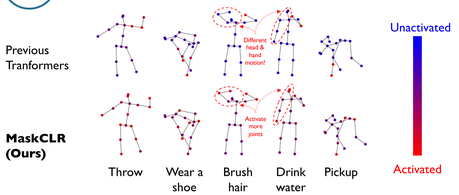
# MaskCLR: Attention-Guided Contrastive Learning for Robust Action Representation Learning

Mohamed Abdelfattah*, Mariam Hassan, Alexandre Alahi

EPFL VITA

CVPR JUNE 17-21, 2024 SEATTLE, WA

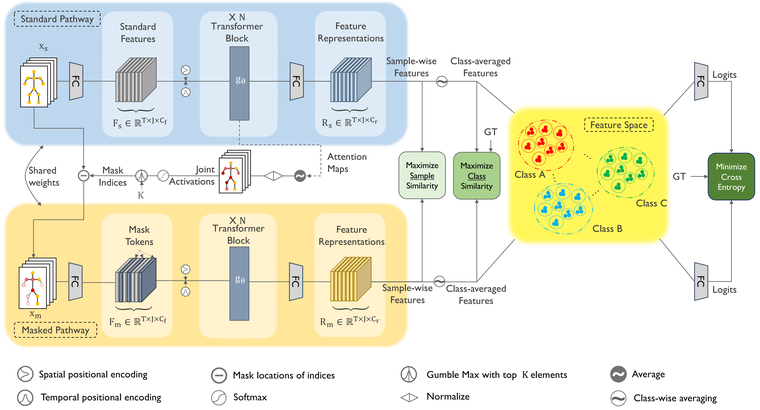
## 01 OVERVIEW



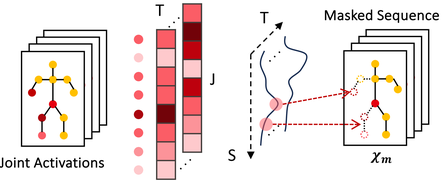

- **Task:** Predicting action class from input skeleton sequence
- **Challenge:** over-reliance on small number of activated joints in transformer self-attention.
- **Approach:** expand the set of discriminative joints with attention-guided masking and contrastive learning.

## 02 METHOD

- Mask the activated joints, re-feed the rest to model.
- Contrast standard and masked representations
- Contrast class representations




## 03 ATTENTION-GUIDED MASKING



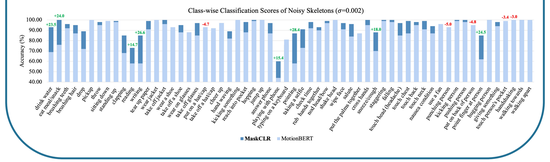

Attention Scores — Probablistic Masking

We mask the most important joints to force the model to explore more discriminative joints.

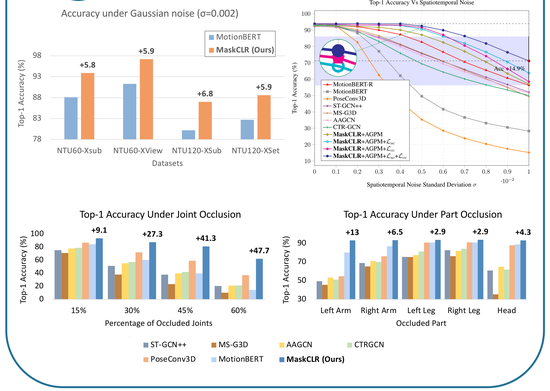
firstname.lastname@epfl.ch
* corresponding author

## 04 RESULTS: MORE ACCURATE MODEL

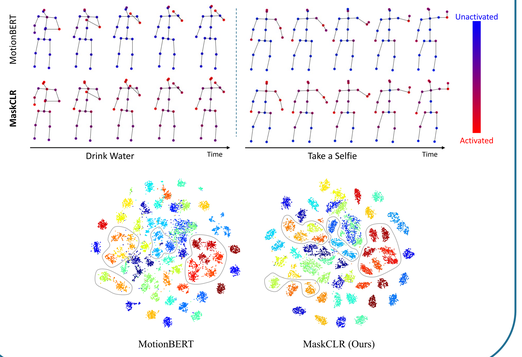
| Method | Backbone | NTU60 XSub | NTU60 XView | NTU120 XSub | NTU120 XSet | Kinetics400 |
|---|---|---|---|---|---|---|
| AimCLR | STTFormer | 83.9 | 90.4 | 74.6 | 77.2 | - |
| CrosSCLR | STTFormer | 84.6 | 90.5 | 75.0 | 77.9 | - |
| SkeletonMAE | STTFormer | 86.6 | 92.9 | 76.8 | 79.1 | - |
| MaskCLR | STTFormer | **90.1** (↑3.5) | **95.4** (↑2.5) | **79.0** (↑2.2) | **80.5** (↑1.4) | - |
| SkeletonMAE | Transformer | 88.5 | 94.7 | 87.0 | 88.9 | - |
| MAMP | Transformer | 93.1 | 97.5 | 90.0 | 91.3 | - |
| MaskCLR | Transformer | **93.5** (↑0.4) | **97.5** | **90.5** (↑0.5) | **91.9** (↑0.6) | - |
| MotionBERT | DSTFormer | 92.8 | 97.1 | 84.8 | 86.4 | 38.8 |
| MaskCLR | DSTFormer | **93.9** (↑1.1) | **97.3** (↑ 0.2) | **87.4** (↑2.6) | **89.5** (↑3.1) | **44.7** (↑5.8) |




## 05 RESULTS: MORE ROBUST MODEL




## 06 QUALITATIVE RESULTS



Website