS dataset does not contain IMU data. MMAct [6] and EPIC-KITCHENS [2] have limited scenes, whereas Ego4D offers diverse environments and an aligned large unlabelled dataset for unsupervised training. See the table below:

Dataset	Modalities			Unlabeled Data	Scene Diversity
	Video	Audio	IMU (head motion)		
MMAct [6]	✓	×	×†	×	4 scenes
EPIC-KITCHENS [2]	✓	✓	×	×	kitchen only
Ego4D [5]	✓	✓	✓	✓	diverse

Table 1. **Datasets comparison.** † IMU is captured from motion of hands and legs.

2. Implementation Details

2.1. Data processing

We follow the standard protocol for processing the input modalities. The number of input video frames is 16. We applied random augmentation [1] and random erasing [8] to the input video during training. Video frames are partitioned into non-overlapping spatiotemporal voxels and projected into an embedding space using a linear layer [3]. Audio is converted into a time-frequency (spectrogram) domain with a dimension of 503×64. We use frequency/time masking [7] as the audio data augmentation during training.

*Equal contribution
†Work done during an internship at Meta Reality Lab.

Modality	Optimizer	Learning Rate	Batch Size	Epoch
video	Adam	1×10^{-3}	256	196
audio	Adam	2×10^{-5}	256	196
IMU	Adam	3×10^{-4}	256	100
Multimodal	Adam	3×10^{-4}	128	50

Table 2. **Hyperparameters for unimodal and multimodal training**.

Audio data is then partitioned into 16×16 non-overlapping patches and linearly projected to an embedding space [4]. IMU data is 2D time series data, whose shape is 1000×6. IMU data is partitioned into non-overlapping windows of length 16 and linearly projected to an embedding space.

2.2. Training Details

Unimodal & multimodal training. We summarize the hyperparameters for unimodal and multimodal training in Tab. 2. During unimodal training, the network of each modality is trained independently. In the multimodal supervised training stage, only the fusion module is trainable, while other parameters are frozen. Modality drop probability p is set to 0.6.

Multimodal Meta-training. In the meta-training stage, we apply SGD optimizer with a learning rate of 2×10^{-4} to update all parameters of the multimodal network. 16 tasks are sampled from ego4d base classes at each training iteration. The model is trained for $40\,000$ iterations in total.

2.3. Few-shot Evaluation Details

All few-shot results reported in our paper are obtained via finetune-based evaluation by default, unless explicitly stated. We perform the evaluation on $10\,000$ episodes, which are randomly drawn from novel classes. Each episode is a 5 way 5 shot task.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF*

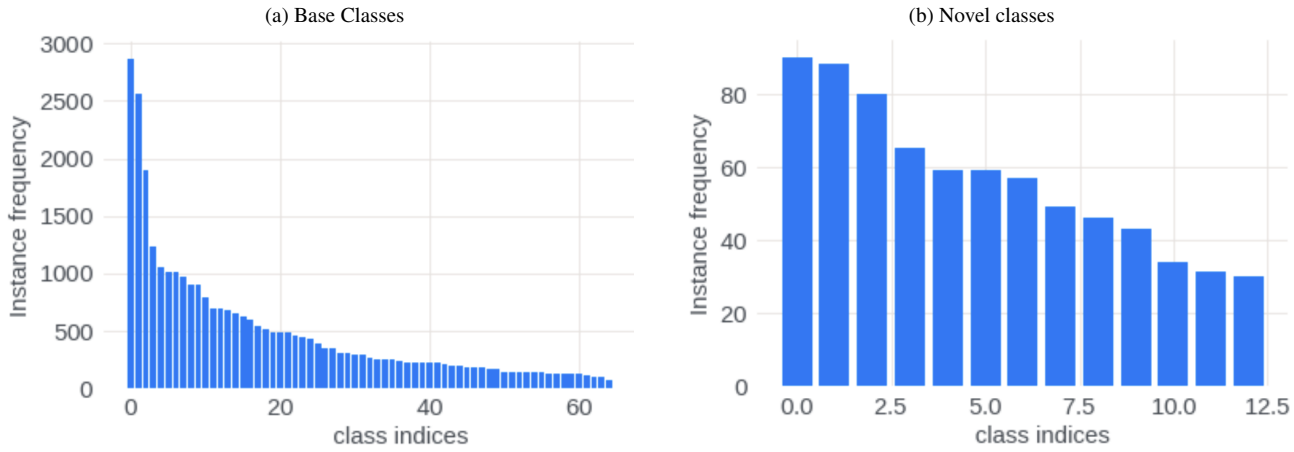

Figure 1. **Number of data points per class**. We show the distribution of data points per label for (a) base classes, and (b) novel classes.

conference on computer vision and pattern recognition workshops, pages 702–703, 2020. 1

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

[3] 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. 1

[4] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 1

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[6] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019. 1

[7] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 1

[8] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 1