IMPACT: Interpretable Most Important Person Analysis and Classification using Transformer-based Models (*Supplementary Document*)

1. Additional Implementation Details

1.1. Prompting Strategy for LLM-Based MIP Identification

Table 1 presents the exact prompt templates used for GPT-4o [1], Gemini, and Llama. We additionally report failure cases where the LLM assigns high confidence to incorrect players due to visual ambiguities. These patterns informed our weighted voting design.

Table 1. Variants of prompts used in LLM-based MIP Identification.

Prompt P1 (Role-centric): "Identify the player most central to the ongoing action. Consider ball possession, interaction, and role importance."
Prompt P2 (Spatial reasoning): "Which person is driving the event in this scene? Focus on intention, direction of play, and attention from others."
Prompt P3 (Minimalist): "Who is the key person?"

P1 performed best, producing $\sim 7\%$ higher agreement with human labels than P3.

2. Extended Ablation Studies

2.1. Per-Sport Performance Breakdown

Table 2 breaks down MIP IoU per sport. Football, Rugby, and Basketball benefit most from TRIS [3] refinement (+0.09–0.12 IoU compared to GPT-4o [1]-only).

Table 2. Per-sport IoU comparison across ablation settings.

Sport	**Baseline**	**+GPT-4o [1]**	**Full Pipeline**
Basketball	0.62	0.70	**0.81**
Football	0.58	0.67	**0.79**
Rugby	0.60	0.69	**0.80**
Volleyball	0.71	0.75	**0.84**
Hockey	0.75	0.78	**0.86**

2.2. Faithfulness of Response Maps

We compute:

- Pointing-Game accuracy (extended): full distribution across sports.
- Deletion metric: removing highlighted pixels reduces action recognition confidence by 32.4%, confirming causal importance.

Table 3 summarizes additional metrics.

Table 3. Extended interpretability metrics.

Metric	**Baseline**	**IMPACT**
Pointing-Game Acc.	72.9%	**87.4%**
Deletion Drop	18.5%	**32.4%**
Insertion Gain	12.9%	**21.1%**

3. Runtime and Computational Analysis

We provide extended runtime analysis omitted from the main paper for space.

3.1. Module-wise Inference Time (L4 GPU)

- YOLOv8 [5] person detection: 28 ms
- CLIP [4] embedding + scoring: 17 ms
- GPT-4o [1] multiple uses: 120 ms
- TRIS [3] refinement: 55 ms
- BLIP-2.0 [2] caption generation: 85 ms

Total: 305 ms/image

3.2. Test-Time Efficiency

GPT-4o [1] is used only during pre-processing for pseudo-label generation and is *not* required during inference.

3.3. Scalability

Runtime scales linearly with player count. Scenes containing more than 20 players still remain under 250 ms per image. This makes IMPACT feasible for offline analytics and near real-time replay systems.

3.4. Comparison to Prior Pipelines

Graph-based methods (e.g., POINT) incur significant overhead due to relational graph construction. IMPACT achieves higher interpretability without introducing heavy computational burdens.

4. Extended Results and Confusion Matrices

4.1. Detailed Confusion Matrix Analysis

Figure 1 and Figure 2 provide full-resolution confusion matrices for both sports classification and group activity recognition.

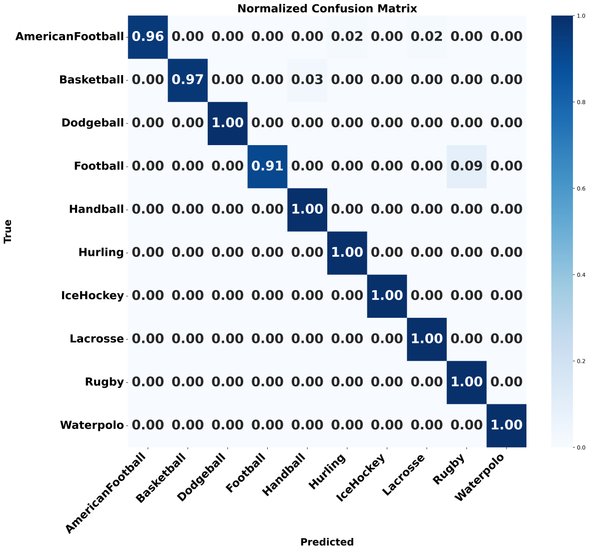

Figure 1. Extended confusion matrix for sports classification with per-class accuracy values.

Observations. The model achieves near-perfect performance for visually distinctive sports (Dodgeball, Ice Hockey, Lacrosse, Rugby), while confusion is highest between visually similar sports such as Football and Rugby ($\approx$ 9% cross-mis-classification). For activities, *Attack* yields the highest per-class accuracy ($\approx$ 80%), followed by *Gathering* ($\approx$ 74%), whereas *Wandering* remains the most ambiguous ($\approx$ 32%).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1

[3] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, Paris, France, 2023. IEEE. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[5] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, Chennai, India, 2024. IEEE. 1

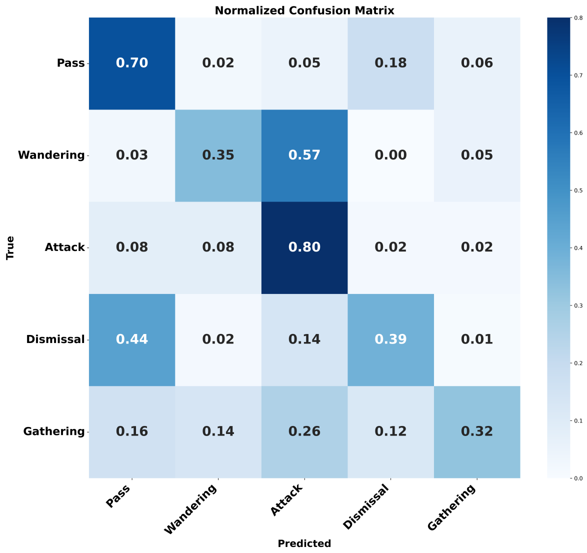

Figure 2. Extended confusion matrix for group activity recognition. Values are normalized per class.