Motion-Relation Modeling for Skeleton-Based Tactical Action Localization in 3-on-3 Basketball

Zihan Ye[1] and Yoshinari Kameda[2][a]
[1]*Master's Program in Intelligent and Mechanical Interaction Systems, University of Tsukuba, Japan*
[2]*Center for Computational Sciences, University of Tsukuba, Japan*

Keywords: Basketball Tactical Analysis, Temporal Action Localization, Skeleton-Based Action Recognition, Graph Attention Networks, Sports Analytics.

Abstract: Basketball tactical action localization requires temporal boundary estimation and reasoning over coordinated player interactions. We present a motion-relation framework that represents players as actor tokens and ordered pairs as directed relation tokens. A GATv2-style encoder aggregates dynamic interactions, while a gated ActionFormer-style head localizes tactical categories and boundaries. We also define Primary-4, a core-action benchmark, and Coarse-4, which maps all 19 annotations to four tactically motivated functional anchors. Across three seeds, the canonical Primary-4 model achieves $58.37 \pm 1.98\%$ mAP@0.5 and $26.25 \pm 0.89\%$ mAP@[0.50:0.05:0.95]. Matched ablations support temporal gating, whereas removing the role proxy does not reduce accuracy. These results establish a compact protocol for skeleton-based tactical localization while exposing limitations in strict boundary alignment, split verification, and role interpretation.

1 INTRODUCTION

Basketball tactics arise from coordinated movements among handlers, screeners, receivers, and defenders. A pick-and-roll depends on their evolving relations, so localization must estimate boundaries while modeling individual motion and inter-player relationships.

Skeletons compactly represent motion, and graphs encode multiple entities. Yet motion can indicate different tactics with possession, screening, defensive reactions, and spacing; pooling or a fixed graph cannot explicitly represent such directed, time-varying interactions.

We represent players as actor tokens and ordered pairs as directed relation tokens, constructing relation-aware event features before localization. Edges contain relative motion and possession cues; soft role proxies remain optional context. A gate fuses a dilated-convolution path for stable local context with an ActionFormer-style multi-scale path for variable-duration actions.

SGA-INTERACT (Yang et al., 2025) contains long-tailed, co-occurring labels with ambiguous boundaries. Primary-4 evaluates four frequent primitives; Coarse-4 preserves every annotation through tactical grouping. Primary-4 omits non-primary labels and does not replace full-label evaluation. Claims remain limited to the processed splits and skeleton inputs.

The main contributions of this paper are as follows:

- A directed motion-relation representation combines actor kinematics, pairwise motion, and possession before localization; soft role proxies are optional context.
- A gated temporal head combines local convolutions with multi-scale localization; three-seed ablations identify gating as the reliable tested contributor.
- Primary-4 and Coarse-4 separate core-action measurement from full-annotation coarse evaluation, while full-threshold and 19-class results expose boundary and label-granularity limits.

2 RELATED WORKS

2.1 Basketball Tactical Datasets

Basketball datasets increasingly support temporal and group-level tactical understanding. TAL3x3 localizes actions in untrimmed 3-on-3 video (Hu et al., 2025), whereas SGA-INTERACT provides multi-player 3D skeleton sequences with fine-grained tactical annota-

[a] https://orcid.org/0000-0001-6776-1267

Ye, Z. and Kameda, Y.
Motion-Relation Modeling for Skeleton-Based Tactical Action Localization in 3-on-3 Basketball.
DOI: 10.5220/0015301100004981
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 14th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2026)*, pages 147-154
ISBN: 978-989-758-863-1; ISSN: 2184-3201
Proceedings Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

tions (Yang et al., 2025). Skeletons compactly describe motion and spatial organization with less dependence on appearance.

These resources contain co-occurrence, ambiguous boundaries, and label imbalance; similar motions may serve different functions with possession and player relations. We use SGA-INTERACT for the main experiments and TAL3x3 for representation transfer. Primary-4, Coarse-4, and the 19-class diagnostic serve distinct purposes; reducing the label space does not solve fine-grained localization.

2.2 Temporal Action Localization

Temporal action localization estimates categories and boundaries in untrimmed sequences. BSN assembles proposals from boundary likelihoods, whereas BMN evaluates start–end combinations (Lin et al., 2018; Lin et al., 2019), establishing the value of explicit boundary evidence.

RTD-Net directly generates proposals with relaxed transformer decoding (Tan et al., 2021); ActionFormer uses local self-attention and an anchor-free multi-scale pyramid (Zhang et al., 2022); and TriDet predicts relative boundaries across temporal granularities (Shi et al., 2023). TemporalMaxer instead obtains competitive context modeling with local max pooling (Tang et al., 2023). Together they emphasize boundaries and scale while indicating that context operators are task dependent.

Generic localizers omit basketball-specific player relations. We place directed relation encoding before an ActionFormer-style path and add a gated dilated-convolution alternative; strict-tIoU results nevertheless identify boundary alignment as the main weakness.

2.3 Group Activity Recognition

Group activity recognition models how actors determine scene-level activity. Space-time region graphs propagate spatial-temporal relations (Wang and Gupta, 2018); Actor-Transformers aggregate actor context (Gavrilyuk et al., 2020); and Progressive Relation Learning and GroupFormer refine or cluster interactions (Hu et al., 2020; Li et al., 2021). They motivate modeling beyond independent actor pooling.

COMPOSER reasons over multi-scale keypoint tokens (Zhou et al., 2022), while Dual-AI learns complementary spatial–temporal interaction paths (Han et al., 2022). Both classify trimmed clips; our untrimmed setting also requires boundaries under overlapping annotations, so we construct ordered player-pair tokens per frame before decoding.

2.4 Skeleton and Graph-Based Relation Modeling

ST-GCN represents joints as a spatio-temporal graph (Yan et al., 2018); 2s-AGCN learns adaptive joint-and-bone graphs (Shi et al., 2019), CTR-GCN refines topology by channel (Chen et al., 2021), and InfoGCN learns information-bottleneck representations (Chi et al., 2022). These methods mainly recognize single-person actions, whereas we localize directed inter-player relations. Skeleton-based boxing also shows sports applicability (Watanabe and Kameda, 2024), but for short two-person actions.

GAT weights neighboring nodes (Veličković et al., 2018), whereas GATv2 makes their ranking query dependent (Brody et al., 2021). In sports, trajectory prediction evaluates off-ball soccer contributions (Teranishi et al., 2022), NETS recognizes basketball activities from tracking (Hauri and Vucetic, 2022), and Skor-xG estimates soccer scoring probability from 3D skeletons (Xu et al., 2025). Unlike these scoring or recognition tasks, we decode player nodes and directed inter-player edges into temporal segments.

3 METHOD

3.1 Overall Architecture

Figure 1 summarizes the pipeline. Skeleton-derived states, ball cues, and center trajectories form actor and ordered-pair tokens. A directed GATv2-style encoder produces frame-level event features. The gated temporal head predicts class logits, actionness, boundary distances, and center-ness scores. Thresholding, boundary refinement, and temporal non-maximum suppression (NMS) then produce action segments.

3.2 Basketball-Specific Motion-Relation Representation

Basketball tactical actions are defined by both individual kinematics and interactions among players. To capture these complementary cues, each frame is represented using actor tokens and directed relation tokens. Soft typed relation cues are computed as optional basketball context and diagnostic outputs rather than as a claimed source of accuracy.

3.2.1 Actor Token

Each player is represented as an actor token. In our implementation, each actor token is a 17-dimensional

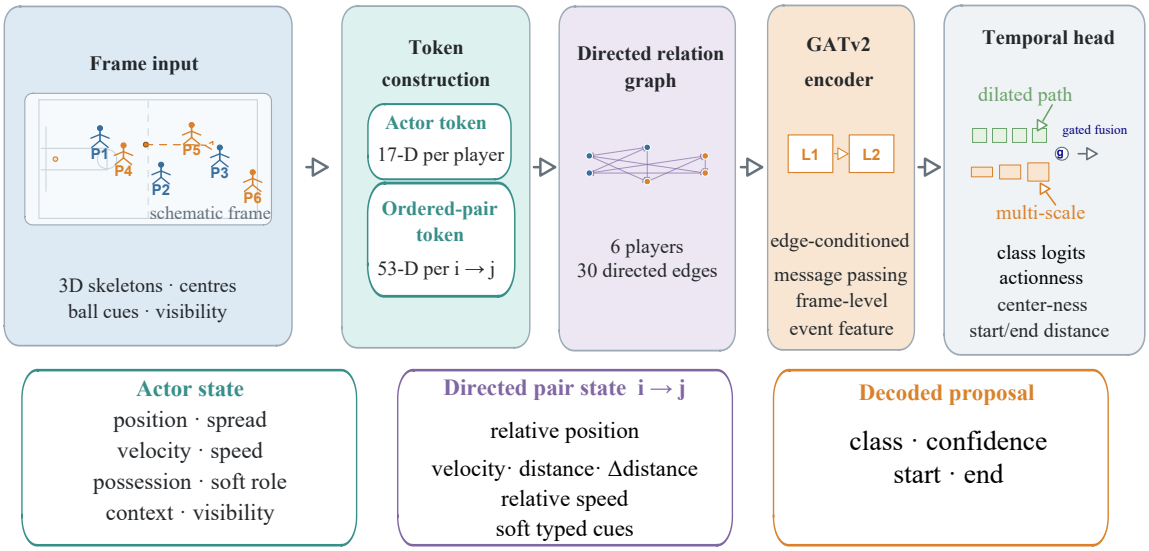


Figure 1: Motion-relation framework. Skeleton-derived player and ball states form actor tokens and ordered-pair tokens. A GATv2-style encoder aggregates pair-conditioned messages. The temporal head fuses dilated-convolution and multi-scale paths through a learnable gate, predicts classes and boundaries, and decodes proposals using NMS. Arrows indicate information flow rather than physical trajectories.

vector that describes the kinematic and soft role state of player i:

$$\mathbf{a}_i = [\mathbf{p}_i, \mathbf{s}_i, \mathbf{v}_i, v_i^m, \omega_i, \mathbf{r}_i, m_i], \quad (1)$$

where $\mathbf{p}_i$ is the player center, $\mathbf{s}_i$ summarizes joint spread, $\mathbf{v}_i$ and $v_i^m = \|\mathbf{v}_i\|_2$ are velocity and speed, $\omega_i \in [0,1]$ is a continuous ball-possession proximity score, $\mathbf{r}_i$ contains soft role proxies, and m_i marks visibility.

The role proxy scores $\mathbf{r}_i$ describe soft basketball states, including ball handler, primary defender, screen support, and receiver/cutter. These scores are heuristic inputs rather than ground-truth role labels. Their contribution is therefore tested separately and interpreted only through matched ablations.

3.2.2 Directed Relation Token

For each ordered player pair (i, j), we construct a directed relation token. This directionality is important because the tactical relationship from player i to player j is not equivalent to the relationship from player j to player i. For example, the relation from a ball handler to a screener differs from the relation from the screener to the ball handler.

Each directed relation token $\mathbf{e}_{ij}$ is formulated as:

$$\mathbf{e}_{ij} = [\mathbf{a}_i, \mathbf{a}_j, \mathbf{p}_{ij}, \mathbf{v}_{ij}, d_{ij}, \Delta d_{ij}, \gamma_{ij}, v_{ij}^r, \Omega_{ij}, \mathbf{R}_{ij}]. \quad (2)$$

where $\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i$ and $\mathbf{v}_{ij} = \mathbf{v}_j - \mathbf{v}_i$ are relative position and velocity vectors. Their scalar counterparts are distance $d_{ij} = \|\mathbf{p}_{ij}\|_2$, distance change Δd_{ij}, and relative speed $v_{ij}^r = \|\mathbf{v}_{ij}\|_2$; γ_{ij} scores whether the pair is approaching. Ω_{ij} combines the two possession scores, while $\mathbf{R}_{ij}$ contains pairwise combinations of the soft role proxies. Thus, velocity and speed are distinct vector and scalar quantities.

In the SGA implementation, the actor token has 17 dimensions, the directed edge attributes have 19 dimensions, and the complete relation token has 53 dimensions. This representation combines individual motion with pairwise tactical interaction information.

3.2.3 Soft Typed Relation Cues

We compute six soft cues: Handler-Defender, Handler-Screen, Handler-Receiver, Receiver-Defender, Player-Ball, and Off-ball Support. They describe possession, screening, receiving, marking, and spacing context. They are inputs and diagnostic scores, not deterministic edges or verified role labels. The three-seed ablation in Section 5 does not support an accuracy benefit from the role proxy.

3.3 GATv2-style Relation Encoder

The relation encoder operates on a dense directed player graph at each frame. Given a frame with $N = 6$ players, players are represented as graph nodes, and all non-self directed player pairs are represented as graph edges.

We use a GATv2-style attention mechanism because the importance of a defender, screener, or receiver should change dynamically with the target

Table 1: Similarity-based tactical label structure.

ID	Tactical Anchor	Fine-grained Labels	Primary
0	Cross Creation	Cross, Isolation, Post, Playmaking, Keep	Cross
1	Pick-and-Roll Screen	PnR, Slip	PnR
2	Handoff	HO, FHO, DHO	HO
3	Off-ball Spacing	Pop, Exit, Down, Flare, Reverse, Basket Cut, Up, Shuffle, Pin	Pop

player state and directed edge attributes at each frame. This dynamic scoring is better aligned with tactical role shifts than a fixed attention ranking over neighbors.

Let $\mathbf{h}_i \in \mathbb{R}^{D_a}$ denote the hidden representation of player i, initialized from the actor token $\mathbf{a}_i$. Let $\mathbf{e}_{j\to i} \in \mathbb{R}^{D_r}$ denote the directed relation token from source player j to target player i. The GATv2-style relation encoder updates player representations by computing attention weights over directed neighbors. The attention coefficient α_{ji} and updated node representation $\mathbf{h}'_i$ are computed as:

$$z_{ji} = \mathbf{q}^\top \mathrm{LeakyReLU}\left(\mathbf{W}_1[\mathbf{h}_i \parallel \mathbf{h}_j \parallel \mathbf{e}_{j\to i}]\right),$$
$$\alpha_{ji} = \frac{\exp(z_{ji})}{\sum_{k\in\mathcal{N}_i}\exp(z_{ki})}. \tag{3}$$

$$\mathbf{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ji}\mathbf{W}_2[\mathbf{h}_j \parallel \mathbf{e}_{j\to i}]\right), \tag{4}$$

where $\parallel$ denotes concatenation, $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable linear transformations, $\mathbf{q}$ is the attention vector, σ represents a non-linear activation function, and $\mathcal{N}_i$ denotes the directed neighborhood of player i.

The frame-level event feature $\mathbf{F}_{\text{event}}$ is obtained by aggregating the updated player representations. The temporal localization head therefore receives a compact representation of the current tactical state, conditioned on individual motion and directed player interactions.

3.4 Gated Temporal Localization Head

Given the sequence of event features $\mathbf{F}_{\text{event}}$, the temporal localization head predicts tactical action categories and boundaries. The head contains two complementary temporal paths. The first uses dilated temporal convolutions for stable local temporal modeling. The second follows an ActionFormer-style multi-scale design to capture actions with different durations. A learnable gate fuses the two paths:

$$\mathbf{F}_{\text{temporal}} = \mathbf{g}\odot\mathbf{F}_{\text{AF}} + (\mathbf{1}-\mathbf{g})\odot\mathbf{F}_{\text{conv}}, \tag{5}$$

where $\mathbf{F}_{\text{AF}}$ denotes the multi-scale attention feature, $\mathbf{F}_{\text{conv}}$ denotes the dilated temporal convolution feature, $\odot$ represents the Hadamard product, and $\mathbf{g} \in [0,1]^C$ is a learnable gating vector.

For each temporal slot, the head predicts class logits, actionness, start distance, end distance, and center-ness. The actionness branch suppresses background frames, the boundary regression branches refine temporal segments, and the center-ness branch favors predictions near the middle of tactical actions.

3.5 Similarity-Based Tactical Label Structure

SGA-INTERACT provides 19 long-tailed and frequently co-occurring categories. Table 1 gives a tactically motivated mapping specified from label semantics and dominant interaction patterns, rather than from model performance. Primary-4 evaluates four frequent primitives and intentionally ignores other labels. Coarse-4 retains every segment by grouping labels according to ball-handler creation, on-ball screening, close ball transfer, or off-ball movement. This grouping is a retrieval-oriented evaluation view, not a replacement for the original taxonomy.

3.5.1 Primary-4 Setting

Primary-4 evaluates Cross, Pick-and-Roll (PnR), Handoff (HO), and Pop. It is a core-action benchmark and does not claim coverage of all annotations.

3.5.2 Coarse-4 Setting

Coarse-4 preserves all 19 labels. Cross Creation represents ball-handler advantage creation, Pick-and-Roll Screen represents on-ball screening, Handoff represents close ball transfer, and Off-ball Spacing represents cuts and screens away from the ball. As an illustrative workflow, a coach could retrieve Pick-and-Roll Screen segments, inspect the localized interval, and compare whether the defender switched, dropped, or remained attached. The present study evaluates retrieval metrics only and does not include a coach study or validate downstream decision quality.

3.6 Proposal Decoding

During inference, class and actionness thresholds select candidates. Adjacent positives are merged,

boundary distances refine each segment, and temporal NMS removes redundant overlaps.

The final output for each detected tactic is represented as:

$$\mathcal{P} = (t_{\text{start}}, t_{\text{end}}, c, s_{\text{conf}}), \tag{6}$$

where t_{start} and t_{end} denote the localized temporal boundaries, c is the predicted tactical class or tactical anchor, and s_{conf} is the final confidence score.

4 EXPERIMENTAL SETUP

4.1 Datasets

We evaluate the proposed framework mainly on SGA-INTERACT (Yang et al., 2025), and further use TAL3x3 (Hu et al., 2025) to validate the general usefulness of actor motion and pairwise relative motion cues.

4.1.1 SGA-INTERACT

SGA-INTERACT is a 3D skeleton-based basketball group-activity dataset. We adapt its annotations to temporal localization. The processed lists contain 1183 training samples (2165 segments) and 506 validation samples (952 segments); clips may contain up to five overlapping segments. We use the supplied processed lists without clip reassignment. We checked the available result export for sample, sequence, video, and game identifiers, but none are present. Consequently, game-level train-validation independence cannot be audited from the current files. We restrict conclusions to the supplied processed split and treat a game-disjoint rerun as required validation rather than an established property.

4.1.2 TAL3x3

TAL3x3 supplies a separate 3-on-3 temporal localization setting. Its 240 training and 77 validation samples retain the original eight labels. We use it only to compare position, actor-motion, and relative-motion representations; labels are not mapped to SGA-INTERACT.

4.2 Evaluation Settings

4.2.1 Primary-4 Setting

Primary-4 evaluates Cross, PnR, HO, and Pop, covering 1001 training and 434 validation segments. Non-primary annotations are ignored, so this setting measures frequent tactical primitives rather than complete dataset performance.

4.2.2 Coarse-4 Setting

Coarse-4 maps all 2165 training and 952 validation segments to Cross Creation, Pick-and-Roll Screen, Handoff, or Off-ball Spacing. It tests functional anchor retrieval while preserving every annotation.

4.2.3 Full 19-Class Diagnostic Setting

The 19-class setting is diagnostic because it contains severe imbalance, a zero-validation label, and frequent co-occurrence. It is not the central performance claim.

4.3 Evaluation Metrics

For a predicted interval $P = [p_s, p_e]$ and ground-truth interval $G = [g_s, g_e]$, temporal intersection-over-union is

$$\text{tIoU}(P,G) = \frac{\max(0, \min(p_e, g_e) - \max(p_s, g_s))}{\max(p_e, g_e) - \min(p_s, g_s)}. \tag{7}$$

For each class c and threshold τ, detections are ranked by confidence and matched one-to-one to ground truth. We compute average precision $\text{AP}_c(\tau)$ from the precision-recall curve and report

$$\begin{aligned} \text{mAP}(\tau) &= \frac{1}{C}\sum_{c=1}^{C} \text{AP}_c(\tau), \\ \text{mAP}_{\text{full}} &= \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \text{mAP}(\tau), \\ \mathcal{T} &= \{0.50, 0.55, \ldots, 0.95\}. \end{aligned} \tag{8}$$

where thresholds increase by 0.05. We report mAP@0.5, mAP@0.7, and mAP_{full}. The CSV also contains a legacy averaged-mAP field from the earlier evaluation script; we label it "legacy mAP" and do not treat it as equivalent to mAP_{full}. Spot mAP measures occurrence detection, and Pred./Sample measures proposal density. Coarse-4 also reports whether each sample contains at least one correct anchor at tIoU 0.5 and 0.7.

4.4 Implementation Details

Models use 17-D actor tokens, 53-D relation tokens, hidden width 256, four attention heads, two GATv2 layers, four temporal layers, and dropout 0.1. Training uses PyTorch, AdamW, learning rate 1.8×10^{-4}, weight decay 10^{-4}, gradient clipping at 5.0, cosine

Table 2: Stage-wise results on Primary-4. The canonical row reports mean±SD over three seeds; other rows are single runs. "Legacy mAP" is the earlier script aggregate and is not mAP@[0.50:0.05:0.95]. Values are percentages.

Stage	Method	Added Component	mAP@0.5	mAP@0.7	Legacy mAP	spot mAP
S1-a	ActionFormer Token Pooling	temporal baseline	45.30	18.65	39.78	53.78
S1-b	GATv2 Relation Encoder	relation modeling	54.87	25.22	48.90	67.92
S2	Gated GATv2-AF	gated temporal head	57.56	27.82	50.50	65.27
S3	S3 Primary-4	typed soft relation cues	55.34	26.76	49.68	68.73
S3-c	S3 Primary-4 Canonicalized	attack-direction canonicalization	**58.37 ± 1.98**	**28.50 ± 2.39**	**51.94 ± 1.22**	67.86 ± 1.61

annealing, and batches of four. Primary-4 and Coarse-4 runs use 45 and 60 epochs unless stated otherwise. The canonical S3 run uses up to 80 epochs, validation every five epochs, and checkpoint selection by validation mAP. S3 Primary-4 is warm-started from S2. Experiments used an NVIDIA GeForce RTX 5060 Ti.

S1 compares token pooling with relation encoding, S2 adds the gated temporal head, and S3 adds soft typed cues and diagnostics. These labels denote model stages, not random seeds. S3 Primary-4 and S3 Coarse-4 use the same architecture with different label spaces.

Decoding uses top-$k = 12$, score threshold 0.03, actionness threshold 0.35, and NMS threshold 0.5. Multi-seed entries report the arithmetic mean and sample standard deviation across seeds 1, 2, and 3. The full-threshold export repeats seed 1 in its metadata, so seed identity was recovered from the run names and cross-checked against the corresponding three canonical runs.

5 EXPERIMENTS AND RESULTS

5.1 Main Results on Primary-4

Table 2 compares system stages, whereas Table 4 tests components. On the same split, the single-run ActionFormer-style token-pooling baseline reaches 45.30% mAP@0.5 and the GATv2 relation encoder reaches 54.87%. Additional baseline seeds remain necessary for uncertainty-matched comparison. Across three seeds, canonicalization reaches 58.37 ± 1.98% mAP@0.5 and 28.50 ± 2.39% mAP@0.7. Figure 3 shows a PnR example.

5.2 Full-Threshold Boundary Analysis

The complete threshold sweep in Figure 2 yields 26.25 ± 0.89% mAP@[0.50:0.05:0.95] across three canonical runs. Performance falls from 58.37 ± 1.98% at tIoU 0.50 to 28.50 ± 2.39% at 0.70 and 0.22 ± 0.08% at 0.95. The monotonic decline shows that the main remaining error is strict temporal boundary alignment rather than only action occurrence.

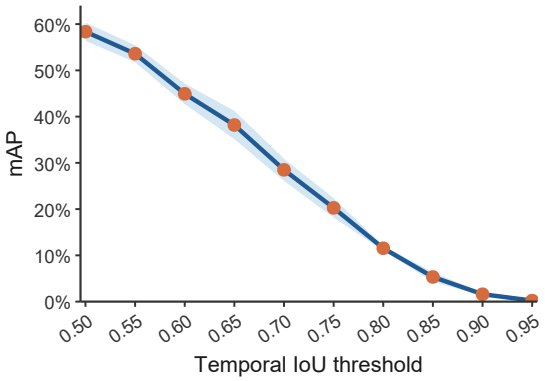


Figure 2: Primary-4 mAP over the complete tIoU range. The line shows the mean across three canonical runs and the shaded band shows one sample standard deviation.

5.3 Results on Coarse-4

Coarse-4 retains every segment and contains denser co-occurrence. Across three seeds, S3 obtains 45.72 ± 3.16% mAP@0.5, 19.22 ± 0.51% mAP@0.7, 43.66 ± 1.90% legacy mAP, and 63.49 ± 2.89% spot mAP. The decline at 0.7 shows weak strict boundary placement. Relaxed hit rates support exploratory retrieval, but not precise segmentation or validated coaching conclusions.

5.4 Validation on TAL3x3

To verify whether the proposed motion-relation representation is useful beyond the processed SGA-INTERACT setting, we conduct a validation experiment on TAL3x3. Table 3 compares different input representations.

Table 3: Validation results on TAL3x3 with different input representations. The mAP column is the legacy evaluation-script aggregate. Values are percentages.

Exp.	Feature	@0.5	@0.7	mAP	spot	Pred.
M0	Position only	19.72	6.90	18.92	36.33	8.60
M1	Actor motion	26.91	12.30	26.07	43.73	8.95
M2	+ pair geometry	24.49	10.62	22.90	40.97	8.84
M3	+ relative motion	**32.41**	10.10	**27.97**	**47.65**	9.06
M4	+ center/quality	24.25	11.12	23.57	41.02	8.94

Actor motion improves mAP@0.5 from 19.72% to 26.91%. Static pair geometry lowers it to 24.49%, whereas relative velocity, distance change, and ap-

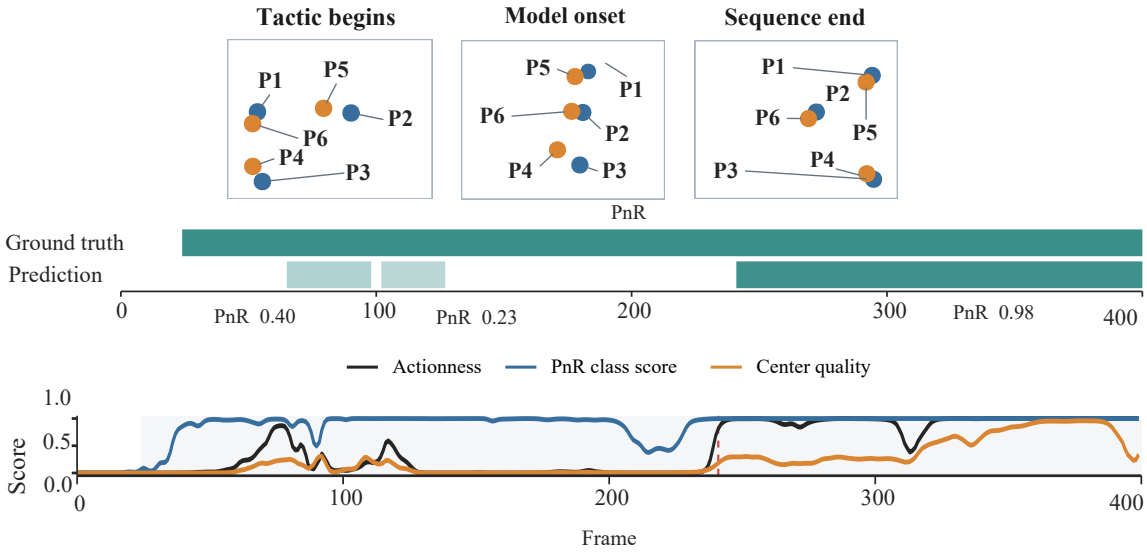


Figure 3: Qualitative PnR localization. Blue and orange distinguish the two three-player groups. The timeline shows ground truth and three proposals; traces show actionness, PnR score, and center quality. The red dashed line marks the highest-confidence onset. This example is illustrative, not dataset-wide evidence.

proach cues raise it to 32.41%. The corresponding mAP@0.7 is only 10.10%, so relative motion mainly improves action retrieval, not strict boundary alignment.

5.5 Ablation Study

Table 4: Matched three-seed ablations on Primary-4. Values are mean±SD in percent. The legacy mAP column is retained for consistency with prior runs.

Variant	@0.5	@0.7	Legacy mAP	spot
Reference	55.34 ± 1.98	26.76 ± 1.50	49.68 ± 0.39	68.73 ± 0.78
w/o Role	58.08 ± 2.56	27.41 ± 2.03	51.70 ± 1.51	68.57 ± 1.61
w/o Typed Cues	54.72 ± 1.93	24.06 ± 0.91	48.51 ± 1.12	67.78 ± 2.82
Ungated	43.81 ± 1.84	22.64 ± 0.84	39.34 ± 1.51	51.94 ± 3.16

Removing the gate lowers mAP@0.5 by 11.53 points relative to the matched reference mean and reduces spot mAP by 16.79 points. Removing the soft typed cues changes mAP@0.5 by −0.62 points, which is small relative to run-to-run variation. Removing the role proxy instead changes mAP@0.5 by +2.74 points. The available evidence therefore supports gating but does not support a general accuracy benefit from handcrafted role cues. We retain those cues only as optional context and diagnostics; a role contribution would require verified supervision or an independently validated downstream analysis.

5.6 Full 19-Class Diagnostic Result

The calibrated 19-class model reaches only 17.12% mAP@0.5 and 6.50% mAP@0.7 (Table 5). Long tails, one absent validation label, co-occurrence, and boundary ambiguity make fine-grained localization unresolved rather than a demonstrated contribution.

Table 5: Diagnostic results on the full 19-class setting. The mAP column is the legacy evaluation-script aggregate. Values are percentages.

Method	mAP@0.5	mAP@0.7	mAP
S5 Head-only from S3	13.89	6.54	13.21
S5 Calibrated	17.12	6.50	15.89

6 CONCLUSIONS

We presented a skeleton-based tactical localizer with directed player relations and gated multi-scale decoding. Three-seed results support temporal gating, but not an accuracy contribution from the handcrafted role proxy. Primary-4 measures frequent primitives; Coarse-4 preserves all functional anchors. Strict boundary localization remains unresolved. Broader coaching claims require game-disjoint splits, matched baselines, multi-label localization, and expert validation.

ACKNOWLEDGEMENTS

Supported by JSPS KAKENHI Grant 26K02944.

AI Tools: ChatGPT assisted translation and language polishing; the authors reviewed all content.

REFERENCES

Brody, S., Alon, U., and Yahav, E. (2021). How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*.

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., and Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13359–13368.

Chi, H.-g., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., and Ramani, K. (2022). Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196.

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

Han, M., Zhang, D. J., Wang, Y., Yan, R., Yao, L., Chang, X., and Qiao, Y. (2022). 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)*, pages 2990–2999.

Hauri, S. and Vucetic, S. (2022). Group activity recognition in basketball tracking data–neural embeddings in team sports (nets). *arXiv preprint arXiv:2209.00451*.

Hu, G., Cui, B., He, Y., and Yu, S. (2020). Progressive relation learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 980–989.

Hu, Q., Yeung, C., Yin, L., and Fujii, K. (2025). Tal3x3: A dataset and algorithm for temporal action localization in 3x3 basketball. *IEEE Access*.

Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., and Yi, S. (2021). Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13668–13677.

Lin, T., Liu, X., Li, X., Ding, E., and Wen, S. (2019). Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3889–3898.

Lin, T., Zhao, X., Su, H., Wang, C., and Yang, M. (2018). Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.

Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., and Tao, D. (2023). Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18866.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12026–12035.

Tan, J., Tang, J., Wang, L., and Wu, G. (2021). Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13526–13535.

Tang, T. N., Kim, K., and Sohn, K. (2023). Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*.

Teranishi, M., Tsutsui, K., Takeda, K., and Fujii, K. (2022). Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction. *arXiv preprint arXiv:2206.01899*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Wang, X. and Gupta, A. (2018). Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417.

Watanabe, S. and Kameda, Y. (2024). Punch type classification and hit judgement using estimated skeletal model in boxing match videos. In *Proceedings of the 12th International Conference on Sport Sciences Research and Technology Support (icSPORTS)*, pages 218–224.

Xu, Y., Bretzner, L., Wang, T., and Maki, A. (2025). Skorxg: Skeleton-oriented expected goal estimation in soccer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5977.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 7444–7452.

Yang, Y., Wang, W., Liu, Y., Dong, L., Wu, H., Zhang, M., Zhong, Z., and Sun, X. (2025). Sga-interact: A 3d skeleton-based benchmark for group activity understanding in modern basketball tactic. *arXiv preprint arXiv:2503.06522*.

Zhang, C.-L., Wu, J., and Li, Y. (2022). Actionformer: Localizing moments of actions with transformers. In *Computer Vision – ECCV 2022*, pages 492–510. Springer.

Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., and Graf, H. P. (2022). Composer: Compositional reasoning of group activity in videos with keypoint-only modality. In *European Conference on Computer Vision (ECCV)*, pages 249–266. Springer.