Integrating Semantic Knowledge for Enhanced Weakly-Supervised Group Activity Recognition

Muhammad Adi Nugroho†, Jinyoung Park†, Yeeun Seong†, and Changick Kim†
† Korea Advanced Institute of Science and Technology, Republic of Korea
E-mail: {madin, jinyoungpark, yeeunseong, changick}@kaist.ac.kr

***Abstract*—Large Vision-Language Models (LVLMs) have recently emerged as powerful tools for jointly modeling vision and language information, enhancing semantic representations. In this work, we introduce a method that utilizes high-level semantic knowledge of LVLM into the Weakly-Supervised Group Activity Recognition (WSGAR) task. Our Semantic Integrating Activity Recognition Network (SInAR-Net) enriches visual representations with language-based semantic information to construct complex spatio-temporal actor relationship. For efficient learning, we transfer the semantic knowledge via pre-extracted text features of contextual information generated through multiple prompt generations. Then, we construct multi-modal relationship between visual actor features and semantic text features using our semantic integration module. We ensure no drastic additional calculation cost by alleviating the use of LVLM in the inference stage, as our semantic integration module intrinsically learned the cross-modal semantic knowledge. Experiments on WSGAR benchmarks demonstrate competitive performance of our method, and ablation studies show the effectiveness of our novel semantic understanding components.**

I. Introduction

Group Activity Recognition (GAR) has gained significant attention, with wide-ranging applications spanning surveillance, sports analytics, and social behavior studies. GAR aims to automatically identify and classify collective behaviors exhibited by multiple individuals in the video sequence. To tackle this, various efforts have been made, including using multi-stage RNNs [1] and graph-based models such as Graph Convolutional Network (GCN) [2]. However, many GAR studies do not sufficiently consider the high cost and time consumption associated with detailed actor-level annotation, such as actor bounding boxes and action classes.

To address these challenges, Weakly Supervised Group Activity Recognition (WSGAR) has emerged as a promising approach. WSGAR leverages coarse video-level labels instead of fine-grained frame-level or individual-level annotations, significantly reducing the annotation burden. Recent advancements in WSGAR have focused on modeling complex interactions between multiple actors and building meaningful features from video frames to capture the essence of collective activities. Yan [3] presented a Social Adaptive Module (SAM) that combines off-the-shelf object detectors with relational graphs, introducing the NBA dataset as a challenging benchmark. Another study, Kim [4], proposed a detector-free approach inspired by DETR [5], utilizing a transformer decoder with learnable queries to directly identify relevant actors from the video data. To address the challenge of motion complexity,

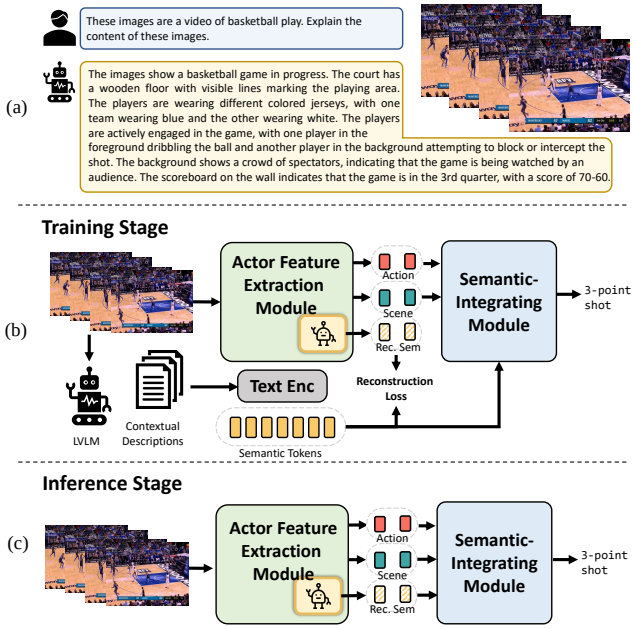

Fig. 1. In (a), we generate contextual information related to the group activity with LVLM. In the training (b), the model learns to integrate the visual features generated from actor feature extraction module, and the language semantic features extracted from the generated contextual descriptions and semantic tokens. In the inference stage (c), the model only requires the video frames as the semantic knowledge has been integrated inside the semantic module, alleviating the expensive cost of LVLM.

Du [6] introduced a 3D convolutional network that decouples camera motion from actor motion, effectively mitigating camera-induced noise and accurately capturing relative actor movements in dynamic scenes.

Recent advancements in GAR have emphasized the importance of generating and utilizing semantic knowledge. Some approaches have introduced detailed semantic information, such as player positions [7] for action captioning. However, annotating player positions across frames is labor-intensive and is not scalable to complex video settings. As another approach for WSGAR, Wu [8] proposed hierarchically organizing group activity labels (e.g., 'score', 'success/failure status'). They generate semantic embeddings through multi-hot encoding, and fuse these embeddings with visual scene data to extract activity-specific features. Yet, such categorical approaches suffer from limited expressiveness, as multi-hot encoding can only represent predefined label combinations and fail to capture the complex contextual relationships in group

979-8-3315-7206-8/25/$31.00

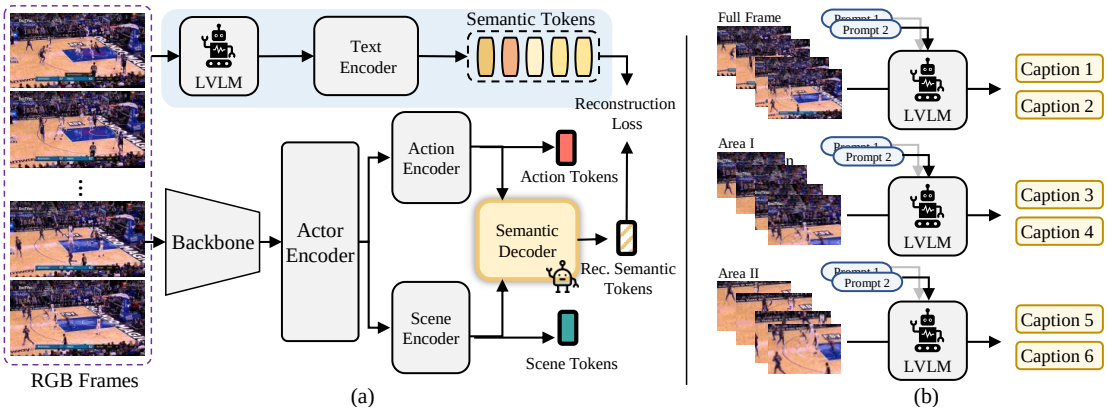

Fig. 2. (a) The actor feature extraction module firstly consist of visual backbone and actor encoder to extract actor features, which are then processed through the action encoder and scene encoder. The action encoder generate tokens that represent individual actor actions, while the scene encoder generate tokens that semantically represent the spatial composition and positioning patterns of multiple actors within each frame. To integrate language semantic knowledge in this level, a semantic decoder facilitates semantic reconstruction loss that encourages the visual tokens to be able to recreate the semantic tokens generated by LVLM. The decoder also generates semantic tokens when no LVLM is available in the inference stage. (b) To enhance generalization, we employ a prompt-based captioning strategy that enables the LVLM to generate diverse textual responses for each input. This multi-caption supervision provides richer semantic signals during training, facilitating more robust learning.

activities. In parallel, the rapid growth of Vision-Language Models (VLMs) that incorporates the high semantic level of natural language processing to vision understanding language processing has opened new avenues for GAR. VLMs such as CLIP [9], demonstrate significant potential in bridging visual and linguistic information. Zhang [10] validated the use of language-driven representations in zero-shot GAR by combining multi-level prompts and label semantics. We take a step further to incorporate the Large Vision-Language Models (LVLMs), which are built upon the success of VLMS and are gaining significant attention artificial intelligence field. LVLMs such as LLaVA [11] represent a substantial leap forward in multimodal understanding, offering sophisticated capabilities backed by large language models with complex semantic understanding. The visual information interpretation and reasoning capabilities of LVLMs are closely aligned with WSGAR which needs to intrinsically understand spatio-temporal inter actor relationships. However, the direct application of LVLMs to WSGAR tasks has been limited due to their substantial computational requirements and large parameter counts.

To address this limitation, we propose SInAR-Net (Semantic Integrating Activity Recognition Network), a novel approach that addresses the limitations of existing WSGAR methods by harnessing the potential of Large Vision-Language Models (LVLMs) in a computationally efficient manner. Our experimental results validate the effectiveness of our approach, showing substantial gains in recognition accuracy across various weakly supervised group activity recognition benchmarks.

II. Proposed Method

The architecture of SInAR-Net is illustrated in Fig. 1. It consists of several sub-modules to extract feature representations from a sequence of RGB frames $\mathbf{X_0} \in R^{T\times 3\times H_0\times W_0}$, which is then classified into one of the activity classes $y = \{0,1\}^C \in R^C$. The overall framework consists of three main stages: (1) acquiring semantic information from LVLM, (2) constructing actor and positioning representations from visual inputs, and (3) integrating semantic and visual representations to infer group activity.

A. Gaining Semantic Information

The main challenge of our work is finding the source of semantic information related to each video, as we only have video-level labels available in WSGAR. One option is to directly apply LVLM into GAR, however, due to the number of parameters, it requires much higher memory and computation costs compared to state-of-the-art GAR models, making both training and inference impractical. Thus, as a workaround, we propose the strategy of pre-extracting features before the training process. We utilize multi-image LVLM, such as LLAVA-NeXT Interleave [12], to generate semantic information in the form of a caption that represents the whole video given a sequence of images and a prompt. We use two methods to increase the variety of captions as shown in Fig. 2(b); (i) utilizing multiple captions, (ii) spatially splitting the video into multiple clips (or tubes in 3D) and generating captions for each clip. Then, we use an off-the-shelf text embedding model to transform the captions into vectorized form and store them in a caption library. In the training loop, we get video semantic tokens s for each video by sampling N_s number of their token sets from the library.

B. Building Actor and Positioning Representations

As illustrated in Fig. 2(a), with the semantic information extracted, we construct our network by first obtaining representations of individual actors within the group activity. Our initial stage follows the scheme of DFWSGAR [4] that uses a 2D backbone F_B to extract feature maps $\mathbf{X_B} \in R^{T\times D\times H_B\times W_B}$. Then, an actor encoder is applied to perform detector-free tokenization of actor features on a frame-by-frame basis, generating actor tokens. To build richer representations, we transform actor tokens into higher-level representations, as individual tokens capture only isolated observations at specific

time steps. Accordingly, we introduce two complementary forms: action tokens and scene tokens. Action tokens Z_{action}, generated by an action encoder, capture the dynamics of individual actor behaviors across the entire sequence. Scene tokens Z_{scene} , produced by a scene encoder, encode the spatial composition and positioning patterns of multiple actors within each frame.

We propose a semantic decoder that distills LVLM knowledge into semantic tokens by utilizing action and scene tokens, enabling rich understanding capabilities without requiring LVLM inference at test time. Specifically, the reconstructed semantic tokens $\tilde{s}$ are obtained via $\tilde{s} = f_d(Z_{action}, Z_{scene})$, where f_d denotes the semantic decoder. To align the decoder output $\tilde{s}$ with the target semantic tokens s, the semantic decoder is trained with a reconstruction loss as follows:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N_{\text{s}}} \sum_{i=1}^{N_{\text{s}}} (s_i - \tilde{s}_i)^2. \tag{1}$$

C. Semantic-Integrating Group Activity Representation

After obtaining semantic and visual representations, we process them using multiple encoders to aggregate them into a representative feature for group activity recognition (Fig. 3).

1) Action to Activity Encoder (A2Actv Encoder): To establish relationships across individual actions, this encoder employs a Multi-Head Self-Attention (MHSA) block. The MHSA mechanism enables global connections between actions regardless of their relative positions in the feature map. Since the global actor-actor interactions are similar to those in the scene encoder, we share MHSA weights between the A2Actv encoder and the scene encoder to improve parameter efficiency.

2) Scene to Activity Encoder (S2Actv Encoder): This encoder uses a stack of 2D convolutional layers to process scene token representations. It is designed to capture the consecutive temporal evolution of spatial configurations by modeling how neighboring actors behave across the activity sequence.

3) Visual-Semantic Encoder: As illustrated in Fig. 3(b), this encoder jointly processes visual tokens alongside semantic tokens. The encoder facilitates information exchange between all token types through cross-modal attention mechanisms. During training, we apply random masking to semantic tokens, replacing masked tokens with a learnable mask token to improve encoder robustness. A class token is also included to aggregate activity-relevant information. From the encoder, we obtain the Visual-Language Actor token (VLA), Visual Language Class token (VLC), and Visual-Language Scene (VLS) token. The resulting tokens are fed into a group activity classifier with three classification heads, each corresponding to the VLA, VLC, and VLS tokens. The outputs of these heads are aggregated to produce the final group activity prediction.

D. Learning Objectives

We train the network with a combination of multiple losses. The first one is the cross-entropy loss of the activity prediction $\mathcal{L}_{\text{CE}}$. The second one is $\mathcal{L}_{\text{gf}}$, an auxiliary loss in the form of per-frame cross-entropy loss intended to make the model

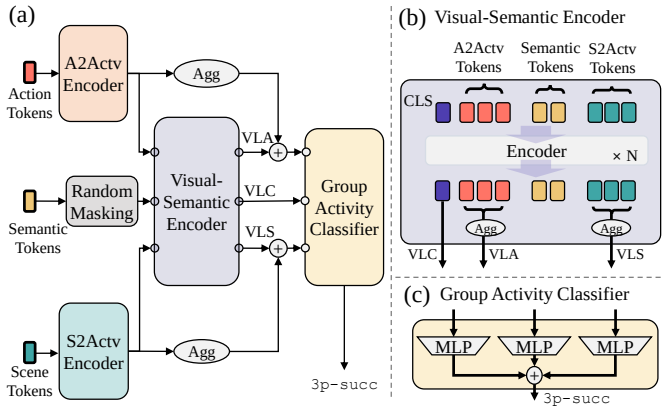

Fig. 3. (a) Overview of the semantic-integrating module with the visual semantic encoders (A2Actv and S2Actv), along with the random masking for the semantic reconstruction objective. Visual-semantic encoder is shown in (b) where the tokens are combined with multi-modal encoder. (c) The group activity classification is performed using these semantic tokens forwarded to a three-headed MLP.

generalize better on different frames. The features used for the prediction in $\mathcal{L}_{\text{gf}}$ are spatially aggregated scene tokens. Lastly, we add the reconstruction loss $\mathcal{L}_{\text{rec}}$ to form the total loss,

$$\mathcal{L}_{\text{sum}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{gf}} + \mathcal{L}_{\text{rec}}. \tag{2}$$

III. Experiments

A. Experiment Setup

1) Datasets: We conducted experiments on two standard WSGAR benchmarks. First, NBA dataset [3] contains 7624 training and 1548 testing clips, each 6 seconds long, with a single group activity label and no individual-level annotations. Second, Volleyball dataset [13] consists of 4830 clips taken from 55 videos. 3494 clips among them are for training and 1337 clips are for testing with frame-level group activity and individual action annotations. We reported Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) for evaluation, where MPCA addresses class imbalance in the NBA dataset.

2) Hyperparameters: We used an ImageNet pretrained ResNet-18 [14] or Inception-v3 [15] with motion augmentation as used in [4] for the backbone. We set $L = 6$ with four number of attention heads and $C = 128$ for the actor encoder. The action encoder consists of three 1D temporal convolutional layers with a kernel size of 5 without padding for the NBA dataset and two 1D convolutional layers with a kernel size of 3 with zero-padding are used for Volleyball. Inside S2Actv Encoder we used two 2D convolution layers for spatio-temporal convolution with a kernel size of 5×3 for NBA and 3×3 for Volleyball. We used segment-based sampling [16] to select T frames from each video in the NBA dataset, and random sampling in the Volleyball dataset. For the NBA dataset, we set $T = 18$ and for the Volleyball dataset $T = 5$, and we resized each frame to $W_0 = 1280$ and $H_0 = 720$. To generate the semantic features, we used LLAVA-NeXT Interleave [12], and a text embedding model SFR-Embedding [17] to convert the caption into tokenized form.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART GAR MODELS AND VIDEO BACKBONES ON THE NBA[3] DATASET

Method	Backbone	#Params	FLOPs	MCA	MPCA
Video backbone					
† TSM [18]	Res18	11.2M	303G	66.6%	60.3%
† VideoSwin[19]	VSwinT	27.9M	478G	64.3%	60.6%
Supervised GAR					
† ARG[2]	Res18	49.5M	307G	59.0%	56.8%
† AT[20]	Res18	29.6M	305G	47.1%	41.5%
† DIN[21]	Res18	26.0M	304G	61.6%	56.0%
KRG-GAR [22]	Res18	-	-	72.4%	67.1%
Weakly supervised GAR					
SAM [3]	Res18	-	-	49.1%	47.5%
† SAM [3]	Res18	25.5M	304G	54.3%	51.5%
Dual-AI (RGB) [23]	Inc-v3	-	-	58.1%	50.2%
DFWSGAR [4]	Res18	17.5M	313G	75.8%	71.2%
LRMM+GCM [6]	Res18	14.2M	306G	<u>77.8%</u>	<u>73.2%</u>
LSGAR [8]	Res18	17.8M	309G	77.1%	72.7%
Bi-Causal [24]	HRNet	-	-	70.3%	64.5%
SInAR-Net	Res18	16.2M	307G	**78.6%**	**74.5%**

Numbers in **bold** indicate the best performance and <u>underlined</u> ones are the second best. † indicates that the results are from Kim [4] reproduction, and others are as reported from the original paper. Res18, VSwinT, and Inc-v3 denote ResNet-18, Video Swin Transformer, and Inception-v3, respectively.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART GAR MODELS ON THE VOLLEYBALL [13] DATASET

Method	Backbone	MCA
Supervised GAR		
† ARG[2]	Res18	91.1%
† AT[20]	Res18	90.0%
† SACRF[26]	Res18	90.7%
† DIN[21]	Res18	93.1%
GroupFormer [27]	Inc-v3	94.1%
Dual-AI (RGB) [23]	Inc-v3	94.4%
KRG-GAR [22]	Inc-v3	95.4%
Bi-Causal [24]	HRNet	96.1%
Weakly supervised GAR		
SAM [3]	Res18	86.3%
† SAM [3]	Inc-v3	-
DFWSGAR [4]	Res18	90.5%
Dual-AI (RGB) [23]	Inc-v3	-
LRMM+GCM [6]	Res18	92.8%
Wu [28]	Res18	90.2%
LSGAR [8]	Res18	92.5%
Bi-Causal [24]	HRNet	93.4%
SInAR-Net	Res18	92.4%

† indicates that the results are from Kim [4] reproduction, and others are as reported from the original paper. Res18 and Inc-v3 denote ResNet-18 and Inception-v3, respectively.

3) Training: Our model was optimized by ADAM [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 18$ for 30 epochs. Weight decay was set to 14 for the NBA dataset. Learning rate was initially set to 16 with linear warmup to 14 for 5 epochs, and linearly decayed after the 6^{th} epoch. We used a mini-batch of size 4.

B. Comparison to State-of-the-art

We compared our method with the state-of-the-art methods in GAR [2], [20], [22], [26] and WSGAR [3], [4], [6], [24] on the NBA and Volleyball dataset. Since the NBA dataset does not provide ground truth bounding boxes, the bounding box proposals provided by SAM [3] are used to accommodate fully supervised learning if needed. Only the RGB frames were available during the inference stage.

Table I shows the comparison with other state-of-the-art GAR models on NBA dataset. Among various works, SInAR-Net performs best with 0.8%p higher MCA and 1.3%p higher MPCA compared to the second best. We also maintain a similar number of parameters to other GAR methods and a much lower number of parameters than LVLM models, such as LLAVA-NEXT Interleave [12] with 7 billion parameters. In the Volleyball dataset, our model also demonstrates competitive performance among models using the same backbone, as shown in Table II. Compared to Volleyball, the NBA dataset is more challenging than Volleyball due to its longer temporal range and more dynamic change of actor positions. The superior performance on this dataset demonstrates its effectiveness in handling complex spatio-temporal relationships.

C. Ablation Studies

1) Effect of Training Losses: We performed an ablation study to observe the effect of different losses in Table III. The ablation study proves the benefit of using $\mathcal{L}_{\text{gf}}$ to increase robustness against variance across frames and $\mathcal{L}_{\text{rec}}$ to ensure reconstructed semantic token carries similar knowledge to the semantic knowledge of LVLM. Both auxiliary losses contributed to performance improvement, and their combination yielded even better results.

2) Semantic-Integrating Module Design: The semantic-integrating module plays an important role in connecting visual action and scene representations with semantic representations. We experimented with various design of the integrator in Table IV. Firstly, we removed the semantic tokens from the calculation, causing the integrator to rely on visual tokens only. Then, we added the semantic tokens but did not add the VLA and VLS tokens from the encoder. Using only the VLC shows improvement in MPCA, but a slight reduce in MCA. Alternatively, in the third row we added only the VLA and VLS and show better MCA yet lower MPCA. This shows that VLA, VLS, and VLC components from the visual-semantic encoder contain different information. Lastly, we combined them all in our SInAR-Net to achieve the best performance.

3) Masking Ratio: In the training stage, we randomly masked the semantic tokens when they were forwarded to the visual-semantic encoder. The intention is to increase the robustness of the encoder, anticipating information loss in the semantic tokens when LVLM is not available in inference. We ablated multiple choices of masking ratio as shown in Table V and found that 0.7 masking ratio gave us the best performance in the NBA dataset. This also shows that the masking operation indeed help increase the robustness of the model.

D. Qualitative Result

We analyzed deeper into the mechanism inside SInAR-Net by examining the attention map inside its actor encoder as shown in Fig. 4. The sequence (a) contains a *failed 2-point shot, defensive rebound* activity. The model correctly focuses

TABLE III
ABLATION ON DIFFERENT LEARNING LOSSES

$\mathcal{L}_{CE}$	$\mathcal{L}_{gf}$	$\mathcal{L}_{rec}$	MCA	MPCA
✓	✗	✗	74.3%	70.1%
✓	✓	✗	76.5%	71.2%
✓	✗	✓	75.1%	71.0%
✓	✓	✓	78.6%	74.5%

TABLE IV
ABLATION ON SEMANTIC-INTEGRATING MODULE DESIGN

Model Type	MCA	MPCA
No Semantic Component	77.0%	72.3%
No VLA & VLS addition	76.9%	73.2%
No VLC	77.6%	72.2%
SInAR-Net	78.6%	74.5%

TABLE V
ABLATION ON MASKING RATIO OF SEMANTIC TOKENS

Masking Ratio	MCA	MPCA
30%	77.6%	72.9%
50%	76.8%	72.6%
70%	78.6%	74.5%

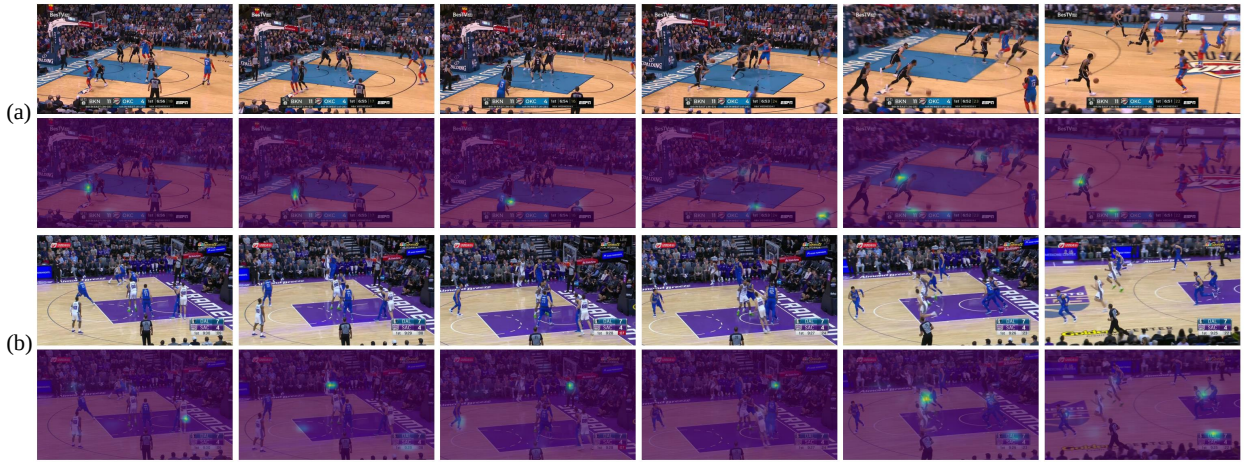

Fig. 4. Visualization of attention maps. The model selectively higlights key players important to the corresponding activity of (a) *failed 2-point shot rebound* and (b) *failed 3-point shot defensive rebound*.

its attention to the offensive player holding the ball in the early stage of the sequence. After the ball was released, the model shifts its attention to the scrimmage under the ring, while also highlighting the referee to recognize any signal. Lastly, after the ball was rebounded and the counter-offensive action started, the model shifts its attention to the players sprinting toward the opposite ring. In the second sequence (b), which contains a *failed 3-point shot and defensive rebound*, the model dynamically shifts its attention to different players based on their relevance to the activity. The model first focuses on the three-point shooter at the top key, then shifts its attention to the referee and the defending player after the rebound.

IV. CONCLUSION

We propose SInAR-Net, which efficiently utilizes the semantic knowledge of Large Vision-Language Models (LVLMs) to improve the performance of Weakly-Supervised Group Activity Recognition (WSGAR). SInAR-Net combines pre-extracted semantic features with visual token representations, effectively leveraging the strength of semantically contextual features with specialized activity oriented features. Experimental evaluation on WSGAR benchmarks shows that SInAR-Net achieves significant improvements in recognition accuracy, particularly for complex and long-term group interactions. Importantly, our approach maintains computational efficiency comparable to existing WSGAR methods. Ablation studies further validate the effectiveness of our learning strategy and visual-semantic fusion modules. These results demonstrate the potential of incorporating LVLM semantic knowledge for enhanced understanding of complex group activities in weakly-supervised settings.

ACKNOWLEDGMENT

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD).

REFERENCES

[1] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 4315–4324.

[2] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 9964–9974.

[3] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Social adaptive module for weakly-supervised group activity recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Aug. 2020, pp. 208–224.

[4] D. Kim, J. Lee, M. Cho, and S. Kwak, "Detector-free weakly supervised group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 083–20 093.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Aug. 2020, pp. 213–229.

[6] Z. Du, X. Wang, and Q. Wang, "Perceiving local relative motion and global correlations for weakly supervised group activity recognition," *Image and Vision Computing*, vol. 137, p. 104 789, 2023.

[7] D. Wu, H. Zhao, X. Bao, and R. P. Wildes, "Sports video analysis on large-scale data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Oct. 2022, pp. 19–36.

[8] L. Wu, M. Tian, Y. Xiang, K. Gu, and G. Shi, "Learning label semantics for weakly supervised group activity recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 6386–6397, 2024.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, Jul. 2021, pp. 8748–8763.

[10] Y. Zhang, B. Sun, J. He, L. Yu, and X. Zhao, "Multi-level neural prompt for zero-shot weakly supervised group activity recognition," *Neurocomputing*, vol. 571, p. 127 135, 2024.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[12] F. Li, R. Zhang, H. Zhang, *et al.*, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, Jul. 2024.

[13] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1971–1980.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826.

[16] L. Wang, Y. Xiong, Z. Wang, *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Oct. 2016, pp. 20–36.

[17] R. Meng, Y. Liu, S. R. Joty, C. Xiong, Y. Zhou, and S. Yavuz, *Sfr-embedding-2: Advanced text embedding with multi-stage training*, 2024. [Online]. Available: https://huggingface.co/Salesforce/SFR-Embedding-2_R.

[18] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 7083–7093.

[19] Z. Liu, J. Ning, Y. Cao, *et al.*, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 3202–3211.

[20] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 839–848.

[21] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 7476–7485.

[22] D. Pei, D. Huang, L. Kong, and Y. Wang, "Key role guided transformer for group activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7803–7818, 2023.

[23] M. Han, D. J. Zhang, Y. Wang, *et al.*, "Dual-ai: Dual-path actor interaction learning for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 2990–2999.

[24] Y. Zhang, W. Liu, D. Xu, Z. Zhou, and Z. Wang, "Bi-causal: Group activity recognition via bidirectional causality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 1450–1459.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2015.

[26] R. R. A. Pramono, Y. T. Chen, and W. H. Fang, "Empowering relational network by self-attention augmented conditional random fields for group activity recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Aug. 2020, pp. 71–90.

[27] S. Li, Q. Cao, L. Liu, *et al.*, "Groupformer: Group activity recognition with clustered spatial-temporal transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 13 668–13 677.

[28] L. Wu, X. Lang, Y. Xiang, C. Chen, Z. Li, and Z. Wang, "Active spatial positions based hierarchical relation inference for group activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2839–2851, 2022.