SPOOQ: ANATOMY-GUIDED QUANTUM SKELETON ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum machine learning (QML) promises compact models that capture correlations costly to represent classically, yet has not been demonstrated on complex spatio-temporal problems such as skeleton-based action recognition. We introduce SpooQ, to our knowledge the first quantum machine learning architecture for skeleton-based action recognition, designed for near-term quantum hardware by encoding human anatomy directly into its circuit structure. On the NW-UCLA and a single-person subset of NTU RGB+D 60 cross-view benchmarks, SpooQ outperforms a parameter-matched state-of-the-art graph convolutional network by 7.0 and 3.5 accuracy points, reaching 84.3% and 68.5%, respectively. Under 1024-shot simulations using noise models derived from the IBM Brisbane and Prague backends, SpooQ retains approximately 80% accuracy and outperforms a flat quantum baseline by 27–38 accuracy points. These results extend QML beyond static benchmarks to a demanding spatio-temporal task and demonstrate robustness under finite-shot, backend-derived noise.

1 INTRODUCTION

Quantum machine learning (QML) is increasingly being explored beyond small benchmark classifiers, with recent work extending hybrid quantum models into language, vision, medical imaging, and temporal sequence modeling Kim et al. (2025); Kong et al. (2025); Aktar et al. (2025); Zhang et al. (2025); Wang & Shang (2025); Li et al. (2025); Park et al. (2026). This growing interest is motivated by a central theoretical appeal: parameterized quantum circuits can encode data into high-dimensional Hilbert spaces and use quantum properties of entanglement, interference, and measurement to expose compact representations of correlations Havlíček et al. (2019); Schuld et al. (2021); Abbas et al. (2021); Cerezo et al. (2021). These properties make QML especially appealing for domains where the relevant signal is distributed across interacting components, including language tokens, graph nodes, and human skeletal joints. Skeleton-based action recognition is a natural instance of this setting: its input is both temporal and anatomical, with joints linked by the human kinematic graph, grouped into functional body regions, and interpreted through coordinated motion over time. Existing quantum temporal models, however, typically apply shared quantum modules to generic temporal windows Park et al. (2026), leaving open how to design quantum architectures that reflect the body's anatomical structure. Realizing such an architecture on near-term hardware is not straightforward. Noisy intermediate-scale quantum (NISQ) devices offer few qubits, shallow feasible circuit depths, and noisy gate operations, so flattening an entire pose into one monolithic circuit both squanders scarce coherent capacity and exposes every encoded feature to accumulated error. The challenge is allocating limited quantum capacity across structured spatio-temporal input.

We introduce SpooQ, an anatomy-guided quantum architecture that answers this question by using the structure of the body as a circuit design principle. Instead of encoding an entire pose or temporal window into a single monolithic quantum circuit, SpooQ decomposes the skeleton into anatomically meaningful regions and assigns compact quantum circuits at three scales: local circuits model body-part configurations, relational circuits compose coordinated regions such as the arms, legs, and upper body, and a global circuit preserves whole-pose information. This hierarchy allocates limited quantum capacity where skeleton data is naturally organized. To make the composition motion-aware, SpooQ interleaves region-preserving temporal mixing between the local and relational levels, so higher-level circuits operate on short-range dynamics rather than static poses. Each circuit is measured through both single-qubit and pairwise observables, allowing SpooQ to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

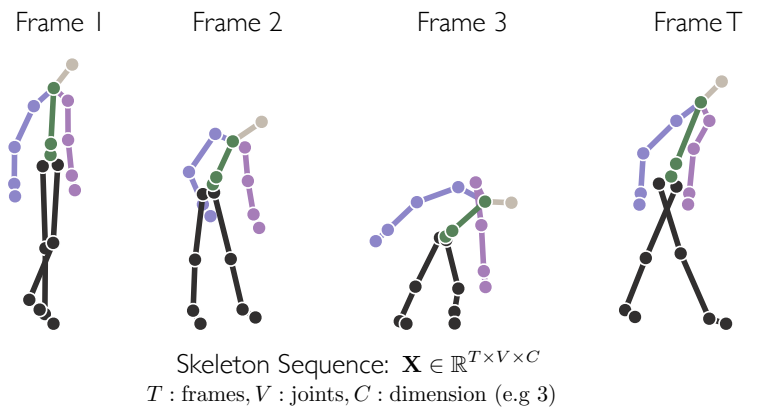

Figure 1: Spatio-Temporal Skeleton for action "pickup".

retain local features while exposing interactions among encoded body components before temporal aggregation. This makes SpooQ a step toward QML architectures that are not only applied to structured data but explicitly shaped by it.

Contributions.

I. To the best of our knowledge, SpooQ is the first anatomy-guided quantum architecture for skeleton-based action recognition, organizing quantum computation around the kinematic structure of the body rather than flattened poses or generic temporal windows.

II. A structure-guided circuit design for NISQ constraints: compact variational circuits assigned across a three-level anatomical hierarchy and read out through single-qubit and pairwise ZZ observables, which expose marginal responses and pairwise joint statistics as classical features.

III. An empirical study showing that SpooQ outperforms flat-quantum, parameter-matched classical, and parameter-matched graph-convolutional baselines, with ablations attributing the largest gains to the quantum circuits and their pairwise readout, and finite-shot IBM backend-derived noise simulations demonstrating greater robustness than a monolithic quantum encoder.

2 Background and Related Work

2.1 Quantum Machine Learning Overview

The basic unit of quantum computation is the *qubit*. Unlike a classical bit, which is either 0 or 1, a qubit can occupy a *superposition*: a weighted combination of both values at once, described by two complex amplitudes whose squared magnitudes give the probability of observing each value. A system of n qubits is likewise described by amplitudes over all 2^n joint bit configurations, so its state occupies a space that grows exponentially with n. Quantum *gates* are operations that transform these amplitudes; gates acting on two or more qubits can *entangle* them, meaning the joint state can no longer be described qubit-by-qubit—information becomes encoded in the correlations among qubits rather than in any qubit alone. Reading information out requires *measurement*, which converts the quantum state back into classical numbers; repeated measurement of an observable quantity yields its *expectation value*, a real number summarizing the state.

In quantum machine learning, a classical input vector is first *encoded* into a quantum state, transformed by a *variational quantum circuit* (VQC), a sequence of gates with trainable rotation parameters, and converted back into classical features by measurement; the parameters are optimized against a standard classical loss, yielding a hybrid quantum–classical system Cerezo et al. (2021). Figure 3 in Appendix B.2 illustrates an example of this pipeline, applied to temporal data.

A common encoding is *amplitude encoding*, which writes a normalized d-dimensional vector into the amplitudes of an $\lceil \log_2 d \rceil$-qubit state, so circuit width scales logarithmically with input dimension. For readout, a single-qubit expectation $\langle Z_i \rangle \in [-1, 1]$ summarizes the state of qubit i, while a pairwise expectation $\langle Z_i Z_j \rangle$ reflects the correlation between qubits i and j created by entangling gates. A VQC with such a readout acts as a trainable feature map whose expressivity derives from entanglement and interference Schuld et al. (2021); Havlíček et al. (2019).

These models must run under noisy intermediate-scale quantum (NISQ) constraints Preskill (2018): few qubits, shallow circuits, and gate errors that compound with circuit width and depth Silver et al. (2022). Encoding a large structured input into one monolithic circuit is therefore costly and fragile; a more practical strategy is to apply small circuits to selected input substructures, preserving task-relevant organization while keeping each quantum computation compact Li et al. (2021).

2.2 Skeleton-Based Action Recognition

Skeleton-based action recognition classifies human activities from sequences of body-joint coordinates. Compared with raw RGB video, skeletons are a compact abstraction of motion, robust to background, lighting, and appearance Shahroudy et al. (2016). Their low native dimensionality (tens of joints, rather than millions of pixels) makes them a natural vision modality for quantum encoding under NISQ constraints. We denote a skeleton sequence as $\mathbf{X} \in \mathbb{R}^{T \times V \times C}$ (Fig. 1), where T is the number of frames, V the number of joints, and C the coordinate dimension; for 3D skeletons, $C = 3$ and each joint observation is $\mathbf{x}_{t,v} \in \mathbb{R}^3$.

Skeleton data is organized at several interacting scales. The human kinematic graph links joints into anatomical parts, which coordinate over time to produce actions such as waving, kicking, or sitting down. Recognizing these actions therefore depends on local joint configurations, interactions between limbs, and whole-body motion. Classical methods exploit this organization through spatial-temporal graph convolutional networks that model the kinematic graph directly Yan et al. (2018). Subsequent work builds on this approach by refining joint dependencies and multi-scale receptive fields Shi et al. (2019); Chen et al. (2021), while part-based models learn representations by first separating body regions and then composing their interactions Thakkar & Narayanan (2018); Lee et al. (2023). This same organization makes skeleton data well suited to NISQ-era quantum models, as it provides a natural decomposition into compact, semantically meaningful substructures. Yet, to our knowledge, existing quantum architectures have not exploited this anatomical structure.

3 Method

SpooQ organizes quantum computation around the anatomical structure of the human body, the same structure classical skeleton models exploit through graph topology and part-based decomposition Lee et al. (2023); Yan et al. (2018); Chen et al. (2021). Rather than encoding each pose into one monolithic circuit, SpooQ assigns compact circuits at three anatomical scales (Fig. 2): *part circuits* encode six local body regions, *region circuits* compose them into coordinated functional groups, and a parallel *body circuit* preserves whole-pose context that decomposition might weaken. Information moves through this hierarchy along two paths: anatomy-preserving temporal mixing between the part and region levels lets regional composition operate on short-range motion rather than static poses, while a skip pathway carries part-level features past regional composition directly into the final fusion. Within each circuit, a relational measurement scheme augments single-qubit responses with pairwise joint statistics before classical fusion. A recurrent classifier aggregates the fused representation over time; the remainder of this section develops each component, beginning with the flat quantum baseline whose limitations motivate the design.

3.1 Flat Quantum Baseline

We first define a flat quantum skeleton model, referred to throughout this work as the *Quantum Baseline*. The model adapts recent hybrid quantum temporal architectures such as HQTCN Park et al. (2026), which apply a shared quantum circuit across generic temporal windows and aggregate the resulting outputs across time, to skeleton data: the full-body pose at each frame is treated as a single coordinate vector and processed by one variational quantum circuit (Fig. 3 in Appendix B.2).

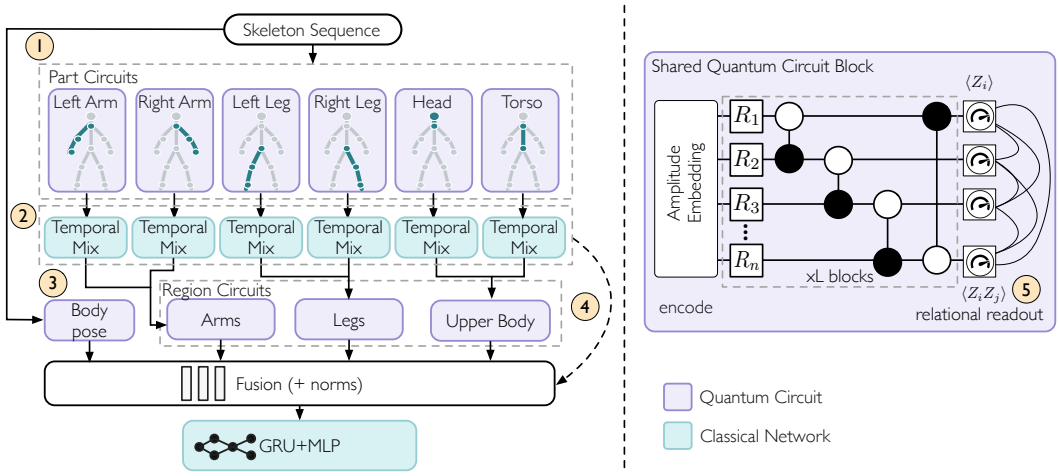

Figure 2: The SpooQ architecture. Each skeleton frame is grouped into six anatomical parts, encoded by dedicated part circuits (1), and temporally mixed within each part stream over a short frame window (2); region circuits then compose paired limbs and the upper body, while a parallel body circuit encodes the raw full pose (3) and a skip pathway (4) carries part features directly to fusion. The fused features and retained input norms are classified by a GRU. Right: all circuits share one template, which amplitude-encodes its input, applies L variational blocks of single-qubit rotations (R, the $R_Z R_Y R_Z$ block of Eq. 4) and a ring of CNOT gates, and measures every single-qubit expectation $\langle Z_i \rangle$ together with every pairwise expectation $\langle Z_i Z_j \rangle$. The latter form the *relational readout* (5), which exposes pairwise joint statistics after circuit evolution as classical features.

Formally, for an input skeleton sequence $X \in \mathbb{R}^{T \times V \times C}$, the complete pose at frame t is flattened into $x_t = \mathrm{vec}(X_{t,:,:}) \in \mathbb{R}^{VC}$ and a shared circuit Q_θ produces a frame-level representation $z_t = Q_\theta(x_t), \qquad t = 1, \ldots, T.$

The sequence $Z = (z_1, \ldots, z_T)$ is passed to a classical recurrent module that aggregates the quantum-encoded poses into the final action representation.

The flat quantum baseline processes each pose as a single vector, but does not explicitly use the skeleton's anatomical structure. In particular, it lacks the grouping of joints into body parts that part-based and hierarchically decomposed classical models exploit to improve representation learning Thakkar & Narayanan (2018); Lee et al. (2023). Local limb configurations, coordination between body regions, and whole-body pose must therefore all be represented within the same compact circuit. Because each frame is processed independently, the quantum encoder also has no direct access to motion information. Temporal structure enters only afterward, through the recurrent module. SpooQ addresses both limitations by replacing the single full-pose encoder with an anatomy-aware hierarchy of compact circuits and interleaving temporal mixing between the part and region levels.

3.2 Anatomy-Aware Circuit Hierarchy

The first level of SpooQ's hierarchy aligns quantum processing with the anatomical organization of the skeleton. Rather than assigning the complete pose to a single circuit, SpooQ groups the joints into six anatomically coherent groups (the left and right arms, left and right legs, head, and torso) and pairs each with a compact *part circuit*. Formally, let $\mathcal{P} = \{P_1, \ldots, P_K\}$ denote the anatomical joint groups, $P_k \subseteq \{0, \ldots, V-1\}$. For frame t, the coordinates of group P_k are extracted as $x_{t,k} = \mathrm{vec}(X_{t,P_k,:})$ and processed by a part circuit $Q_k^{(P)}$ with parameters shared across time:

$$h_{t,k}^{(P)} = Q_k^{(P)}(x_{t,k})\,. \tag{1}$$

Each circuit receives only one region's joints, so its output retains that regional correspondence.

Static part configurations, however, do not describe how a body part moves. Before composing the part-level outputs, SpooQ therefore applies a short-range temporal mixing operation to each

anatomical feature stream,

$$\tilde{h}_{t,k}^{(P)} = M_k\Big(h_{t-r,k}^{(P)}, \ldots, h_{t+r,k}^{(P)}\Big), \tag{2}$$

where M_k consists of a depthwise temporal convolution, which lets each feature channel observe short motion patterns over nearby frames, followed by a lightweight channel-mixing layer that recombines the resulting channels within the stream. The entire operation is *anatomy-preserving*: features are mixed over time and across channels only within their own body-part stream, never across part boundaries, so motion information enters the hierarchy without blurring the anatomical separation established by the part circuits.

The motion-aware part representations are then composed by *region circuits*. Let $\mathcal{R}$ denote the set of functional relations between parts; for each relation $r = (i, j) \in \mathcal{R}$, a region circuit combines the corresponding part features:

$$h_{t,r}^{(R)} = Q_r^{(R)}\Big(\Big[\tilde{h}_{t,i}^{(P)};\, \tilde{h}_{t,j}^{(P)}\Big]\Big). \tag{3}$$

Concretely, an arms circuit composes the two arm streams, a legs circuit the two leg streams, and an upper-body circuit the head and torso streams. The hierarchy thus progresses from local anatomical encoding, through short-range motion, to coordinated regional composition.

3.3 Body Circuit

Decomposing the skeleton into parts and regions makes local structure explicit, but no single circuit in the hierarchy observes the pose as a whole. SpooQ therefore includes a parallel *body circuit* that encodes the complete pose at each frame. The full joint set is flattened into $x_t^{(B)} = \mathrm{vec}(X_{t,:,:}) \in \mathbb{R}^{VC}$ and processed by a body circuit $Q^{(B)}$: $g_t = Q^{(B)}(x_t^{(B)})$. Unlike the region circuits, which operate on features composed from related parts, the body circuit observes raw whole-pose coordinates directly. Its output preserves coarse full-body configuration, arm and leg placement relative to the torso, overall posture, and body orientation, complementing the anatomically structured features of the hierarchical pathway. The body representation joins the part and region features in the final fusion. All three pathways share a single circuit construction, which we describe next.

3.4 Circuit Construction and Relational Readout

We now describe the circuit design shared by the part, region, and body pathways, including how inputs are encoded and features are read out. For an input vector $u \in \mathbb{R}^d$, the circuit uses $n = \lceil \log_2 d \rceil$ qubits. We zero-pad u to dimension 2^n and normalize it to prepare an amplitude-encoded state. Since normalization removes magnitude information, circuits that take joint coordinates as input also retain the original norm as a classical feature.

After state preparation, the circuit applies L variational blocks, each consisting of trainable rotations R_Z, R_Y, R_Z, on every qubit followed by a ring U_{ent} of CNOT entangling gates:

$$U_\theta = \prod_{\ell=1}^{L}\left[U_{\mathrm{ent}} \prod_{i=1}^{n} R_Z(\theta_{\ell,i,1}) R_Y(\theta_{\ell,i,2}) R_Z(\theta_{\ell,i,3})\right]. \tag{4}$$

This shared construction lets the qubit count adapt to the dimensionality of each anatomical input while keeping a consistent design across the hierarchy.

Single-qubit expectations $\langle Z_i \rangle$ summarize the marginal response of each qubit, while pairwise expectations $\langle Z_i Z_j \rangle$ capture second-order joint statistics across qubits. SpooQ therefore measures both:

$$q(u) = [\, \langle Z_i \rangle \mid i = 1, \ldots, n\,;\ \langle Z_i Z_j \rangle \mid 1 \le i < j \le n \,]. \tag{5}$$

We call the pairwise component the *relational readout*: it exposes dependencies between qubit responses after variational and entangling evolution as classical features.

Each boundary of the circuit discards something by default, and SpooQ compensates at both: encoding discards magnitude, so norms are retained classically; measurement of independent observables discards correlations, so they are read out explicitly.

3.5 Fusion and Temporal Classification

At each frame, SpooQ holds representations at three anatomical scales: part features $\tilde{h}_{t,k}^{(P)}$, region features $h_{t,r}^{(R)}$, and the body feature g_t. Alongside their use in regional composition, the part features pass directly to fusion through a *part skip pathway*. This gives the classifier access to both fine-grained local information and the regional features derived from it. The frame representation is

$$f_t = \mathrm{Concat}\left(\left\{ \tilde{h}_{t,k}^{(P)} \right\}_{k=1}^{K}, \left\{ h_{t,r}^{(R)} \right\}_{r \in \mathcal{R}}, g_t,\ s_t \right), \tag{6}$$

where s_t collects the retained norms of the raw part and body coordinate vectors. Through this fusion the classifier jointly observes local body-part structure, inter-region coordination, and whole-body context.

The sequence of fused representations $F = (f_1, \ldots, f_T)$ is aggregated by a gated recurrent unit, and the final hidden state is mapped to class logits by an MLP :$H = \mathrm{GRU}(F)$, $\hat{y} = \mathrm{MLP}(H_T)$. The circuits, temporal mixing, and classifier are optimized jointly with the cross-entropy objective.

SpooQ distributes the modeling of spatial and temporal structure across its part, region, and body pathways, so no single circuit has to capture all of this information.

4 Experimental Setup

Datasets and preprocessing. We evaluate SpooQ using data from two established skeleton-based action-recognition benchmarks selected to probe complementary regimes: NW-UCLA evaluates cross-view recognition in a small-data setting, while the single-person-filtered NTU60 split tests scaling to a substantially larger and more diverse 60-class corpus under a consistent single-skeleton input graph. Northwestern-UCLA Multiview Action3D (NW-UCLA) Wang et al. (2014) contains 1,494 sequences spanning 10 action classes performed by 10 subjects and captured from three viewpoints; following the standard cross-view protocol, we train on camera views 1 and 2 and evaluate on view 3. Each skeleton comprises 20 three-dimensional joints, yielding a 60-dimensional full-pose vector per frame. NTU RGB+D 60 (NTU60) Shahroudy et al. (2016) covers 60 action classes captured from three cameras with 25 joints per skeleton; we adopt the cross-view (XView) protocol, training on cameras 2 and 3 and evaluating on camera 1.

Because SpooQ's anatomical decomposition is defined over a single kinematic graph, we restrict training and evaluation to clips containing exactly one valid tracked skeleton; multi-person clips are removed rather than merged or averaged, avoiding artificial hybrid poses. All 60 action classes are retained—the mutual-action classes (A50–A60) survive through clips in which only a single skeleton was tracked—leaving 30,683 training and 15,296 evaluation clips in the filtered XView split; the full split composition and exact clip lists are provided in the supplementary material. Our NTU60 results are therefore not comparable to published XView numbers and we make no state-of-the-art claim on NTU60; all methods are trained and evaluated under the identical filtered protocol. We sample 64 frames for NTU60 and 52 for NW-UCLA, without temporal windowing. Quantum inputs are ℓ_2-normalized and zero-padded to a power of two; their pre-normalization norms are retained as classical features.

Implementation. All quantum circuits are implemented in PennyLane (Bergholm et al., 2018) and trained using analytic simulation (`default.qubit`). Every circuit uses $L = 4$ variational blocks; the quantum circuits contribute 516 and 552 trainable parameters on NW-UCLA and NTU60, respectively. Exact circuit dimensions, temporal mixing, and classifier specifications are given in Appendices A and B. For noise evaluation, we reuse the three trained NW-UCLA checkpoints without retraining in 1024-shot Qiskit Aer simulations (Javadi-Abhari et al., 2024) with noise models derived from the IBM Brisbane and Prague backends; the noise configuration is detailed in Appendix C. Code, filtered-split clip lists, anatomical group definitions, and all training configurations are included in the supplementary material.

Baselines. We compare SpooQ against three baselines chosen to separate the contributions of quantum processing, anatomical structure, and classical graph modeling.

Table 1: Accuracy on NW-UCLA cross-view and the single-person-filtered NTU60 XView split (60 classes; see Datasets). Mean ± std over three seeds; best per column in bold. Classical SpooQ and CTR-GCN are parameter-matched to SpooQ; the Quantum Baseline is smaller.

Method	NW-UCLA	NTU60
Quantum Baseline	70.9 ± 1.5	61.9 ± 0.6
CTR-GCN	77.3 ± 0.2	65.0 ± 1.2
Classical SpooQ	79.0 ± 0.4	66.6 ± 0.2
SpooQ	$\mathbf{84.3 \pm 1.4}$	$\mathbf{68.5 \pm 1.0}$

Quantum Baseline. The flat quantum model described earlier: the full pose at each frame is amplitude-encoded into a single shared variational circuit of depth 4 (7 qubits for the 75-dimensional NTU60 pose), followed by a GRU with hidden size 22 and an MLP classifier with hidden width 64. It uses the same variational circuit template as SpooQ, but a different classifier configuration and fewer total parameters (5,230 and 8,954 on NW-UCLA and NTU60); the comparison is therefore not parameter-matched.

Classical SpooQ. A parameter-matched classical counterpart of SpooQ (within 2% in total trainable parameters) that retains the full anatomical hierarchy, interleaved temporal mixing, and GRU classifier, but replaces each quantum circuit with a classical MLP with matching input and output dimensions. This tests whether SpooQ's gains derive from the hierarchy and temporal design alone.

CTR-GCN. A state-of-the-art classical graph convolutional architecture Chen et al. (2021) scaled to SpooQ's parameter budget by architectural knobs rather than naive truncation: channel width and stage count are reduced (e.g., base width 5 with five stages on NTU60) while preserving CTR-GCN's characteristic interleaving of graph-spatial and temporal-convolutional blocks. For context, full-scale CTR-GCN ($\sim$1.4M parameters) reports 96.5% on NW-UCLA and 96.8% on NTU60 XView under the unfiltered protocol with four-stream ensembling Chen et al. (2021); our comparison probes the low-parameter regime relevant to NISQ-scale models and makes no state-of-the-art claim. At similar trainable-parameter counts, SpooQ, Classical SpooQ, and CTR-GCN use 13,380, 13,593, and 13,449 parameters on NW-UCLA, and 16,610, 16,912, and 15,781 on NTU60.

Training and evaluation. SpooQ, the Quantum Baseline, and Classical SpooQ use AdamW; CTR-GCN uses SGD with Nesterov momentum. All models train for 65 epochs with identical augmentation. Unless otherwise noted, results report mean and standard deviation of test accuracy over three random seeds. Full training details appear in Appendix E.

5 Results

5.1 Main Results

Across both datasets, SpooQ achieves the best accuracy among the evaluated models, surpassing the flat Quantum Baseline, a parameter-matched CTR-GCN, and its parameter-matched classical counterpart, and reaching 84.3% on NW-UCLA and 68.5% on the filtered NTU60 cross-view split (Table 1). These gains separate three effects. Against the flat quantum encoder, SpooQ's anatomy-aware hierarchy improves accuracy by 13.4 points on NW-UCLA and 6.6 points on NTU60: how quantum capacity is allocated across the skeleton matters more than the presence of quantum circuits alone. Against CTR-GCN, SpooQ gains 7.0 and 3.5 points at the same parameter budget. Finally, Classical SpooQ, which retains the full anatomical hierarchy but replaces every circuit with a parameter-matched classical block, reaches 79.0% and 66.6%: the anatomical design is itself a strong architecture, exceeding CTR-GCN on both datasets, and the quantum circuits add a further 5.3 and 1.9 points on top of it. Because CTR-GCN is a strong graph-convolutional architecture purpose-built for skeleton data, and Classical SpooQ shares SpooQ's every component except the circuits, these results together suggest that compact quantum feature maps, when aligned with anatomical structure, extract discriminative information that parameter-matched classical models miss at this budget.

Table 2: Component ablation on NW-UCLA (cross-view). Mean $\pm$ std over three seeds; Δ is the change from full SpooQ. The w/o-pairwise ablation shrinks the fused dimension from 124 to 50; the classical product readout restores it to 124 by appending products $\langle Z_i \rangle \langle Z_j \rangle$.

Configuration	Accuracy (%)	Δ
SpooQ	$\mathbf{84.3 \pm 1.4}$	—
w/o pairwise readout	78.0 ± 2.6	-6.3
classical product readout	82.1 ± 0.8	-2.2
w/o temporal mixing	83.8 ± 0.8	-0.5
w/o channel mixing	83.7 ± 1.8	-0.6
w/o body circuit	84.0 ± 1.8	-0.3
w/o norm features	84.0 ± 1.2	-0.3

Notably, SpooQ's margins are largest on NW-UCLA, whose training set is more than an order of magnitude smaller than NTU60: the gain over Classical SpooQ is 5.3 points on NW-UCLA versus 1.9 on NTU60. This pattern is consistent with theoretical results on quantum learning models, which are predicted to generalize from few training samples due to their favorable capacity per trainable parameter Caro et al. (2022); Abbas et al. (2021).

5.2 Component Ablation

Table 2 examines the contributions of SpooQ's components on NW-UCLA. The largest effects arise from the circuit-based feature maps and the relational readout. First, replacing every quantum circuit with a parameter-matched classical block, while keeping the anatomical hierarchy, temporal mixing, fusion, and classifier unchanged, reduces accuracy by 5.3 points (Classical SpooQ; see also Table 1, where the same model is evaluated on both datasets). This difference is notable given the small parameter footprint of the substituted modules: the quantum circuits account for only 516 of SpooQ's $13{,}380$ trainable parameters, or less than 4% of the model. The classical counterpart preserves the circuit input and output dimensions and all downstream components, while using a matched total parameter budget (13,593 versus 13,380). This controlled substitution therefore indicates that the circuit-based feature maps contribute materially beyond the surrounding anatomical architecture, relative to the classical replacements considered here. Second, removing the pairwise measurements reduces accuracy by 6.3 points, but also shrinks the fused representation from 124 to 50 dimensions, conflating the information carried by the pairwise correlators with the effect of reduced feature width. We therefore introduce a width-matched control that replaces each measured $\langle Z_i Z_j \rangle$ with the factorized product $\langle Z_i \rangle \langle Z_j \rangle$. This preserves the feature width, multiplicative structure, parameter count (13,380), and all downstream components, while removing the connected component $\langle Z_i Z_j \rangle - \langle Z_i \rangle \langle Z_j \rangle$ of the measured correlator. The control recovers accuracy to 82.1 ± 0.8, but remains 2.2 points below full SpooQ. Thus, approximately 4.1 points of the original reduction are recovered by restoring second-order feature capacity, while the remaining 2.2-point gap is associated with measured joint expectations that are not captured by their factorized products.

The remaining components make smaller, complementary contributions. Removing the full per-part temporal block reduces accuracy by 0.5 points, removing only its channel mixer by 0.6, removing the body circuit by 0.3, and removing the retained norm features by 0.3. Because these differences lie within the observed seed-level variability (± 0.8–1.8), we do not treat them as established individual effects. Rather, they appear to provide modest and partially overlapping forms of context: short-range motion, cross-channel interaction, whole-body configuration, and coordinate magnitude. Their combined presence yields the highest mean accuracy, but none individually accounts for the gap between SpooQ and its classical counterpart.

Taken together, the circuit-substitution and readout controls provide the strongest evidence for the proposed design. Replacing the circuits with matched classical blocks reduces accuracy by 5.3 points, while the width-matched readout control shows that second-order feature capacity explains only part of the benefit of the pairwise measurements. These results suggest that SpooQ benefits from both its anatomical hierarchy and the circuit-based relational features.

Table 3: NW-UCLA accuracy under 1024-shot IBM backend-derived noise simulation. Accuracy is mean $\pm$ std over three seeds; Δ is SpooQ minus the Quantum Baseline.

Evaluation setting	Quantum Baseline	**SpooQ**	Δ
Analytic, noiseless	70.9 ± 1.5	$\mathbf{84.3 \pm 1.4}$	$+13.4$
Brisbane-derived	42.6 ± 4.3	$\mathbf{80.1 \pm 2.6}$	$+37.5$
Prague-derived	53.2 ± 4.5	$\mathbf{80.4 \pm 2.5}$	$+27.2$

5.3 Robustness under IBM Backend Noise

Table 3 evaluates the three trained NW-UCLA checkpoints without noisy retraining using 1024-shot Qiskit Aer simulations with IBM Brisbane- and Prague-derived noise models. Each checkpoint is tested on the complete 463-clip evaluation split using optimization-level-2 transpilation. Amplitude initialization is decomposed before execution, exposing both state-preparation and variational gates to the noise model. Simulations use packed logical circuits of at most seven qubits, without full-device topology mapping, and are not physical-hardware runs. SpooQ retains $80.1 \pm 2.6\%$ and $80.4 \pm 2.5\%$ accuracy under the Brisbane- and Prague-derived models, only 4.2 and 3.9 points below its noiseless result. The Quantum Baseline falls to $42.6 \pm 4.3\%$ and $53.2 \pm 4.5\%$, corresponding to drops of 28.3 and 17.7 points. Consequently, SpooQ's advantage over the flat quantum encoder grows from 13.4 points under noiseless simulation to 37.5 and 27.2 points, respectively.

This pattern is consistent with SpooQ's architectural motivation. The Quantum Baseline concentrates feature extraction in a single full-pose circuit, so noise affects the entire frame representation. In contrast, SpooQ distributes computation across compact part-, region-, and body-level circuits whose outputs are combined through classical fusion. This decomposition limits degradation within anatomical streams while preserving complementary pose representations, helping SpooQ retain more of its noiseless accuracy.

6 Conclusion

We introduced SpooQ, to our knowledge the first quantum machine learning architecture for skeleton-based action recognition. SpooQ's central idea is to treat human anatomy as a circuit design principle: rather than encoding poses into a single monolithic circuit, it assigns compact variational circuits to local body parts, functional regions, and global pose context, and exposes pairwise joint statistics among encoded features through relational measurements.

Our experimental results provide supporting evidence for this design. SpooQ outperforms the compact flat quantum baseline and parameter-matched classical and graph-convolutional baselines on NW-UCLA and NTU60. Ablations further indicate that the improvements cannot be explained by the anatomy-aware hierarchy alone. Replacing the variational quantum circuits with parameter-matched classical counterparts while keeping the hierarchy, temporal mixing, and classifier fixed reduces accuracy by 5.3 points on NW-UCLA, while removing the pairwise correlation readout leads to an even larger drop. Together, these results suggest that the proposed quantum feature maps and relational measurement scheme contribute substantially beyond the surrounding architecture. Finally, under 1024-shot simulations using noise models derived from the IBM Brisbane and Prague backends, SpooQ retains approximately 80% accuracy, while the flat Quantum Baseline falls to 42.6% and 53.2%, respectively. This is consistent with the hypothesis that distributing quantum computation across compact, anatomy-guided circuits yields a more noise-tolerant allocation of coherent circuit capacity.

Beyond action recognition, these results suggest a broader lesson for near-term quantum learning: the scarce resource on NISQ devices is not parameters but coherent circuit capacity, and allocating it along the natural structure of the data yields models that are simultaneously more accurate and more robust. Future work includes validating SpooQ on quantum hardware, extending the hierarchy to multi-person interactions and larger label spaces such as NTU RGB+D 120, and applying structure-guided circuit design to other domains with natural decompositions, such as molecular graphs.

AI USE STATEMENT

We used ChatGPT to assist with rewriting and language editing, suggest additional ablation experiments, and identify relevant references. We used Codex for debugging and code review. We did not use generative AI to generate data. The authors reviewed the AI-assisted text, references, and code and independently ran the experiments. We take responsibility for the final content of this work, including all text, claims, results, and code produced with AI assistance.

REPRODUCIBILITY STATEMENT

We provide supplementary source code for SpooQ and the three baselines, including training and evaluation scripts, dependency specifications, and data-preparation instructions. The package includes input checksums and exact clip lists for the filtered NTU60 split; datasets and trained checkpoints are not included. Appendices A and B specify the anatomical groups, circuit dimensions, and baseline architectures. Appendix E documents optimization, learning-rate schedules, and augmentation. Appendix D describes the single-person NTU60 filtering protocol, and Appendix C details the backend-derived noise simulations. Unless otherwise noted, results report mean and standard deviation over three random seeds.

REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1:403–409, 2021. doi: 10.1038/s43588-021-00084-1.

Shamminuj Aktar, Andreas Bärtschi, Abdel-Hameed A. Badawy, and Stephan Eidenbenz. Quantum graph transformer for nlp sentiment classification. *arXiv preprint arXiv:2506.07937*, 2025. URL https://arxiv.org/abs/2506.07937.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, B. AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13:4919, 2022.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021. doi: 10.1038/s42254-021-00348-9.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13359–13368, 2021.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with Qiskit, 2024.

Sang Hyub Kim, Jonathan Mei, Claudio Girotto, Masako Yamada, and Martin Roetteler. Quantum large language model fine-tuning. *arXiv preprint arXiv:2504.08732*, 2025. URL https://arxiv.org/abs/2504.08732.

Desheng Kong, Xiangshuo Cui, Jiaying Jin, Jing Xu, and Donglin Wang. Hybrid quantum transformer for language generation. *arXiv preprint arXiv:2511.10653*, 2025. URL https://arxiv.org/abs/2511.10653.

Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10410–10419, 2023. doi: 10.1109/ICCV51070.2023.00958.

Guangxi Li, Zhixin Song, and Xin Wang. VSQL: Variational shadow quantum learning for classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8357–8365, 2021.

Yangyang Li, Zhengya Qi, Yuelin Li, Haorui Yang, Ronghua Shang, and Licheng Jiao. A distributed hybrid quantum convolutional neural network for medical image classification. *arXiv preprint arXiv:2501.06225*, 2025. URL https://arxiv.org/abs/2501.06225.

Junghoon Justin Park, Maria Pak, Sebin Lee, Samuel Yen-Chi Chen, Shinjae Yoo, Huan-Hsin Tseng, and Jiook Cha. Hybrid quantum temporal convolutional networks. *arXiv preprint arXiv:2602.23578*, 2026. URL https://arxiv.org/abs/2602.23578.

John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. The effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021. doi: 10.1103/PhysRevA.103.032430.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016. doi: 10.1109/CVPR.2016.115.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2019.

Daniel Silver, Tirthak Patel, and Devesh Tiwari. QUILT: Effective multi-class classification on quantum computers using an ensemble of diverse quantum classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8324–8332, 2022.

Kalpit Thakkar and P. J. Narayanan. Part-based graph convolutional network for action recognition. *arXiv preprint arXiv:1809.04983*, 2018.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656, 2014.

Mingzhu Wang and Yun Shang. Hybrid vision transformer and quantum convolutional neural network for image classification. *arXiv preprint arXiv:2510.12291*, 2025. URL https://arxiv.org/abs/2510.12291.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12328.

Hui Zhang, Qinglin Zhao, Mengchu Zhou, and Li Feng. HQViT: Hybrid quantum vision transformer for image classification. *arXiv preprint arXiv:2504.02730*, 2025. URL https://arxiv.org/abs/2504.02730.

This supplementary material provides implementation details that support the main paper's experimental and reproducibility claims. We first define the anatomical joint groups and circuit sizes used by SpooQ on NW-UCLA and NTU60. We then specify the classical and quantum baselines, including the parameter-matched Classical SpooQ replacement blocks and the reduced CTR-GCN configurations. Finally, we document the single-person NTU60 filtering protocol and the noise-model setup used for the robustness experiments.

A Anatomical Groups and Circuit Specifications

A.1 Joint Groups

The six anatomical groups used by SpooQ's part circuits are defined by the following joint indices, as used in our implementation.

NW-UCLA (20 joints).

Group	Joint indices
Left arm	$\{4, 5, 6, 7\}$
Right arm	$\{8, 9, 10, 11\}$
Left leg	$\{12, 13, 14, 15\}$
Right leg	$\{16, 17, 18, 19\}$
Head	$\{2, 3\}$
Torso	$\{0, 1, 2\}$

On this skeleton the head and torso groups share the neck joint (index 2), so the groups form a cover of the joint set rather than a strict partition; the shared joint's coordinates enter both circuits.

NTU60 (25 joints).

Group	Joint indices
Left arm	$\{4, 5, 6, 7, 21, 22\}$
Right arm	$\{8, 9, 10, 11, 23, 24\}$
Left leg	$\{12, 13, 14, 15\}$
Right leg	$\{16, 17, 18, 19\}$
Head	$\{2, 3\}$
Torso	$\{0, 1, 20\}$

On NTU60 the six groups are disjoint. The relation set used by the region circuits is identical on both datasets: $\mathcal{R} = \{(\text{left arm}, \text{right arm}),\ (\text{left leg}, \text{right leg}),\ (\text{head}, \text{torso})\}$.

A.2 Per-Circuit Dimensions

Table 4 lists, for every circuit in the hierarchy, the input dimension d and qubit count $n = \lceil \log_2 d \rceil$.

B Baseline Specifications

B.1 Classical SpooQ

Classical SpooQ keeps the same anatomical hierarchy, temporal mixing, feature fusion, GRU, and classifier as SpooQ, but replaces each quantum circuit by a small MLP with the same input and output dimensions. For a block with input dimension d, output dimension r, and hidden width h, the replacement is

$$\mathrm{Linear}(d, h) \to \tanh \to \mathrm{Linear}(h, r).$$

Both linear layers use biases. No batch normalization, dropout, or other normalization is used inside these replacement blocks. The classical baseline uses $h = 2$ for every block, with total trainable

Table 4: Per-circuit specifications. All circuits use $L = 4$ variational blocks, contributing $3nL$ trainable parameters each, for 516 quantum parameters in total on NW-UCLA and 552 on NTU60. The fused per-frame representation has 124 dimensions on NW-UCLA and 141 on NTU60, including 7 retained norm features.

Dataset	Circuit	Input d	Qubits n
NW-UCLA	Arm (part, $\times 2$)	12	4
NW-UCLA	Leg (part, $\times 2$)	12	4
NW-UCLA	Head (part)	6	3
NW-UCLA	Torso (part)	9	4
NW-UCLA	Arms (region)	20	5
NW-UCLA	Legs (region)	20	5
NW-UCLA	Upper body (region)	16	4
NW-UCLA	Body	60	6
NTU60	Arm (part, $\times 2$)	18	5
NTU60	Leg (part, $\times 2$)	12	4
NTU60	Head (part)	6	3
NTU60	Torso (part)	9	4
NTU60	Arms (region)	30	5
NTU60	Legs (region)	20	5
NTU60	Upper body (region)	16	4
NTU60	Body	75	7

parameters matched to SpooQ within 2% on both datasets. The output dimension r is always the quantum readout width $n + \binom{n}{2}$, so the fused feature vector and GRU input dimension are identical to SpooQ: 124 on NW-UCLA and 141 on NTU60.

Table 5: Classical SpooQ replacement-block parameter counts for the parameter-matched baseline. Counts include both linear-layer biases and exclude the shared temporal mixer, GRU, and classifier.

Dataset	Block	Input/output	Params
NW-UCLA	Arm part, each	$12 \to 10$	56
NW-UCLA	Leg part, each	$12 \to 10$	56
NW-UCLA	Head part	$6 \to 6$	32
NW-UCLA	Torso part	$9 \to 10$	50
NW-UCLA	Arms region	$20 \to 15$	87
NW-UCLA	Legs region	$20 \to 15$	87
NW-UCLA	Upper body	$16 \to 10$	64
NW-UCLA	Body	$60 \to 21$	185
NW-UCLA	All replacements	–	729
NTU60	Arm part, each	$18 \to 15$	83
NTU60	Leg part, each	$12 \to 10$	56
NTU60	Head part	$6 \to 6$	32
NTU60	Torso part	$9 \to 10$	50
NTU60	Arms region	$30 \to 15$	107
NTU60	Legs region	$20 \to 15$	87
NTU60	Upper body	$16 \to 10$	64
NTU60	Body	$75 \to 28$	236
NTU60	All replacements	–	854

The final SpooQ and Classical SpooQ models both use no temporal windowing, a one-layer GRU with hidden size 24, and a classifier $\mathrm{Linear}(\cdot, 32) \to \mathrm{ReLU} \to \mathrm{Linear}(32, C)$, where C is the number of classes. The interleaved temporal mixer is a residual depthwise 1D convolution with kernel size 5 applied separately to each level-1 body-part feature stream, followed by a 1×1 channel mixing convolution; each part has a learned scalar residual strength initialized to zero. This mixer contributes 934 parameters on NW-UCLA and 1,254 on NTU60. Total trainable parameters are

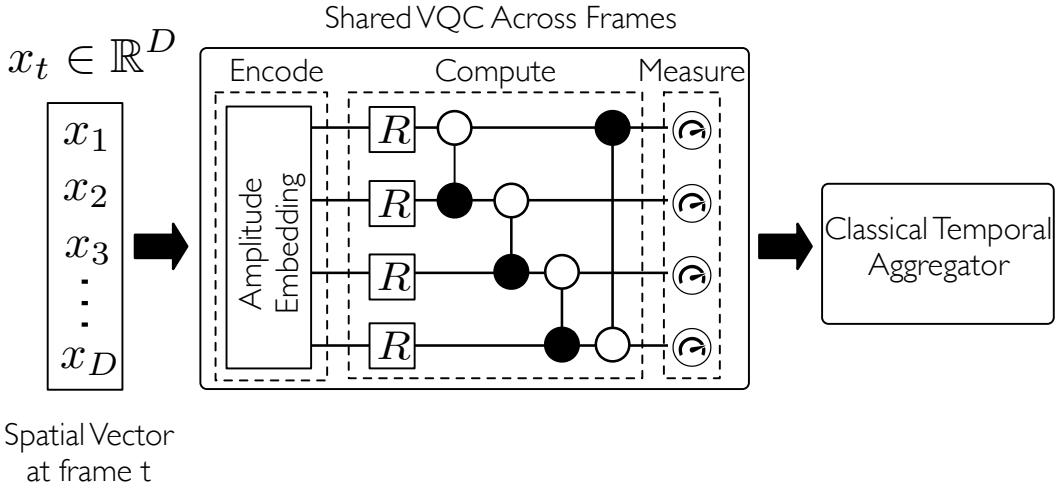

Figure 3: Flat quantum skeleton baseline. Each frame is flattened into a single spatial vector and amplitude-encoded into a shared variational quantum circuit. Trainable R rotations and entangling gates produce quantum measurement features for each frame, which are then aggregated by a classical temporal model for action recognition.

13,380 for SpooQ and 13,593 for Classical SpooQ on NW-UCLA, and 16,610 for SpooQ and 16,912 for Classical SpooQ on NTU60.

B.2 Quantum Baseline

The Quantum Baseline removes the anatomical hierarchy and uses a single flat amplitude-encoded circuit per frame. The full skeleton coordinates are centered, flattened, ℓ_2-normalized, zero-padded to the next power of two, and passed through the same $L = 4$ variational template as SpooQ. The circuit readout consists of all single-qubit Z expectations and all pairwise ZZ expectations, plus the retained input norm. Thus NW-UCLA uses a 60-dimensional input, 6 qubits, 21 circuit observables, and a 22-dimensional GRU input; NTU60 uses a 75-dimensional input, 7 qubits, 28 circuit observables, and a 29-dimensional GRU input. The classifier is the same GRU/MLP structure as above with hidden sizes 22 and 64. The resulting trainable parameter counts are 5,230 on NW-UCLA and 8,954 on NTU60, with 72 and 84 quantum-circuit parameters, respectively.

B.3 Reduced CTR-GCN

Our CTR-GCN baseline is a dependency-light PyTorch port of the original CTR-GCN blocks. The graph convolution, channel-wise topology refinement, multi-scale temporal convolution, residual structure, and initialization match the original implementation. The changes are limited to removing MMCV/MMEngine dependencies, adding an NW-UCLA 20-joint graph, supporting the filtered single-person NTU60 input, and exposing width/depth knobs for parameter matching.

Table 6: Reduced CTR-GCN configurations used for the paper-facing comparisons. Stage numbers follow the implementation indexing, where stage 1 is the first CTR-GCN block.

Dataset	Params	Channels	Strides
NW-UCLA	13,449	$6, 12, 12, 12, 12$	$1, 2, 1, 1, 1$
NTU60	15,781	$5, 10, 10, 10, 10$	$1, 2, 1, 1, 1$

Each CTR-GCN block performs CTR graph convolution followed by multi-scale temporal convolution and a residual connection. The temporal module has four branches: two temporal-convolution branches with dilations 1 and 2, one max-pooling branch, and one 1×1 branch. The second stage

uses temporal stride 2; all other stages use stride 1. The input tensor has one person stream for all reported results: $[N, 1, T, V, 3]$. The final representation is obtained by adaptive average pooling over time and joints, averaging over the person dimension, and a linear classifier.

No-pairwise ablation. When the pairwise readout is removed, the quantum circuits still run the same variational template but return only the n single-qubit expectations. The pairwise ZZ features are not replaced, so the fused feature dimension shrinks from 124 to 50 on NW-UCLA; the GRU input dimension shrinks accordingly. The dimension-matched product-readout control instead appends classical products $\langle Z_i \rangle \langle Z_j \rangle$ so the fused width remains 124.

B.4 Composition of the Filtered NTU60 Split

Table 7 lists, per mutual-action class, the number of clips containing exactly one valid tracked skeleton that survive filtering. Surviving mutual-action clips constitute 1.1% of the filtered training split and 1.2% of the filtered evaluation split. The equal A59 counts across splits are coincidental (103 of 623 training and 103 of 316 evaluation A59 clips survive); the released clip lists allow direct verification. In addition, 399 training and 356 evaluation clips are removed from the single-person classes (A1–A49) in which a spurious second skeleton was tracked, distributed broadly across all 49 classes.

Table 7: Surviving single-skeleton clips per mutual-action class (A50–A60) in the filtered NTU60 XView split.

Split	A50	A51	A52	A53	A54	A55
Train	33	27	21	18	14	15
Eval	14	9	10	3	3	3

Split	A56	A57	A58	A59	A60
Train	27	23	17	103	27
Eval	5	6	3	103	19

C Noise-Model Provenance

The finite-shot robustness experiments use Qiskit 1.2.4, Qiskit Aer 0.14.2, and qiskit-ibm-runtime 0.34.0. They evaluate noiselessly trained NW-UCLA checkpoints; no model is retrained under noise. The reported fake-backend runs use 1,024 shots and Qiskit transpiler optimization level 2 unless otherwise stated.

Table 8: Fake backends used to instantiate Aer noise models for the main robustness table.

Backend	Qubits
FakeBrisbane	127
FakePrague	33

For each backend, the simulator noise model is created with `qiskit_aer.noise.NoiseModel.from_backend`. The resulting model contains calibrated errors for `sx`, `x`, reset, measurement, and the backend's native two-qubit gate (`ecr` for Brisbane and `cz` for Prague). Readout error is enabled. Thermal relaxation and gate-duration effects are included through Aer's backend noise-model construction from the calibration properties.

For state-preparation-noise experiments, the amplitude encoding is implemented by Qiskit's `initialize` operation and transpiled to the selected backend basis, so the decomposed state-preparation gates are subject to the same backend-derived instruction errors as the variational gates. Independent logical circuits are packed side-by-side before measurement to reduce simulator overhead. Each individual NW-UCLA quantum circuit uses at most six qubits; the packed Qiskit circuits

use at most seven logical qubits because a 3-qubit and a 4-qubit circuit may be placed in the same sampled circuit.

We used the script's `noise_only` fake-backend mode for the paper-facing fake-backend results. In this mode the Aer simulator uses the backend-derived noise model and basis gates, but the fake backend's coupling map is not supplied to the transpiler.

D SINGLE-PERSON FILTERING OF NTU RGB+D 60

SpooQ's anatomical decomposition assigns joints to part circuits over a single kinematic graph, so the model is defined for one skeleton per clip. We therefore restrict the NTU RGB+D 60 cross-view (XView) split to clips containing exactly one valid tracked skeleton. Multi-person clips are removed entirely rather than merged or averaged, which would create artificial hybrid poses with no anatomical interpretation.

The filtering affects the two class groups differently. For the mutual-action classes (A50–A60), which are typically performed by two subjects, only clips in which the capture system tracked a single skeleton are retained: 325 training and 178 evaluation clips in total. For the single-person classes (A1–A49), clips in which a spurious second skeleton was tracked (e.g., background objects or reflections detected as a person) are removed: 399 training and 356 evaluation clips. Table 9 summarizes the resulting split.

	Training	Evaluation
A1–A49: clips removed (spurious 2nd skeleton)	399	356
A50–A60: clips retained (single skeleton)	325	178
Filtered XView split (all 60 classes)	30,683	15,296

Table 9: Composition of the single-person-filtered NTU60 XView split.

Because the filtered split differs from the standard XView protocol, results on it are not comparable to published NTU60 XView numbers, and we make no state-of-the-art claim on NTU60. All methods in this paper (SpooQ, the Quantum Baseline, Classical SpooQ, and CTR-GCN) are trained and evaluated on the identical filtered split.

E TRAINING AND EVALUATION DETAILS

SpooQ, the Quantum Baseline, and Classical SpooQ are trained end-to-end with AdamW (learning rate 10^{-3}, weight decay 10^{-4}), batch size 16, and cross-entropy loss with label smoothing 0.1, for 65 epochs with a 5-epoch warmup and a $0.1\times$ step decay at epoch 50. Each training epoch repeats the training set five times with fresh augmentation draws—random temporal crop, yaw rotation of $\pm 30^{\circ}$, uniform scaling in $[0.9, 1.1]$, left–right flip with probability 0.5, and Gaussian coordinate noise ($\sigma = 0.005$)—while evaluation uses deterministic temporal sampling with no augmentation; all models, including CTR-GCN, use identical augmentation. CTR-GCN follows its original recipe at the same budget and schedule: SGD with Nesterov momentum 0.9, learning rate 0.1, and weight decay 10^{-4}. Unless otherwise noted, all results report the mean and standard deviation of test accuracy over three random seeds.