Robust Pose-Sequence Retrieval with Contextual Metric Learning

Literature Review and Feasibility Assessment

Giacomo Cappelletto
Prepared for Professor Brian Kulis

July 10, 2026

Executive Summary

I propose to test whether contextual metric learning can make pose-sequence retrieval more robust to imperfect skeleton inputs. The literature supports proceeding, with an important qualification: contextual loss has been shown to resist label noise and overfitting in image retrieval, but that result does not imply robustness to corrupted pose coordinates. At the same time, recent work has separately established (i) skeleton action retrieval on NTU RGB+D, and (ii) robust action representations under joint noise, occlusion, frame masking, and pose-estimator shift. The open and falsifiable question is whether a neighborhood-based retrieval objective adds robustness beyond a strong temporal encoder and corruption-aware training.

The project is feasible because the public MotionBERT one-shot pipeline already provides normalized sequence embeddings, class-balanced sampling, supervised contrastive training, and cosine nearest-neighbor evaluation. The missing pieces are bounded: expose the contextual neighborhood parameter, implement a proper query/gallery evaluator, add reproducible pose corruptions, and compare losses under matched conditions. I recommend a staged study on NTU120: first validate the retrieval and corruption protocol with frozen features, then compare contextual and standard objectives, and only then add full fine-tuning and a MaskCLR-style robustness baseline. A positive result would identify a useful retrieval objective; a null result would still distinguish the value of neighborhood learning from the value of robust augmentation and encoder design.

1 Research Proposition

Primary question.

> Does contextual neighborhood optimization preserve pose-sequence retrieval quality better than standard metric-learning objectives as skeleton inputs become noisy, incomplete, or distribution-shifted?

I would frame the project as retrieval research rather than as action classification or sports analysis. A pose sequence x is mapped to a normalized embedding $z = f_\theta(x)/\|f_\theta(x)\|_2$; a query is useful when relevant motions rank ahead of irrelevant motions under cosine similarity. In the first study, relevance can be defined by action class so that the protocol is objective and reproducible. Fine-grained technique, rehabilitation status, or movement quality would require a different dataset and expert labels, and should not be claimed from NTU class retrieval alone.

The proposed contribution is deliberately narrow:

- a retrieval-first robustness benchmark for temporal skeleton embeddings;
- a controlled transfer of contextual loss from images to pose sequences; and

- an analysis that separates the effect of the loss from the effects of the encoder and corruption augmentation.

It is not a claim of a new motion architecture. This is a strength for an initial project: the central hypothesis can be rejected without first solving dataset construction, pose estimation, or model design.

Why the result would matter. Classification and retrieval ask different questions. A classifier can retain acceptable top-1 accuracy while the ordering of its embedding space deteriorates, and it cannot return examples that support inspection or few-shot decisions. Robust retrieval is useful for example-based motion search, one-shot action recognition, dataset auditing, and later domain-specific systems. The project also tests a broader scientific hypothesis: whether local neighborhood structure provides a useful inductive bias when observations, rather than labels, are corrupted.

2 What the Literature Establishes

2.1 Contextual metric learning

Liao et al. (2023) introduce a contextual loss for supervised image retrieval. Given a balanced mini-batch of normalized embeddings, the method combines cosine similarity with shared and reciprocal neighborhood structure, and optimizes the resulting contextual similarity against label agreement. Their full objective includes contextual, contrastive, and similarity-regularization terms. The neighborhood size k is not arbitrary: with M samples per class in a balanced batch, the analysis sets $k = M$. The contextual calculation has cubic time and space complexity in batch size, although it is expressed through highly optimized matrix operations and was practical at the batch sizes studied in the paper.

The paper reports reduced overfitting and improved robustness to *label noise*. That evidence motivates the transfer, but it does not establish robustness to joint jitter, missing frames, or pose-estimator shift. Input corruption can move an entire neighborhood coherently or destroy it, so improvement on noisy skeletons is a hypothesis to test rather than a consequence of the image result.

2.2 Pose and skeleton retrieval

Pose embedding is already an established idea. Earlier work learns embeddings for matching static poses (Mori et al., 2015), while VIPE and Pr-VIPE learn view-invariant probabilistic pose spaces and explicitly study cross-view retrieval and partial keypoints (Sun et al., 2019; Liu et al., 2020b). Normalized pose features further show that anthropometry and viewpoint can be factored out before metric learning for video alignment and pose-similarity tasks (Liu et al., 2021). These papers make an important design point: nuisance normalization and corruption robustness should be evaluated separately.

At the sequence level, SL-DML and Skeleton-DML formulate one-shot action recognition as nearest-neighbor search in a learned skeleton embedding space (Memmesheimer et al., 2020, 2022). Self-supervised skeleton representation learning also treats action retrieval as a downstream task; representative examples include HiCo (Dong et al., 2023) and the recent heterogeneous-skeleton framework of Wang et al. (2025), which reports cosine-based retrieval on NTU60 and NTU120. MotionBERT supplies a strong temporal encoder and a public one-shot NTU120 embedding pipeline (Zhu et al., 2023; Zhu and collaborators, nd). Thus, sequence retrieval itself is not the novelty. The relevant question is retrieval *under controlled input corruption*, with the training objective isolated.

2.3 Robust skeleton representations

Robustness to imperfect skeletons is also well motivated. RA-GCN and PR-GCN address incomplete or noisy skeletons in action recognition (Song et al., 2020; Shi et al., 2020); PoseC3D reports improved robustness to pose-estimation noise and cross-dataset transfer by using heatmap volumes rather than raw coordinate graphs (Duan et al., 2022). Most importantly, MaskCLR uses attention-guided masking and multi-level contrastive learning to improve transformer representations under Gaussian joint noise, body-part and joint occlusion, frame masking, and a change of pose estimator (Abdelfattah et al., 2024). It evaluates top-1 classification rather than retrieval, but it is the closest robustness-specific comparator and should be treated as such.

2.4 Position of the proposed study

Table 1 states the gap conservatively. I did not identify a study that holds a temporal pose encoder and evaluation protocol fixed while testing whether the contextual objective of Liao et al. (2023) improves retrieval degradation curves under pose corruption. This is a bounded literature finding, not proof that no such work exists; it should be rechecked before any publication-level novelty claim.

Table 1: Closest prior work and the remaining experimental question.

Line of work	What it establishes	What remains for this project
Contextual metric learning (Liao et al., 2023)	Neighborhood-aware ranking can improve image retrieval and resist label noise.	Does the objective help when temporal pose inputs, rather than labels, are corrupted?
Skeleton metric learning (Memmesheimer et al., 2022)	Skeleton sequences can support one-shot nearest-neighbor recognition.	Full ranking metrics and controlled robustness curves are not the focus.
Skeleton action retrieval (Dong et al., 2023; Wang et al., 2025)	NTU60/120 embeddings are routinely evaluated by cosine retrieval.	The effect of pose corruption and a supervised contextual objective is not isolated.
Robust action representations (Duan et al., 2022; Abdelfattah et al., 2024)	Noise, occlusion, frame loss, and pose-source shift materially affect skeleton models and can be mitigated.	Evaluation is primarily classification; retrieval ordering may behave differently.
Pose invariance (Liu et al., 2020b, 2021)	Viewpoint, occlusion, and body shape can be handled through probabilistic embeddings or normalization.	Static-pose retrieval and alignment do not answer class-level temporal retrieval under corruption.
Motion-language retrieval (Messina et al., 2023; Bensabath et al., 2024)	Modern motion retrieval increasingly aligns motion with language across datasets.	Cross-modal semantic retrieval is a different task and should be a later extension, not the first experiment.

3 Proposed Study

3.1 Dataset and retrieval protocol

NTU RGB+D 120 is the most defensible first benchmark: it contains 114,480 samples across 120 actions with official cross-subject and cross-setup protocols (Liu et al., 2020a). The MotionBERT action pipeline uses 2D HRNet detections distributed through the OpenMMLab skeleton tooling, so it tests the intended setting of estimated rather than perfect pose. NTU60 is an acceptable plumbing pilot, but NTU120 offers more classes and direct compatibility with the one-shot code and prior metric-learning work.

The existing MotionBERT one-shot protocol has one reference example for each of 20 held-out action classes and reports cosine 1-nearest-neighbor accuracy. It is a useful smoke test for loss integration, but a one-reference gallery cannot support a meaningful full-ranking mAP study. The headline experiment should therefore use a fixed, disjoint query/gallery construction from the held-out portion of the official cross-subject and cross-setup splits, with multiple relevant gallery sequences per action. The split manifest should be saved, the query clip must never appear in the gallery, and model selection must use a separate validation partition.

The primary deployment model is a noisy query against a clean reference archive. The main evaluation should corrupt only the query and keep the gallery fixed. A secondary experiment can corrupt both query and gallery to model an archive built from estimated poses.

3.2 Models, objectives, and controls

MotionBERT is the lowest-risk first encoder because its public code already contains an `ActionHeadEmbed` that returns normalized embeddings, a class-balanced `MPerClassSampler`, supervised contrastive training, and cosine one-shot validation. The contextual repository supplies the loss-side reference implementation (Liao, nd). The 224-pixel path passes k from the samples-per-class setting, while the 256-pixel path hard-codes $k = 4$; the adaptation should expose $k = M$ explicitly in one tested implementation.

The core comparison is:

1. cross-entropy with the normalized penultimate feature;
2. supervised contrastive loss;
3. multi-similarity or triplet loss as a standard metric-learning control;
4. contextual loss;
5. the contextual-plus-contrastive hybrid; and
6. a MaskCLR-style corruption-aware contrastive model using the same backbone.

Two controls are essential. First, every loss must use the same data split, encoder initialization, embedding dimension, batch composition, augmentation policy, tuning budget, and evaluator (Musgrave et al., 2020). Second, loss and augmentation must be separated in a factorial comparison: train each central objective both without corruption augmentation and with the same masking/jitter augmentation. Otherwise, a robustness gain could be incorrectly attributed to contextual learning when it came from seeing corrupted inputs during training.

I would start with a frozen MotionBERT encoder and train only the projection head. This validates the sampler, loss, evaluator, and corruption code cheaply. Full encoder fine-tuning is justified only after the comparison is stable.

3.3 Corruption suite

The suite should begin from published protocols rather than an arbitrary list. MaskCLR varies Gaussian noise, removes specific body parts, masks increasing fractions of joints, masks frames, and changes the pose estimator (Abdelfattah et al., 2024). I propose the following primary corruptions:

- Gaussian coordinate noise, calibrated after normalization to body scale;
- random joint masking at several severities and complete limb/body-part occlusion;
- random and contiguous frame dropout, plus temporal jitter; and
- confidence masking, with pose-estimator shift included when alternate predictions are available.

Each corruption should be deterministic given a stored seed, applied after the same preprocessing stage, and visually checked at every severity to confirm that the action remains recognizable. Left-right swaps and large geometric transformations should be secondary ablations because they can change the action semantics rather than merely degrade observation quality.

3.4 Metrics and interpretation

For each clean or corrupted query, rank the fixed gallery by cosine similarity and report Recall@1/5/10, mAP, mAP@R or R-precision, and the area under the performance-versus-corruption curve. Report both absolute performance and clean-to-corrupted degradation. Per-class results are important because some actions depend on a small set of joints and may fail differently from whole-body actions.

Results should be averaged over at least three training seeds, with confidence intervals over queries or seeds. Hyperparameters and stopping points must be selected on validation data, not on the test corruption curves. The study succeeds if it produces an interpretable comparison, not only if contextual loss wins:

- **robust gain with similar clean retrieval:** evidence that contextual neighborhoods help under input noise;
- **clean gain but no robustness gain:** evidence for better ranking, but not for the robustness hypothesis;
- **gain only with corruption augmentation:** neighborhood learning and augmentation are complementary; and
- **no gain over SupCon or MaskCLR:** evidence that robust data exposure or encoder design matters more than the contextual objective.

4 Feasibility, Risks, and Milestones

I recommend three go/no-go milestones:

1. **Protocol validation:** reproduce MotionBERT one-shot accuracy, create the held-out query/gallery split, and obtain monotone corruption curves from frozen embeddings. Stop and repair the protocol if relevance or severity is unstable.
2. **Loss isolation:** train matched projection heads with SupCon, contextual, and hybrid objectives; verify that k, M, and batch composition agree; report clean and corrupted retrieval with fixed seeds.

3. **Robustness claim:** fine-tune the encoder only if milestone 2 is stable, add the corruption-aware MaskCLR comparison, and run the full controlled ablation. A second backbone such as PoseC3D or ST-GCN++ is validation, not a prerequisite.

Table 2: Feasibility evidence and remaining risks.

Component	Evidence already available	Residual work or risk
Data	NTU120 has standard protocols; MotionBERT documents preprocessed HRNet skeleton files.	Confirm access and licensing before implementation; save an explicit query/gallery manifest.
Encoder	MotionBERT exposes a pretrained temporal representation and normalized action-embedding head.	Its documented environment targets an older Python/CUDA stack; pin a reproducible environment before modifying code.
Loss	The contextual implementation is compact and encoder-agnostic once it receives normalized embeddings and labels.	Make $k = M$ explicit, add shape/gradient tests, and preserve balanced batches.
Evaluator	The one-shot script already computes cosine 1-nearest-neighbor accuracy.	Build Recall/mAP evaluation with multiple positives, self-match exclusion, and fixed clean/noisy query handling.
Robustness	MaskCLR provides published noise, occlusion, frame-masking, and pose-source protocols.	Normalize severity carefully and separate loss effects from augmentation effects.
Compute	A frozen-backbone projection-head pilot is inexpensive relative to raw-video training.	The contextual term scales as $O(B^3)$ in batch size. Profile first; reduce classes per batch while preserving M. Gradient accumulation does not recreate within-batch neighborhoods.

Sports-specific validation should come after these milestones. SportsPose offers dynamic 3D athletic poses (Ingwersen et al., 2023), and SU-EMD demonstrates one-shot skeleton metric learning across marker-based and markerless strength-exercise data (Deyzel and Theart, 2023). These sources support eventual application value, but they do not yet provide labels for technique quality or rowing faults. Moving to such claims would require domain-specific annotations and a separate validation study.

5 Assessment and Questions for Discussion

The proposed study is both feasible and scientifically useful if it remains a controlled retrieval experiment. Its strongest feature is not guaranteed superiority of contextual loss; it is the ability to distinguish three possible sources of robustness–neighborhood-aware optimization, corruption-aware training, and encoder choice–using public data and largely existing code. The project also has a clear stopping point: a reproducible NTU benchmark and a fair loss comparison are complete contributions even if sports-domain transfer is deferred.

The main conceptual risk is that action-class relevance is coarse. Contextual neighborhoods may improve class retrieval without capturing subtle movement quality. I would therefore seek guidance on two decisions before implementation:

1. Is class-based NTU retrieval an appropriate first scientific target, or should relevance be continuous or fine-grained from the outset?
2. Should the first deliverable be a focused single-backbone comparison, with MaskCLR as the robustness control, or a broader benchmark?

My recommendation is to proceed with the narrow version. It has a credible literature gap, a minimal implementation path, clear negative controls, and a result that remains informative whether or not contextual loss improves robustness.

References

Abdelfattah, M., Hassan, M., and Alahi, A. (2024). MaskCLR: Attention-guided contrastive learning for robust action representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18678–18687.

Bensabath, L., Petrovich, M., and Varol, G. (2024). A cross-dataset study for text-based 3d human motion retrieval. *arXiv preprint arXiv:2405.16909*.

Deyzel, M. and Theart, R. P. (2023). One-shot skeleton-based action recognition on strength and conditioning exercises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5169–5178.

Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., and Wang, X. (2023). Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 525–533.

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

Ingwersen, C. K., Mikkelstrup, C. M., Jensen, J. N., Hannemose, M. R., and Dahl, A. B. (2023). Sportspose: A dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5219–5228.

Liao, C. (n.d.). metric-learning-using-contextual-similarity. `https://github.com/Chris210634/metric-learning-using-contextual-similarity`. Accessed 2026-07-10.

Liao, C., Tsiligkaridis, T., and Kulis, B. (2023). Supervised metric learning to rank for retrieval via contextual similarity optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21012–21033.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. (2020a). Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701.

Liu, J., Shi, M., Chen, Q., Fu, H., and Tai, C.-L. (2021). Normalized human pose features for human action video alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11521–11531.

Liu, T., Sun, J. J., Zhao, L., Zhao, J., Yuan, L., Wang, Y., Chen, L.-C., Schroff, F., and Adam, H. (2020b). View-invariant, occlusion-robust probabilistic embedding for human pose. *arXiv preprint arXiv:2010.13321*.

Memmesheimer, R., Häring, S., Theisen, N., and Paulus, D. (2022). Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3702–3710.

Memmesheimer, R., Yang, H., Pfeiffer, S., and Paulus, D. (2020). Sl-dml: Signal-to-image representation and deep metric learning for skeleton-based action recognition. *arXiv preprint arXiv:2004.11085*.

Messina, N., Sedmidubsky, J., Falchi, F., and Rebok, T. (2023). Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2420–2425.

Mori, G., Pantofaru, C., Kothari, N., Leung, T., Toderici, G., Toshev, A., and Yang, W. (2015). Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*.

Musgrave, K., Belongie, S., and Lim, S.-N. (2020). A metric learning reality check. *arXiv preprint arXiv:2003.08505*.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2020). Pose refinement graph convolutional network for skeleton-based action recognition. *arXiv preprint arXiv:2010.07367*.

Song, Y.-F., Zhang, Z., He, X., Cheng, J., and Lu, H. (2020). Richly activated graph convolutional network for robust skeleton-based action recognition. *arXiv preprint arXiv:2008.03791*.

Sun, J. J., Zhao, J., Chen, L.-C., Schroff, F., Adam, H., and Liu, T. (2019). View-invariant probabilistic embedding for human pose. *arXiv preprint arXiv:1912.01001*.

Wang, H., Ma, X., Kuang, J., and Gui, J. (2025). Heterogeneous skeleton-based action representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19154–19164.

Zhu, W. and collaborators (n.d.). Motionbert. `https://github.com/Walter0807/MotionBERT`. Accessed 2026-07-06.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., and Wang, Y. (2023). Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099.