Inertial Strengthened CLIP model for Zero-shot Multimodal Egocentric Activity Recognition

Mingzhou He, Haojie Wang, Shuchang Zhou, Qingbo Wu*, King Ngi Ngan, Fanman Meng, Hongliang Li
School of Information and Communication Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
* {Correspondence: qbwu@uestc.edu.cn}

***Abstract*—The popular CLIP model has empowered various zero-shot learning tasks by unifying them into the vision-language alignment framework. However, due to the dynamic subject-object interactions and complex motion variations, the egocentric videos become more diverse and widen the gap between the vision and language data, which limits the applicability of the CLIP model. Thanks to the widely used Inertial Measurement Unit (IMU) in wearable devices, this paper proposed an Inertial Strengthened CLIP (IS-CLIP) model to refine the visual representation of egocentric videos, which achieves highly effective zero-shot multimodal egocentric activity recognition. Our IS-CLIP is composed of two modules, i.e., the Subject-object Interaction Refinement Module (SIRM) and the Subject-motion Guided Aggregation Module (SGAM). On the one hand, SIRM embeds the motion patterns of IMU into the visual representation of each frame to enhance the action-related features. On the other hand, the SGAM maps the IMU to the frame-wise weights to aggregate the visual representations of all frames, which reduces the differences between the videos of the same action with different speeds, directions, and temporal locations. Experiments on the UESTC-MMEA-CL dataset show that the proposed IS-CLIP outperforms many state-of-the-art methods in the zero-shot multimodal egocentric activity recognition task.**

***Index Terms*—egocentric vision, multimodal activity recognition, IMU, zero-shot learning, CLIP**

I. Introduction

Egocentric activity recognition (EAR) has become a significant research direction, with the widespread application of wearable devices in fields such as health monitoring, human-robot interaction, etc [1], [2]. In recent years, numerous outstanding fully supervised deep learning methods for egocentric activity recognition have emerged [3]–[5], leveraging wearable device sensors such as cameras, inertial sensors, and optical flow. For example, Huang et al. [6] introduce the knowledge graph and present a knowledge-driven egocentric activity recognition framework. Hao et al. [7] extend a single inertial sensor to two-hand inertial sensors and propose a two-branch late-fusion framework. With the development of technology, practical application scenarios have proposed zero-shot demand, which requires that the model be trained on base categories while being capable of recognizing novel categories [8], to reduce the labor-intensive data acquisition and annotation. However, none of the previously mentioned methods possess zero-shot recognition capabilities.

Contrastive Language Image Pre-training (CLIP) [9], which is trained on hundreds of millions of image-caption pairs collected from the Internet, has good feature representation

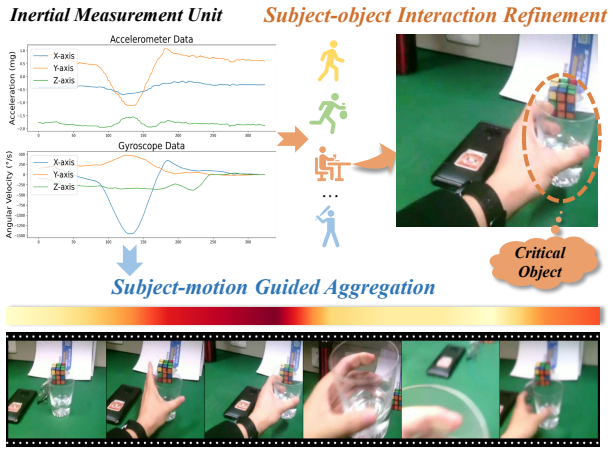

Fig. 1. The IMU provides supplementary information on subject motion, which is utilized to refine CLIP's visual representation. (a) Subject-object Interaction Refinement: In hand movements mode, we should focus on the cup in hand rather than the irrelevant phone. (b) Subject-motion Guided Aggregation: Utilize the IMU's temporal features to emphasize video frames with significant motion.

for images and texts. There are many CLIP-based image recognition works [10]–[12] that show amazing zero-shot performance. Recently, some research has applied CLIP to video input and achieved zero-shot third-person video activity recognition. ActionCLIP [13] proposed the "pre-train, adapt, and fine-tune" paradigm for activity recognition and expanded text labels to enhance the matching between video and text. OpenVCLIP [14] models spatial-temporal relationships in videos and proposes the interpolated weight optimization method to maintain the generalization of CLIP after training. FROSTER [15] adopts a residual feature distillation approach to ensure that CLIP retains its generalization while effectively adapting to activity recognition.

However, zero-shot methods based on videos typically encounter two significant challenges when applied to egocentric activity recognition. First, egocentric videos capture complex external environments rather than the action subject. Numerous irrelevant objects appear in the video, severely impacting CLIP's visual representation. Second, changes in first-person video may reflect environmental object movement rather than the subject's motion, causing significant interference with activity recognition. Inertial sensors on wearable devices provide additional motion data of the action's subject, offering crucial

compensation to mitigate these interferences and achieve more accurate visual-language alignment. To our knowledge, there is currently a lack of research on integrating IMU data into zero-shot egocentric activity recognition based on CLIP.

Therefore, we introduce the IMU modality and utilize IMU's subject motion information to strengthen the CLIP's visual representation. To this end, we design the Subject-object Interaction Refinement Module (SIRM) and Subject-motion Guided Aggregation Module (SGAM). Specifically, egocentric videos usually contain multiple background objects, most of which are irrelevant to the activity. IMU time-frequency features correspond to typical motion mode, such as hand and head movements. The SIRM utilizes IMU time-frequency features as query vectors to guide the focus on objects related to behavior patterns. On the other hand, the SGAM module leverages the temporal features of IMU to emphasize video frames with significant motion, reducing the interference from frames without subject motion and the differences between instances of the same action caused by the environment. In addition, our method employs feature distillation [16] to more effectively embed the knowledge of CLIP into the multimodal model.

Overall, we are the first to propose a zero-shot multimodal (video and IMU) egocentric activity recognition framework based on CLIP, called IS-CLIP. We introduce IMU data to refine the CLIP visual representation of video to enhance the zero-shot performance. Our method achieves state-of-the-art performance in two zero-shot settings on the UESTC-MMEA-CL [17] dataset, establishing an exemplary benchmark for zero-shot multimodal egocentric activity recognition.

II. Method

A. Preliminaries

First, we introduce the procedure of zero-shot recognition based on CLIP and apply it to a single video modality. Zero-shot recognition aims to classify the unseen category set $\mathcal{C}_{\mathcal{N}}$ through the classification network trained on the seen category image set $\mathcal{C}_{\mathcal{B}}$, where $\mathcal{C}_{\mathcal{B}} \cap \mathcal{C}_{\mathcal{N}} = \emptyset$. In this context, the traditional classification head relying on fully connected layers loses its effectiveness. Thanks to CLIP's rich image-text knowledge and image-text matching paradigm [9], CLIP is often used as a classifier for various tasks to easily achieve zero-shot capabilities [18]–[20]. Specifically, an image $\mathcal{I}_i$ is passed through the visual encoder $\mathcal{V}(\cdot)$ to obtain a visual embedding v_i. On the other side, constructed text inputs like "a photo of a $[Class]$" are fed into the language encoder $\mathcal{L}(\cdot)$ to obtain text embeddings $\{t_1, t_2, ...t_n\}$, where n is the number of classes. Then, the cosine similarity between the visual embedding v_i and each text embedding $t_j, j \in [1, n]$ is calculated by:

$$cos(v_i, t_j) = \frac{v_i \cdot t_j^T}{\|v_i\| \cdot \|t_j\|}, j \in [1, n] \tag{1}$$

and the category c_i of the image $\mathcal{I}_i$ is defined as:

$$c_i = \underset{j \in [1,n]}{\arg\max} \cos(v_i, t_j) \tag{2}$$

In this way, zero-shot recognition can be achieved by just appending semantic inputs of novel categories. For video input, the method of equally spaced frame extraction is usually used, and the input is actually a set of time-ordered images $\{\mathcal{I}_{t_1}, \mathcal{I}_{t_2}, ...\mathcal{I}_{t_N}\}$. All of them passed through the visual encoder to obtain visual embeddings $\{v_{t_1}, v_{t_2}, ...v_{t_N}\}$. These embeddings are then averaged into one visual embedding and matched with the semantic embeddings for classification according to Eq. (1)(2). This CLIP-based single video modality activity recognition method should be considered as our baseline.

B. Framework Pipeline

The overall framework of zero-shot multimodal egocentric activity recognition is illustrated in Fig. 2. The input of the model is N frames extracted from the video and IMU data of the same time period. IMU data includes accelerometer and gyroscope, which are one-dimensional time series data. In order to better extract the time domain and frequency domain features of IMU data, we performed the short-time Fourier transform (STFT) [21] on the IMU data. The STFT divides a long-time signal into several short periods and then applies the Fourier transform to each period. In this way, the spectrum of the signal in each time period can be obtained, thereby describing the time-frequency characteristics. The formula for STFT is defined as:

$$\mathrm{STFT}(x(t))(\tau, \omega) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j\omega t}\, dt \tag{3}$$

where $x(t)$ is the time signal, $w(t-\tau)$ is the window function, τ is the time offset, and ω is the frequency.

In the visual branch, we built two identical ViT-B/16 visual encoders and initialized them with CLIP pre-trained weights. In order to maintain the generalization of CLIP, we adopt knowledge distillation. One of the encoders is regarded as the teacher model, and its parameters are completely frozen. The other one is the student model whose last two blocks in the transformer and the last fully connected layer are fine-tuned during training. N frames are respectively passed through the teacher encoder and the student encoder to obtain visual features f_v^t and $f_v \in \mathbb{R}^{N \times 512}$, and feature distillation is performed between them. The distillation loss $\mathcal{L}_{kd}$ is defined as:

$$\mathcal{L}_{kd} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{f_{vi}^t \cdot f_{vi}}{\|f_{vi}^t\| \|f_{vi}\|} \tag{4}$$

In the IMU branch, the accelerometer and gyroscope spectrograms obtained by STFT are sent to the IMU encoder to extract their respective time-frequency features. Note that the two sensors share one encoder to reduce the number of parameters. The two sensor features are then concatenated and reduced dimension through a linear layer to obtain the IMU feature $f_{IMU} \in \mathbb{R}^{1 \times 512}$.

Next, we need to consider how to integrate visual and IMU features. The presence of motion-irrelevant objects in first-person videos and pseudo-motion caused by these objects can

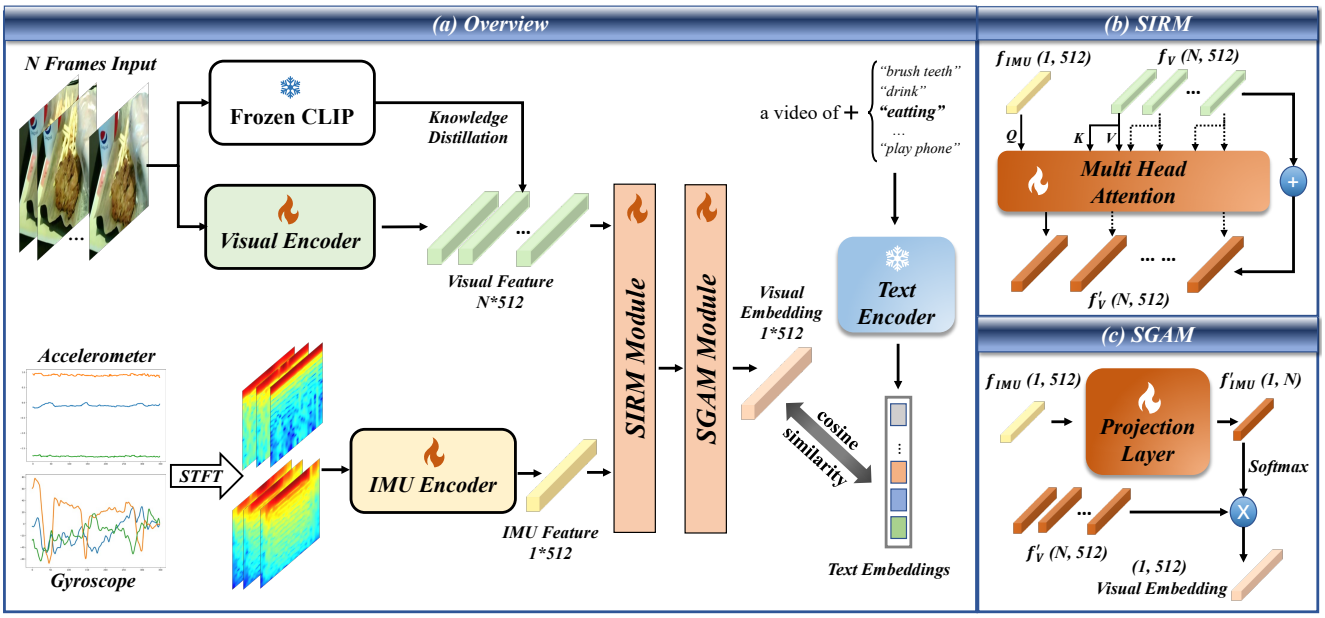

Fig. 2. (a) An overview of our proposed zero-shot multimodal egocentric activity recognition framework. The input is temporally ordered frames and IMU data (accelerometer and gyroscope). Among them, STFT stands for short-time Fourier transform, (b) SIRM: Subject-object Interaction Refinement Module. (c) SGAM: Subject-motion Guided Aggregation Module. Note that N in the figure represents N frames of an instance rather than the batch size.

interfere with CLIP's visual representation. The supplementary information on subject motion provided by IMU data can mitigate these interferences. Meanwhile, CLIP has never encountered IMU data, directly fusing IMU and visual features would harm generalization. For these reasons, we design the SIRM and SGAM modules, which use IMU features f_{IMU} to refine visual representation capabilities, resulting in enhanced visual embeddings $v \in \mathbb{R}^{1\times512}$.

Finally, the visual embedding is matched with the text embedding for category prediction, as Eq. (1)(2). On the text processing side, it is exactly the same as vanilla CLIP, with all parameters frozen. The overall loss $\mathcal{L}$ of the framework is defined as:

$$\mathcal{L} = \mathcal{L}_{CrossEntropy} + \mathcal{L}_{kd} \tag{5}$$

C. Subject-object Interaction Refinement Module (SIRM)

In egocentric videos, there are usually many objects that are irrelevant to the behavior, which are mixed into the visual features to affect the image-text matching of CLIP. The main advantage of IMU data is that it can provide subject's motion information, which can reflect the basic motion mode, and we should focus on the objects related to motion in vision. Inspired by this, we present the Subject-object Interaction Refinement Module (SIRM). We use the IMU feature $f_{IMU} \in \mathbb{R}^{1\times512}$ as query vectors and visual features $f_v \in \mathbb{R}^{N\times512}$ as both 'key' and 'Value' for multi-head attention calculation. Note that f_{IMU} performs attention calculations with each of the N dimensions in f_v. The calculation process is as:

$$Q_i = f_{IMU}W_{Q_i},\ K_i = f_vW_{Ki},\ V_i = f_vW_{Vi} \tag{6}$$

$$head_i = \text{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_k}}\right)V_i \tag{7}$$

$$f_v^{'} = \text{Concat}(\text{head}_i)W_O,\ i \in [1,8] \tag{8}$$

Among them, W_Q, W_K, W_V, W_O are weight matrices, and i is the number of heads. In this way, the interaction between IMU and visual features generates the weighted mask, which is applied to enhance key object attention to refine the visual features. Furthermore, we also add residual connections to the visual features before and after refinement, thereby preserving the generalization capabilities of CLIP.

D. Subject-motion Guided Aggregation Module (SGAM)

Among the N frames extracted at equal intervals from the video, not all frames are in the period of subject movement. Furthermore, changes between frames may be caused by the motion of unrelated objects in the video rather than the movement of the subject. The video-based single-modal method cannot effectively identify critical frames, and directly averaging visual features will dilute essential information. The IMU's temporal features can accurately reflect the strength of the subject's movement. Therefore, we propose the Subject-motion Guided Aggregation Module (SGAM), as shown in Fig. 2(c). We map the IMU features $f_{IMU} \in \mathbb{R}^{1\times512}$ to N dimensions in the time series through a projection layer (MLP) and apply softmax to obtain the temporal weighted mask. Then, weighted fusion is performed on the N temporally ordered visual features $f_v^{'} \in \mathbb{R}^{N\times512}$ so that the final visual embedding v contains more information from periods of the subject's intense activity.

III. Experiments

We mainly report the results of our method on zero-shot egocentric activity recognition and compare them with other

TABLE I
ZERO-SHOT EGOCENTRIC ACTIVITY RECOGNITION RESULTS UNDER TWO EXPERIMENT SETTINGS. THE FINE-TUNED CLIP IS OUR BASELINE, WHICH USES KNOWLEDGE DISTILLATION. HM IS THE HARMONIC MEAN.

Method	Publication	Prompt Augment	Base(16) Novel(16)			Base(24) Novel(8)		
			Acc^{novel}	Acc^{base}	HM	Acc^{novel}	Acc^{base}	HM
ActionCLIP [13]	Arxiv2021	✓	35.0	99.1	51.7	44.7	98.8	61.5
XCLIP [22]	ECCV2022	×	41.6	99.7	58.7	54.6	98.8	70.4
VPT [23]	ECCV2022	×	31.1	94.6	46.8	39.1	94.5	55.3
Text4Vis [24]	AAAI2023	×	43.9	98.8	60.8	55.2	98.2	70.7
ViFi-CLIP [25]	CVPR2023	×	55.3	98.2	70.7	63.7	96.9	76.9
Open-VCLIP [14]	ICML2023	✓	63.9	96.6	76.9	75.9	95.3	84.5
FROSTER [15]	ICLR2024	✓	64.0	97.0	77.1	71.4	97.1	82.3
CLIP [9]	ICML2021	×	58.5	55.2	56.8	68.0	44.5	53.8
Fine-tuned CLIP	-	×	60.5	97.6	74.7	74.2	96.6	83.9
IS-CLIP(ours)	-	×	**66.3**	97.7	**79.0**	**77.8**	97.2	**86.4**

state-of-the-art CLIP-based activity recognition methods. Further, we conduct an in-depth analysis of the proposed modules' effectiveness. Finally, we conduct ablation experiments on both modalities and method components.

Dataset and metric. We conduct experiments on the UESTC-MMEA-CL [17] dataset, which is a multimodal dataset of first-person perspective activity recognition acquired by smart glasses. Alongside first-person perspective video from the camera, it also includes data streams from acceleration and gyroscope sensors. The dataset consists of 32 categories of human daily activities, with 4,553 samples in the training set and 1,316 samples in the test set. We adopt two zero-shot settings on this dataset, i.e., the first 16 classes as base classes and the first 24 classes as base classes, separately. Since we focus on zero-shot performance, the critical evaluation metric is the accuracy of novel classes.

Implementation details. Our model adpot ViT-B/16 as the visual encoder and initialize with CLIP pretrained weights. We train the model on base catrgories for 50 epochs on an RTX4090 GPU, and the batch size is 16. We utilize different optimizers and learning rates for various model parameters. For the visual encoder of CLIP, we employ the Adam optimizer with a learning rate of 1e-5. SGD is the optimizer for other parameters with a learning rate of 1e-3.

A. Zero-shot Egocentric Activity Recognition

To validate the superiority of our multimodal framework, comparative experiments are conducted under two different zero-shot settings. The experimental results are presented in Table I. All comparison methods use only the video modality as input, as they inherently lack the capability to adapt to inertial measurement unit (IMU) data input. These methods are all based on CLIP, thus possessing certain zero-shot recognition capabilities. Therefore, comparing our method with these advanced methods is reasonable. We adopt the same fine-tuned method as described in this paper, including knowledge distillation, to obtain a robust baseline, which is referred to as Fine-tuned CLIP in Table I.

The experimental results show that our method outperforms the baseline by 5.8% and 3.6% in novel categories' accuracy (Acc) under the two settings, respectively. This demonstrates that our proposed modules effectively adapt to IMU input and leverage the advantages of IMU data to enhance zero-shot detection performance. Compared to the second-best method, our novel categories' Acc is higher by 2.3% and 1.9%, achieving state-of-the-art performance. Notably, we don't use prompt augmentation as some other methods do. This further showcases the superiority of the multimodal approach. Additionally, we also achieved the best performance in the harmonic mean metric, indicating that our multimodal method improves the overall network performance rather than boosting zero-shot recognition at the expense of base class performance.

B. Analysis of Visual Feature

Our goal is to embed CLIP knowledge into a multimodal framework based on video and IMU data, to achieve zero-shot egocentric activity recognition that is superior to single video modality. To this end, we design the SIRM and SGAM modules to adapt to the input of IMU data and enhance zero-shot capabilities. To verify the effectiveness of our method, two aspects should be considered: (1) whether the two modules enhance the representation of visual features. (2) whether the addition of IMU modality enhances or impairs CLIP's image-text matching ability. In this subsection, we examine the effect of our method at the feature level.

Firstly, we project the visual embeddings of our method and the baseline into 2D space using the t-SNE algorithm to observe their distribution. As shown in Fig. 3, the visual embeddings of our method are significantly more separable after the two modules guided by IMU data. Among them, the distance between classes is increased dramatically, and the distribution within the class is more compact. More importantly, the problematic categories that were initially mixed are also successfully distinguished. These results prove that our

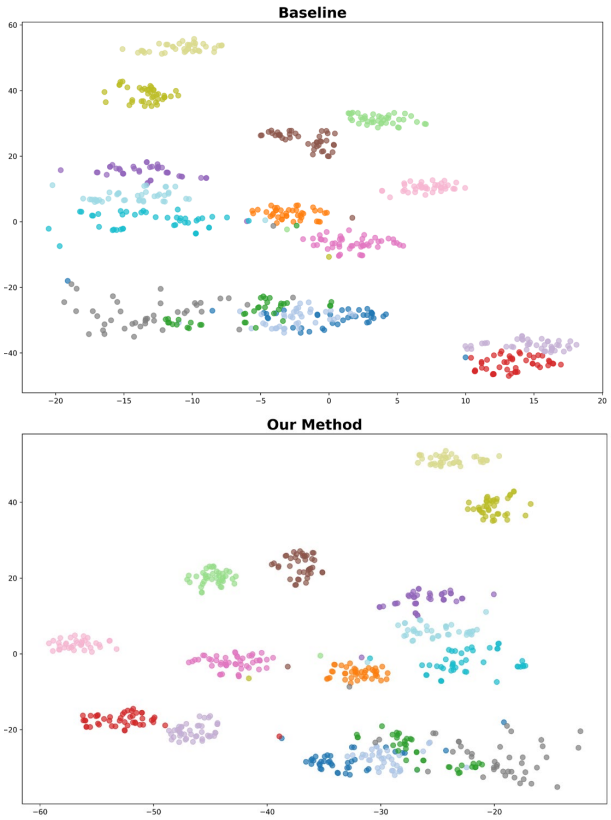

Fig. 3. t-SNE plots of our method and baselines' visual embeddings on novel categories. Our method shows better separability.

method exactly enhances the visual representation. In addition, we count the cosine similarity of each novel category instance's visual embeddings and their corresponding text embeddings, and show the mean of category dimension. As shown in Fig. 4, the results show that the similarity between visual and semantic embedding has increased in almost all categories. This verifies that the integration of IMU data does not destroy the knowledge of CLIP, and the ability of zero-shot recognition is improved as expected after being refined by our method.

C. Ablation Study

We conduct ablation experiments in (Base:16, Novel:16) zero-shot setting of UESTC-MMEA-CL dataset. Table II shows that SIRM and SGAM modules can effectively improve the accuracy of novel categories. Among them, the improvement brought by adding SGAM is better, and the improvement in the base categories is also more apparent. This is because the recognition method based only on video frame extraction naturally lacks temporal information regardless of whether it is zero-shot. However, temporal features can be easily obtained through IMU data. The employment of SIRM will slightly reduce the performance of base categories, which may be because the feature representation learned based on full supervision is more accurate than the refinement guided by IMU.

Furthermore, we also conduct ablation experiments on accelerometer and gyroscope modalities, as shown in Table III.

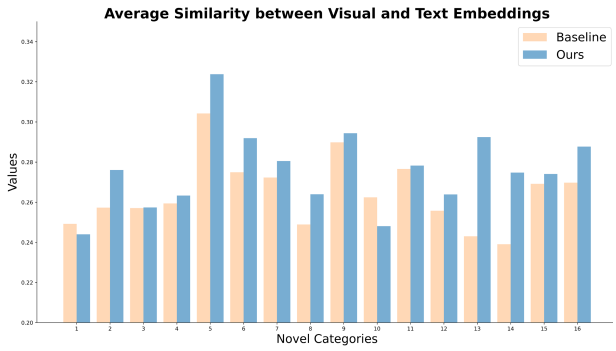

Fig. 4. Comparison of the average cosine similarities of visual and text embeddings on novel categories.

The results show that the integration of both modalities can effectively improve the performance of the novel categories, which further proves the superiority of our proposed multimodal framework.

TABLE II
ABLATION STUDY OF PROPOSED MODULES.

SIRM	SGAM	*Base*	*Novel*	*HM*
		97.6	60.5	74.7
✓		97.1	63.9	77.0
	✓	**98.6**	64.6	78.1
✓	✓	97.7	**66.3**	**79.0**

TABLE III
ABLATION STUDY OF ACCELEROMETER AND GYROSCOPE MODALITIES.

Accelerometer	Gyroscope	*Base*	*Novel*	*HM*
		97.6	60.5	74.7
✓		97.7	64.1	77.4
	✓	**98.3**	65.3	78.5
✓	✓	97.7	**66.3**	**79.0**

IV. CONCLUSIONS

In this paper, we pioneeringly proposed a multimodal zero-shot egocentric activity recognition framework based on video and Inertial Measurement Unit (IMU). We embed the knowledge of CLIP into this framework to achieve zero-shot recognition capabilities and design two modules to adapt IMU input to enhance zero-shot performance. Our method achieves state-of-the-art zero-shot egocentric activity recognition performance, to establish a benchmark for subsequent research based on video and IMU in this field.

ACKNOWLEDGMENT

This work was supported in part by the National Science and Technology Major Project (No. 2021ZD0112001), the National Natural Science Foundation of China (No. U23A20286), the Independent Research Project of Civil Aviation Flight Technology and Flight Safety Key Laboratory (FZ2022ZZ06), and the Natural Science Foundation of Sichuan Province (2023NSFSC1972).

References

[1] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106 970, 2021.

[2] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2620–2628.

[3] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic attention with privileged information for egocentric action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 249–12 256.

[4] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.

[5] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1868–1877.

[6] Y. Huang, X. Yang, J. Gao, J. Sang, and C. Xu, "Knowledge-driven egocentric multimodal activity recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 4, pp. 1–133, 2020.

[7] Y. Hao, A. Kanezaki, I. Sato, R. Kawakami, and K. Shinoda, "Egocentric human activities recognition with multi-modal interaction sensing," *IEEE Sensors Journal*, 2024.

[8] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 384–400.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[11] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.

[12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

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

[14] 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 *International Conference on Machine Learning*, PMLR, 2023, pp. 36 978–36 989.

[15] X. Huang, H. Zhou, K. Yao, and K. Han, "Froster: Frozen clip is a strong teacher for open-vocabulary action recognition," *arXiv preprint arXiv:2402.03241*, 2024.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] L. Xu, Q. Wu, L. Pan, *et al.*, "Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 2430–2443, 2024. DOI: 10.1109/TMM.2023.3295899.

[18] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[19] Y. Zhong, J. Yang, P. Zhang, *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 793–16 803.

[20] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 084–14 093.

[21] S. Kumawat, M. Verma, Y. Nakashima, and S. Raman, "Depthwise spatio-temporal stft convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4839–4851, 2021.

[22] B. Ni, H. Peng, M. Chen, *et al.*, "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–18.

[23] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*, Springer, 2022, pp. 105–124.

[24] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 2847–2855.

[25] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned clip models are efficient video learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6545–6554.