LiGAR: LiDAR-Guided Hierarchical Transformer for Multi-Modal Group Activity Recognition
Supplementary Material

1. Additional Experiments

1.1. Impact of Different Backbone Architectures

To demonstrate the versatility of LiGAR and investigate its performance with more advanced architectures, we conducted experiments using three different backbone networks: ResNet-18, ViT-B/16, and ViT-B/8. This comparison aims to show LiGAR's adaptability to different feature extractors and analyze the impact of more sophisticated architectures on its performance.

1.1.1 Experimental Setup

We evaluated LiGAR's performance on three datasets: JRDB-PAR [2], Volleyball [4], and NBA [6]. The backbone architectures used are ResNet-18 [3], a conventional CNN architecture; ViT-B/16 [1], a Vision Transformer with patch size 16x16; and ViT-B/8 [1], a Vision Transformer with patch size 8x8 offering higher resolution. All models were trained using the same hyperparameters and data augmentation techniques to ensure a fair comparison. We used the standard evaluation metrics for each dataset: F1-score for JRDB-PAR, Mean per-Class Accuracy (MCA) and Merged MCA (M-MCA) for Volleyball, and MCA and Mean Per Class Accuracy (MPCA) for NBA.

1.1.2 Performance Comparison Across Datasets

Table 1 presents the performance of LiGAR with different backbone architectures across the three datasets.

Table 1. Performance comparison of LiGAR with different backbones

Backbone	JRDB-PAR (F1)	Volleyball (MCA/M-MCA)	NBA (MCA/MPCA)
ResNet-18 [3]	59.3	93.1 / 95.4	87.4 / 79.4
ViT-B/16 [1]	61.5	94.2 / 96.1	88.9 / 81.2
ViT-B/8 [1]	63.2	95.0 / 96.8	90.1 / 82.7

The results demonstrate several key findings. ViT-B/8 consistently outperforms the other backbones across all datasets, with the improvement being particularly notable on the complex JRDB-PAR dataset, showing a 3.9 percentage point increase in F1-score compared to ResNet-18. The superior performance of ViT-B/8 over ViT-B/16 highlights the importance of higher resolution feature maps for group activity recognition, as the smaller patch size allows for more fine-grained spatial information to be captured. LiGAR's ability to benefit from more advanced backbone architectures demonstrates its flexibility and potential for future improvements as new architectures emerge. Interestingly, the performance gap between backbones is more pronounced on the JRDB-PAR dataset, suggesting that more complex scenes benefit more from advanced architectures.

1.1.3 Computational Requirements Analysis

While ViT-B/8 provides the best performance, it's important to note that it also has the highest computational cost. The choice of backbone may depend on the specific application requirements and available computational resources. To illustrate this trade-off, Table 2 provides a comparison of the computational requirements for each backbone architecture.

Table 2. Computational requirements of different backbones

Backbone	Parameters (M)	FLOPs (G)	Inference Time (ms)
ResNet-18 [3]	11.7	1.8	80
ViT-B/16 [1]	86.6	55.5	45
ViT-B/8 [1]	80.2	17.6	15

As expected, the ViT architectures, particularly ViT-B/8, have significantly higher computational requirements compared to ResNet-18. This trade-off between performance and computational cost should be considered when deploying LiGAR in real-world applications.

1.1.4 Discussion

This experiment demonstrates that LiGAR can effectively leverage the strengths of different backbone architectures, with Vision Transformers providing significant performance gains. The results suggest that future work could explore even more advanced backbones or hybrid architectures to further improve LiGAR's performance, while also considering the computational requirements for practical deployments. The flexibility of LiGAR in adapting to different backbone architectures highlights its potential for continued improvement as new architectures are developed in the field of computer vision.

2. Additional Ablation Studies

In the main paper, we presented comprehensive ablation studies on the JRDB-PAR dataset to evaluate the effectiveness of LiGAR's design choices and components. To provide a fair analysis across all datasets and demonstrate the generalizability of our findings, we extend these experiments to the Volleyball and NBA datasets. These additional studies aim to quantify the impact of different modality combinations, the contribution of each key component, and the effect of hierarchical processing levels on LiGAR's performance across diverse group activity recognition scenarios.

2.1. Impact of Modality Combinations

We first investigated the influence of different modality combinations on LiGAR's performance for the Volleyball and NBA datasets. Table 3 presents these results alongside the JRDB-PAR results from the main paper for comparison.

Table 3. Impact of modality combinations across datasets

Modalities	JRDB-PAR*	Volleyball		NBA	
	$\mathcal{F}_g$	MCA	M-MCA	MCA	MPCA
RGB only	51.2	74.8	76.1	62.7	57.1
RGB + Text	53.8	82.4	82.9	70.2	68.4
RGB + LiDAR	56.5	87.1	88.3	79.5	73.2
RGB + LiDAR + Text	**59.3**	**93.1**	**95.4**	**87.4**	**79.4**

*Results from the main paper

2.2. Contribution of LiGAR Components

Next, we examined the individual contribution of each key component in LiGAR for the Volleyball and NBA datasets. Table 4 shows these results alongside the JRDB-PAR results from the main paper.

Table 4. Ablation study on LiGAR components across datasets

Model Variant	JRDB-PAR*	Volleyball		NBA	
	$\mathcal{F}_g$	MCA	M-MCA	MCA	MPCA
LiGAR w/o MLT	55.4	80.8	82.6	70.5	64.9
LiGAR w/o CMGA	57.1	88.2	90.1	80.2	76.6
LiGAR w/o AFM	58.3	86.3	89.6	79.7	74.1
Full LiGAR	**59.3**	**93.1**	**95.4**	**87.4**	**79.4**

*Results from the main paper

2.3. Impact of Hierarchical Processing

Lastly, we evaluated the effect of different hierarchical processing levels on LiGAR's performance for the Volleyball and NBA datasets. Table 5 presents these results alongside the JRDB-PAR results from the main paper.

Table 5. Impact of different hierarchical levels across datasets

Hierarchical Levels	JRDB-PAR*	Volleyball		NBA	
	$\mathcal{F}_g$	MCA	M-MCA	MCA	MPCA
Single-level	54.6	70.2	72.3	65.2	58.3
Two-level	57.7	82.3	86.2	80.4	71.6
Three-level (Full LiGAR)	**59.3**	**93.1**	**95.4**	**87.4**	**79.4**

*Results from the main paper

2.4. Analysis

The extended ablation studies on the Volleyball and NBA datasets corroborate and strengthen the findings from the JRDB-PAR dataset presented in the main paper. Across all three datasets, we observe consistent trends that validate the effectiveness of LiGAR's design choices. The impact of modality combinations, as shown in Tab. 3, demonstrates the complementary nature of RGB, LiDAR, and textual information, with the full multi-modal configuration consistently achieving the best performance. Improvements range from 18.3 to 24.7 percentage points over the RGB-only baseline across datasets, underscoring the importance of integrating diverse data sources for robust group activity recognition.

The component-wise ablation, as shown in Tab. 4, reveals the crucial role of each LiGAR module across all datasets. The Multi-Scale LiDAR Transformer (MLT) proves to be the most critical component, with its removal causing the largest performance drop in all cases. This highlights the significance of multi-scale LiDAR processing in capturing complex spatial relationships, even in scenarios where LiDAR data is not directly available during inference, such as in the Volleyball and NBA datasets. The Cross-Modal Guided Attention (CMGA) and Adaptive Fusion Module (AFM) also contribute substantially to the model's performance, emphasizing the importance of effective cross-modal feature alignment and dynamic modality fusion.

The hierarchical processing experiment, as shown in Tab. 5, demonstrates the benefits of LiGAR's multi-level approach across all datasets. The three-level architecture consistently outperforms single and two-level variants, with performance gains ranging from 12.9 to 22.2 percentage points compared to the single-level model. This suggests that the hierarchical approach enables LiGAR to capture group activities at various granularities, from individual actions to overall scene dynamics, regardless of the specific activity domain.

Interestingly, while the performance improvements follow similar trends across datasets, they are more pronounced on the JRDB-PAR dataset compared to Volleyball and NBA. This could be attributed to the greater complexity and diversity of activities in the JRDB-PAR dataset, which may benefit more from LiGAR's sophisticated multi-modal and multi-scale processing approach. However, the consistent trends across all three datasets validate the generalizability of LiGAR's design principles across different group activity recognition scenarios.

In conclusion, these extended ablation studies provide strong evidence for the effectiveness and generalizability of LiGAR's multi-modal, multi-scale, and hierarchical approach to group activity recognition. The results demonstrate that each component and design choice contributes meaningfully to the model's performance across diverse datasets, with their combination yielding state-of-the-art results in various group activity recognition contexts.

3. More Visualizations

Figure 1 illustrates t-SNE visualizations of feature representations learned by LiGAR across different modality combinations for the Volleyball, NBA, and JRDB-PAR datasets. The progression from RGB-only to the full multi-modal configuration (RGB + LiDAR + Text) shows a clear improvement in feature discrimination and cluster formation across all datasets.

In the RGB-only scenario, classes are largely intermixed, indicating poor separability based solely on visual features. Adding textual information (RGB + Text) begins to show some improvement in cluster formation. The incorporation of LiDAR data (RGB + LiDAR) leads to a significant leap in feature discrimination, with distinct clusters starting to form. The full multi-modal configuration achieves the most distinct and compact clusters across all datasets, particularly evident in the complex JRDB-PAR scenario.

This visual analysis corroborates our quantitative results, demonstrating LiGAR's ability to leverage complementary information from multiple modalities effectively. The consistent improvement pattern across datasets underscores the generalizability and effectiveness of LiGAR's multi-modal approach in capturing the nuanced dynamics of group interactions, especially in complex, real-world scenarios with varied activities.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Ruize Han, Haomin Yan, Jiacheng Li, Songmiao Wang, Wei Feng, and Song Wang. Panoramic human activity recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 244–261. Springer, 2022. 1, 4

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016. 1, 4

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[6] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *ECCV*, pages 208–224. Springer, 2020. 1, 4

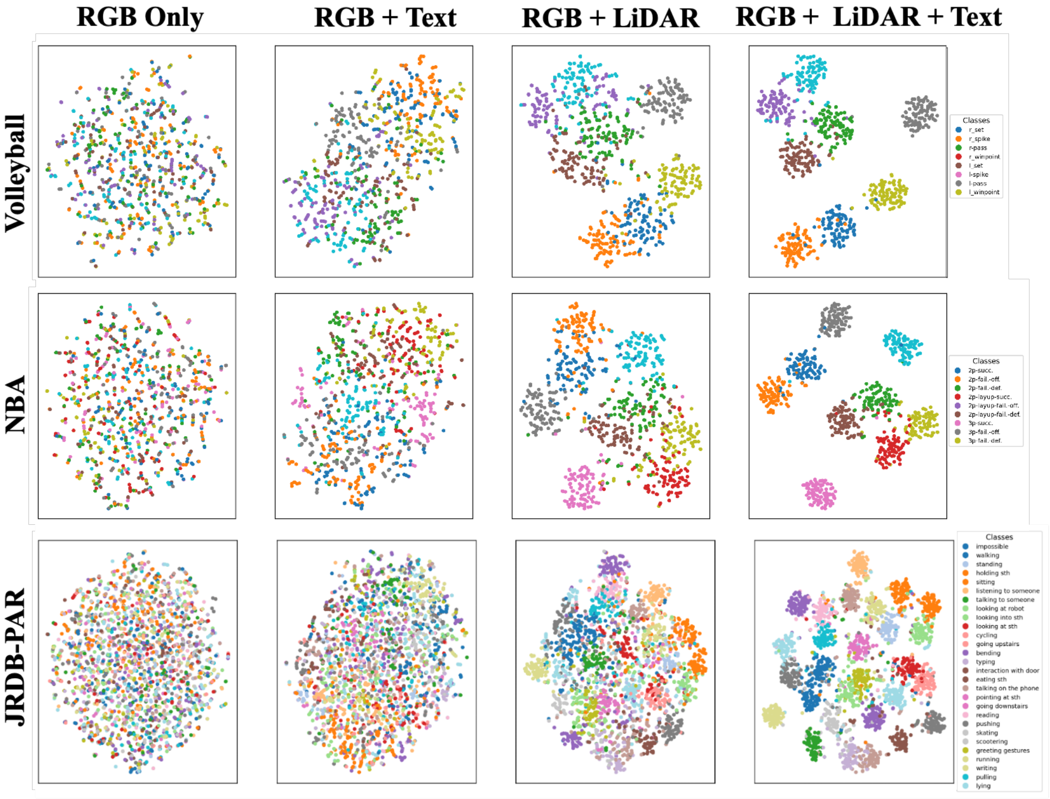

Figure 1. t-SNE [5] visualization of video representation on the Volleyball [4], NBA [6] and JRDB-PAR [2] datasets learned by LiGAR model for different combinations of modalities. **Best viewed in zoom and color.**