JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Meta

TEXAS The University of Texas at Austin

MMG-EGO4D: MULTI-MODAL GENERALIZATION IN EGOCENTRIC ACTION RECOGNITION

TUE-PM-225

XINYU GONG, SREYAS MOHAN, NAINA DHINGRA, JEAN-CHARLES BAZIN, YILEI LI, ZHANGYANG WANG, RAKESH RANJAN

Meta Reality Lab
Visual Informatics Group @ The University of Texas at Austin (VITA)

Egocentric Action Recognition: Ubiquitous in Daily Life

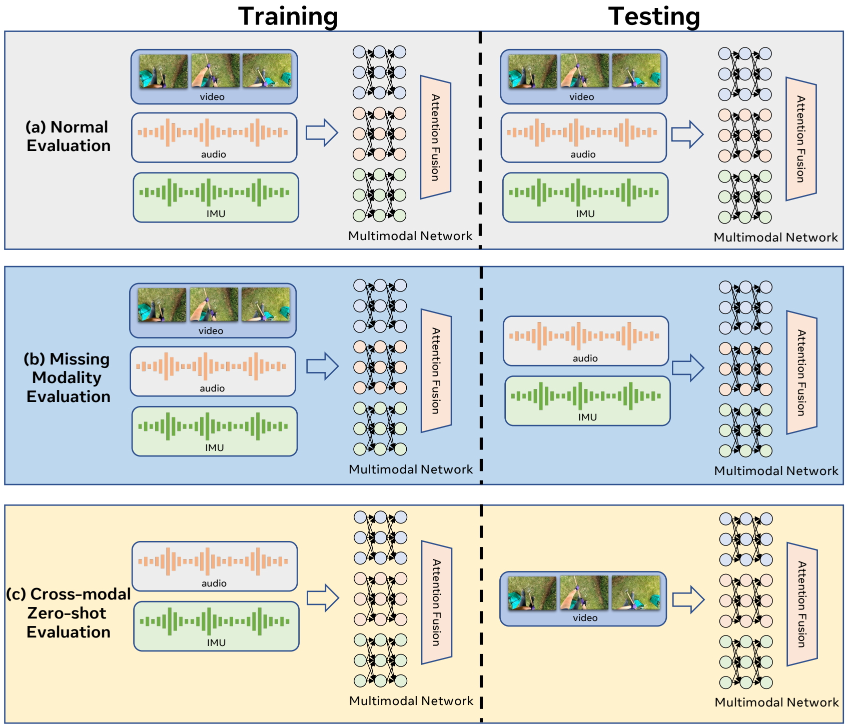
An egocentric view example of “using phone”.

Multi-Modal Generalization (*MMG*)

MMG investigates system generalizability under limited or absent data modalities.

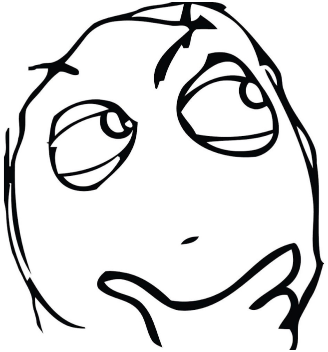
- **missing modality generalization**: During inference, some modalities present at training are absent.
- **cross-modal zero-shot generalization**: Training and inference modalities are disjoint.

Settings: Few-shot & Many-shot



Intriguing, isn't it?

But what are its practical applications?



Motivation

Missing Modality Evaluation

User Case 1: Users only allow partial device access (camera, microphone, etc.).

Solution: Deploy a multimodal model robust to different input modalities.

1. Why partial access? *Reasons include user privacy and location-specific restrictions (e.g., libraries, where audio isn't available).*
2. Why using a single model to deal with different input modalities? *Storage efficiency.*

Motivation

Cross-Modal Zero-Shot Evaluation

User case 2: Customizing user devices to learn specific visual actions.

Solution: Implement an efficient few-shot learner, capable of learning novel visual representations from inexpensive, locally sourced modalities.

1. Why learn locally? *To respect user privacy.*
2. Why use cheap data (audio, IMU) to train? *Training with video data is very expensive.*
3. Why few-shot examples? *It's impractical to ask users collect more data.*
4. Why use video data for inference? *Video data is highly informative, thus promoting accurate predictions.*

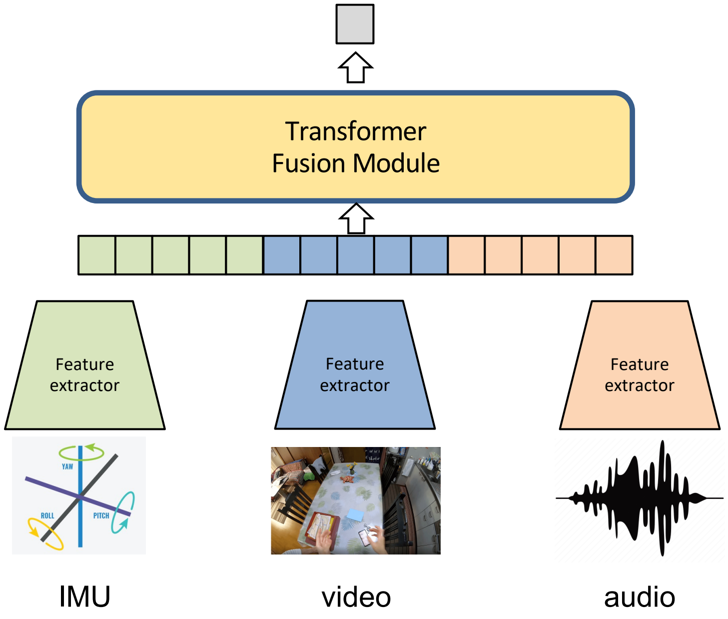
Modality	video	audio	IMU
Memory per second of data (KB)	593.92	62.76	9.44
Typical model FLOPs (G)	70.50	42.08	1.65

MMG-Ego4D Dataset

The MMG-Ego4D dataset includes data points across three modalities—video, audio, and inertial motion (IMU)—derived from the Ego4D dataset.

- **Task**: MMG egocentric action recognition under both many-shot and few-shot settings.
- **Data** : 167-hour unlabeled and 35-hour of labeled temporal-aligned Video-Audio-IMU data.
- **Label**: Consisting of 79 classes in total, with 65 base classes and 14 novel classes designated for few-shot tasks. Each sample has a single label.

A Streamlined Multimodal Transformer Architecture



Model	FLOPs (G)	Param (M)	Modality	5 Way 5 Shot Accuracy	Top-1 Accuracy
MViT-B [15]	70.50	36.50	video	58.89	52.40
AST [26]	42.08	87.03	audio	31.06	39.48
IMU Transformer	1.65	15.55	IMU	40.07	29.78

Overview of Training Pipeline

For Many-shot scenarios:

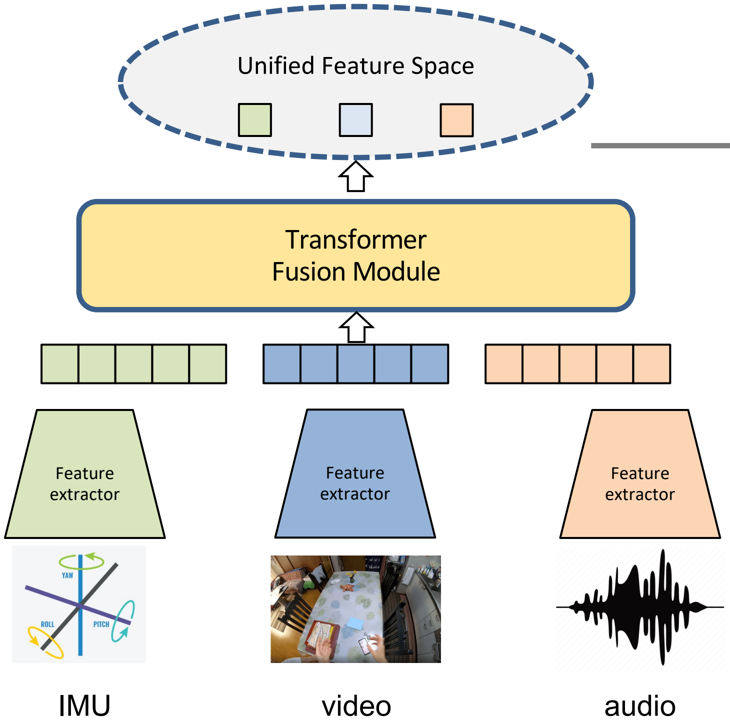
- Regular/Missing-Modal: (1) Unimodal supervised pre-training, (2) Multimodal supervised training
- Cross-Modal Zero-Shot: (1) Multimodal unsupervised pre-training, (2) Multimodal supervised training

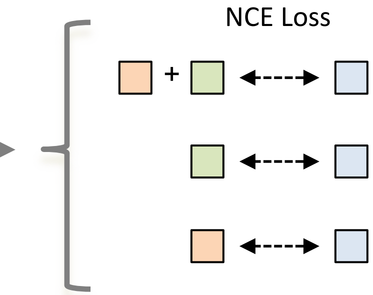
For Few-shot scenarios :

- All tasks: (1) Unimodal supervised pre-training, (2) Multimodal supervised training, (3) Multimodal meta-training

Setting	**Task**	**Multimodal unsupervised pre-train**	**Unimodal supervised pre-train**	**Multimodal supervised train**	**Multimodal meta-train**
Many-shot	Regular	-	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$	-
	Missing Modal	-	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$	-
	Zero-Shot	$\mathcal{L}_{\text{align}}$	-	$\mathcal{L}_{\text{CE}}$	-
Few-shot	Regular	-	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{proto}}$
	Missing Modal	-	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{proto}}$
	Zero-Shot	-	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{proto}}$

Multimodal Alignment Contrastive Loss

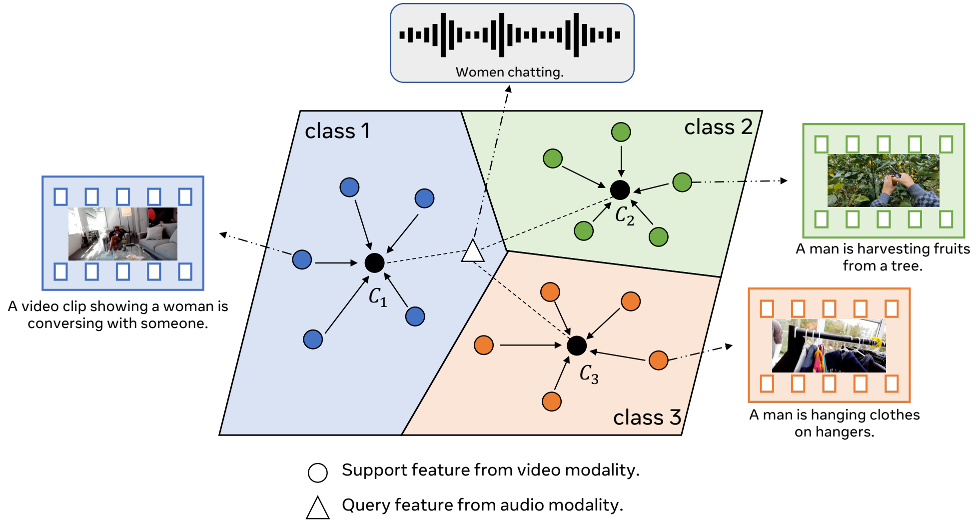




Positive pairs: *pairs of different modalities from the same temporal location.*
Negative pairs: *pairs from different temporal locations.*

$$\mathcal{L}_{\text{align}}(\boldsymbol{z}_{\text{video}}, \boldsymbol{z}_m) = \sum_{m\in\{\text{audio, IMU}\}} -\log\left(\frac{\exp(\boldsymbol{z}_{\text{video}}^\top \boldsymbol{z}_m/\tau)}{\exp(\boldsymbol{z}_{\text{video}}^\top \boldsymbol{z}_m/\tau)+\sum_{z'\in\mathcal{N}}\exp(\boldsymbol{z}_{\text{video}}^\top \boldsymbol{z}'_m/\tau)}\right)$$

Cross-modal Prototypical Loss



We meta-train the multimodal system using cross-modal prototypical loss.

Core idea: *Calculate the distance between support and query embeddings of different modalities.*

Results from Multimodal Few-Shot Evaluation

Model	FLOPs (G)	Param (M)	Modality	5 Way 5 Shot Accuracy
MViT-B [15]	70.50	36.50	video	58.89
AST [26]	42.08	87.03	audio	31.06
IMU Transformer	1.65	15.55	IMU	40.07

Unimodal Results

Eval. Setting	Support Modalities			Query Modalities			5 Way 5 Shot Accuracy
	Video	Audio	IMU	Video	Audio	IMU	
Regular	✓	✓	✓	✓	✓	✓	63.00
Missing Modality	✓	✓	✓	✓	✓		61.76
	✓	✓	✓		✓	✓	50.77
	✓	✓	✓	✓		✓	62.79
	✓	✓	✓	✓			62.68
	✓	✓	✓		✓		43.65
	✓	✓	✓			✓	47.48
Zero-Shot Evaluation		✓		✓			46.90
			✓	✓			42.07
		✓	✓	✓			50.80
	✓				✓		44.01
	✓					✓	46.56
	✓				✓	✓	49.37

Multimodal Results

Key Takeaways:

1. *Our model exhibits robustness when some modalities are absent during evaluation.*
2. *Training with affordable data and evaluating with expensive but informative data yields superior results compared to training and evaluation using only affordable data.*

Results from Multimodal Many-Shot Evaluation

Model	FLOPs (G)	Param (M)	Modality	Top-1 Accuracy
MViT-B [15]	70.50	36.50	video	52.40
AST [26]	42.08	87.03	audio	39.48
IMU Transformer	1.65	15.55	IMU	29.78

Unimodal Results

Eval. Setting	Train Modalities			Test Modalities			Top-1 Accuracy
	Video	Audio	IMU	Video	Audio	IMU	
Regular	✓	✓	✓	✓	✓	✓	55.66
Missing Modality	✓	✓	✓	✓	✓		55.47
	✓	✓	✓		✓	✓	37.07
	✓	✓	✓	✓		✓	54.57
Zero-Shot Evaluation		✓		✓			30.98
			✓	✓			20.00
		✓	✓	✓			25.03
	✓				✓		43.43
	✓					✓	35.67
	✓				✓	✓	41.02

Multimodal Results

Key Takeaways:

1. *Our model demonstrates robustness in scenarios where some modalities are absent during evaluation.*
2. *Substantial room for improvement exists in the domain of cross-modal zero-shot evaluation.*

Ablation Study

Eval. Setting	Train/Support Modal.			Test/Query Modal.			Fusion Module	Contrastive Alignment	Top-1 Accuracy	Cross-Modal Proto. Loss	5 Way 5 Shot Accuracy
	Video	Audio	IMU	Video	Audio	IMU					
Regular	✓	✓	✓	✓	✓	✓	Attention	✓	**55.66**	✓	**63.00**
							Attention	×	52.18	✓	61.16
							MLP	✓	52.79	✓	58.67
							Attention	✓	-	×	62.37
Missing Modality	✓	✓	✓		✓	✓	Attention	✓	**37.07**	✓	**50.77**
							Attention	×	21.32	✓	40.87
							MLP	✓	32.89	✓	49.00
							Attention	✓	-	×	50.03
Zero-shot Cross-Modal		✓	✓	✓			Attention	✓	**25.03***	✓	**51.40**
							Attention	×	2.37	✓	33.93
							MLP	✓	24.54*	✓	51.08
							Attention	✓	-	×	50.80

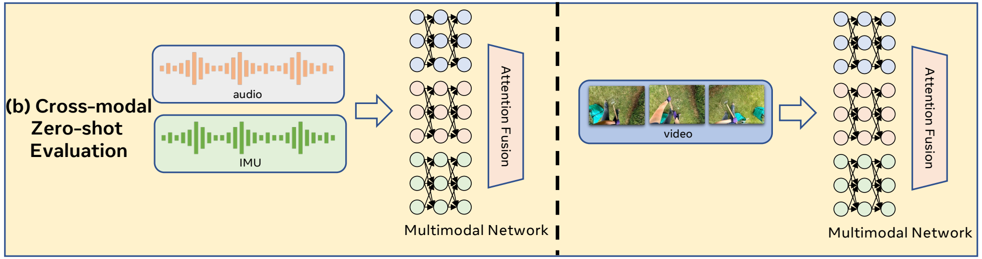
Table 6. **Ablation study of each design component under many-shot & few-shot settings**. Our proposed components improve the performance under all evaluation settings. Note that cross-modal prototypical loss is only applied under the few-shot setting. *Different from other settings, the cross-modal contrastive alignment loss is applied at the unsupervised multimodal pre-training stage in the supervised zero-shot cross-modal setting.

Qualitative Examples

query video

Train on support data

Test on query data



Setting: Few-shot

Support modalities: audio + IMU

Query Modality: video

Ground truth: Mop the floor

Model Prediction: Mop the floor

THANKS!