Multivariate Gaussian Representation Learning for Medical Action Evaluation

Luming Yang[1*†], **Haoxian Liu**[2*], **Siqing Li**[3], **Alper Yilmaz**[1†]

[1]The Ohio State University
[2]Hong Kong University of Science and Technology
[3]Southern University of Science and Technology
yang.7670@osu.edu, hliueu@connect.ust.hk, lisq2022@mail.sustech.edu.cn, yilmaz.15@osu.edu

Abstract

Fine-grained action evaluation in medical vision faces unique challenges due to the unavailability of comprehensive datasets, stringent precision requirements, and insufficient spatiotemporal dynamic modeling of very rapid actions. To support development and evaluation, we introduce CPREVAL-6K, a multi-view, multi-label medical action benchmark containing 6,372 expert-annotated videos with 22 clinical labels. Using this dataset, we present GAUSSMEDACT, a multivariate Gaussian encoding framework, to advance medical motion analysis through adaptive spatiotemporal representation learning. Multivariate Gaussian Representation projects the joint motions to a temporally scaled multi-dimensional space, and decomposes actions into adaptive 3D Gaussians that serve as tokens. These tokens preserve motion semantics through anisotropic covariance modeling while maintaining robustness to spatiotemporal noise. Hybrid Spatial Encoding, employing a Cartesian and Vector dual-stream strategy, effectively utilizes skeletal information in the form of joint and bone features. The proposed method achieves 92.1% Top-1 accuracy with real-time inference on the benchmark, outperforming the baseline by +5.9% accuracy with only 10% FLOPs. Cross-dataset experiments confirm the superiority of our method in robustness.

Code — https://github.com/HaoxianLiu/GaussMedAct

Introduction

Cardiac arrest claims over 436,000 lives annually in the US alone, where high-quality cardiopulmonary resuscitation (CPR) can double survival rates. Recent advances in medical vision systems have underscored the critical need for fine-grained and rapid motion understanding in time-sensitive clinical scenarios, particularly CPR (Patil, Halperin, and Becker 2015; Masterson et al. 2024). The quality of CPR directly determines survival outcomes in cardiac arrest emergencies (Shinozaki et al. 2016), where compression depth and frequency are strongly correlated with survival rate (Shinozaki et al. 2016; Daudre-Vignier et al. 2023). Using feedback techniques can effectively improve results (Yeung et al. 2009). However, the current manual CPR assessment suffers from fundamental limitations revealed in our experiments. The results showed that human evaluators achieve only 74.8% accuracy in detecting critical errors such as incomplete chest recoil and frequency deviations (tested in 23 certified practitioners). Existing computer vision systems fail to capture centimeter motion deviations (*e.g.* 5 cm compression depth (Stiell et al. 2012)) and millisecond-level frequency variations (*e.g.* 115 ± 5 bpm requirements) (Travers et al. 2010). These challenges stem from persistent gaps in visual computing: modeling anatomical causality in spatiotemporal dynamics, preserving clinical interpretability, and achieving medical-grade temporal precision.

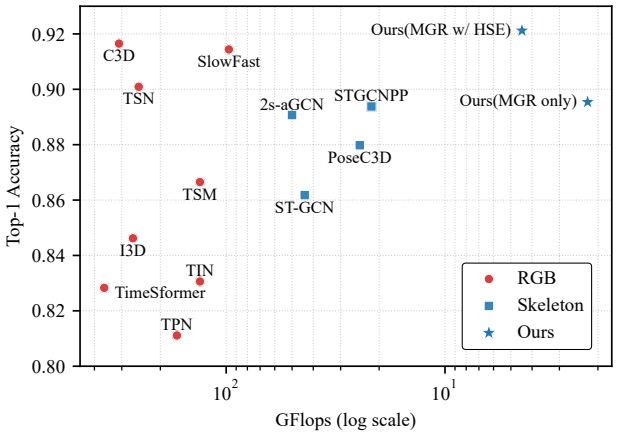

Figure 1: **Efficiency-Accuracy Trade-off Comparison.** Each point represents a model, with coordinates indicating computational complexity (GFLOPs) in log scale and top-1 accuracy. The proposed model is Pareto optimal.

Current action recognition approaches that focus on images (RGB-based) or graphs (skeleton-based) face limitations. Methods based on RGB (*e.g.* TimeSformer (Bertasius et al. 2021)) that rely on pre-trained backbones suffer from drawbacks such as fundamentally lacking anatomical modeling capabilities while incurring prohibitive computational latency. Skeleton-based methods (*e.g.* ST-GCN (Yu, Yin, and Zhu 2018)) discard motion semantics through rigid temporal pooling operations and remain vulnerable to noise.

*These authors contributed equally.
†Correspondence to {yilmaz.15, yang.7670}@osu.edu

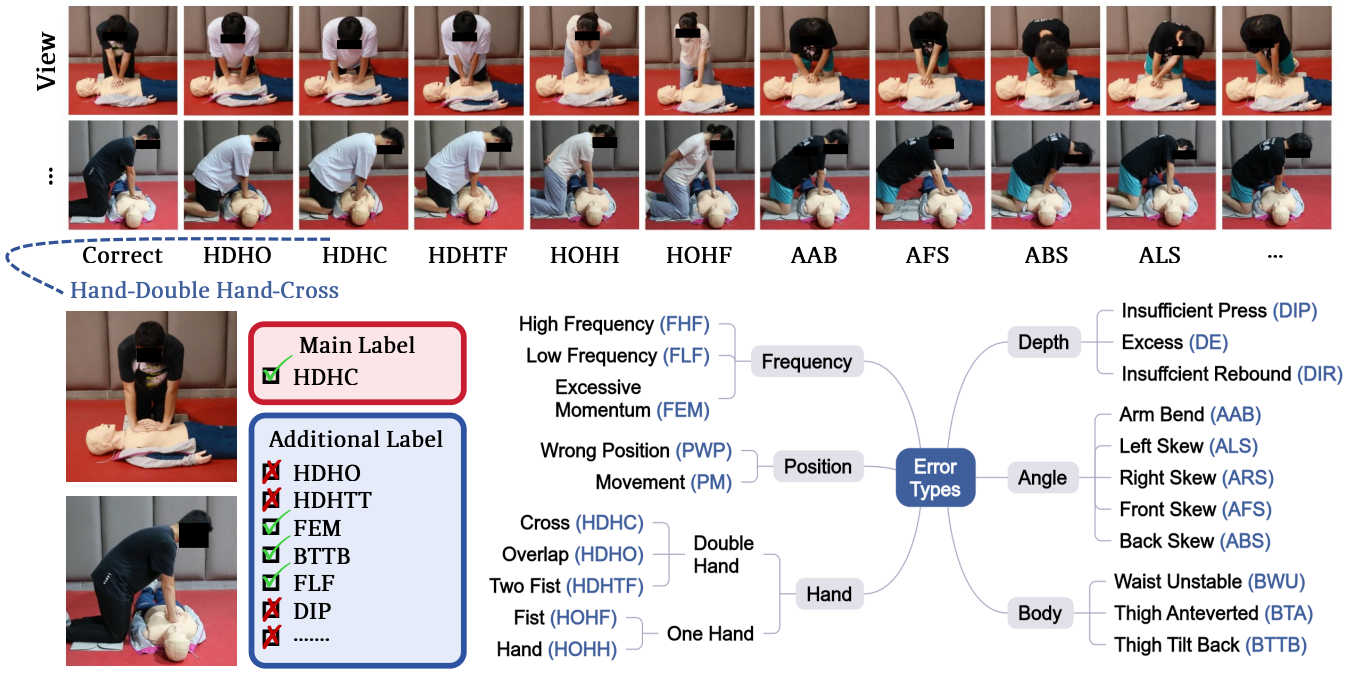

Figure 2: **Dataset Overview.** A multi-view CPR dataset with hierarchical error annotation, comprising 6 primary error classes and 21 fine-grained sub-classes. Each instance includes one primary error label and multiple secondary labels for compound error analysis.

A good solution requires a balance between accuracy, explainability, and latency (Gruber et al. 2012). These limitations directly affect the reliability of any method in the real world, while suboptimal CPR guidance techniques do not improve lifesaving effectiveness (Nassar and Kerber 2017).

Recent advances in Gaussian Splatting have demonstrated efficient and high-quality rendering in computer graphics, suggesting that sparse Gaussian distributions can effectively represent complex 3D point clouds (Kerbl et al. 2023; Chen and Wang 2024). However, its potential remains unexplored for spatiotemporal representations. This gap presents an opportunity given two key observations: First, Gaussian Mixture Models (Reynolds et al. 2009) naturally align with the probabilistic nature of motion point-cloud distributions. Second, conjecture on spatiotemporal interest points reveals that human motion dynamics can be decomposed into temporal components (Laptev 2005). Building on these insights, we propose a Multivariate Gaussian Representation (MGR) for robust spatiotemporal skeleton learning, which models temporal evolution of keypoints as probability distributions to achieve compact and noise-resistant action representations.

Our design is further motivated by fundamental principles of motion perception. Psychological studies show that sparse 2D point-light displays can convey strong action impressions through basic kinematic patterns (Johansson 1973). Complementary motion semantics exist in both absolute joint positions and relative bone kinematics (Shi et al. 2019). To fully exploit this dual nature, we introduce a Hybrid Spatial Encoding (HSE) through a Cartesian-Vector dual-stream architecture reconciles absolute anatomical positioning with relative kinematic patterns - an approach particularly effective for modeling rapid human motion.

Our key contributions can be summarized as follows:

- CPREVAL-6K: The largest multi-view clinical CPR dataset featuring synchronized RGB-skeleton streams and expert-validated multi-label annotations. Our benchmark contains 6,372 chest compression clips in 22 categories, each video hierarchically annotated with a primary critical error and secondary factors.
- GAUSSMEDACT: An end-to-end framework that combines MGR and HSE. By generating precision action token tensors, GAUSSMEDACT enables multiple downstream tasks including real-time classification and the generation of evaluation reports.

CPREval-6k Dataset

The proposed dataset CPREVAL-6K comprises 6,372 manually annotated videos documenting manual chest compression procedures, captured from multiple viewpoints. Through consultations with emergency physicians and team expertise, we identified 21 different error categories in 6 medically critical aspects of chest compression techniques, including 'hand position', 'arm angle', 'body posture', 'compression depth', 'operation frequency', and 'body positioning', as shown in Figure 2. Each CPR video is annotated with one primary error designation and multiple secondary error labels. See supplementary material for details of the dataset.

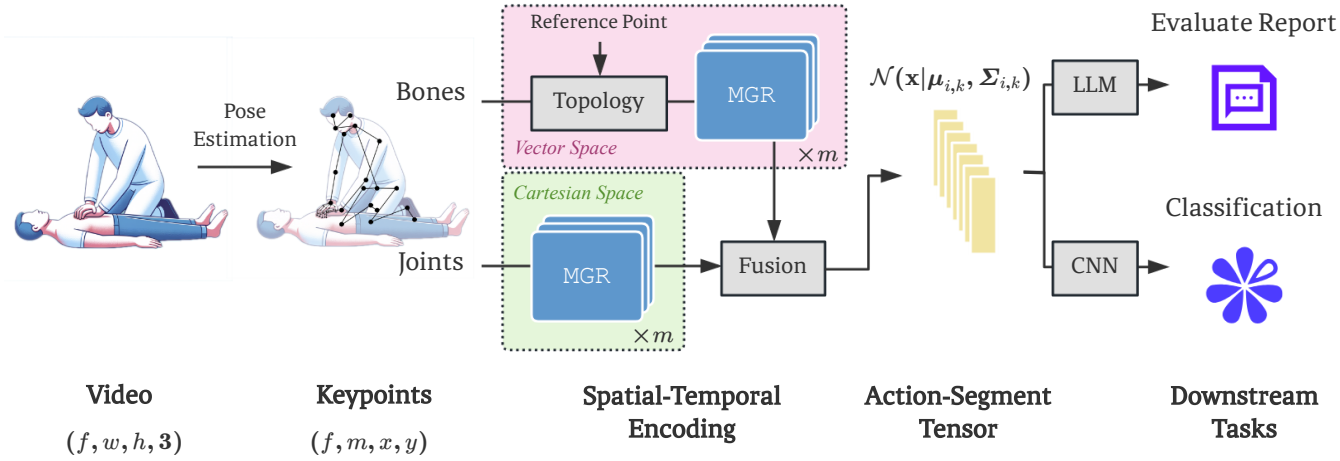

Figure 3: **Schematic of the GAUSSMEDACT Pipeline.** Input data undergoes cartesian and vector based dual-stream encoding and pass through MGR to generate gaussians. Through feature fusion, action tokens are generated for downstream tasks.

Data Collection. We implemented a multiview camera recording system with synchronized initiation to ensure temporal alignment between perspectives. To validate dataset robustness, 6 volunteers who had not received professional CPR training and 6 American Heart Association (AHA) Heartsaver® CPR-certified volunteers were recruited for standardized error simulation. The participants performed equal repetitions of each predefined error category, establishing a balanced distribution for the experimental validity. The entire dataset collection process is under the supervision of professionals to ensure that their actions meet the data collection standards.

Hierarchical Annotation. During data verification, we observed the involuntary co-occurrence of secondary errors during primary error simulations, necessitating our hierarchical annotation scheme of primary-secondary error labeling. A quality assurance protocol (QAP) involved 10 certified annotators who passed inter-rater reliability testing through rigorous Inter-Annotator Agreement evaluation. Annotation guidelines were strictly enforced under the supervision of a certified instructor who conducted final label verification, ensuring consistency and eliminating conspicuous annotation errors.

Data Analysis. The annotated dataset exhibits a significant class imbalance in the primary labels. Among the primary error labels, the *Depth-Insufficient Press* type accounts for 9.4%, making it the most frequent primary error type. This distribution arises from the Explicit Error Prioritization annotation protocol, where each sample is assigned only one dominant primary error along with multiple secondary errors.

In the overall error category distribution, which includes both primary and secondary error labels, *Depth-Insufficient Press* emerges as the most prevalent error type, accounting for 29.1% of all annotated error instances. This is consistent with clinical studies (Bobrow et al. 2013) that highlight challenges in monitoring chest compression quality.

The analysis of secondary labels under the primary label reveals frequent co-occurrence patterns between specific error types. Preliminary correlation analysis identifies statistically significant relationships. For example, *Freq-Excessive Momentum* and *Depth-Excess* exhibit a Pearson correlation coefficient of 0.52, while *Wrong Position* and *Position-Movement* show a correlation of 0.39, which aligns with biomechanical intuition.

To systematically uncover hierarchical error dependencies, we implement an enhanced Apriori algorithm ($min\_support = 0.025$, $min\_confidence = 0.25$) for the mining of association rules. The results are listed in the supplementary material. We have the following key findings:

- **Strong Association:** Rule 5 (M:*Position-Movement* $\rightarrow$ A:*Position-Wrong Position*) demonstrates an exceptional association, achieving 77.6% confidence and 17.4× lift. This quantifies the linkage where compression point instability directly induces positional errors.
- **Kinematic Chain Coupling:** Upper-limb anomalies (M:*Angle-Arm Bend*) correlate with both posterior thigh tilt (A:*Body-Thigh Tilt Back*, 38.8% confidence) and insufficient compression depth (A:*Depth-Insufficient Press*, 38.4% confidence), revealing whole-body kinetic chain interactions during CPR execution.
- **Hierarchical Error Propagation:** M$\rightarrow$A rules exhibit higher confidence (38.8–77.6%) compared to A$\rightarrow$A co-occurrence rules (26.1–32.7%), indicating that primary errors exert strong causative drives on secondary errors.

Statistical data indicate that manual chest compressions constitute a coordinated technical action of the entire body. A single error in execution can propagate additional errors, ultimately resulting in suboptimal CPR outcomes. This highlights the critical importance of mastering the correct posture of CPR to ensure effective performance.

GaussMedAct

We propose GAUSSMEDACT, a dual-stream spatiotemporal encoding framework for medical action recognition (see

Figure 3). The overall pipeline is summarized as follows.

- **Spatial dimension.** Human key points are extracted using pose estimation, and their characteristics are decoupled into complementary fluids from joint and bone. These two streams are processed separately and fused at a later stage to capture spatial dependencies and mitigate *collinearity* issues.
- **Temporal dimension.** MGR is introduced to process joints and bones independently. MGR captures action tokens along the temporal dimension and encodes them into Gaussian distributions, thereby modeling temporal dynamics and alleviating noise from pose estimation.
- **Feature Fusion.** Different fusion strategies used to integrate dual-stream features.
- **Downstream Tasks.** Two downstream tasks are introduced, including report generation and action classification. The use of label smoothing loss enhances the generalization performance.

Multivariate Gaussian Representation

Rationale. In the field of skeleton-based action recognition, GCN architectures (*e.g.* ST-GCN (Yu, Yin, and Zhu 2018), CTR-GC (Chen et al. 2021)) typically rely on local convolution in temporal modeling. However, these methods often lack the ability to capture the temporal dynamics of human actions.

Inspired by the breakthroughs in Gaussian splatting (Kerbl et al. 2023) from the field of computer graphics, which has demonstrated remarkable performance in rendering tasks, we re-examined the temporal modeling problem for Gaussian splatting. In graphics rendering, Gaussian splatting can represent a highly dense point cloud in a vast spatial domain using only a small number of Gaussian distributions. This insight informs us that an intricate set of original spatial points can, in fact, be effectively described using significantly fewer key points.

We extend this idea to temporal encoding. In many human actions, the movements of keypoints are inherently continuous and can be represented by a compact set of temporal segments. Critical transition points, such as the start and end of accelerations or directional changes, constitute only a small subset relative to the overall temporal sampling space. This observation aligns with early discussions on spatiotemporal interest points (STIP) (Laptev 2005). Building on these insights, we combined the innovations of STIP with Gaussian splatting, proposing MGR, enabling efficient and expressive temporal encoding.

Input Construction. For each joint i, its temporal trajectory is defined as a spatiotemporal point set $\mathcal{X}_i = \{\mathbf{x}_{i,t}\}_{t=1}^{T}$, where $\mathbf{x}_{i,t} = (x_{i,t}, y_{i,t}, t) \in \mathbb{R}^3$ contains 2D coordinates and normalized timestamps. To balance spatial and temporal scales, we introduce a time-axis scaling factor α:

$$\mathbf{x}_{i,t} \leftarrow (x_{i,t}, y_{i,t}, \alpha \cdot t) \quad (\alpha \in \mathbb{R}^+). \tag{1}$$

Gaussian Modeling. We conjecture that $\mathcal{X}_i$ is generated by a mixture of K Gaussian distributions (Reynolds et al. 2009). The probability density function is:

$$p(\mathbf{x}|\boldsymbol{\theta}_i) = \sum_{k=1}^{K} \pi_{i,k} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}), \tag{2}$$

where $\boldsymbol{\theta}_i = \{\pi_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\}_{k=1}^{K}$, with mixture weights $\pi_{i,k}$ satisfying $\sum_{k=1}^{K} \pi_{i,k} = 1$. Parameters are optimized using the Expectation Maximization (EM) algorithm.

Action Token. For each Gaussian component k, we extract compressed features from $\boldsymbol{\mu}_{i,k}$ and $\boldsymbol{\Sigma}_{i,k}$. Each Gaussian represents an action token. Although we did not use higher dimensions in our architecture (discussed in Section HSE), MGR can be extended to more dimensions, such as incorporating limb angles, states, etc. For the x-y-t or r-θ-t point set, we have a 3D Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}, \tag{3}$$

where, $\boldsymbol{\mu}$ represents the average position of the joints within this segment, while the covariance matrix $\boldsymbol{\Sigma}$ represents the scaling (motion) and rotation (direction) across dimensions.

Covariance Decomposition. To achieve a representation that is both differentiable and interpretable, we transform the covariance matrix into scale and rotation components that have physical significance (Kerbl et al. 2023).

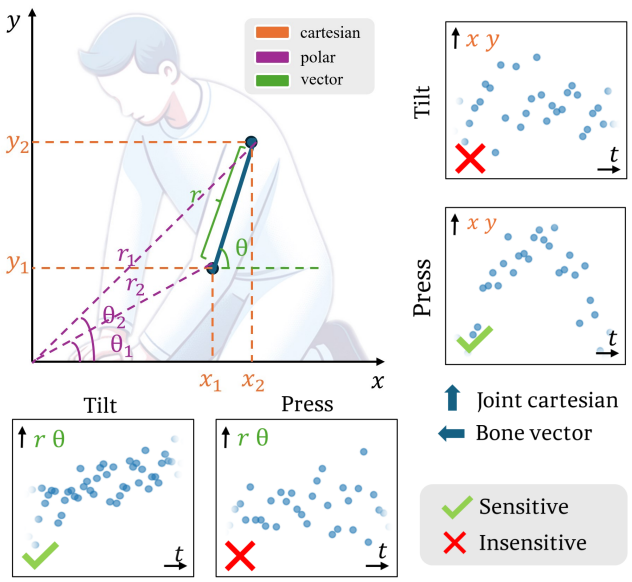

Figure 4: **Feature Discriminability of Hybrid Spatial Encoding.** The figure illustrates three information modes: cartesian-based, polar-based, and vector-based. Using chest compression and limb tilt as prototypes, the analysis reveals distinct signal-formative capabilities: Sensitive modes generate structured point clusters that fit to kinematic functions, while insensitive modes exhibit noise distributions. The hybrid architecture orchestrates dual-stream processing to adaptively harness these geometric discriminators.

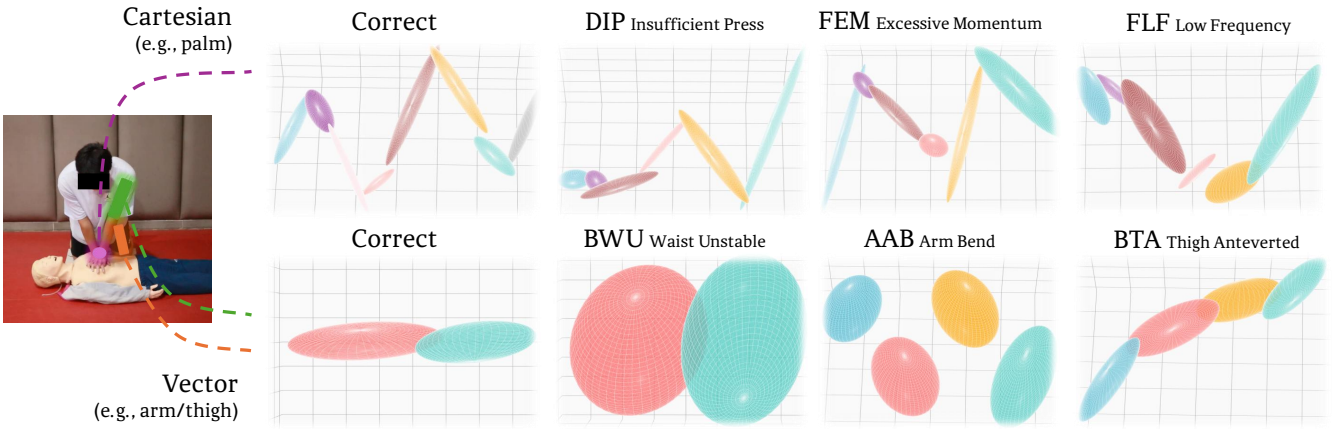

Figure 5: **Dynamics Visualization of MGR.** Each ellipsoid represents a Gaussian distribution (mean μ, covariance Σ) of a single joint/bone dynamics over time. From the perspective of stream-specific strengths, Joint Cartesian Stream excels at capturing global trajectory consistency (*e.g.*, palm trajectory); Bone vector better encodes kinematic transitions (*e.g.*, arm bend).

$$\mathbf{\Sigma}_{i,k} = \mathbf{R}_{i,k}\mathbf{S}_{i,k}\mathbf{S}_{i,k}^{\top}\mathbf{R}_{i,k}^{\top}, \tag{4}$$

where $\mathbf{S}_{i,k} = \text{diag}(s_{i,k}^x, s_{i,k}^y, s_{i,k}^t) \in \mathbb{R}^{3\times3}$ is derived from a 3D vector $\mathbf{s}_{i,k} \in \mathbb{R}^3$, $\mathbf{R}_{i,k} \in \mathbb{R}^{3\times3}$ is computed from a unit quaternion $\mathbf{q}_{i,k} \in \mathbb{R}^4$.

This decomposition disentangles the motion dynamics into scale (magnitude of movement) and orientation (direction of movement), aligning with human biomechanical constraints. Finally, we concatenate $\boldsymbol{\mu}_{i,k}$, $\mathbf{s}_{i,k}$, and $\mathbf{q}_{i,k}$ into a compact 10D tensor:

$$\mathbf{f}_{i,k} = [\boldsymbol{\mu}_{i,k};\ \mathbf{s}_{i,k};\ \mathbf{q}_{i,k}] \in \mathbb{R}^{3+3+4=10}. \tag{5}$$

Hybrid Spatial Encoding

Rationale. Human action recognition, especially in medical scenarios, is based on modeling anatomical structures from sparse keypoints. Research has shown that 2D point and line combinations provide a strong impression of the type of action (Johansson 1973). Both absolute joint positions and relative bone kinematics encode complementary motion semantics (Shi et al. 2019). The features have two streams of information including direct use of Cartesian coordinates as input (coordinate-based encoding, (x, y, t)); and angles or vectors as input (vector-based encoding (r, θ, t)). In terms of anatomy, if bones are projected into a 2D space, information such as angles can be expressed in polar coordinates. These features are critically important in the recognition of medical actions. See Figure 4 for the sensitivity analysis of the two streams.

Let us discuss the limitations of prior encoding schemes:

- Joint Cartesian *w/* Bone Cartesian: Widely adopted in GCNs and variants using a dynamic graph or attention mechanism (Yu, Yin, and Zhu 2018; Wang et al. 2019; Ye et al. 2020a,b). This approach forces networks to learn positions, resulting in inefficient feature redundancy.
- Joint Polar *w/* Bone Vector: Despite their anatomical grounding, angular wraparound artifacts (*e.g.*, $-\pi \leftrightarrow \pi$ discontinuities) induce gradient instability, as evidenced by our polar-based model in ablation studies; see supplementary material.

We propose the HSE solution that takes advantage of the joint Cartesian coordinates (x, y, t) and the bone vector (r, θ, t), and decouples absolute localization from relative kinematics, contextualizing fine-grained bone dynamics.

Semantics Disentangling. Joint stream (Cartesian encoding) is represented in the xyt-space to preserve absolute spatial-temporal positions. Bone stream (vector encoding) is parameterized by dynamic vector features in $r\theta t$-space.

$$\mathbf{J}_i = [x_i, y_i, t] \in \mathbb{R}^3, \mathbf{B}_{ij} = [\Delta r_{ij}, \theta_{ij}, t] \in \mathbb{R}^3, \tag{6}$$

where $\Delta r_{ij} = \|\mathbf{J}_j - \mathbf{J}_i\|_2$ and $\theta_{ij} = \arctan\left(\frac{y_j - y_i}{x_j - x_i}\right)$. Concatenating raw joint and bone features (x, y, r, θ, t) as input to MGE may risk *multicollinearity*, as Cartesian coordinates and polar parameters are geometrically interdependent ($x = r\cos\theta, y = r\sin\theta$). Our MGE module first processes joints and bones in isolated embedding spaces, disentangling absolute and relative semantics. Subsequent fusion operates on decorrelated high-level features.

Feature Fusion Strategy. We adopt fusion with multiple variants: Cross-Attention (Eq. 7), Interleaved Concatenation (Eq. 8), and others. Different fusion strategies can work differently on various data complexities and dataset types.

$$\mathbf{F} = \text{LayerNorm}(\mathbf{F}_J + \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})). \tag{7}$$

$$\mathbf{F} = \text{Interleave}(\mathbf{F}_J, \mathbf{F}_B) \in \mathbb{R}^{2M\times K\times 10}, \tag{8}$$

where $\text{Interleave}(\mathbf{F}_J, \mathbf{F}_B)$ rearranges features in an alternating pattern $(j_1, b_1, j_2, b_2, \ldots, j_M, b_M)$ for each skeleton. See the supplementary material for details.

Temporal Dynamics with MGR

As illustrated in Figure 5, our temporal encoding module generates compact yet discriminative representations

Model	Modality	Backbone	Dual-stream	Pre-trained	Accuracy			GFLOPs
					Top-1	Top-5	Mean	Per sample
TSM (Lin, Gan, and Han 2019)	RGB	ResNet-50	✗	✗	0.8665	0.9820	0.8469	131.83
TSN (Wang et al. 2018)	RGB	ResNet-50	✗	Kinetics-400	0.9009	0.9867	0.8899	251.12
TPN (Yang et al. 2020)	RGB	ResNet-50	✗	✗	0.8111	0.9727	0.7235	168.00
TIN (Shao, Qian, and Liu 2020)	RGB	ResNet-50	✗	✗	0.8306	0.9766	0.7951	131.83
C3D (Tran et al. 2015)	RGB	3D ConvNet	✗	Sports-1M	0.9165	0.9906	0.9069	308.92
I3D (Carreira and Zisserman 2017)	RGB	ResNet-50	✗	✗	0.8462	0.9789	0.8331	266.80
SlowFast (Feichtenhofer et al. 2019)	RGB	ResNet-50	✓	Kinetics-400	0.9144	0.9874	0.9107	97.27
TimeSformer (Bertasius et al. 2021)	RGB	ViT	✗	✗	0.8283	0.9781	0.8043	360.89
ST-GCN (Yu, Yin, and Zhu 2018)	Skeleton	GCN	✗	✗	0.8618	0.9758	0.8356	43.76
2s-aGCN (Shi et al. 2019)	Skeleton	GCN	✓	✗	0.8907	0.9703	0.8812	50.01
STGCNPP (Duan et al. 2022a)	Skeleton	GCN	✗	✗	0.8938	0.9742	0.8777	21.69
PoseC3D (Duan et al. 2022b)	Skeleton	3D ConvNet	✗	✗	0.8798	0.9727	0.8582	24.51
InfoGCN (Chi et al. 2022)	Skeleton	GCN	✗	✗	0.8977	0.9789	0.8825	38.47
SkateFormer (Do and Kim 2024)	Skeleton	Transformer	✗	✓	0.9047	0.9789	0.8907	42.17
HDGCN (He et al. 2023)	Skeleton	GCN	✗	✓	0.9063	0.9797	0.8911	39.50
Ours *(MGR w/ HSE)*	Skeleton	CNN	✓	✗	0.9212	0.9836	0.9082	4.45
Ours *(MGR only)*	Skeleton	CNN	✗	✗	0.8954	0.9602	0.8836	2.23

Table 1: **Model Performance Across Modalities.** Blue indicates the best result and Orange indicates the second-best result.

through MGR. The temporal evolution of actions (*e.g.*, chest compression) is encoded as compact sequences of Gaussian components that represent motion primitives. Remarkably, complex 60 frame motions can only be represented by ≈ 6 Gaussians, demonstrating MGR's ability to distill high-level motion primitives. Gaussian parameters (mean μ, covariance Σ) create separable clusters in feature space, enabling classifiers to achieve 92.1% accuracy with minimal fine-tuning (see Table 1).

This analysis confirms that Gaussian-based representation learning bridges the gap between raw pose dynamics and specific semantics, an advantage for medical applications that require both precision and interpretability.

Downstream Tasks

Loss Function. Taking into account the fine-grained nature of the medical model and the fact that some labels have few or uneven samples, we use MixUp (Zhang et al. 2017) to comprehensively improve the effectiveness of the model in processing medical data. Through linear interpolation of inputs and labels, the model can be generalized by generating synthetic samples. For a given pair of samples (x_i, y_i) and (x_j, y_j), the mixed input and label are computed as:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j, \tag{9}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ for $\alpha \in (0, +\infty)$ is sampled from a Beta distribution, and $\alpha > 0$ controls the interpolation strength. The MixUp loss is defined as:

$$\mathcal{L}_{\text{MixUp}} = \lambda \mathcal{L}(f(\tilde{x}), y_i) + (1-\lambda)\mathcal{L}(f(\tilde{x}), y_j). \tag{10}$$

By combining pairs of inputs and labels, smoother decision boundaries and reduced overfitting are encouraged, which helps to enhance the robustness of the model. For other loss strategies such as CE and Label Smoothing (Müller, Kornblith, and Hinton 2019).

Classifier. The fused features $\mathbf{F}_{\text{fused}}$ are mapped to category scores through multilayer CNN with spatiotemporal pooling and network outputs $\mathbf{y}_{\text{pred}}$. See the supplementary material for the detailed architecture.

Evaluation Report Generation. For generating a report of the evaluation, MGR-encoded skeletal tensors undergo spatiotemporal tokenization to bridge visual and textual modalities. Specifically, we first discretize continuous kinematic features (depth, frequency, posture angles) into clinically-grounded linguistic descriptors ("insufficient 4cm compression","optimal 110bpm rhythm") using quantization bins derived from AHA guidelines, then made final inferences (metrics→manifestation→consequence) through a logical chain reasoning mechanism based on dataset analysis. Effectiveness score calculation and causal analysis is shown in the supplementary material.

Experiments

All experiments were implemented in PyTorch and use the MMPose framework (Contributors 2020). To generate a pose input for GAUSSMEDACT from RGB modality, we adopted RTMpose (Jiang et al. 2023). The image sequences are uniformly sampled to 32-frame clips with a spatial resolution of 224×224. To ensure fairness, all models shared identical training-test splits (80%-20% random partition) and underwent rigorous adjustments for confounding factors, as detailed in Table 1. All skeleton baselines were trained using the exact same skeleton data. We trained models using early stopping with a maximum of 300 epochs. Memory-intensive models were deployed on NVIDIA A100 GPUs, while other models utilized NVIDIA A6000 GPUs.

Comparative Analysis

We conduct comprehensive evaluations of multimodal approaches on the CPREVAL-6K dataset, comparing our skeleton-based GAUSSMEDACT with state-of-the-art methods that use RGB and skeleton modalities. Four metrics are adopted: Top-1/5 accuracy, class-wise mean accuracy, and

Model	Modality	Epoch	Accuracy	
			Top-1	Top-3
TSN-pretrained	RGB	50	0.9067	0.9921
TSN	Flow	50	0.8304	0.9851
STGCN-best	Skeleton	50	0.9246	0.9970
PoseC3D	Skeleton	240	0.9208	0.9922
Ours	Skeleton	100	0.9524	0.9950

Table 2: **Cross-dataset Evaluation on *Coach*.** The dataset is a medical action dataset with 14 categories. Blue indicates the best result and Orange indicates the second-best.

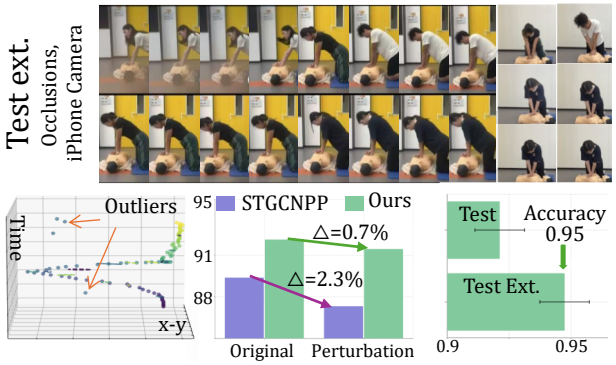

Figure 6: **Test Extention for Real Scene.** GAUSSMEDACT shows inherent robustness to perturbations comparing to STGCNPP. In practical usage scenarios, the accuracy is higher than CPREVAL-6K.

computational complexity (GFLOPs). The top-1 accuracy receives the primary emphasis due to its clinical relevance as it critically affects CPR effectiveness. There may be multiple relatively reasonable answers (*e.g.*, when both primary and secondary labels exist, and the model identifies the secondary category). In such cases, using the Top-5 accuracy metric would provide a fairer evaluation of the model's performance. Class-wise mean accuracy ensures balanced performance across CPR errors, particularly for low-frequency but high-risk categories (*e.g.*, depth-excess). As shown in Table 1 and Figure 1, three key observations emerge:

Modality. Skeleton-based models demonstrate significantly lower computational requirements (6.13× fewer GFLOPs on average) compared to RGB counterparts. Even when accounting for pose extraction costs (4.03-5.45 GFLOPs for RTMpose), the total computation remains substantially lower than RGB methods. However, RGB approaches achieve higher accuracy ceilings (91.65% vs. 89.38% for prior skeleton methods), indicating that the RGB modality has specific advantages in feature richness.

Superiority of GAUSSMEDACT. Our method establishes new state-of-the-art performance with 92.12% Top-1 accuracy, surpassing all existing models, including RGB-based approaches. In particular, this is achieved with only 4.45 GFLOPs, which is 10% of the computational cost required by ST-GCN (43.76 GFLOPs). The accuracy improvement over RGB models confirms the underexploited potential of skeleton data in medical action recognition when coupled with effective representation learning.

Efficiency-Performance Trade-off. The *MGR-only* variant (89.54% Top-1) already outperforms all skeleton baselines while requiring only half the computational complexity of full model, validating the efficacy of MGR. The complete model further improves performance by 2.58%, demonstrating synergistic effects between MGR and HSE.

Robustness

Cross-Dataset Evaluation. We conduct rigorous cross-dataset validation using the recently released medical CPR-Coach benchmark (Wang et al. 2024), which contains 14 classes. Following the official (60/40) split protocol, we compare GAUSSMEDACT with four approaches in different modality. Training configurations strictly follow original papers for all methods. All models use the best results reported under different settings in the original paper. As shown in Table 2, our method achieves 2.78% absolute improvement in Top-1 accuracy while maintaining real-time performance.

Perturbation and real scene. To evaluate the model's robustness, particularly against occlusion and sensor variations, we introduced an additional test set comprising 114 real-world training videos. These videos, captured primarily from beginners, feature realistic scenarios including occlusions and diverse mobile phone sensors. Other experiments are provided in the supplementary materials.

Social Impact

Beyond technical, our framework has demonstrated tangible social impact through deployment in CPR training programs. Partnering with training centers, GAUSSMEDACT has been integrated into standardized courses across several institutions. Quantitative evaluations show a 32% improvement in practical assessment scores among trainees compared to traditional methods under identical conditions, while instructors highlights enhanced diagnosis precision and actionable feedback. These real-world validations underscore the potential of AI-assisted medical training to improve resuscitation quality and patient outcomes.

Conclusions

This paper addresses critical dataset and methodology gaps in medical action evaluation. We introduce CPREVAL-6K, a much-needed fine-grained medical dataset that incorporates expert-validated hierarchical annotations which capture subtle error patterns. Through comprehensive comparative studies, we demonstrate that our proposed framework using spatiotemporal Gaussian mixture representation in decoupled joint and bone spaces outperforms both RGB- and skeleton-based models in accuracy while achieving significant computational cost reduction. The framework also exhibits robustness to cross-dataset generalization. These contributions establish a new foundation for the real-time CPR evaluation, with applications that extend to other medical assessments.

References

Bertasius, G.; et al. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Bobrow, B. J.; Vadeboncoeur, T. F.; Stolz, U.; Silver, A. E.; Tobin, J. M.; Crawford, S. A.; Mason, T. K.; Schirmer, J.; Smith, G. A.; and Spaite, D. W. 2013. The influence of scenario-based training and real-time audiovisual feedback on out-of-hospital cardiopulmonary resuscitation quality and survival from out-of-hospital cardiac arrest. *Annals of emergency medicine*, 62(1): 47–56.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, G.; and Wang, W. 2024. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the ICCV*, 13359–13368.

Chi, H.-g.; Ha, M. H.; Chi, S.; Lee, S. W.; Huang, Q.; and Ramani, K. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20186–20196.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose. Accessed: 22 December 2025.

Daudre-Vignier, C.; Bates, D. G.; Scott, T. E.; Hardman, J. G.; and Laviola, M. 2023. Evaluating current guidelines for cardiopulmonary resuscitation using an integrated computational model of the cardiopulmonary system. *Resuscitation*, 186: 109758.

Do, J.; and Kim, M. 2024. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, 401–420. Springer.

Duan, H.; Wang, J.; Chen, K.; and Lin, D. 2022a. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7351–7354.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022b. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2969–2978.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the CVPR*, 6202–6211.

Gruber, J.; Stumpf, D.; Zapletal, B.; Neuhold, S.; and Fischer, H. 2012. Real-time feedback systems in CPR. *Trends in Anaesthesia and Critical Care*, 2(6): 287–294.

He, M.; Chen, J.; Gong, M.; and Shao, Z. 2023. HDGCN: Dual-channel graph convolutional network with higher-order information for robust feature learning. *IEEE Transactions on Emerging Topics in Computing*, 12(1): 126–138.

Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; and Chen, K. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14: 201–211.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Laptev, I. 2005. On space-time interest points. *International journal of computer vision*, 64: 107–123.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7083–7093.

Masterson, S.; Norii, T.; Yabuki, M.; Ikeyama, T.; Nehme, Z.; Bray, J.; et al. 2024. Real-time feedback for CPR quality–A scoping review. *Resuscitation Plus*, 19: 100730.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Nassar, B. S.; and Kerber, R. 2017. Improving CPR performance. *Chest*, 152(5): 1061–1069.

Patil, K. D.; Halperin, H. R.; and Becker, L. B. 2015. Cardiac arrest: resuscitation and reperfusion. *Circulation Research*, 116(12): 2041–2049.

Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663): 3.

Shao, H.; Qian, S.; and Liu, Y. 2020. Temporal interlacing network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11966–11973.

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.

Shinozaki, K.; Nonogi, H.; Nagao, K.; and Becker, L. B. 2016. Strategies to improve cardiac arrest survival: a time to act. *Acute Medicine & Surgery*, 3(2): 61.

Stiell, I. G.; Brown, S. P.; Christenson, J.; Cheskes, S.; Nichol, G.; Powell, J.; Bigham, B.; Morrison, L. J.; Larsen, J.; Hess, E.; et al. 2012. What is the role of chest compression depth during out-of-hospital cardiac arrest resuscitation? *Critical Care Medicine*, 40(4): 1192–1198.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the CVPR*, 4489–4497.

Travers, A. H.; Rea, T. D.; Bobrow, B. J.; Edelson, D. P.; Berg, R. A.; Sayre, M. R.; Berg, M. D.; Chameides, L.; O'Connor, R. E.; and Swor, R. A. 2010. Part 4: CPR overview: 2010 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 122(18_suppl_3): S676–S684.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755.

Wang, S.; Wang, S.; Yang, D.; Li, M.; Kuang, H.; Zhao, X.; Su, L.; Zhai, P.; and Zhang, L. 2024. CPR-Coach: Recognizing Composite Error Actions based on Single-class Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18782–18792.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Yang, C.; Xu, Y.; Shi, J.; Dai, B.; and Zhou, B. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 591–600.

Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; and Tang, H. 2020a. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACMMM*, 55–63.

Ye, J.; He, J.; Peng, X.; Wu, W.; and Qiao, Y. 2020b. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 649–665. Springer.

Yeung, J.; Meeks, R.; Edelson, D.; Gao, F.; Soar, J.; and Perkins, G. D. 2009. The use of CPR feedback/prompt devices during training and CPR performance: a systematic review. *Resuscitation*, 80(7): 743–751.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.