Beyond Binary Contrast: Modeling Continuous Skeleton Action Spaces with Transitional Anchors

Supplementary Material

Contents

1. Theoretical Analysis

To deeply understand the behavior of the proposed soft alignment strategy, we provide a rigorous theoretical derivation showing its connection to mutual information maximization, supported by established contrastive learning principles [10].

Recall that for each query-key pair $(\boldsymbol{q}, \boldsymbol{k})$, we first retrieve the top-K most similar memory samples $\mathcal{N}_K = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K\} \subseteq \mathcal{M}$ with respect to $\boldsymbol{k}$. The similarity vectors restricted to these high-confidence neighbors are

$$\boldsymbol{p}_k = \big[\mathrm{sim}(\boldsymbol{k}, \boldsymbol{m}_j)\big]_{j=1}^{K}, \quad \boldsymbol{p}_q = \big[\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_j)\big]_{j=1}^{K}, \tag{1}$$

and the alignment loss is the asymmetric KL divergence

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{k}) = \mathrm{KL}\Big(\mathrm{softmax}(\boldsymbol{p}_k/\tau_k) \,\|\, \mathrm{softmax}(\boldsymbol{p}_q/\tau_q)\Big) = -\sum_{j=1}^{K} \tilde{p}_{k,j} \log \tilde{p}_{q,j}, \tag{2}$$

where

$$\tilde{p}_{k,j} = \frac{\exp(\mathrm{sim}(\boldsymbol{k}, \boldsymbol{m}_j)/\tau_k)}{\sum_{l=1}^{K} \exp(\mathrm{sim}(\boldsymbol{k}, \boldsymbol{m}_l)/\tau_k)}, \quad \tilde{p}_{q,j} = \frac{\exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_j)/\tau_q)}{\sum_{l=1}^{K} \exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_l)/\tau_q)}, \tag{3}$$

and $\tau_k < \tau_q$ ensures the target distribution $\tilde{p}_k$ is sharper than $\tilde{p}_q$.

Substituting the softened probabilities into Eq. (2) yields

$$\begin{aligned} \mathcal{L}(\boldsymbol{q}, \boldsymbol{k}) &= -\sum_{j=1}^{K} \tilde{p}_{k,j} \left[\frac{\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_j)}{\tau_q} - \log \sum_{l=1}^{K} \exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_l)/\tau_q) \right] \\ &= \log \sum_{l=1}^{K} \exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_l)/\tau_q) - \frac{1}{\tau_q} \sum_{j=1}^{K} \tilde{p}_{k,j}\, \mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_j). \end{aligned} \tag{4}$$

In the limiting case $\tau_k \to 0^+$, the target distribution $\tilde{p}_k$ collapses to a **sharp distribution** centered around the samples with maximum similarity to $\boldsymbol{k}$. If there are multiple samples tied for maximum similarity, $\tilde{p}_k$ will become a uniform distribution over these samples. Assuming a unique nearest neighbor in $\mathcal{N}_K$, *i.e.*, $\boldsymbol{m}_* = \arg\max_{\boldsymbol{m}\in\mathcal{N}_K} \mathrm{sim}(\boldsymbol{k}, \boldsymbol{m})$, we have

$$\lim_{\tau_k \to 0^+} \mathcal{L}(\boldsymbol{q}, \boldsymbol{k}) = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_*)/\tau_q)}{\sum_{j=1}^{K} \exp(\mathrm{sim}(\boldsymbol{q}, \boldsymbol{m}_j)/\tau_q)}. \tag{5}$$

This recovers the standard InfoNCE objective over $\mathcal{N}_K$ with m_* as the positive sample [10].

More importantly, minimizing $\mathcal{L}(\boldsymbol{q}, \boldsymbol{k})$ is equivalent to maximizing a lower bound on the mutual information between $\boldsymbol{q}$ and the discrete neighborhood structure induced by $\boldsymbol{k}$. Define a random variable $J \in \{1, \ldots, K\}$ indexing samples in $\mathcal{N}_K$, with uniform prior $P(J = j) = 1/K$. Interpreting $\tilde{p}_{k,j} \approx P(J = j \mid \boldsymbol{k})$ and $\tilde{p}_{q,j} \approx P(J = j \mid \boldsymbol{q})$, the expected loss satisfies

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\boldsymbol{q}, \boldsymbol{k})] &= \mathbb{E}\Big[\mathrm{KL}\big(P(J \mid \boldsymbol{k}) \parallel P(J \mid \boldsymbol{q})\big)\Big] \\ &= I(J; \boldsymbol{k}) - I(J; \boldsymbol{q}) + \mathrm{const} \\ &= H(J) - I(J; \boldsymbol{q}) + \mathrm{const}, \end{aligned} \tag{6}$$

where the constant absorbs terms independent of the encoder parameters. Thus,

$$I(J; \boldsymbol{q}) = H(J) - \mathbb{E}[\mathcal{L}(\boldsymbol{q}, \boldsymbol{k})] + \mathrm{const}\,. \tag{7}$$

Since $H(J) = \log K$ is fixed, minimizing the soft alignment loss directly maximizes the mutual information $I(J; \boldsymbol{q})$ between the query $\boldsymbol{q}$ and the semantic neighborhood index J. As J encodes the high-confidence neighbors of $\boldsymbol{k}$, this forces $\boldsymbol{q}$ and $\boldsymbol{k}$ to align in their local semantic structures within the embedding space.

Consequently, our method provides a smooth, differentiable relaxation of hard nearest-neighbor contrastive learning that: (i) automatically filters noisy samples via top-K selection, (ii) maintains a tight theoretical connection to mutual information maximization, and (iii) gracefully reduces to InfoNCE in the low-temperature limit.

2. Experimental Setting Details

2.1. Datasets

- **NTU RGB+D (NTU-60)** [11] is a large-scale dataset containing 56,880 video samples with 60 action classes and 25 joints per skeleton. Captured from 40 subjects using three cameras, it includes both individual activities and two-person interactions.
- **NTU RGB+D 120 (NTU-120)** [7] extends NTU-60 and is the largest skeleton-based action recognition dataset, comprising 114,480 video samples with 120 action classes. It features data from 106 subjects across 32 distinct setups and multiple camera views.
- **PKU-MMD** [8] is a multi-modal dataset for 3D human action understanding, containing nearly 20,000 action instances across 51 categories, with 25 joints per skeleton. It consists of two parts: Part I offers an easier setup for action recognition, while Part II presents greater challenges due to view variations and increased skeleton noise.

2.2. Training Strategy

Linear Evaluation. The unsupervised setting evaluates the feature representation by a linear evaluation mechanism. The linear evaluation mechanism applies a linear classifier to the online encoder $f_q(\cdot)$ with frozen pretrained weights to classify the features extracted from it to evaluate the feature representation and utilizes the action recognition accuracy as a measure of the quality of the representation. We train for 100 epochs with learning rate set to 3 (multiplied by 0.1 at epoch 80).

Transfer Learning. To explore the generalization ability, we evaluate the performance of transfer learning. In transfer learning, we exploit self-supervised task pre-training on the source data. Then we utilize the linear evaluation mechanism to evaluate the performance on the target dataset. To evaluate the transferability of learned features, we pretrain on NTU X-sub dataset and performs linear evaluation on PKUMMD part I dataset and PKUMMD part II dataset. We train the classifier $\phi(\cdot)$ for 100 epochs with learning rate set to 3 (multiplied by 0.1 at epoch 80).

Skeleton-Based Action Retrieval. This protocol evaluates the semantic structure of the learned embedding space. For a query skeleton sequence, we extract its feature with the frozen pre-trained encoder, retrieve the nearest neighbor from the training set based on cosine similarity, and assign that neighbor's label to the query.

Calibration Analysis. Model calibration measures the alignment between a model's predicted confidence scores and its actual accuracy. This can be quantified by the Calibration Error [9], defined as:

$$\mathbb{E}_{(\boldsymbol{x}_i, y_i)\sim\mathcal{D}}\left[|P(\hat{y}_i = y_i \mid \hat{p}) - \hat{p}|\right] \tag{8}$$

To our knowledge, calibration remains underexplored in self-supervised human action recognition. We introduce it to evaluate the reliability of learned representations. Following the linear evaluation protocol, we compute the Expected Calibration Error (ECE) [2] and Adaptive ECE (AECE) [1] of the downstream classifier, hypothesizing that a better feature space yields a well-calibrated model. Specifically, we partition the predictions into M disjoint bins $\{B_m\}_{m=1}^{M}$. For each bin B_m, the accuracy $\text{acc}(B_m)$ and average confidence $\text{conf}(B_m)$ are computed as:

$$\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m}\mathbf{1}(\hat{y}_i = y_i), \quad \text{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m}\hat{p}_i, \tag{9}$$

where $\hat{p}_i$ is the predicted confidence for sample i and $\mathbf{1}(\cdot)$ is the indicator function. ECE is then calculated as the weighted average of the absolute difference between these two metrics:

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{10}$$

where N is the total number of samples. While ECE utilizes bins of uniform width, AECE adopts an adaptive binning strategy where bin boundaries are determined to ensure an equal number of samples in each bin (*i.e.*, $|B_m| \approx N/M$). In our experiments, we set the number of bins to $M = 15$.

Fusion Strategy. For all protocols evaluating our three-stream (i.e., Joint, Motion, and Bone) architecture, the final embedding is formed by concatenating the feature vectors from each stream. This combined vector is then passed through a final linear projection layer to produce the output for the specific downstream task.

Data Augmentation $\mathcal{T}$. For skeleton sequence, we employ *Shear* and *Crop* as our augmentation strategies.

Shear is a linear transformation applied to the spatial dimensions, which randomly slants the 3D joint coordinates. This transformation is defined by the matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix} \tag{11}$$

where a_{ij} are shear factors randomly sampled from a uniform distribution $U(-\beta, \beta)$, with β denoting the shear amplitude. Following SkeletonCLR [4], we set $\beta = 0.5$. The skeleton sequence is then multiplied by $\mathbf{A}$ along the channel dimension. The transformation $\mathbf{A}$ is then applied to the skeleton sequence along the spatial dimension.

Crop is an augmentation on the temporal dimension that symmetrically pads some frames to the sequence and then randomly crops it to the original length, which increases the diversity while maintaining the distinction of original samples. The padding length is defined as T/γ, where γ is the padding ratio and here we set $\gamma = 6$.

3. Task-Specific Evaluation Details

3.1. Robustness Analysis

Table 1. Comparison of transfer learning performance on PKU-MMD Part I, with models pretrained on NTU X-Sub benchmark.

Method	Transfer to PKU-MMD I	
	NTU-60	NTU-120
3s-AimCLR [3]	85.6	87.0
3s-ActCLR [5]	90.0	91.1
3s-ActCLR+ [6]	91.6	-
3s-TranCLR	**92.4**	**92.5**

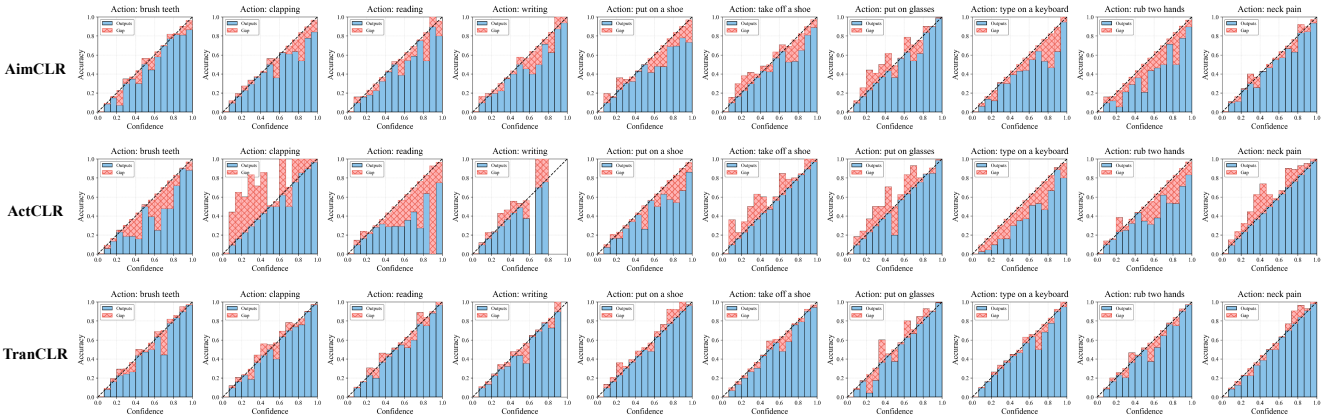

Figure 1. **Class-wise reliability diagrams for human action.** Compared to the binary contrastive objective, TranCLR reduces the calibration gap and shows improved reliability, particularly for fine-grained actions with ambiguous motion boundaries.

As shown in Tab. 1, TranCLR demonstrates exceptional robustness in transfer learning tasks. Pretrained on the NTU X-Sub benchmarks, the model achieves outstanding performance when applied to the PKU-MMD Part I dataset, with accuracies of 92.4% and 92.5% for NTU-60 and NTU-120, respectively. These results significantly outperform prior methods such as 3s-AimCLR and 3s-ActCLR, highlighting the superior domain adaptation capabilities of TranCLR. This robustness is attributed to the model's ability to generalize across different datasets, thanks to its multi-level geometric manifold calibration mechanism. By constructing a smooth and continuous action manifold, TranCLR effectively mitigates overfitting to the source dataset's specific characteristics, allowing it to transfer learned features to new, unseen domains. This ability to maintain strong performance across diverse datasets demonstrates TranCLR's robustness, making it highly adaptable to real-world scenarios where data distributions can vary significantly.

3.2. Confidence Calibration Analysis

To obtain a more detailed perspective on prediction reliability, we include class-wise reliability diagrams for ten representative action categories, as shown in Fig. 1. These categories cover both easily distinguishable motions such as "clapping" or "rub hands" as well as subtle and fine-grained actions including "reading" and "writing." Each diagram presents the relationship between predicted confidence on the horizontal axis and actual accuracy on the vertical axis, allowing a direct visual assessment of how well each model aligns probability estimates with true outcomes.

Across the majority of categories, TranCLR produces reliability curves that more closely follow the ideal diagonal identity line. This indicates that its predicted probabilities match the true correctness likelihood more faithfully across the full confidence spectrum, including the challenging mid-confidence regions where many contrastive learning methods tend to exhibit unstable or erratic behavior. Compared with binary contrastive objective, TranCLR shows a markedly reduced calibration gap, with misalignment substantially narrowed even in fine-grained action classes where motion boundaries are inherently ambiguous. The improved reliability achieved by TranCLR can be attributed to the continuous geometric structure shaped by the transitional anchor design and the multi-level manifold calibration mechanism. By regulating feature distances and encouraging smoother transitions between related actions, TranCLR avoids the abrupt decision boundaries that commonly lead to inflated confidence scores. As a result, the model is able to express uncertainty more appropriately when encountering ambiguous or transitional motion cues, producing probability estimates that remain stable and trustworthy across both simple and complex categories. The class-wise reliability diagrams provide strong qualitative evidence that TranCLR not only enhances recognition performance but also delivers significantly improved confidence calibration, reinforcing its usefulness in applications that require reliable uncertainty estimation.

4. Qualitative Analysis

4.1. Feature Distribution Visualization

Fig. 2 shows the global feature distribution achieved by the t-SNE method during the linear evaluation, using different loss calibrations on the NTU-60 dataset. We observe that the baseline model results in a relatively dispersed feature space, with significant overlap between categories, indicating that the model struggles to establish clear boundaries between action

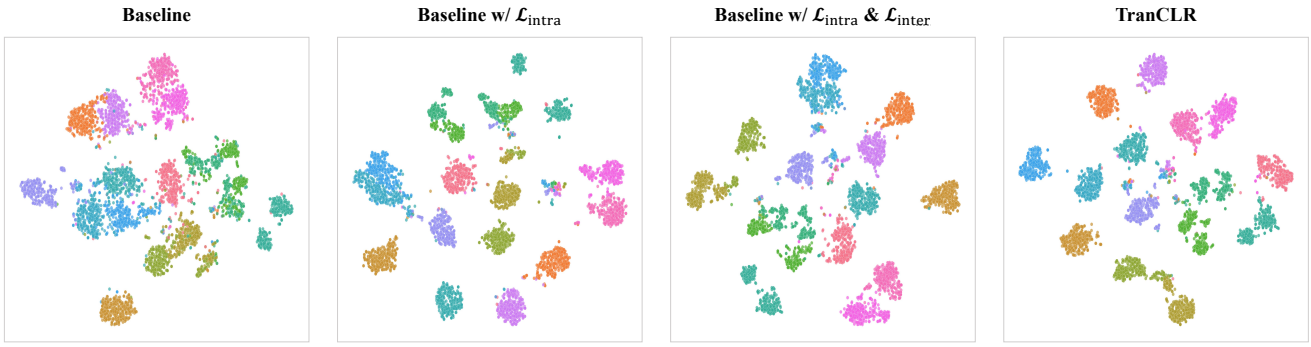

Figure 2. **Feature distribution visualization on NTU-60.** TranCLR achieves the most distinct action separation and coherent feature organization compared to baseline and other methods.

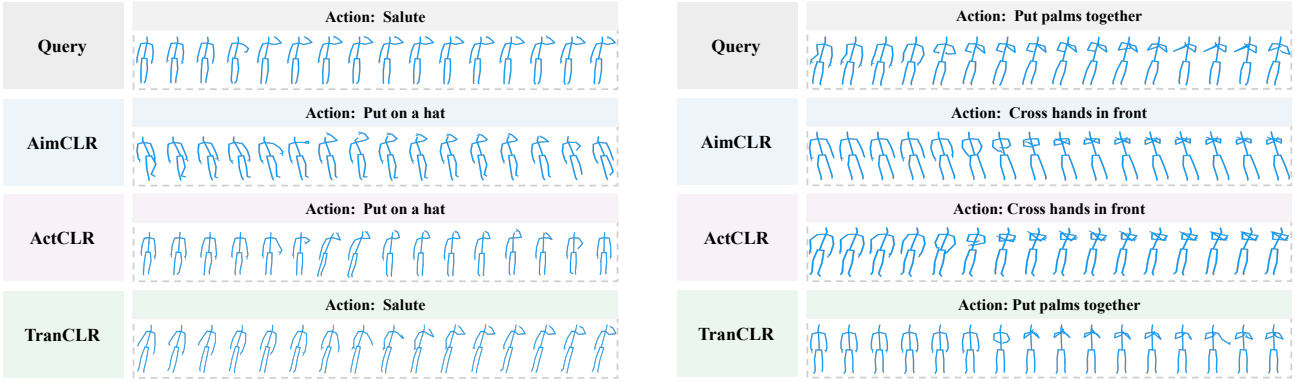

Figure 3. **Comparison of Top-1 retrieved action samples.** TranCLR consistently retrieves semantically coherent sequences, even with subtle motion differences, while traditional methods often produce less relevant results, indicating fragmented feature clustering.

classes. Introducing the $\mathcal{L}_{\text{intra}}$ loss improves the feature space structure, leading to tighter clustering within each class. However, there remains considerable overlap between different categories, suggesting that while $\mathcal{L}_{\text{intra}}$ helps reduce the distance between samples of the same class and improves intra-class consistency, its impact on inter-class separation is more limited. When both $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$ losses are incorporated, the feature space shows improved separation between categories, but overlap still exists, especially among finer action classes. This suggests that although the addition of these losses enhances feature representation, it may not fully capture subtle action distinctions.

In contrast, TranCLR achieves much clearer separation between categories, with actions becoming distinctly clustered into separate regions. While fine-grained actions continue to present some challenges in terms of full separation, the overall class boundaries are significantly more distinct compared to other methods. TranCLR's ability to organize actions along a continuous manifold, coupled with its innovative anchor generation and multi-level geometric calibration, contributes to this superior feature organization. Overall, the t-SNE visualizations demonstrate the clear superiority of TranCLR in constructing a more discriminative and coherent feature space, highlighting the advantages of our method in action recognition tasks.

4.2. Action Retrieval Sample

To qualitatively illustrate the effectiveness of TranCLR, we visualize retrieved skeleton sequences for a set of query actions, comparing them with top-1 results from previous contrastive learning methods. As shown in Fig. 3, TranCLR consistently retrieves actions that are semantically more coherent and closely aligned with the query, even when subtle differences exist in motion style or execution speed. This confirms that our learned embedding space preserves fine-grained motion characteristics while maintaining smooth transitions across similar actions. In contrast, baseline methods occasionally retrieve visually or semantically less relevant sequences, reflecting fragmented or overly rigid feature clusters. These qualitative results complement our quantitative findings, highlighting that the MGMC mechanism effectively balances discriminability with topological continuity, producing a feature manifold where related actions are meaningfully grouped while preserving generalization to diverse motion patterns.

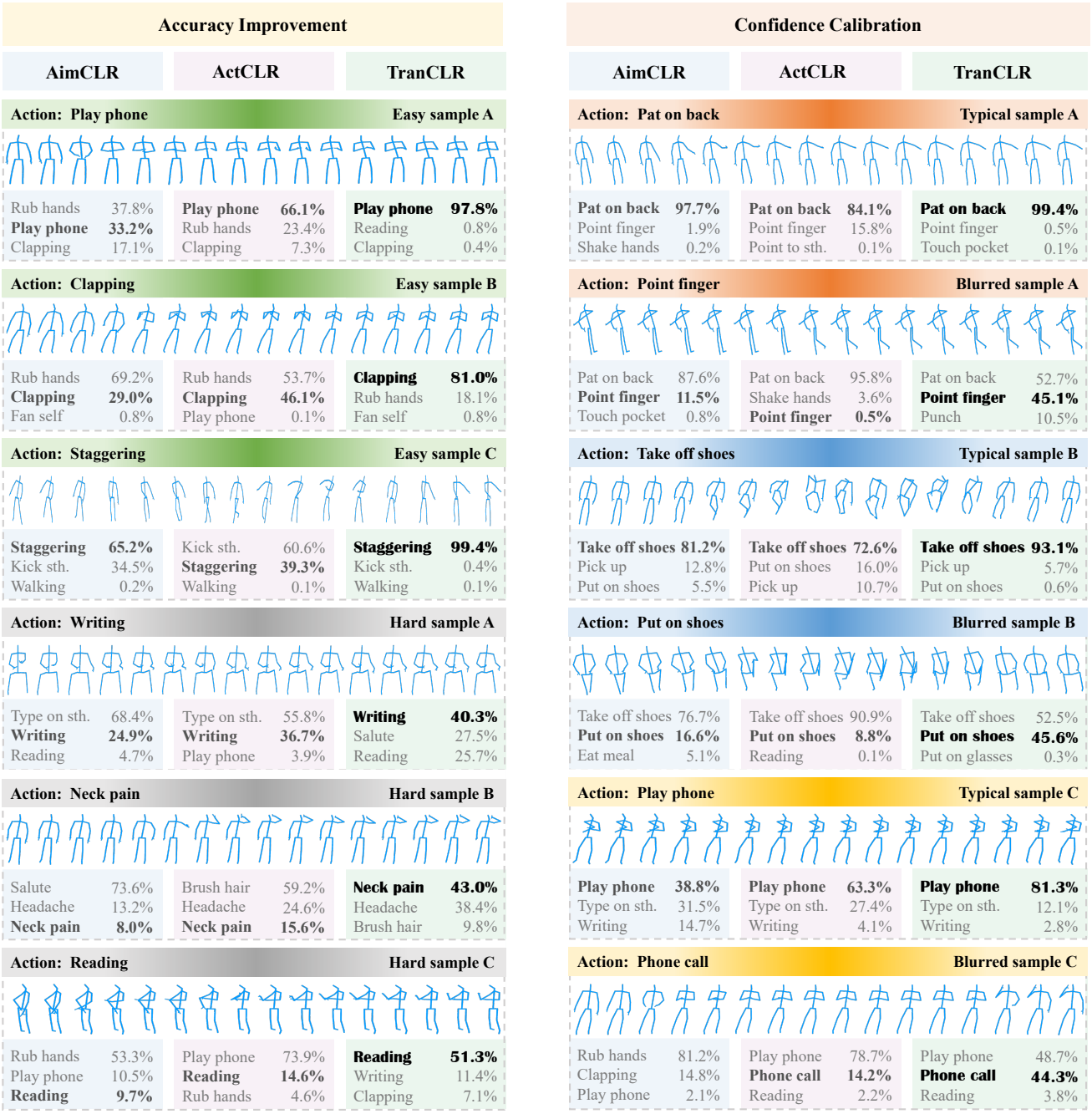

Figure 4. **Visualization of difficulty-aware action examples.** Our method achieves higher accuracy on challenging samples and better-calibrated confidence across diverse motion scenarios. The top-3 predictions with confidence scores are shown for each sample.

4.3. Difficulty-Aware Visualization of Model Performance and Confidence Calibration

To further examine model behavior under varying levels of difficulty, we visualize various representative action samples and categorize them into four groups: easy samples, hard samples, typical samples, and blurred samples. As shown in Fig. 4, we compare TranCLR with mainstream contrastive learning baseline models AimCLR and ActCLR. This difficulty-aware visualization reveals clear distinctions in accuracy and confidence calibration across different action scenarios.

For easy samples, TranCLR consistently yields the highest and most stable top-1 confidence. In Easy Sample B "Clapping", both AimCLR and ActCLR misclassify the action as "Rub hands", assigning only 25–50% confidence to the correct

class. In contrast, TranCLR produces an 81.0% confidence score for "Clapping", indicating decisive and accurate prediction on unambiguous motions. Similar patterns are observed in "Play phone" and "Staggering", where TranCLR maintains high confidence and accuracy for simple actions. For hard samples, conventional contrastive models often produce highly confident yet incorrect predictions. In Hard Sample B, AimCLR misclassifies "Neck Pain" as "Salute" with 73.6% confidence. In contrast, TranCLR correctly identifies the action while assigning a more moderate confidence level (around 40–50%), reflecting a cautiously correct behavior. It is crucial for actions with subtle and hard motion cues. Similar observations arise in the "Writing" and "Reading" examples: whereas other models output strongly confident misclassifications, TranCLR maintains reasonable confidence and accurate recognition.

Beyond easy and hard cases, the typical–blurred sample pairs provide a more sensitive probe for examining confidence calibration, as they consist of semantically related actions that differ mainly in their level of visual ambiguity. The Typical samples correspond to actions that are generally easy to recognize. In these cases, all models correctly predict the ground-truth class, yet TranCLR provides the most calibrated and appropriately high confidence. For instance, in Typical Sample C "Play phone", baseline models produce noticeably flatter top-3 distributions and sometimes assign nontrivial confidence to unrelated classes. TranCLR, by contrast, yields a well-structured probability hierarchy: the correct label dominates with high confidence, while secondary classes receive clearly lower scores. This indicates that TranCLR is not only confident but also correctly aligned with the true likelihood. Blurred samples share motion primitives with their typical counterparts but introduce greater visual ambiguity. While most models still include the correct label within their top-2 predictions, their confidence structures diverge significantly. AimCLR and ActCLR often display confidence collapse, assigning inflated probability to incorrect or weakly related actions and producing unstable decision boundaries. In contrast, TranCLR maintains a coherent confidence profile, consistently assigning the higher probability to the true class and distributing residual probabilities smoothly among semantically related alternatives. Even under visual degradation, TranCLR preserves stable semantic relationships and avoids the brittle, discontinuous behavior characteristic of binary contrastive models.

Collectively, these difficulty-aware examples demonstrate the superior accuracy improvement and confidence calibration achieved by TranCLR. Unlike contrastive baselines that suffer from overconfidence, fragmented intra-class clusters, and brittle decision boundaries, TranCLR remains confidently correct on easy samples, reliably accurate on hard samples, and semantically consistent across typical–blurred action pairs. This coherence enables TranCLR to better preserve the continuous nature of human motion, resulting in more reliable predictions and more trustworthy confidence estimates across a wide range of real-world scenarios.

5. Ablation Study Details

Table 2. Effectiveness of the Soft Alignment Strategy.

Alignment Strategy	**NTU-60**		
	X-Sub	X-View	Avg.
Baseline	74.9	79.9	77.4
InfoNCE (hard)	80.9	85.2	83.1
Soft Alignment (Ours)	**83.8**	**87.9**	**85.9**

Impact of Soft Alignment. As discussed in Section 3.4 of the main text, simultaneously optimizing intra-sample compactness ($\mathcal{L}_{\text{intra}}$) and inter-sample continuity ($\mathcal{L}_{\text{inter}}$), along with other training objectives, may introduce conflicting gradients that lead to instability when using rigid targets. Tab. 2 validates our proposed solution. Replacing the standard hard InfoNCE loss with our Soft Alignment yields a significant performance boost (*e.g.*, $+2.8\%$ on average for NTU-60). This improvement confirms that softening the target distributions via knowledge distillation allows the model to flexibly accommodate both local compactness and manifold smoothness, effectively resolving the tension between the multi-level calibration objectives.

Evaluation of Compositional Similarity Metrics. In the Cross-Anchor Relational Consistency module, the weighting coefficient λ_{cross} is critical for accurately reflecting the geometric overlap between anchors. We compare three strategies in Tab. 3: a static weight, a geometric mean approximation, and our proposed minimum operation. The Minimum strategy achieves the best average performance (+1.0% over Fixed and +0.6% over Geometric Mean on NTU-60). This aligns with our theoretical derivation in Section 3.3 of the main text, where the $\min(\cdot)$ operator mathematically represents the exact intersection of parentage in the compositional space, providing the most precise supervisory signal for refining global manifold consistency.

Table 3. Ablation on the Compositional Similarity Metric.

Composition Metric	NTU-60		
	X-Sub	X-View	Avg.
Baseline	74.9	79.9	77.4
Fixed ($\lambda_{\text{cross}} = 0.5$)	83.1	86.6	84.9
Geometric Mean ($\sqrt{\cdot}$)	83.3	87.3	85.3
Minimum (Ours)	**83.8**	**87.9**	**85.9**

6. Limitation and Future Works

While TranCLR demonstrates strong performance across accuracy, generalization, and calibration, there are several limitations to consider. First, although our method has been validated on common human motion datasets, its generalization and robustness could benefit from evaluation on an even wider range of datasets, since transitional anchor generation relies on action pairs from the pre-training data. Second, the current transitional anchors are not physically precise motions; they primarily serve as manifold regularizers to enrich the latent topology. Developing more realistic transition anchors without increasing computational cost could further enhance the model's ability to capture action continuity beyond binary contrasts. These considerations suggest promising directions for future research, enabling the continued advancement of self-supervised skeleton-based action representation learning.

One potential avenue is the integration of multi-modal data, such as RGB video and depth, to enrich motion representations and leverage cross-modal correlations for more robust and comprehensive embeddings. Another direction is the development of physically grounded or generative transitional anchors, which could incorporate kinematic constraints or motion priors to produce more realistic intermediate states and improve interpretability beyond interpolation and substitution. Exploring adaptive and dynamic topology learning is also promising, where the density and placement of transitional anchors are modulated according to local manifold complexity or uncertainty, enabling better modeling of highly nonlinear or abrupt actions. Additionally, optimizing the computational efficiency of the multi-level geometric manifold calibration could facilitate real-time deployment and resource-constrained applications, including edge devices or interactive systems. By pursuing these directions, future work can enhance the realism, generalization, and adaptability of self-supervised skeleton-based representation learning, ultimately contributing to broader and more effective applications in human-computer interaction, robotics and motion analysis.

References

[1] Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *CVPRW*, 2020. 3

[2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. 2017. 3

[3] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. *AAAI*, 2022. 3

[4] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3D human action representation learning via cross-view consistency pursuit. In *CVPR*, 2021. 3

[5] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *CVPR*, 2023. 3

[6] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Self-supervised skeleton representation learning via actionlet contrast and reconstruct. 2025. 3

[7] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. 2020. 2

[8] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM MM*, 2020. 2

[9] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015. 3

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 1, 2

[11] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016. 2