Teaching Robots New Activities with a Sentence: On-Device Multimodal Action Recognition for Elderly-Centered HRI*

1st Youhan Li
DIBRIS Department, RICE laboratory
University of Genoa
Genoa, Italy
youhan.li@edu.unige.it

2nd Jinhua Xu
DIBRIS Department, SmartLab
University of Genoa
Genoa, Italy
jinhua.xu@edu.unige.it
DIAG Department
Sapienza University of Rome
Rome, Italy
jinhua.xu@uniroma1.it

3rd Antonio Sgorbissa
DIBRIS Department, RICE laboratory
University of Genoa
Genoa, Italy
antonio.sgorbissa@unige.it

4th Carmine Recchiuto†
DIBRIS Department, RICE laboratory
University of Genoa
Genoa, Italy
carmine.recchiuto@unige.it

***Abstract*—Can we use a sentence to teach the robot to recognise a new human action without retraining? Socially assistive robots (SARs) and companion robots should be able to recognise the activities of daily living (ADLs) performed by older adults. However, conventional closed-set models cannot identify unseen activities or adapt to new scenarios without retraining. In addition, many existing approaches are not specifically designed for older adults.**

To address these limitations, we propose a tri-modal fusion encoder that integrates RGB video appearance, skeletal motion pose, and object context (VPOCLIP). Contrastive learning is used to align the fused representation with the frozen CLIP text embedding space. As a result, new activities can be introduced through a single natural-language description without retraining.

Experiments on the elderly subset of ETRI-Activity3D reach 69.3% top-1 accuracy for unseen actions on the 50/5 split and 78.1% on the closed 55-class seen set, with an on-device latency of about 45 ms per clip on a consumer laptop GPU (RTX 3060). Overall, VPOCLIP provides a potentially effective source of human-state perception for socially assistive robots, allowing robot behaviour to be adjusted according to the user's ongoing activity and surrounding context.

***Index Terms*—human action recognition, zero-shot learning, multimodal fusion, socially assistive robots, vision-language models, activities of daily living**

I. Introduction

As populations age, demand for socially assistive robots (SARs) is expected to increase, particularly in home-care settings [1]. Reliable and continuous perception of older adults' activities is therefore important for enabling SARs to adapt their assistance to users' changing needs. Existing behaviour-recognition systems, however, face two practical constraints. First, on-robot deployment requires low-latency inference while preserving users' privacy, which limits dependence on cloud-based processing [2]. Second, caregiving applications are inherently subject-specific because behaviour patterns vary across individuals, health conditions, and living environments. After deployment, caregivers may also need to introduce previously unseen activities that were not included in the original training set. Retraining the entire recognition model whenever a new action is added would impose substantial computational and operational costs [3]. This tension motivates the development of efficient, privacy-preserving recognition frameworks that support incremental adaptation without repeated full-model retraining.

Multimodal perception may improve the accuracy and robustness of human action recognition compared with single-modality approaches, including graph convolutional network (GCN)-based skeleton recognition [4]–[6] and video-based methods [7]–[9]. Nevertheless, many existing multimodal frameworks remain closed-set and cannot incorporate new action classes without retraining [10]. CLIP-based zero-shot human action recognition offers a promising alternative by enabling recognition of previously unseen actions through vision-language alignment [11]–[14]. Although such methods have achieved competitive results on offline benchmarks and high-performance GPUs [15], their practical use in socially assistive robots remains limited. Most studies give insufficient attention to edge deployment, do not provide a complete real-

†Corresponding author: carmine.recchiuto@unige.it

time perception pipeline, and rarely consider actions specific to older adults [16]. Consequently, a clear research gap remains in developing a unified framework that balances recognition accuracy, real-time deployability, and extensibility to new actions. To address this gap, the proposed framework makes the following contributions:

1) A three-stream multimodal architecture is developed to fuse skeleton, video, and object information and align the resulting representation with the CLIP text embedding space for older adults' activities of daily living.
2) A describe-to-add mechanism allows a new action to be introduced through a textual description without retraining the full model. Its open-vocabulary performance is evaluated under multiple seen and unseen class splits.
3) A complete real-time pipeline is implemented, from webcam input and person detection to skeleton extraction, video feature encoding, and action recognition. The system runs locally on a GPU at 5-10 frames per second and is intended for integration with socially assistive robots.
4) Experiments on the older-adult subset of ETRI-Activity3D evaluate recognition accuracy, generalisation to unseen actions, and real-time performance.

II. Related Work

Human action recognition has been studied extensively for assistive settings, yet most established methods rely on supervised learning over predefined action categories, and they typically require additional labelled data and model retraining once new behaviours are introduced [4]–[8], [17], [18]. This limits their adaptability in real-world scenarios, where caregivers may need to customize the behaviors monitored by the robot.

Zero-shot action recognition addresses this limitation by using semantic information, such as natural-language descriptions [19], [20], to recognize unseen actions. However, existing approaches are often computationally intensive or primarily evaluated offline, which makes their real-time deployment on robotic platforms challenging [21], [22]. Moreover, relatively limited attention has been given to connecting zero-shot recognition outputs with human-adaptive robot responses. In contrast, this study develops a lightweight framework that allows non-technical users to define target behaviors through natural language and supports real-time, locally deployed action recognition for adaptive social robot interaction.

III. Method

VPOCLIP builds on three complementary streams: YOLOv5m provides object-level context that captures human–object interactions [23], X3D encodes RGB appearance and the surrounding environment [9], and CTR-GCN extracts skeleton-based motion dynamics [6]. Together they yield a tri-modal representation that makes recognition robust across the visually similar activities of daily living. More importantly, we align this tri-modal representation with a frozen CLIP text encoder, which grounds visual features in natural language and enables the model to recognize actions directly from textual descriptions.

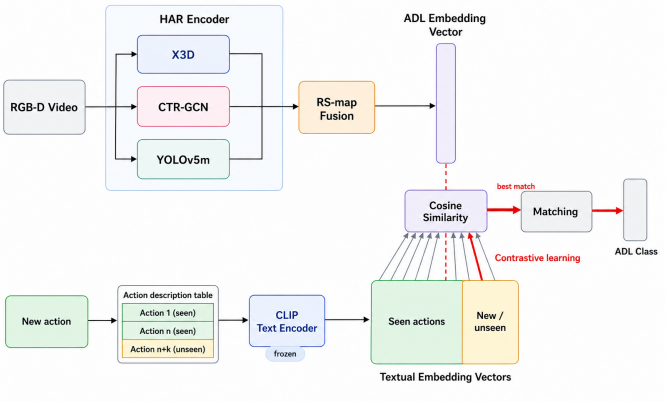

Fig. 1. Overview of the proposed VPOCLIP framework, which integrates multi-branch RGB-D action encoding, RS-map feature fusion, and CLIP-based semantic matching for real-time recognition of both seen and unseen activities of daily living.

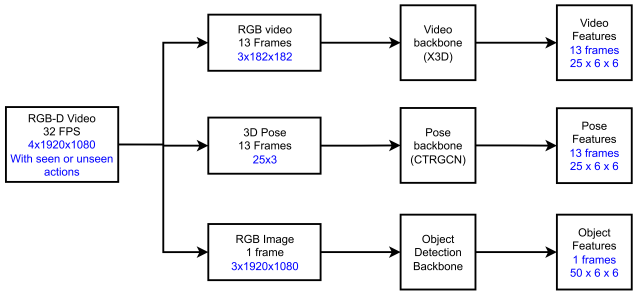

Fig. 2. Tri-modal feature extraction streams of VPOCLIP.

A. Tri-modal feature streams

As shown in Fig. 2, the input video is processed by three parallel streams: RGB appearance, 3D skeleton pose, and YOLO-based object detection. Each stream is encoded by its own backbone into video, pose, and object feature maps, which are then passed to the fusion module for alignment with the CLIP semantic space.

B. Fusion and embedding

Fig. 3 illustrates the two-stage spatial mid-fusion. The video and pose feature maps, each of size $13 \times 25 \times 6 \times 6$, are first concatenated along the channel dimension into a $13 \times 50 \times 6 \times 6$ representation, which is then concatenated with the $1 \times 50 \times 6 \times 6$ object feature map along the temporal dimension, yielding a unified tensor of size $14 \times 50 \times 6 \times 6$. Finally, cascaded spatial and temporal convolutions followed by a fully-connected projection reduce this tensor to a 512-dimensional ADL embedding, which is L2-normalized and aligned with the CLIP semantic space. The method is from [24].

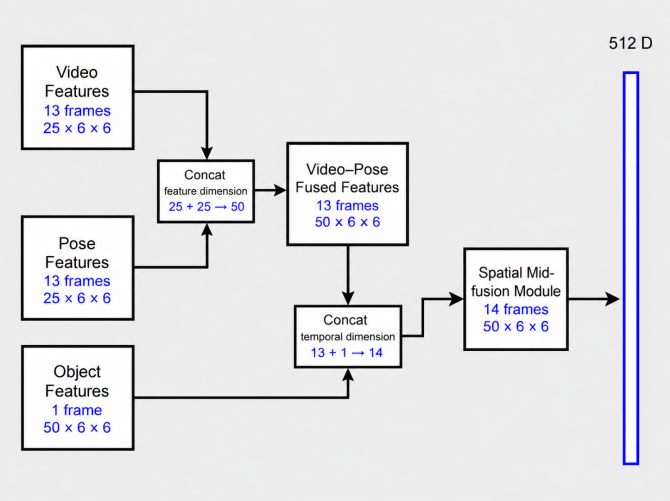

Fig. 3. Two-stage spatial mid-fusion and reduction to the 512-D ADL embedding.

C. Language-aligned contrastive training

Each seen action is described by a short sentence in an action description table. A frozen CLIP text encoder maps these descriptions to 512-dimensional text embeddings. Both the ADL embeddings and the text embeddings are L2-normalized, and an InfoNCE contrastive loss pulls each visual embedding toward the text embedding of its ground-truth action. Training thus aligns the tri-modal visual representation with the CLIP language space, while the text encoder remains unchanged.

D. Open-vocabulary inference and real-time pipeline

At inference time, all action descriptions are encoded offline once and cached as a text embedding bank. The visual encoder produces an L2-normalized ADL embedding, and the predicted class is the $\arg\max$ of its cosine similarity to the bank. To add a new action, a caregiver simply appends a one-sentence description (up to 25 words) to the table; it is encoded by the frozen CLIP text encoder and becomes immediately recognizable, with no retraining or parameter updates.

In the real-time pipeline, a 30-FPS RGB-D camera feeds the three streams, with MediaPipe providing skeleton keypoints. At every camera frame, 13 frames are sampled from a 2–4 s sliding history window and passed through the model, and a majority vote over the last ten predictions gives the final label. The interface displays the current top-1 action and the measured end-to-end latency in real time.

IV. Experiments

A. Setup

All experiments were carried out on ETRI-Activity3D [16], a large RGB-D corpus of daily activities recorded mostly from elderly subjects, from which we retain 55 action classes under a subject-disjoint 70/10/20 train/validation/test partition. Because no zero-shot protocol has, to our knowledge, been published for this dataset, we define four seen/unseen class splits (55/0, 50/5, 45/10 and 40/15) and release the class lists to allow replication. Class prototypes are obtained by encoding one natural-language description per action with a frozen CLIP ViT-B/32 text encoder [11]; recognition then reduces to cosine matching in the shared embedding space. For the generalized setting we adopt calibrated stacking [25], scaling unseen-class scores by a single coefficient $\gamma = 1.18$ selected by sweeping the harmonic mean. Training used one V100 GPU; latency was profiled on the same card and, separately, on a consumer laptop (RTX 3060).

TABLE I
Zero-shot action recognition results as reported in the original publications. Datasets and class draws differ across rows, so the numbers situate our benchmark rather than rank the methods.

Method	Benchmark	ZSL Top-1	GZSL (S/U/H)
SynSE [20]	NTU-60 55/5	64.19%	–
SMIE [27]	NTU-60 55/5	65.08%	–
STAR [26]	NTU-60 55/5	81.4%	–
ReViSE [28]	PKU-MMD 46/5	71.75%	52.06 / 60.34 / –
Ours	ETRI 50/5	69.28%	54.3 / 53.4 / 53.8
Ours	ETRI 45/10	53.45%	49.0 / 51.0 / 50.0
Ours	ETRI 40/15	36.07%	52.0 / 34.6 / 41.6

B. Zero-shot and generalized zero-shot recognition

Direct comparison with prior zero-shot skeleton methods is complicated by the fact that each line of work evaluates on different datasets and, at times, different class draws for nominally identical splits. The re-evaluation in [26], for instance, credits SynSE and SMIE with considerably higher accuracy than their original papers report, which suggests that protocol details matter as much as the methods themselves. Table I should therefore be read as context rather than as a ranking; our ETRI numbers are, at present, the only ones available for this dataset.

On the 50/5 split the model recognises unseen actions with 69.28% top-1 accuracy, which sits between the figures originally reported for SynSE and ReViSE on their respective benchmarks. Accuracy degrades as the unseen set grows, falling to 36.07% at 40/15; a similar trend is visible across the NTU-based literature and is arguably inherent to prototype matching with an expanding candidate set. Without calibration the generalized setting collapses almost entirely (harmonic means below 3%), since seen prototypes absorb nearly every unseen sample. A single shared γ restores the balance reported in Table I. We would caution that γ was tuned on the evaluated sets, as the protocol provides no independent unseen validation data; the resulting sensitivity, roughly ± 0.04 around the optimum, indicates that additive calibration might be preferable in deployment.

C. Eldercare-oriented subset

Service robots for older adults rarely need all 55 categories at once. We therefore repeated the generalized evaluation on a reduced vocabulary of 13 seen classes per split, drawn from eldercare-relevant actions such as eating, taking medicine, using a gas stove and falling, together with the full unseen set

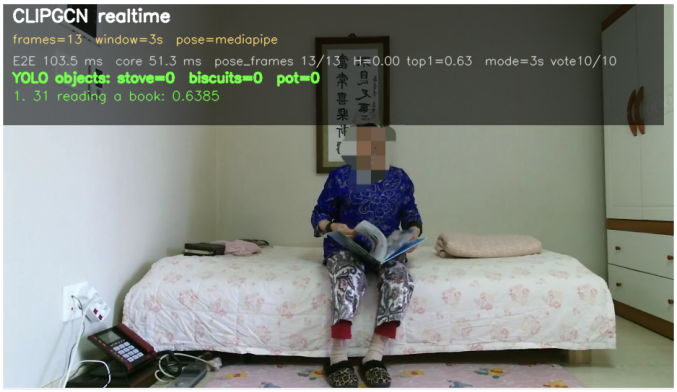

Fig. 4. Real-time recognition on a consumer laptop (RTX 3060). The system processes webcam input at roughly 100 ms per 13-frame clip while matching both seen and unseen action prototypes.

of the corresponding split. Under this configuration the 50/5 model reaches S $= 60.0\%$, U $= 55.6\%$ and H $= 57.7\%$, with the correct unseen class always contained in the top five predictions. The 45/10 and 40/15 variants attain harmonic means of 47.6% and 41.9%. These figures appear more representative of the intended use case than the full-vocabulary ones, though we acknowledge that a smaller candidate set naturally inflates absolute accuracy.

D. Efficiency

The fusion head that combines the three streams carries 13.0M trainable parameters; together with YOLOv5m [29] (21.1M), X3D-S [9] (3.1M) and CTR-GCN [6] (1.5M) the deployed stack totals roughly 38.6M parameters. The 151M-parameter CLIP text encoder is only used offline to precompute prototypes and never runs at inference time. Measured end to end at batch size one on a V100, a 13-frame clip is processed in 58.8 ms (detection 13.1, video backbone 15.8, skeleton backbone 27.9, fusion and matching 2.0), or about 17 clips per second; the live demo on a laptop RTX 3060 runs at approximately 45 ms per clip, excluding pose-estimation preprocessing. For comparison, embedded two-stream pipelines relying on optical flow have been reported to require several hundred milliseconds per inference [30]. Supervised accuracy is not sacrificed for this budget: on the closed 55-class test set the fused model reaches 78.1% top-1, on par with its strongest single stream and with a fine-tuned X3D baseline, while adding a zero-shot capability that closed-set classifiers lack by construction.

V. Discussion: Implications for HRI

This study supports the practical deployment of socially assistive robots in three ways. First, caregivers and family members can define target behaviors through natural-language descriptions, enabling zero-shot recognition without model retraining. Second, the lightweight architecture provides sufficiently low end-to-end latency for local robotic deployment. Third, recognition results can directly trigger human-adaptive behaviors, such as confirming medication dosage after detecting medication intake or proactively checking on a user after detecting fall-related actions.

However, several limitations remain. When seen and unseen classes are recognized jointly, dominant seen classes may suppress unseen classes and reduce zero-shot accuracy. Real-time predictions may also fluctuate during transitions between actions, which suggests an excessive sensitivity to short-term changes. In addition, selecting an appropriate scaling threshold between seen and unseen classes is difficult. Future work should investigate temporal smoothing and adaptive, class-specific thresholds.

VI. Conclusion

We presented VPOCLIP, a tri-modal recognition framework that couples appearance, skeletal motion, and object context with a frozen CLIP text encoder, so that activities of daily living can be recognised, and later extended, through natural language. The design targets a specific gap in assistive robotics: caregivers often need to add behaviours after deployment, yet retraining a closed-set model for every new action is rarely practical. Because the fused representation is aligned with language rather than a fixed label set, a new activity becomes recognisable once a single sentence is appended to the description table, with no gradient updates. On the elderly subset of ETRI-Activity3D the framework recognised unseen actions at 69.3% top-1 on the 50/5 split while retaining closed-set accuracy on par with its strongest single stream, and the full stack runs on a consumer laptop at roughly 45 ms per clip. The generalized setting remains the harder case, since seen prototypes tend to dominate unless the score calibration is chosen with care, and short-term prediction jitter during action transitions still limits the smoothness of the live system. We therefore regard temporal aggregation and adaptive, class-specific thresholds as the most promising next steps, together with a closer coupling between the recognised activity and the robot's adaptive response.

References

[1] S. A. Olatunji, V. Falcon, A. Ramesh, and W. A. Rogers, "Considerations for designing socially assistive robots for older adults," *Frontiers in Robotics and AI*, vol. 12, Art. no. 1622206, 2025.

[2] D. Yoon, J. Kim, and D. Lee, "Resource-efficient RGB-only action recognition for edge deployment," *arXiv preprint arXiv:2602.10818*, 2026.

[3] J. Park, M. Kang, and B. Han, "Class-incremental learning for action recognition in videos," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 13698–13707.

[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2018, pp. 7444–7452.

[5] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12026–12035.

[6] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 13359–13368.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 20–36.

[9] C. Feichtenhofer, "X3D: expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 203–213.

[10] Y. Zhang and Y. Wang, "A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications," *EURASIP J. Image and Video Processing*, vol. 2025, Art. no. 15, 2025.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.

[12] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: a new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[13] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *Proc. European Conf. Computer Vision (ECCV)*, 2022, pp. 1–18.

[14] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned CLIP models are efficient video learners," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6545–6554.

[15] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang, "Open-VCLIP: transforming CLIP to an open-vocabulary video model via interpolated weight optimization," in *Proc. Int. Conf. Machine Learning (ICML)*, vol. 202, 2023, pp. 36978–36989.

[16] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: a large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2020, pp. 10990–10997.

[17] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 568–576.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[19] A. Zhu, Q. Ke, M. Gong, and J. Bailey, "Part-aware unified representation of language and skeleton for zero-shot action recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18761–18770.

[20] P. Gupta, D. Sharma, and R. K. Sarvadevabhatla, "Syntactically guided generative embeddings for zero-shot skeleton action recognition," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2021, pp. 439–443.

[21] K. Zhou, S. Zhang, Z. You, J. Hu, M. Tan, and F. Liu, "Zero-shot skeleton-based action recognition with prototype-guided feature alignment," *IEEE Trans. Image Processing*, vol. 34, pp. 4602–4617, 2025.

[22] J. Do and M. Kim, "Bridging the skeleton-text modality gap: diffusion-powered modality alignment for zero-shot skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2025, pp. 12757–12768.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[24] F. Robinson and G. Nejat, "A deep learning human activity recognition framework for socially assistive robots to support reablement of older adults," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2023, pp. 6160–6167.

[25] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 52–68.

[26] Y. Chen, J. Guo, T. He, X. Lu, and L. Wang, "Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition," in *Proc. 32nd ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 778–786.

[27] Y. Zhou, W. Qiang, A. Rao, N. Lin, B. Su, and J. Wang, "Zero-shot skeleton-based action recognition via mutual information estimation and maximization," in *Proc. 31st ACM Int. Conf. Multimedia (ACM MM)*, 2023, pp. 5302–5310.

[28] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 3591–3600.

[29] G. Jocher *et al.*, "Ultralytics YOLOv5," 2020. [Online]. Available: `https://github.com/ultralytics/yolov5`

[30] R. Wang, Z. Wang, P. Gao, M. Li, J. Jeong, Y. Xu, Y. Lee, C. M. Baum, L. T. Connor, and C. Lu, "Real-time human action recognition on embedded platforms," *arXiv preprint arXiv:2409.05662*, 2024.