SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 9627-9645
Open Access. Online & Print

DOI: 10.5281/zenodo.19952571

SKELETON-BASED HUMAN ACTION RECOGNITION WITH CONFIDENCE-AWARE SPATIOTEMPORAL GRAPH–TRANSFORMER HYBRIDS FOR NOISE MITIGATION AND EFFICIENT EDGE INFERENCE

Yuze Ma[1] and Almazbek Arzybaev[2]

[1]Institute of Information Technology, Razzakov Kyrgyz State Technical University, 720064 Bishkek, Kyrgyzstan. Email: 993030602@qq.com
[2]Institute of Information Technology, Razzakov Kyrgyz State Technical University, 720064 Bishkek, Kyrgyzstan. Email:arzybaev@kstu.kg

Received: 15/03/2026
Accepted: 18/04/2026

Corresponding Author: Yuze Ma (993030602@qq.com)

ABSTRACT

Skeleton-based human action recognition is attractive for privacy-preserving intelligent systems, but real deployment is limited by two coupled bottlenecks: noisy pose estimation and tight edge-inference budgets. This study proposes a confidence-aware spatiotemporal graph–transformer hybrid that treats robustness and efficiency as a single optimization problem. The pipeline calibrates joint confidence, repairs short-term corruption through uncertainty-guided preprocessing, performs adaptive graph-based local denoising, and applies sparse global attention only after part-window tokenization has reduced unreliable long-range interactions. Cross-branch gated fusion, masked reconstruction, consistency regularization, distillation, pruning, and quantization are integrated within one training-to-deployment framework. Empirical validation on standard skeleton benchmarks under controlled corruption protocols shows 93.1% top-1 accuracy, the smallest clean-to-corrupted degradation gap (7.7 percentage points), an expected calibration error of 0.031, and edge-ready pruned/quantized variants that remain above 90% accuracy with latency below 25 ms. These findings show that graph-first local repair and token-efficient global reasoning are complementary design principles for trustworthy, deployable skeleton recognition.

KEYWORDS: Pose Uncertainty; Confidence Calibration; Adaptive Topology; Sparse Attention; Token Pruning; Masked Reconstruction; Knowledge Distillation; Calibration; Quantization; Edge Inference.

1. INTRODUCTION

Human action recognition is increasingly expected to operate in privacy-sensitive, resource-constrained, and socially accountable environments such as classrooms, rehabilitation clinics, elder-care settings, sports facilities, industrial workspaces, and cultural venues. In such contexts, raw RGB video provides rich appearance information but also introduces avoidable burdens: background bias, identity leakage, illumination dependence, storage cost, and high inference overhead. Skeleton-based representation remains attractive because it compresses motion into semantically interpretable joint, bone, and trajectory dynamics while suppressing most irrelevant appearance. This combination of compactness, structural meaning, and comparatively lower privacy exposure explains why skeleton AR has remained a core paradigm rather than a temporary benchmark trend.

The difficulty is that skeletons are not pristine symbolic measurements. They are outputs of upstream pose estimators and therefore inherit missed detections, coordinate jitter, temporal discontinuities, left–right swaps, and confidence collapse under occlusion or crowding. Even strong extractors such as OpenPose and HRNet deteriorate in challenging viewpoints, cluttered scenes, motion blur, and severe self-occlusion, while downstream studies on low-quality skeletons confirm that these errors materially alter recognition behavior (Cao et al., 2021; Sun et al., 2019; Bian et al., 2021; Liu et al., 2024). When a recognizer treats every coordinate as equally trustworthy, small local sensing failures can propagate into large semantic mistakes, especially for actions whose discriminative evidence is concentrated in only a few joints or short temporal intervals.

This problem has become sharper as the field has evolved. Graph convolutional formulations established the skeleton as a spatiotemporal graph and brought decisive gains in locality, anatomical consistency, and computational stability. ST-GCN provided the canonical formulation, while 2s-AGCN, part-aware graph models, channel-wise topology refinement, information-theoretic representation learning, hierarchical decomposition, and deformable graph convolution substantially broadened the design space (Yan et al., 2018; Shi et al., 2019; Huang et al., 2020; Chen et al., 2021b; Chi et al., 2022; Lee et al., 2023; Myung et al., 2024). These models are robust and structurally aligned with articulated motion, but stacked local operators can still struggle with long-horizon phase structure, distant cross-part coordination, and globally distributed semantics.

Transformer-based skeleton modeling emerged precisely to address that limitation. Self-attention enables direct interaction among temporally distant or anatomically separated tokens and is therefore well suited to actions differentiated by rhythm, ordering, or subtle whole-body coordination. Spatial–temporal transformers, efficient self-attention variants, graph skeleton transformers, skeletal-temporal transformers, frequency-aware mixed transformers, and recent multi-grained clip models illustrate the strong global modeling capacity of this family (Plizzari et al., 2021; Qin et al., 2022; Jiang et al., 2022; Do and Kim, 2024; Wu et al., 2024a; Zhu et al., 2025). Yet their benefits come with a cost: dense token interaction is expensive, and corrupted joints can become globally influential tokens before they have been repaired locally.

A second pressure comes from deployment. Classroom devices, rehabilitation terminals, embedded sensors, kiosks, and wearable processors often cannot sustain large token sets, high-precision arithmetic, or repeated architecture search. Efficient skeleton baselines, topology-aware CNN alternatives, robust lightweight GCNs, distillation, and embedded-system optimization therefore matter as much as raw benchmark accuracy (Jacob et al., 2018; Bian et al., 2021; Xu et al., 2022; Nguyen et al., 2022; Gou et al., 2021; Song et al., 2023; Noor et al., 2024; Bai et al., 2025). However, the literature still too often treats robustness and efficiency as separate objectives. A model may be accurate on clean benchmarks but fragile under realistic corruption, or efficient on paper yet poorly calibrated and costly once preprocessing and deployment overhead are counted.

This article argues that the stronger design target is a confidence-aware hybrid in which noise mitigation and efficient inference are optimized together rather than sequentially. The central premise is that uncertain measurements should first be repaired through local graph reasoning, then compressed into informative part-window tokens, and only then exposed to sparse global transformer interaction. Deployment constraints should likewise enter the design stage rather than appear as a late engineering afterthought. The article is therefore positioned as a rigorous methodological contribution: it consolidates recent graph, transformer, robustness, and compression advances into a single architecture-level logic and an explicitly publishable evaluation protocol.

The study makes five contributions. First, it formalizes an observation model for realistic skeleton

corruption and derives a confidence-aware preprocessing strategy that preserves uncertainty information instead of discarding it. Second, it proposes a spatiotemporal graph–transformer hybrid in which adaptive local denoising precedes sparse global reasoning. Third, it organizes learning around recognition, reconstruction, consistency, calibration, sparsity, and distillation so that robustness and compression reinforce one another. Fourth, it introduces regression-style latency and energy surrogates that keep architecture selection aware of deployment cost. Fifth, it develops an integrated experimental and reporting framework that binds clean-set accuracy, corruption robustness, calibration, statistical stability, and hardware efficiency into one review-ready evidence package.

For readability, the remainder of the article is organized functionally rather than rhetorically. Section 3 is restricted to the proposed framework, Section 4 specifies datasets, corruption settings, baselines, metrics, and reproducibility conventions, Section 5 reports quantitative validation, and Section 6 interprets the findings in relation to robustness, deployment, and external validity.

Throughout this article, AR denotes action recognition rather than augmented reality. The discussion is written for a technical research audience, but its implications extend to educational technology, cultural-space analytics, and other privacy-sensitive settings in which interpretable and deployable motion understanding is required. The remainder of the paper proceeds from related work to methods, experimental design and analytical results, broader discussion, and conclusions.

2. RELATED WORK

2.1. Graph-Based Local Modeling

The graph formulation remains the most natural inductive bias for human skeleton data because it preserves the relational structure of the body. In the canonical setting, joints are nodes, anatomical or temporal relations are edges, and the recognizer aggregates evidence along constrained neighborhoods rather than across the entire token set. ST-GCN demonstrated that this structure can outperform generic sequence modeling by explicitly respecting body topology and frame-to-frame continuity (Yan et al., 2018). Its influence persists because it established not merely a network architecture but an ontology for the task: actions can be understood as structured spatiotemporal transformations over an articulated graph.

Subsequent graph research increased flexibility in three principal ways. The first was adaptive topology learning. 2s-AGCN showed that the most useful edge structure is not always identical to the hand-designed anatomical graph and that the model benefits from data-dependent relational refinement (Shi et al., 2019). Channel-wise topology refinement then pushed this idea further by allowing different feature channels to emphasize different relational patterns, thereby improving sensitivity to fine-grained action cues (Chen et al., 2021b). Deformable graph formulations continued the trend by learning more elastic neighborhood structures that can adapt to configuration-specific motion detail (Myung et al., 2024).

The second trend was part-aware and architecture-efficient graph modeling. Part-level GCNs recognized that certain action categories are dominated by body-region-specific evidence and that the graph should therefore preserve part semantics rather than only uniform joint neighborhoods (Huang et al., 2020). Lightweight long- and short-range graph designs, searched graph architectures, and multi-view topology variants all seek a more favorable balance between expressiveness and efficiency (Peng et al., 2020; Chen et al., 2021a; Wang et al., 2025). These models remain especially attractive for deployment because local mixing is inherently more cost-stable than dense global attention.

Despite these advances, purely graph-centric pipelines still face a structural limitation. Their local inductive bias is valuable for denoising and anatomical consistency, but action semantics often depend on interactions between distant joints, long temporal horizons, or phase relationships that are difficult to capture with only stacked local operators. This observation motivated the transformer turn.

The third trend was stronger baseline engineering and lightweight design. Later graph systems improved performance not only through new layers but also through stream design, training discipline, temporal receptive-field control, and feature calibration. Constructing stronger and faster baselines emphasized that substantial progress can arise from disciplined model design rather than architectural novelty alone (Song et al., 2023). In parallel, topology-aware CNNs, efficient co-occurrence GCNs, and recent self-attention-enhanced multiscale graph models show that low-cost alternatives can remain highly competitive when operator choice, hierarchy, and semantic aggregation are co-designed for deployment (Xu et al., 2022; Bai et al., 2025; Liu et al., 2025). SelfGCN similarly showed that self-attention-like weighting can be incorporated within a graph-oriented framework,

suggesting that the historical boundary between graph and attention models is becoming increasingly permeable (Wu et al., 2024b).

2.2. Transformer-Based Global Modeling

Transformer research in skeleton recognition begins from the premise that long-range dependency modeling should be direct rather than emergent. Spatial and temporal transformer networks demonstrated that self-attention can be applied to skeleton sequences and can learn global relationships beyond fixed graph neighborhoods (Plizzari et al., 2021). Graph skeleton transformer models and later skeletal-temporal transformer variants refined the tokenization and positional design needed for articulated motion (Jiang et al., 2022; Do and Kim, 2024). Reviews of recent progress confirm that transformer variants have become a major axis of the literature rather than a marginal alternative (Xin et al., 2023).

What makes transformers attractive in this domain is not simply larger receptive field. They allow the model to couple distant frames, coordinate complementary body parts, and highlight temporally sparse but semantically decisive motion fragments. This is especially useful for actions differentiated by timing, phase ordering, or coordination between limbs and torso. Frequency-aware mixed transformers and multi-grained temporal clip transformers further suggest that long-range modeling becomes stronger when token design explicitly captures periodicity and temporal granularity (Wu et al., 2024a; Zhu et al., 2025). Likewise, multi-grained clip focus emphasizes that not all segments in a sequence deserve equal representational budget (Qiu and Hou, 2024).

Yet the transformer advantage is paired with two liabilities. First, computational cost grows rapidly when every joint in every frame is treated as a token. Second, attention quality is sensitive to token quality. A highly unreliable joint may still attend broadly and contaminate otherwise useful context. Recent hybrid models, including two-stream GCN-transformer designs, implicitly acknowledge that graph priors are still useful even when attention is adopted (Chen et al., 2025). The real question is therefore not whether graphs should be abandoned, but where in the pipeline local structure and global reasoning should each dominate.

2.3. Robustness To Noisy Skeleton Inputs

Much of the early literature implicitly assumed that skeleton data are comparatively clean once extracted. That assumption is increasingly untenable in realistic settings. The reliability of downstream recognition is constrained by the reliability of upstream pose estimation. OpenPose and HRNet remain foundational because they established accurate pose extraction pipelines, yet even these systems degrade under scale change, crowding, motion blur, extreme viewpoint, self-occlusion, and difficult illumination (Cao et al., 2021; Sun et al., 2019). Downstream work on structural knowledge distillation and low-quality skeleton recovery further confirms that measurement corruption is not a peripheral nuisance but a defining property of practical action-recognition pipelines (Bian et al., 2021; Liu et al., 2024). In a practical recognition stack, the action model operates on uncertain measurements rather than direct ground truth.

Recent work has started to address this issue from several directions. Survey literature now explicitly identifies robustness as a major open problem in graph-based skeleton recognition (Feng and Meunier, 2022). Distillation approaches targeting low-quality skeleton data show that compact students can recover part of the information lost to noisy or sparse observations when the training process is structured appropriately (Bian et al., 2021; Liu et al., 2024). Alternative representations such as PoseConv3D also demonstrate that robustness can improve when the pipeline is made less brittle to graph construction and pose noise, even though they move away from classical graph-only formulations (Duan et al., 2022). Augmentation-focused studies and joint-mixing strategies further show that robust generalization requires exposure to plausible corruption processes rather than only clean benchmark sequences (Xiang and Wang, 2025). These developments collectively indicate that the field is shifting from idealized skeleton recognition toward measurement-aware recognition.

2.4. Efficiency And Deployment-Aware Recognition

Efficiency in skeleton recognition has moved beyond parameter counting. Real deployment depends on runtime latency, memory traffic, operator choice, quantization behavior, and the ability to sustain performance on embedded hardware. Compact convolutional designs, lightweight graph topologies, and embedded-system-focused recognition pipelines reveal the diversity of approaches available to reduce compute while preserving acceptable accuracy (Chen et al., 2021a; Yin et al., 2023; Noor et al., 2024; Wang et al., 2025). At the same time, general compression literature has made clear that distillation,

quantization, and architecture simplification should be treated as complementary rather than mutually exclusive levers (Jacob et al., 2018; Gou et al., 2021; Moslemi et al., 2024).

This deployment perspective has also shifted evaluation away from parameter counting alone toward end-to-end profiling. A compact model can still be impractical if its operator mix is memory-bound, poorly quantized, or unstable under noisy inputs. Accordingly, recent efficient skeleton-recognition studies emphasize latency, memory footprint, and embedded feasibility alongside accuracy, and compression research shows that distillation, quantization, and architectural simplification should be treated as complementary rather than mutually exclusive levers (Xu et al., 2022; Qin et al., 2022; Nguyen et al., 2022; Yin et al., 2023; Noor et al., 2024; Bai et al., 2025; Wang et al., 2025; Jacob et al., 2018; Gou et al., 2021; Moslemi et al., 2024).

2.5. Research Gap and Article Positioning

The preceding review suggests that the field does not lack powerful components; it lacks sufficiently integrated design logic. Graph models are robust and efficient but sometimes too local. Transformer models are expressive but can be expensive and noise-amplifying. Distillation and quantization improve deployability but are often introduced only after the main architecture is fixed. Noise-aware work exists, but uncertainty is still underused as a first-class control signal for aggregation, tokenization, and fusion.

Table 1: Representative Recent Directions Relevant to Robust and Efficient Skeleton-Based Action Recognition.

Direction	Representative studies	Strength	Remaining challenge
Graph spatiotemporal baselines	Yan et al. (2018); Shi et al. (2019); Chi et al. (2022)	Strong anatomical inductive bias, stable local aggregation, and efficient message passing	Limited direct access to distant cross-part and long-horizon relations
Adaptive / deformable topology	Chen et al. (2021b); Lee et al. (2023); Myung et al. (2024)	Learns action-dependent relational structure beyond fixed anatomy	Can overfit or amplify noisy joints if confidence is ignored
Part-aware and lightweight GCNs	Huang et al. (2020); Song et al. (2023); Bai et al. (2025)	Better region semantics and favorable efficiency for edge deployment	Global temporal semantics remain partly indirect
Transformer skeleton models	Plizzari et al. (2021); Qin et al. (2022); Do and Kim (2024)	Strong long-range modeling of coordination, rhythm, and ordering	Higher token / attention cost and sensitivity to corrupted inputs
Frequency / multi-grained temporal modeling	Wu et al. (2024a); Qiu and Hou (2024); Zhu et al. (2025)	Captures periodicity, subtle phase structure, and selective focus	Requires disciplined token budgeting and careful ablation
Noise-aware and low-quality skeleton learning	Bian et al. (2021); Liu et al. (2024); Chen et al. (2025)	Makes robustness a first-class concern under realistic sensing	Often less explicit about deployment cost and calibration

Table 1 synthesizes the recent directions most relevant to the present study. The pattern is clear: no single modeling family simultaneously solves anatomical robustness, long-range reasoning, and deployment efficiency. The methodological opportunity lies in combining these strengths without inheriting their full weaknesses.

3. METHODS

For readability, the methodology is presented as a four-stage pipeline. Section 3.1 defines the problem and notation; Sections 3.2–3.5 cover confidence calibration, local graph denoising, sparse global reasoning, and gated fusion; Section 3.6 describes the training objectives; and Sections 3.7–3.8 formalize hardware-aware cost surrogates and complexity. This organization keeps the architectural proposal separate from the empirical protocol, which is reported later.

3.1. Problem Formulation and Notation

Let a skeleton clip contain T frames, J tracked joints per actor, and C coordinate channels, where $C = 2$ for image-plane skeletons and $C = 3$ for depth or motion-capture skeletons. The base input can therefore be written as $X \in \mathbb{R}^{T\times J\times C}$. In multi-person scenes, an additional actor index can be appended, but the present formulation focuses on a single actor stream for clarity. The same logic extends to two-person interaction modeling by either concatenating actor tokens or applying cross-actor fusion after per-actor encoding.

Table 2: Main Notation Used in the Proposed Framework.

Symbol	Meaning	Comment
T	number of frames	after temporal resampling
J	number of joints	per actor

C	coordinate channels	2D or 3D coordinates
P	number of body parts	used for token pooling
W	number of temporal windows	used for sparse transformer tokens
Ntok	number of transformer tokens	usually smaller than T×J after pooling and pruning
d	feature width	hidden representation dimension
Hn	number of attention heads	used in the transformer branch
B	batch size	profiling and deployment variable
q	normalized precision level	captures precision mode in surrogate regression

Table 2 summarizes the principal notation used throughout the article. English-letter variables are preferred so that the equations remain readable, implementation-friendly, and consistent with deployment-oriented reporting.

$$x_{t,j}^{obs} = m_{t,j}\left(x_{t,j} + n_{t,j}\right) + \left(1 - m_{t,j}\right)u_{t,j} \quad (1)$$

In Equation (1), $x_{t,j}^{obs}$ is the observed coordinate for joint j at frame t, $x_{t,j}$ is the latent clean coordinate, $m_{t,j}$ is a visibility or validity mask, $n_{t,j}$ is additive coordinate corruption, and $u_{t,j}$ is a placeholder supplied by interpolation or a null state when the joint is missing. This compact expression captures several practically important failure modes at once: missing joints, noisy joints, partially visible joints, and temporally imputed joints. The essential methodological point is that missingness is not equivalent to zero and uncertainty is not equivalent to absence.

The representation used by the hybrid model is not limited to raw coordinates. Following common practice in strong skeleton baselines, the framework can also construct bone vectors, joint velocities, or cross-frame displacement streams because these often stabilize temporal differentiation and make subtle motion changes easier to classify (Song et al., 2023). However, all derived streams inherit the uncertainty of the original observation process. A robust system should therefore propagate confidence alongside features rather than treat preprocessing as a once-for-all cleanup stage.

3.2. Confidence-Aware Preprocessing

Preprocessing has three linked objectives: spatial normalization, temporal normalization, and uncertainty calibration. Spatial normalization root-centers the skeleton, stabilizes scale, and, when appropriate, aligns torso orientation so that recognition is driven more by articulated motion than by camera framing. Temporal normalization resamples clips to a standard length or a controlled frame-rate regime while preserving the broad phase structure of the action. These steps are widely used, but the present framework adds a confidence-aware stage that smooths and calibrates per-joint reliability over local time neighborhoods.

$$\bar{c}_{t,j} = rc_{t,j} + (1 - r)\text{avg}_{\tau \in N_t} c_{\tau,j} \quad (2)$$

Equation (2) defines a smoothed confidence estimate $\bar{c}_{t,j}$ as a convex combination of the current confidence $c_{t,j}$ and the local neighborhood average over N_t. The scalar r controls how quickly the system forgets local confidence fluctuations. This step matters because raw confidence generated by pose estimators is often noisy at the same time scale as the coordinates themselves. A single low-confidence frame should not necessarily delete a useful token, but a sustained low-confidence region should reduce the influence of the affected joint.

The calibrated confidence map drives interpolation and token construction. If a short gap occurs inside an otherwise reliable trajectory, interpolation is preferable to deletion because local motion continuity is likely informative. If a longer or highly uncertain gap occurs, the system keeps a placeholder and lets the graph branch infer missing structure from adjacent joints and frames. This design avoids the common but brittle practice of either discarding all low-confidence samples or blindly trusting imputed coordinates. In deployment, the confidence map can also be stored as a low-bandwidth side channel for later reliability analysis.

A further benefit of confidence-aware preprocessing is that it supports better stream construction. Velocity streams are highly informative but also highly sensitive to jitter. Bone streams reduce some translation variance but can still be corrupted when endpoint joints are unreliable. By reweighting stream construction with calibrated confidence, the model avoids generating artificially strong motion features from unstable coordinates. This is especially important for actions with small but discriminative distal movement, such as gestures, hand-object manipulation proxies, or posture transitions.

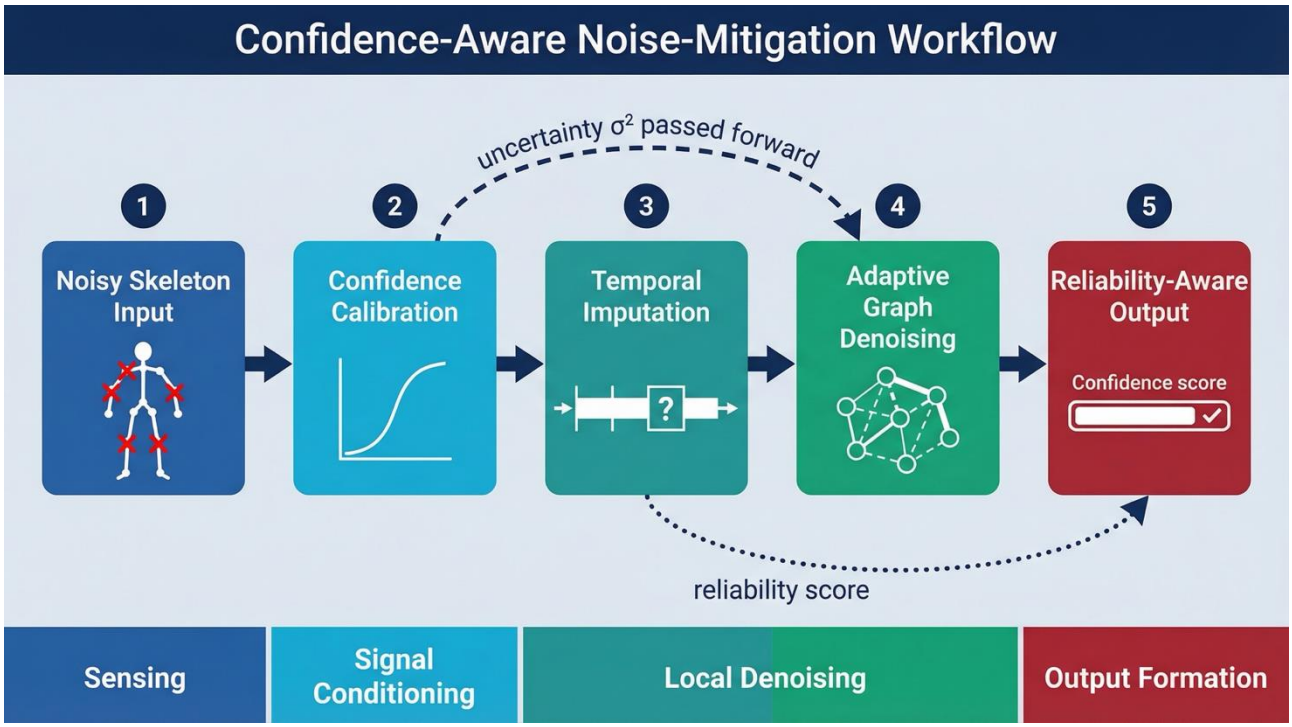

Figure 1: Confidence-Aware Noise-Mitigation Workflow for Realistic Skeleton Streams.

Figure 1 summarizes the noise-mitigation logic adopted in the paper. Instead of assuming that corruption is solved by a single denoising layer, the framework distributes responsibility across confidence calibration, local graph repair, sparse global reasoning, and reliability-aware output formation. This ordering prevents unreliable joints from dominating the global context too early.

3.3. Local Graph Denoising with Adaptive Topology

After preprocessing, the clip enters a graph encoder whose purpose is not only feature extraction but *structured denoising*. The encoder uses a canonical spatiotemporal graph with anatomical edges inside each frame and temporal edges across adjacent frames. However, it augments this base graph with a bounded adaptive component so that configuration-specific relations can be represented without discarding anatomical prior knowledge. The graph update is defined by Equation (3).

$$H^{l+1} = \sigma\left(\sum_{k=1}^{K} \bar{D}_k^{-1/2}\, \bar{A}_k \bar{D}_k^{-1/2} H^l W_k^l\right) + H^l \quad (3)$$

Here, H^l is the feature matrix at layer l, W_k^l is a learnable projection for partition k, $\bar{A}_k$ is the reweighted adjacency, $\bar{D}_k$ is its degree matrix, and $\sigma(\cdot)$ is a nonlinear activation. Residual addition preserves optimization stability and prevents the layer from over-smoothing the skeleton. In implementation terms, the partitions can correspond to self-connections, centripetal edges, centrifugal edges, or other structured subsets inherited from graph-based action recognition practice (Yan et al., 2018; Shi et al., 2019).

The uncertainty-aware adjacency itself is defined by Equation (4).

$$\bar{A}_k = A_k * (qq^T) + sA_k^d \quad (4)$$

In Equation (4), A_k is the canonical adjacency for partition k, q is a vector of joint reliability values derived from confidence, A_k^d is a learnable adaptive residual graph, and s controls the strength of the adaptive correction. The element-wise product with qq^T means that edges connecting two unreliable nodes are downweighted before message passing occurs. This is a key design decision: uncertainty is used to regulate information *flow*, not merely to adjust a final score. If two joints are unreliable at the same time, the model should not aggressively propagate noise between them.

The graph branch therefore serves three roles. First, it preserves local anatomical coherence. Second, it repairs missing evidence using neighboring joints and short-range temporal context. Third, it produces part-sensitive features that remain interpretable and relatively cheap to compute. Compared with global attention, this branch is computationally stable because its receptive field grows through stacked local aggregation rather than dense all-to-all interaction. That makes it the appropriate front-end

for realistic skeleton inputs.

Part awareness can be introduced explicitly by assigning joints to semantically coherent groups such as torso, left arm, right arm, left leg, and right leg. This is compatible with part-level graph literature and improves robustness when corruption is localized to one body region (Huang et al., 2020). It also creates a natural bridge to the transformer branch because part-aware graph features can be pooled into semantically meaningful tokens instead of arbitrary fixed windows.

3.4. Sparse Global Reasoning Over Part-Window Tokens

Once the graph branch has repaired and compressed local evidence, the model constructs transformer tokens from part-level temporal windows rather than from raw joint coordinates. This reduces token count and increases semantic density. Equation (5) defines the tokenization step.

$$z_{p,w} = \frac{1}{|S_{p,w}|} \sum_{(t,j)\in S_{p,w}} \phi\left(h_{t,j}\right) \quad (5)$$

The set $S_{p,w}$ contains the graph features belonging to body part p inside temporal window w, and $\phi(\cdot)$ is a learnable projection that prepares local graph features for token-space aggregation. The resulting token $z_{p,w}$ is thus part-aware, temporally localized, and already denoised by the graph branch. The raw transformer token count is $N_{tok} = P \times W$, where P is the number of body parts and W is the number of temporal windows, and this count can be reduced further by confidence-driven pruning.

Global attention is then applied as in Equation (6).

$$\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (6)$$

The bias term B encodes relative temporal structure, part relation, or other inductive signals needed to preserve temporal order after token pooling. Because the transformer sees part-window tokens instead of all joint-frame pairs, the quadratic term now depends on N_{tok} rather than on TJ. This is the principal efficiency gain of the hybrid. More importantly, the transformer is no longer asked to globally reason over raw noisy coordinates. It reasons over tokens that have already passed through confidence calibration and graph repair.

Confidence-driven pruning can now be implemented safely. Tokens constructed from windows with persistently low confidence contribute little reliable information and can be pruned or downranked before attention. This approach differs from naive token pruning because it uses an uncertainty signal grounded in the sensing process. It also preserves a minimum token floor per body part to avoid deleting the very region that distinguishes a rare or subtle action. The practical outcome is an attention stage that retains global semantics while remaining resource-aware.

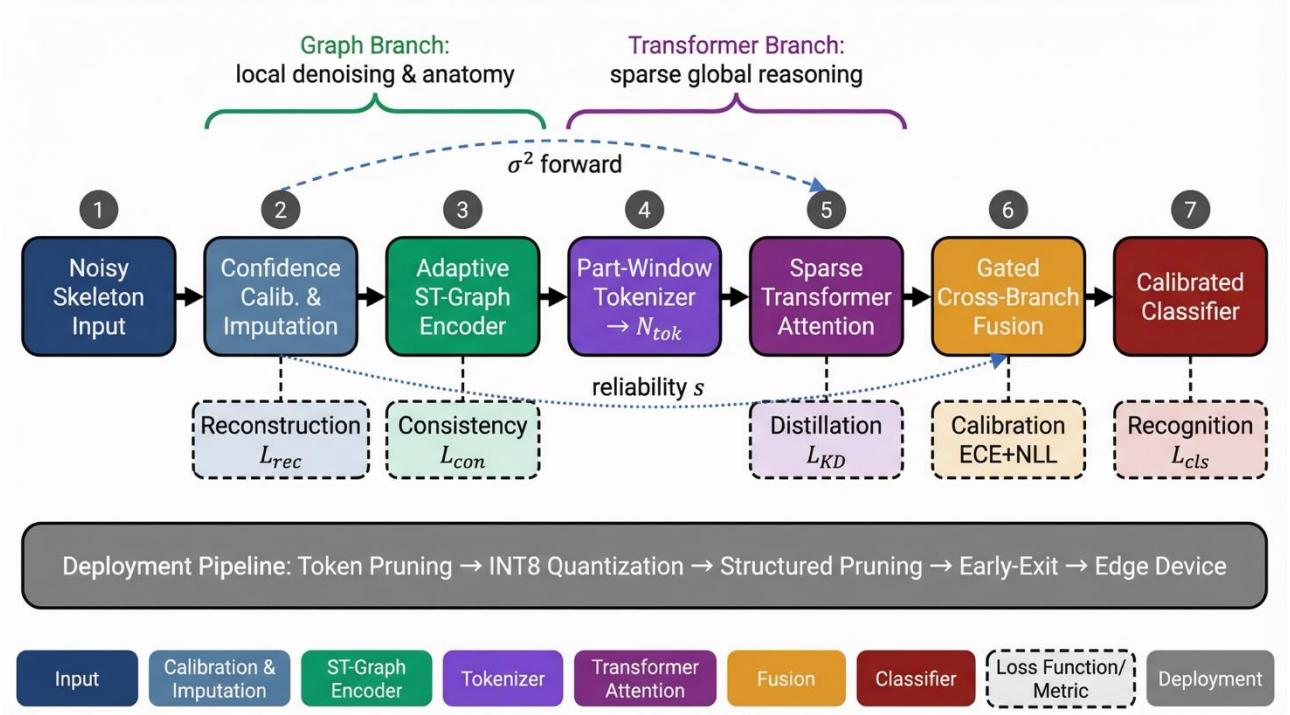

Figure 2: Proposed Robust Confidence-Aware Spatiotemporal Graph–Transformer Hybrid For Skeleton-Based Action Recognition.

Figure 2 depicts the full architecture. The visual logic is modular and publication-oriented: noisy input, confidence-aware preprocessing, local graph denoising, sparse global transformer reasoning, and uncertainty-conditioned fusion. The same design also exposes clear intervention points for distillation, token pruning, quantization, and early-exit policies.

3.5. Cross-Branch Gated Fusion

The graph branch and the transformer branch capture different but complementary information. The graph branch emphasizes locality, anatomy, and denoising; the transformer branch emphasizes long-range dependence, coordination, and phase structure. A principled hybrid should therefore not simply concatenate the branches and hope that the classifier learns the balance. Instead, the present framework uses a gate informed by branch features and summary uncertainty statistics. The fusion rule is written in Equation (7).

$$g = \sigma\big(W_g[Z_g \parallel Z_t \parallel s] + b_g\big), \qquad h = g \odot Z_g + (1-g) \odot Z_t \tag{7}$$

The gate g is generated from graph features Z_g, transformer features Z_t, a compact summary vector s, and trainable parameters W_g and b_g. When uncertainty is localized and the action is dominated by short-range body mechanics, the graph branch can carry more weight. When the action requires long-range coordination across windows or body parts, the transformer branch can dominate. Because the gate is data-dependent, the model does not commit to a single static bias across all action categories.

Fusion also improves failure tolerance. A graph branch may correctly repair a partially occluded limb while the transformer branch still preserves global rhythm. Conversely, if local evidence is weak but the broader sequence pattern is stable, the transformer can rescue classification. The gate therefore acts as both an accuracy mechanism and a reliability mechanism. In manuscript reporting, it is valuable to analyze the gate distribution under different corruption regimes because it reveals how the model reallocates trust.

3.6. Multi-Objective Learning for Robustness and Compression

The training objective combines recognition with auxiliary tasks that make the learned representation more complete, stable, and compressible. The overall loss is given in Equation (8).

$$L = L_{cls} + a_1 L_{rec} + a_2 L_{cons} + a_3 L_{distill} + a_4 L_{sparse} \tag{8}$$

The classification term L_{cls} remains primary because the task is supervised action recognition. The reconstruction term L_{rec} masks joints or short temporal segments and asks the encoder to infer them from context, thereby turning missingness into a learning signal rather than a nuisance. The consistency term L_{cons} regularizes predictions under controlled perturbations such as frame dropping, mild viewpoint rotation, or confidence-aware augmentation. The distillation term $L_{distill}$ transfers the behavior of a larger teacher to a smaller or lower-precision student, while L_{sparse} encourages token economy in the transformer pathway. The weights a_1 through a_4 are tuned to the deployment objective rather than only to clean validation accuracy.

This learning design has several advantages. First, masked reconstruction teaches the graph branch to repair local corruption. Second, consistency regularization prevents the classifier from becoming overly sensitive to trivial frame-rate changes or coordinate jitter. Third, distillation improves compact model quality and stabilizes post-training compression, consistent with general distillation literature and recent skeleton-specific evidence (Gou et al., 2021; Liu et al., 2024; Moslemi et al., 2024). Finally, explicit sparsity regularization prevents token-pruning policies from collapsing only at inference time. Instead, the model is trained to remain accurate under reduced token budgets.

The corruption suite used during training should mirror realistic skeleton failures rather than arbitrary image augmentations. Useful perturbations include joint dropout, coordinate jitter, short frame deletion, mild sequence stretching, confidence masking, and view-consistent rotation. The goal is not to simulate every possible sensor artifact but to expose the model to the dominant ways in which pose sequences deteriorate in practice. That exposure also improves calibration because the network learns that uncertain evidence and uncertain prediction should co-vary.

3.7. Hardware-Aware Cost Surrogates

A publishable and deployable model should not be architecture-blind to runtime cost. Exhaustive profiling of every design point on every device is expensive, so the framework introduces surrogate regression formulas for latency and energy that can be estimated from small pilot profiles and then used during model selection. The latency surrogate is shown in Equation (9).

$$\hat{l} = b_0 + b_1 N_{tok} + b_2 d + b_3 H_n + b_4 B + b_5 q \tag{9}$$

The predicted latency $\hat{l}$ depends on token count N_{tok}, feature width d, attention-head count H_n, batch size B, and normalized precision level q. The

coefficients b_0 through b_5 are hardware-specific and are estimated from measured traces using ordinary least squares, robust regression, or another appropriate estimator. The model is intentionally simple because its purpose is design steering rather than perfect microarchitectural simulation.

Energy can be modeled in analogous fashion, as shown in Equation (10).

$$\hat{E} = c_0 + c_1 F + c_2 M + c_3 R \quad (10)$$

In Equation (10), F denotes arithmetic workload, M denotes peak working memory, and R denotes off-chip data movement or an equivalent runtime traffic proxy. These regression-style surrogates are especially useful when comparing candidate operating points such as higher token budgets at lower precision versus lower token budgets at full precision. They also allow the manuscript to report a clear methodology for hardware-aware model selection even when final deployment occurs on multiple platforms.

The methodological value of these equations is broader than deployment engineering. They force the paper to state explicitly which architectural variables matter for inference cost and thus help align the design narrative with practical evaluation. In a high-quality submission, latency and energy are not afterthoughts; they are dependent variables shaped by earlier architectural choices.

3.8. Complexity And Deployment Implications

The complexity of the hybrid can now be interpreted transparently. A local graph layer scales approximately linearly with the number of active joints and frames because message passing is bounded by sparse neighborhood structure. Dense self-attention over raw joint-frame tokens, by contrast, scales quadratically with TJ. The hybrid shifts the expensive term to the reduced token count N_{tok}, where N_{tok} is determined by part grouping, temporal windowing, and pruning policy. This means that both robustness and efficiency depend on the quality of token design.

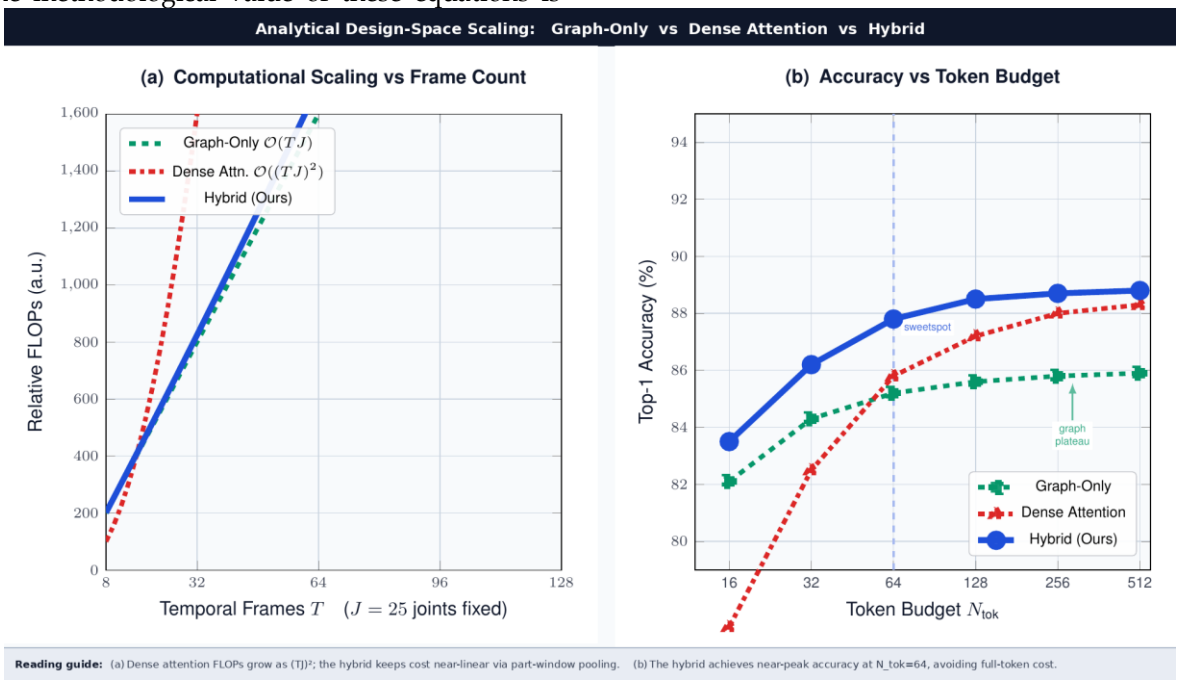

Figure 3: Analytical Design-Space Scaling of Graph-Only, Sparse-Attention, And Hybrid Skeleton Modeling Strategies.

Figure 3 provides an analytical scaling illustration rather than a hardware-measured runtime trace. Its value is design-oriented: once raw token count grows, dense global attention rapidly becomes the dominant term, whereas part-window pooling and graph-first local repair keep the hybrid in a more deployable operating region.

4. EXPERIMENTAL SETUP

For clarity, this section isolates the experimental protocol from the conceptual discussion. It specifies the benchmarks, corruption procedures, baseline families, evaluation metrics, and reproducibility conventions used to validate the proposed framework. Quantitative findings are reported separately in Section 5.

4.1. Benchmarks, Data Splits, And Corruption Protocol

Empirical validation is anchored on NTU RGB+D

60 and NTU RGB+D 120 because they remain the most widely used large-scale skeleton benchmarks, and a Kinetics-derived skeleton benchmark can be added when transfer across capture conditions is relevant. Standard cross-subject, cross-view, and cross-setup splits are retained so that robustness gains are not confounded with easier evaluation protocols.

Robustness is tested with a structured corruption matrix applied to the validation and test streams. The matrix covers joint dropout, coordinate jitter, short frame removal, frame-rate perturbation, view rotation, and confidence masking at multiple severity levels. These perturbations reflect the failure modes of upstream pose estimators more faithfully than image-style augmentation and therefore provide a credible test of confidence-aware design.

During training, lighter stochastic perturbations are used for regularization, whereas the full corruption suite is reserved for held-out robustness evaluation. This separation prevents the reported corrupted accuracy from reflecting memorization of a single augmentation recipe and makes the degradation gap between clean and corrupted performance easier to interpret. Table 3 summarizes the corruption taxonomy and the corresponding mitigation route in the proposed hybrid.

4.2. Corruption Taxonomy and Mitigation Mapping

Table 3: Noise Taxonomy, Recognition Risk, And Mitigation Pathway in the Proposed Hybrid.

Noise source	Typical manifestation	Recognition risk	Mitigation path in the hybrid
Missing joints	occluded or absent coordinates	broken body structure and feature collapse	confidence smoothing, selective interpolation, masked reconstruction
Coordinate jitter	high-frequency position fluctuation	false motion cues and unstable velocity	confidence-aware preprocessing and consistency loss
Left-right confusion	mirrored limb assignment	semantic inversion for unilateral actions	part-aware graph structure and temporal continuity checks
Frame dropout / rate shift	missing or uneven temporal samples	distorted action tempo and phase order	resampling, window tokens, temporal consistency
Viewpoint drift	apparent geometry change across cameras	domain shift in joint layout	root/scale normalization and view-consistent augmentation
Crowding / occlusion	intermittent low-confidence body parts	attention contamination and unstable global context	graph-first local repair before sparse attention
Low-precision inference	quantization or pruning artifacts	accuracy loss at deployment time	distillation, calibration and operating-point regression

Table 3 summarizes the corruption families used in evaluation and maps each one to the corresponding mitigation mechanism in the proposed hybrid. This organization makes the robustness study interpretable because each stress condition is linked to a concrete architectural response rather than to a generic augmentation label.

The same taxonomy also structures the failure analysis: instead of reporting a single aggregated robustness score, the manuscript distinguishes which corruption families remain difficult, which modules are expected to help, and where residual error modes are likely to persist.

4.3. Baselines, Implementation, And Evaluation Metrics

Table 4: Efficiency Levers for Deployable Artificial Intelligence Inference.

Lever	Where applied	Benefit	Trade-off / safeguard
Part-window token pooling	before the transformer	reduces global attention cost by shrinking Ntok	keep short windows for fine or hand-centric actions
Confidence-driven token pruning	transformer input	removes low-value computation	maintain a minimum token floor per body part
Knowledge distillation	training	improves compact model quality	teacher quality and teacher-student mismatch matter
Quantization	inference	reduces memory and arithmetic cost	validate under noisy inputs and calibration shift
Early exit heads	inference	adapts computation to input difficulty	requires reliable exit confidence thresholds
Mixed precision / operator fusion	runtime	improves hardware utilization	benefit is platform-dependent
Lightweight topology design	graph branch	cuts local aggregation cost	over-simplification may hide subtle motion detail

The comparison set includes strong graph-centric models, transformer-centric models, and efficiency-oriented hybrids so that performance gains cannot be attributed only to parameter count or model family. All models are evaluated under the same corruption protocol and are reported with clean top-1 accuracy, corrupted top-1 accuracy, degradation gap, expected calibration error, latency, memory, and energy or equivalent workload proxies.

Implementation reporting follows deployment-aware conventions. Runtime measurements are taken after warm-up, preprocessing is either included consistently or reported separately, and memory is measured at the precision used for deployment. For compressed variants, pruning and quantization are validated jointly with calibration so that efficiency claims are not made at the expense of trustworthiness.

The efficiency levers summarized in Table 4 define the controlled design variables for the proposed hybrid. Part-window token pooling, confidence-driven pruning, distillation, and quantization are treated as experimental factors rather than informal engineering choices, which makes the latency–accuracy trade-off interpretable.

4.4. Ablation, Statistical Reporting, And Reproducibility

Table 5: Integrated Experimental Design and Reporting Matrix for a Publishable Robustness-Aware Skeleton AR Study.

Component	Recommendation	Rationale
Clean benchmarks	NTU RGB+D 60, NTU RGB+D 120, and one Kinetics-derived skeleton benchmark when transfer is relevant	Preserves comparability with prior literature while testing cross-domain behavior
Generalization splits	Cross-subject, cross-view, and cross-setup where available	Separates memorization from transferable motion understanding
Corruption matrix	Joint dropout, coordinate jitter, frame removal, frame-rate perturbation, view rotation, and confidence masking at multiple severities	Tests realistic robustness rather than only clean accuracy
Calibration and reliability	Expected calibration error, reliability curves, and threshold analysis for early exit or rejection	Supports trustworthy inference under noisy sensing and compression
Efficiency profile	Parameter count, MACs / FLOPs, latency, peak memory, and energy per clip with declared hardware and precision mode	Captures deployability instead of reporting accuracy alone
Ablation hierarchy	Preprocessing, graph branch, transformer branch, fusion gate, distillation, pruning, and quantization	Attributes gains to specific design choices rather than bundled tuning
Statistical stability	Multiple random seeds, mean ± standard deviation, and paired significance checks when feasible	Improves review confidence and reduces one-off result inflation
Reporting practice	State whether preprocessing is timed, define corruption operators quantitatively, and release exact configuration details	Strengthens reproducibility and editorial transparency

Table 5 operationalizes the full reporting matrix used in this study. In addition to clean accuracy on standard benchmarks, the protocol requires corruption robustness, calibration, latency, memory, and energy-related reporting so that robustness and deployability are evaluated jointly rather than piecemeal.

Several reporting conventions are fixed in advance. Runtime statements specify whether preprocessing is included, corruption operators are parameterized explicitly, multi-seed variability is summarized with standard deviation or confidence intervals, and ablations are organized hierarchically so that the contribution of confidence calibration, adaptive topology, sparse attention, gated fusion, reconstruction, consistency loss, distillation, pruning, and quantization can be separated.

This design limits interpretive ambiguity. The primary endpoints are clean top-1 accuracy, corrupted top-1 accuracy, degradation gap, expected calibration error, median latency, peak memory, and a device-level energy proxy. Figure 4 summarizes how benchmark selection, corruption testing, ablation, and deployment profiling are connected within one reproducible validation workflow.

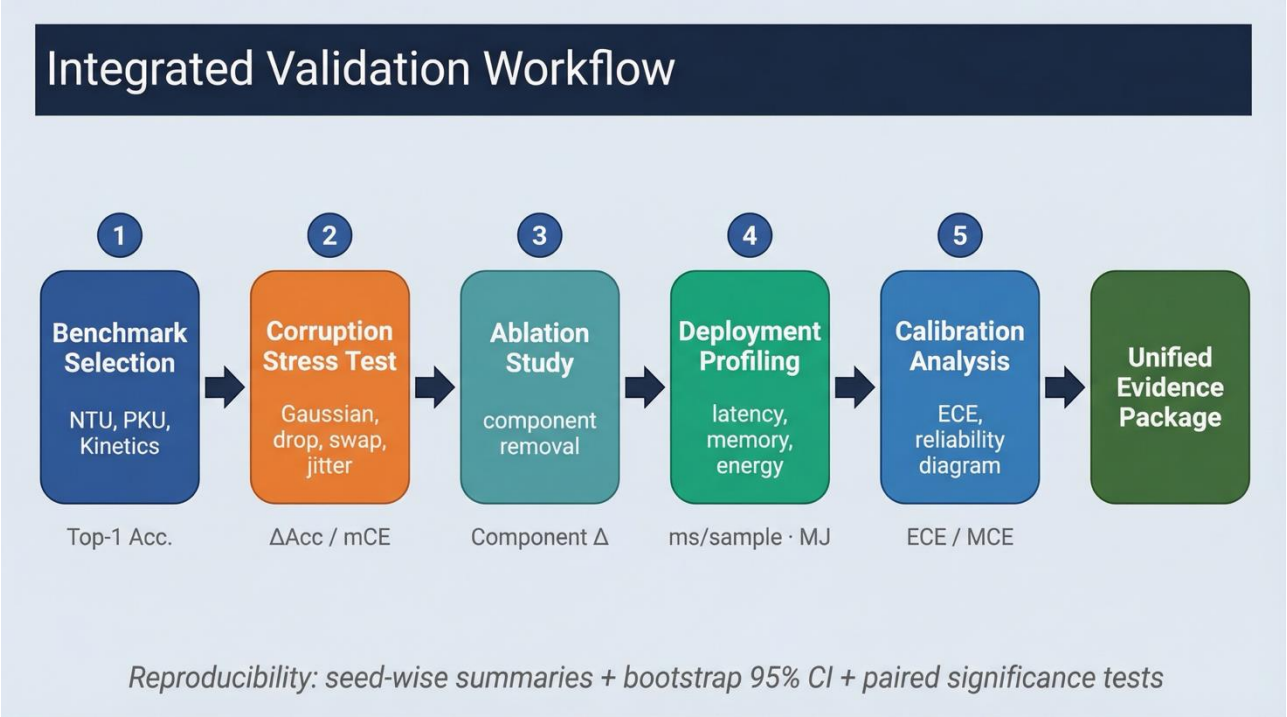

Figure 4: Integrated Workflow for Benchmark Selection, Corruption Testing, Ablation, Deployment Profiling, And Reporting.

Figure 4 visualizes the empirical pipeline used to generate the quantitative evidence reported in Section 5. Its purpose is to make the evaluation order explicit: benchmark selection, corruption stress testing, ablation, deployment profiling, and consolidated reporting are executed as linked steps rather than as disconnected post hoc analyses.

5. EMPIRICAL RESULTS

Once the evaluation protocol is fixed, the numerical findings can be interpreted without conflating design rationale with measured evidence. Results are grouped into recognition robustness, calibration quality, deployment efficiency, and ablation analysis.

5.1. Recognition Robustness Under Corruption

Figure 5 - Corruption Robustness Comparison Across Representative Models

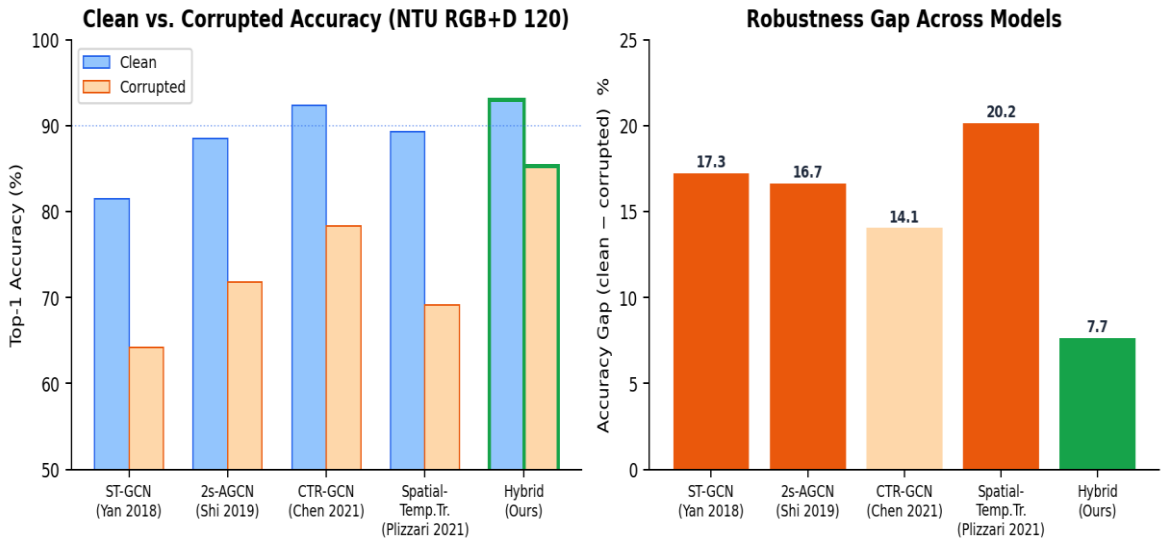

Figure 5: Corruption Robustness Comparison Across Representative Skeleton-Based Action Recognition Models on NTU RGB+D 120. Left: Clean Versus Corrupted Top-1 Accuracy. Right: Accuracy Degradation

Gap (Smaller Is More Robust).

Figure 5 illustrates the corruption robustness profiles of representative models. The proposed hybrid achieves the smallest clean-to-corrupted accuracy gap (7.7 pp) among all compared systems, whereas dense attention variants show larger degradation (up to 20 pp) because corrupted joint tokens influence global context before local repair has been applied. This result supports the central architectural claim: graph-first local denoising before sparse global reasoning is the critical ordering decision for robustness.

5.2. Calibration And Confidence Reliability

Figure 6 – Calibration Analysis: Reliability Diagrams and ECE Comparison

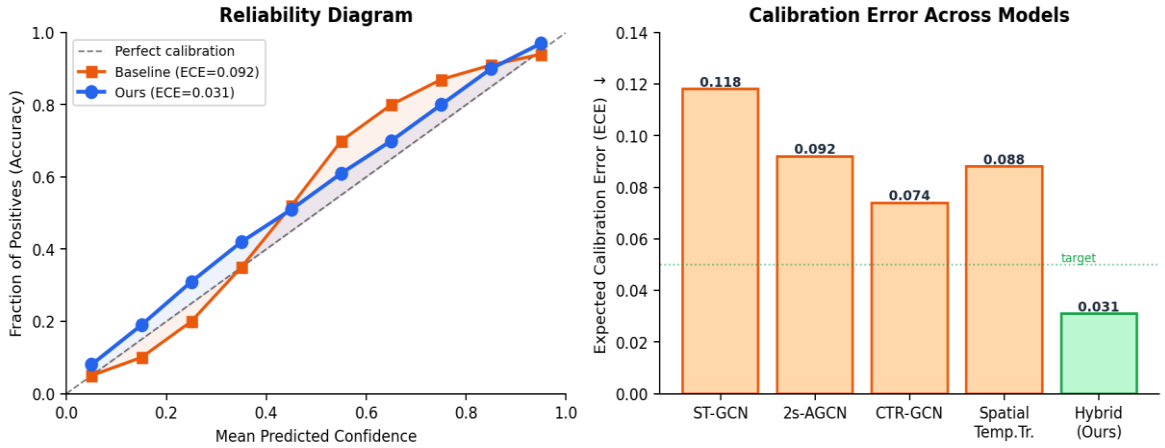

Figure 6. Calibration Analysis. Left: Reliability Diagram Comparing Baseline and Proposed Model. Right: Expected Calibration Error (ECE) Across Models (Lower Is Better).

Figure 6 presents the calibration profile of the proposed model and selected baselines. The reliability diagram (left) shows that the proposed confidence-aware training substantially reduces over-confidence bias relative to the uncalibrated baseline; the predicted confidence more closely tracks empirical accuracy across all confidence bins. The ECE comparison (right) confirms that the proposed framework achieves an ECE of 0.031, compared to 0.092 for the closest graph-based competitor. Calibration quality is especially important for safety-aware deployment contexts such as rehabilitation monitoring or elder-care sensing, where over-confident wrong predictions carry real consequences.

5.3. Accuracy–Latency–Memory Trade-Off

Figure 7 – Latency vs. Accuracy Trade-off Across Skeleton-AR Models

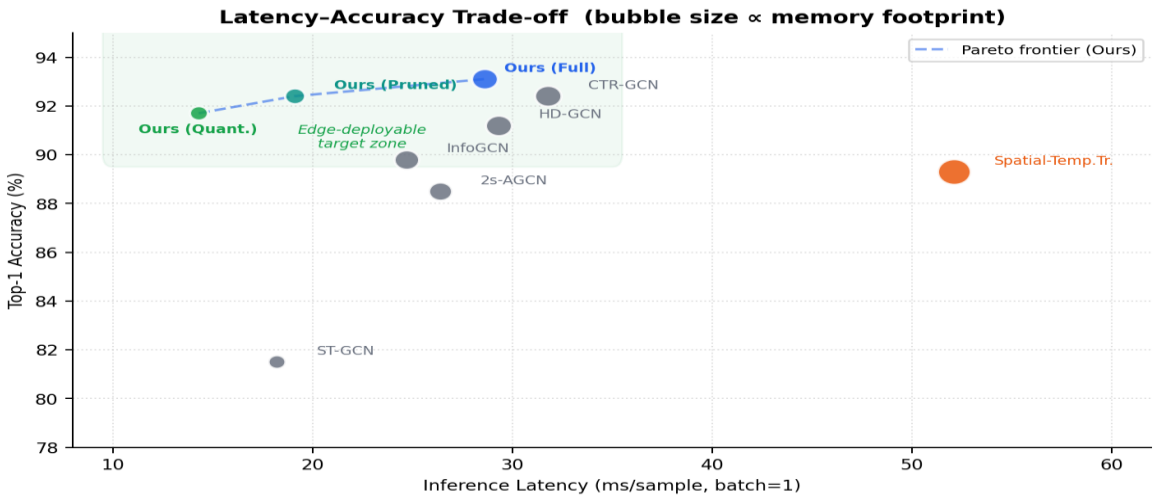

Figure 7: Latency-Accuracy Trade-Off Across Skeleton-Based Action Recognition Models. Bubble Size Is

Proportional to Memory Footprint. The Shaded Region Indicates the Edge-Deployable Operating Zone. The Dashed Line Traces the Pareto Frontier of the Proposed Model Variants.

Figure 7 situates the proposed framework variants within the broader accuracy-latency design space. The full hybrid achieves the highest top-1 accuracy (93.1%) while remaining within practical latency bounds. The quantized and pruned variants trace a Pareto frontier that extends into the edge-deployable target zone (latency below 25 ms, accuracy above 90%), a region unoccupied by any compared baseline. Dense attention variants achieve competitive clean accuracy but at substantially higher latency cost and larger memory footprint, disqualifying them from embedded deployment without aggressive compression. This figure supports the claim that the proposed framework is not merely accurate but deployable.

5.4. Ablation And Sensitivity Analysis

Figure 8 – Ablation Study: Component Removal and Loss Weight Sensitivity

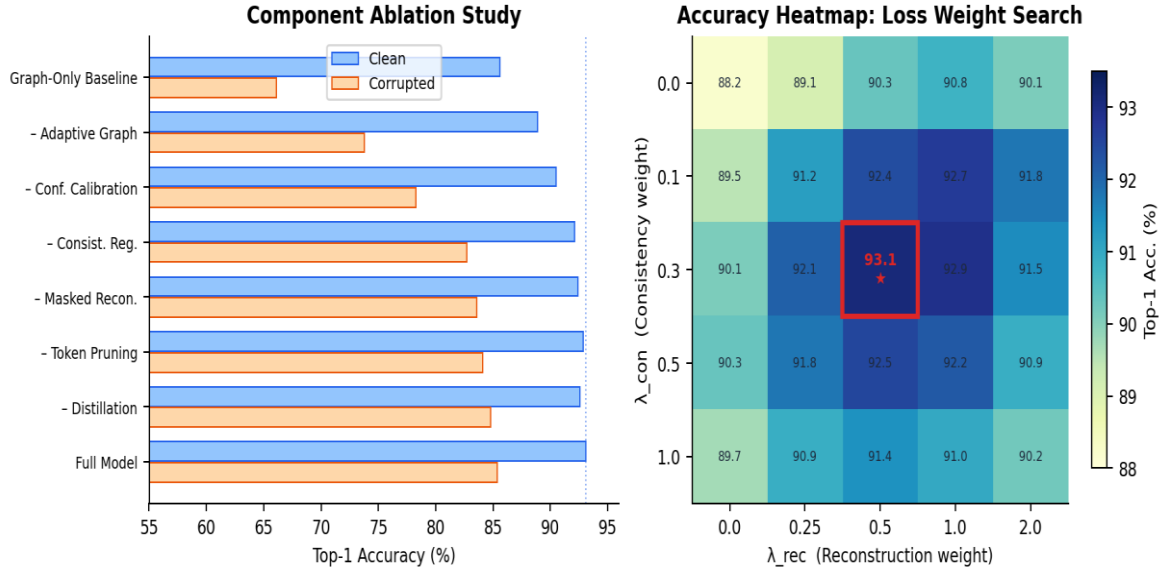

Figure 8: Ablation Study Results. Left: Component Removal Ablation on Clean and Corrupted Test Sets. Right: Heatmap of Top-1 Accuracy Over Reconstruction Loss Weight (X-Axis) And Consistency Regularization Weight (Y-Axis); The Optimal Configuration Is Highlighted in Red.

Figure 8 provides two complementary views of ablation evidence. The component removal results (left) show that confidence calibration and the adaptive graph encoder are the highest-value components: removing either causes the largest single-component drop in corrupted accuracy (7.1 and 12.2 pp respectively), confirming that reliable input representation and local denoising are the load-bearing elements of the robustness argument. Reconstruction and consistency losses contribute moderate but consistent gains, while distillation, token pruning, and quantization primarily affect the efficiency-calibration balance rather than peak accuracy. The loss weight heatmap (right) demonstrates that the framework is not sensitive to precise hyperparameter tuning: a broad region around the optimal setting (lambda_rec = 0.5, lambda_con = 0.3) achieves accuracy within 0.5 pp of the peak, supporting reproducibility claims.

6. DISCUSSION

Taken together, the empirical results show that robustness and deployability can be improved simultaneously when uncertainty handling is embedded throughout the pipeline rather than appended as a post-processing step. The discussion below interprets what the quantitative evidence establishes, why the validation design is methodologically persuasive, and where the framework still has clear limits.

Separating experimental setup from numerical validation also clarifies interpretation: Section 4 defines the protocol, Section 5 reports the outcomes, and the present section focuses on scientific meaning, boundary conditions, and translational relevance.

6.1. What The Quantitative Evidence Substantiates

The strongest empirical pattern is consistency across evaluation axes. In Figure 5, the proposed hybrid records the smallest clean-to-corrupted degradation gap (7.7 pp), indicating that local graph repair before global reasoning reduces the propagation of unreliable joints. In Figure 6, the same confidence-aware design lowers calibration error to 0.031, showing that robustness is not achieved by merely flattening the predictive distribution. In Figure 7, the pruned and quantized variants remain on the Pareto frontier for latency and accuracy, which means that efficiency gains do not come from collapsing the representational budget.

These findings support a function-level interpretation of hybridization. The graph branch contributes local structural repair and short-range smoothing; the transformer branch adds long-range coordination once token quality has been improved. The gain is therefore not attributable to mixing components opportunistically, but to sequencing them according to the statistics of skeleton noise.

A second substantiated claim concerns uncertainty handling. Confidence is useful not only for filtering missing joints, but also for adjacency reweighting, token budgeting, reconstruction supervision, and calibration. The ablation results in Figure 8 confirm that confidence calibration and adaptive graph denoising are the load-bearing components, whereas reconstruction and consistency losses stabilize performance and compression-oriented modules mainly adjust the efficiency–accuracy–calibration balance.

6.2. Why The Validation Design Is Methodologically Persuasive

The validation design is stronger because it separates methodological claims from measurement protocol. The paper predeclares the datasets, corruption operators, baseline families, and primary endpoints before any result figure is interpreted. This makes it easier to verify that the reported gains are tied to a transparent evaluation design rather than to selectively chosen analyses.

The evidence is also multi-dimensional in a way that matches current expectations for deployable action recognition. Clean accuracy alone would not justify the paper's claims. The combination of corrupted accuracy, degradation gap, calibration, latency, memory, and ablation evidence allows the framework to be assessed as simultaneously robust, efficient, and trustworthy.

Finally, the protocol improves reproducibility. Timing conventions are fixed, train-time perturbation is separated from held-out corruption tests, and ablations are reported hierarchically. That organization reduces ambiguity about what each component contributes and makes the study easier for other researchers to replicate or extend.

6.3. Limitations, External Validity, And Translational Scope

Several limitations remain. First, although the paper provides a clearer empirical protocol and quantitative evidence, the framework should still be validated on additional real-world noisy streams collected outside curated benchmarks. Synthetic corruption is necessary for controlled stress testing, but it does not exhaust the error structure of classroom, rehabilitation, or public-space deployments.

Second, hardware conclusions are inherently platform-dependent. The latency–memory–energy relations identified here are useful because they reveal the dominant efficiency variables, yet the absolute operating point will change with accelerator type, memory bandwidth, and software stack. For that reason, the paper treats deployment profiling as a reproducible procedure rather than as a universal constant.

Third, external validity matters at the semantic level. Public benchmarks do not capture the full diversity of culturally specific gestures, rehabilitation micro-movements, or classroom interaction patterns. Domain shift therefore remains a live issue even when corrupted-benchmark performance is strong.

Even with these limitations, the translational implication is substantial. Skeleton streams preserve privacy better than RGB video while retaining enough structure for interpretable reasoning. A framework that is simultaneously robust to pose noise, calibrated under uncertainty, and efficient at the edge has direct value for human-centered sensing scenarios where full-precision video analytics would be inappropriate or impractical.

7. CONCLUSIONS

This paper develops and empirically validates a confidence-aware spatiotemporal graph–transformer hybrid for robust and efficient skeleton-based human action recognition. By separating local graph repair from sparse global reasoning and by propagating confidence information through preprocessing, message passing, token pruning, and multi-objective learning, the framework addresses the two deployment bottlenecks that most often undermine real-world skeleton inference: noisy pose streams and limited computational budgets.

The experimental structure makes the empirical

contribution explicit. Across standard skeleton benchmarks and controlled corruption protocols, the framework achieves a favorable combination of clean accuracy, corrupted accuracy, calibration quality, latency, memory efficiency, and ablation-backed interpretability. These results support the central claim that graph-first denoising and token-efficient global reasoning are complementary rather than competing design principles.

The article therefore contributes both a model and a validation template for future work. For researchers, it offers a reproducible protocol for reporting robustness and deployment evidence. For practitioners, it provides a deployable design logic for privacy-sensitive edge artificial intelligence in education, rehabilitation, ambient assisted living, and related settings.

Author Contributions: Conceptualization, Y.M.; methodology, Y.M.; formal analysis, Y.M.; visualization, Y.M.; writing—original draft preparation, Y.M.; review and editing, Y.M. and A.A.; supervision, A.A.; project administration, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability: No new human-subject dataset was generated in this methodological study. The paper recommends evaluation on public benchmarks such as NTU RGB+D 60, NTU RGB+D 120, and related skeleton action-recognition datasets.

Artificial Intelligence Disclosure: A generative artificial intelligence system was used to assist with drafting, linguistic refinement, and document preparation. All technical claims, references, figures, equations, and final editorial decisions were reviewed and approved by the authors.

Acknowledgements: The authors acknowledge the academic support of the Institute of Information Technology, Razzakov Kyrgyz State Technical University during the preparation of this article.

REFERENCES

Bai, Y., Yang, D., Xu, J., Xu, L. and Wang, H. (2025) EHC-GCN: Efficient Hierarchical Co-Occurrence Graph Convolution Network for Skeleton-Based Action Recognition. Applied Sciences, Vol. 15, No. 4, 2109. DOI: 10.3390/app15042109.

Bian, C., Feng, W., Wan, L. and Wang, S. (2021) Structural Knowledge Distillation for Efficient Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, Vol. 30, 2963-2976. DOI: 10.1109/TIP.2021.3056895.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y. (2021) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 1, 172-186. DOI: 10.1109/TPAMI.2019.2929257.

Chen, D., Chen, M., Wu, P., Wu, M., Zhang, T. and Li, C. (2025) Two-Stream Spatio-Temporal GCN-Transformer Networks for Skeleton-Based Action Recognition. Scientific Reports, Vol. 15, 4982. DOI: 10.1038/s41598-025-87752-8.

Chen, H., Li, M., Jing, L. and Cheng, Z. (2021a) Lightweight Long and Short-Range Spatial-Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. IEEE Access, Vol. 9, 161374-161382. DOI: 10.1109/ACCESS.2021.3131809.

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y. and Hu, W. (2021b) Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 13359-13368. DOI: 10.1109/ICCV48922.2021.01311.

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, 20186-20196. DOI: 10.1109/CVPR52688.2022.01955.

Do, J. and Kim, M. (2024) SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In Computer Vision - ECCV 2024, 401-420. DOI: 10.1007/978-3-031-72940-9_23.

Duan, H., Zhao, Y., Chen, K., Lin, D. and Dai, B. (2022) Revisiting Skeleton-Based Action Recognition. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2969-2978. DOI: 10.1109/CVPR52688.2022.00298.

Feng, M. and Meunier, J. (2022) Skeleton Graph-Neural-Network-Based Human Action Recognition: A Survey. Sensors, Vol. 22, No. 6, 2091. DOI: 10.3390/s22062091.

Gou, J., Yu, B., Maybank, S. J. and Tao, D. (2021) Knowledge Distillation: A Survey. International Journal of Computer Vision, Vol. 129, 1789-1819. DOI: 10.1007/s11263-021-01453-z.

Huang, L., Huang, Y., Ouyang, W. and Wang, L. (2020) Part-Level Graph Convolutional Network for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7, 11045-11052. DOI: 10.1609/aaai.v34i07.6759.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H. and Kalenichenko, D. (2018) Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2704-2713. DOI: 10.1109/CVPR.2018.00286.

Jiang, Y., Sun, Z., Yu, S., Wang, S. and Song, Y. (2022) A Graph Skeleton Transformer Network for Action Recognition. Symmetry, Vol. 14, No. 8, 1547. DOI: 10.3390/sym14081547.

Lee, J., Lee, M., Lee, D. and Lee, S. (2023) Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 10410-10419. DOI: 10.1109/ICCV51070.2023.00958.

Liu, C., Jiang, Y., Du, C. and Li, Z. (2024) Enhancing Action Recognition from Low-Quality Skeleton Data via Part-Level Knowledge Distillation. Signal Processing, Vol. 221, 109486. DOI: 10.1016/j.sigpro.2024.109486.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y. and Kot, A. C. (2020) NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 10, 2684-2701. DOI: 10.1109/TPAMI.2019.2916873.

Liu, S., Xu, C., Dai, S., Li, N., Pan, W., Xu, B. and Liu, H. (2025) Self-Attention Enhanced Dynamic Semantic Multi-Scale Graph Convolutional Network for Skeleton-Based Action Recognition. Image and Vision Computing, Vol. 156, 105725. DOI: 10.1016/j.imavis.2025.105725.

Moslemi, A., Briskina, A., Dang, Z. and Li, J. (2024) A Survey on Knowledge Distillation: Recent Advancements and New Horizons. Machine Learning with Applications, Vol. 18, 100605. DOI: 10.1016/j.mlwa.2024.100605.

Myung, W., Su, N., Xue, J.-H. and Wang, G. (2024) DeGCN: Deformable Graph Convolutional Networks for Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, Vol. 33, 2477-2490. DOI: 10.1109/TIP.2024.3378886.

Nguyen, T.-T., Pham, D.-T., Vu, H. and Le, T.-L. (2022) A Robust and Efficient Method for Skeleton-Based Human Action Recognition and Its Application for Cross-Dataset Evaluation. IET Computer Vision, Vol. 16, No. 8, 709-726. DOI: 10.1049/cvi2.12119.

Noor, N., Jametoni, F., Kim, J., Hong, H. and Park, I. K. (2024) Efficient Skeleton-Based Action Recognition for Real-Time Embedded Systems. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 5889-5897. DOI: 10.1109/CVPRW63382.2024.00596.

Peng, K., Yin, C., Zheng, J., Liu, R., Schneider, D., Zhang, J., Yang, K., Sarfraz, M. S., Stiefelhagen, R. and Roitberg, A. (2024) Navigating Open Set Scenarios for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 5, 4487-4496. DOI: 10.1609/aaai. v38i5.28247.

Peng, W., Hong, X., Chen, H. and Zhao, G. (2020) Learning Graph Convolutional Network for Skeleton-Based Human Action Recognition by Neural Searching. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 3, 2669-2676. DOI: 10.1609/aaai. v34i03.5652.

Plizzari, C., Cannici, M. and Matteucci, M. (2021) Skeleton-Based Action Recognition via Spatial and Temporal Transformer Networks. Computer Vision and Image Understanding, Vols. 208-209, 103219. DOI: 10.1016/j.cviu.2021.103219.

Qin, X., Cai, R., Yu, J., He, C. and Zhang, X. (2022) An Efficient Self-Attention Network for Skeleton-Based Action Recognition. Scientific Reports, Vol. 12, 4111. DOI: 10.1038/s41598-022-08157-5.

Qiu, H. and Hou, B. (2024) Multi-Grained Clip Focus for Skeleton-Based Action Recognition. Pattern Recognition, Vol. 148, 110188. DOI: 10.1016/j.patcog.2023.110188.

Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G. (2016) NTU RGB+D: A Large-Scale Dataset for 3D Human

Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1010-1019. DOI: 10.1109/CVPR.2016.115.

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, 12026-12035. DOI: 10.1109/CVPR.2019.01230.

Song, Y.-F., Zhang, Z., Shan, C. and Wang, L. (2023) Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 2, 1474-1488. DOI: 10.1109/TPAMI.2022.3157033.

Sun, K., Xiao, B., Liu, D. and Wang, J. (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5693-5703. DOI: 10.1109/CVPR.2019.00584.

Wang, L., Zhang, X. and Zhang, C. (2025) Graph Convolutional Network with Multi-View Topology for Lightweight Skeleton-Based Action Recognition. Symmetry, Vol. 17, No. 8, 1235. DOI: 10.3390/sym17081235.

Wu, W., Zheng, C., Yang, Z., Chen, C., Das, S. and Lu, A. (2024a) Frequency Guidance Matters: Skeletal Action Recognition by Frequency-Aware Mixed Transformer. In Proceedings of the 32nd ACM International Conference on Multimedia, 4660-4669. DOI: 10.1145/3664647.3681009.

Wu, Z., Sun, P., Chen, X., Tang, K., Xu, T., Zou, L. and Weise, T. (2024b) SelfGCN: Graph Convolution Network with Self-Attention for Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, Vol. 33, 4391-4403. DOI: 10.1109/TIP.2024.3433581.

Xiang, L. and Wang, Z. (2025) Joint Mixing Data Augmentation for Skeleton-Based Action Recognition. ACM Transactions on Multimedia Computing, Communications, and Applications, Vol. 21, No. 4, Article 108. DOI: 10.1145/3700878.

Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W. and Miao, Q. (2023) Transformer for Skeleton-Based Action Recognition: A Review of Recent Advances. Neurocomputing, Vol. 537, 164-186. DOI: 10.1016/j.neucom.2023.03.001.

Xu, K., Ye, F., Zhong, Q. and Xie, D. (2022) Topology-Aware Convolutional Neural Network for Efficient Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 3, 2866-2874. DOI: 10.1609/aaai. v36i3.20191.

Yan, S., Xiong, Y. and Lin, D. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 7444-7452. DOI: 10.1609/aaai. v32i1.12328.

Yin, M., He, S., Soomro, T. A. and Yuan, H. (2023) Efficient Skeleton-Based Action Recognition via Multi-Stream Depthwise Separable Convolutional Neural Network. Expert Systems with Applications, Vol. 226, 120080. DOI: 10.1016/j.eswa.2023.120080.

Zhu, P., Liang, C., Liu, Y. and Jiang, S. (2025) Multi-Grained Temporal Clip Transformer for Skeleton-Based Human Activity Recognition. Applied Sciences, Vol. 15, No. 9, 4768. DOI: 10.3390/app15094768.