Grounding Foundational Vision Models with 3D Human Poses for Robust Action Recognition

Nicholas Babey
Arizona State University
nbabey@asu.edu

Tiffany Gu
Emory University
tiffany.gu2@emory.edu

Yiheng Li
University of California, Berkeley
leo3219@berkeley.edu

Cristian Meo
LatentWorlds AI
Delft University of Technology
Algoverse AI Research

Kevin Zhu
Algoverse AI Research
kevin@algoverse.us

Abstract

For embodied agents to effectively understand and interact within the world around them, they require a nuanced comprehension of human actions grounded in physical space. Current action recognition models, often relying on RGB video, learn superficial correlations between patterns and action labels, so they struggle to capture underlying physical interaction dynamics and human poses in complex scenes. We propose a model architecture that grounds action recognition in physical space by fusing two powerful, complementary representations: V-JEPA 2's contextual, predictive world dynamics and CoMotion's explicit, occlusion-tolerant human pose data. Our model is validated on both the InHARD and UCF-19-Y-OCC benchmarks for general action recognition and high-occlusion action recognition, respectively. Our model outperforms three other baselines, especially within complex, occlusive scenes. Our findings emphasize a need for action recognition to be supported by spatial understanding instead of statistical pattern recognition.

1 Introduction

The ability to recognize and understand human actions is the cornerstone of embodied AI, enabling applications from collaborative robotics to assistive technologies [1, 2, 3]. However, understanding requires not only labeling an activity but also grounding the action in the occupied 3D space [4, 5, 6, 7]. A model must differentiate between "giving a high-five" and "reaching for an object," actions that might appear visually similar but are defined by distinct spatial and postural dynamics [8].

Current action recognition approaches are primarily divided into two dominant methodologies: RGB-based models that utilize video pixel data to capture contextual appearance information and skeleton-based models that rely on human 3D joint data to capture movement dynamics [9, 10, 11, 12, 13, 14, 15]. Self-supervised video models, like V-JEPA 2 [16], have demonstrated prowess in understanding and predicting world states from visual data alone [17]. However, these RGB video models learn spatial relationships from pixel presentations, so their understanding is limited in occluded scenes where key limbs are hidden from view [18, 19]. Conversely, skeleton-based models like CoMotion [20], provide explicit representations of human posture by tracking detailed 3D skeletons through visual noise and occlusion. Yet, this approach is limited by its lack of rich, contextual information like environmental cues and human-object interactions [21, 22]. Due to the limitations of these methodologies, neither approach alone is sufficient for a spatially-grounded understanding of human action.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SPACE in Vision, Language, and Embodied AI.

The limitations of these two paradigms and their complementary nature emphasizes the need for a unified model architecture that pairs contextual video understanding with precise geometric skeletons. This project is motivated by the hypothesis that fusing these complementary data streams will lead to a more robust and accurate action recognition system. This grounded understanding of how a body is configured and acting within a 3D space is critical for any embodied agent that must navigate human-centric environments.

To achieve this fusion architecture, we integrate a stream of V-JEPA 2's contextual visual features with a stream of CoMotion's 3D skeletal poses through a cross-attention mechanism. This mechanism enables each feature stream to inform one another, enforcing a holistic understanding of the action space [23, 24]. We evaluate our model's action recognition capabilities on the Industrial Human Action Recognition Dataset (InHARD) [25] and UCF101's [26] high occlusion subset UCF-19-Y-OCC [27] against strong unimodal and state-of-the-art baselines. Our experiments, including a targeted ablation study on different fusion mechanisms, validate our fusion architecture's ability to ground human action within its spatial context.

We summarize our contributions as follows: (1) A novel modality fusion that achieves physically grounded action recognition by synergizing predictive representations of a high-level world model with precise geometric data from a 3D human pose tracker, (2) A clear demonstration that fusing an implicit world model with an explicit skeletal model achieves a more spatially aware representation of human actions in highly occluded environments, and (3) A contribution to the embodied AI community by providing a model that better understands the geometric and physical nature of human interactions, a crucial step towards developing more intelligent and capable agents.

2 Related works

Recently, video understanding has been reshaped with the emergence of more sophisticated architectures, such as 3D convolutions [11, 28] and temporal transformers [17, 29], that have enhanced spatio-temporal feature extraction [30]. A recent paradigm shift has been towards a large-scale self-supervised pre-training framework that has enabled vast learning of general all-purpose features [31, 32, 33]. Models that are structured on this self-supervised framework [34, 35] use these features for action recognition [36]. While these models excel at general motion understanding, their reliance on pixel data can make them susceptible to challenges posed by cluttered environments where precise spatial reasoning is required [37, 38, 39].

Additionally, the limitations of single-modality action recognition approaches have motivated the integration of different data types [40, 20, 41] to enhance spatial understanding. These multi-modal approaches have commonly fused visual data with other input, such as depth maps [40], language models [42, 43] and even skeletal data [44, 45]. While these fusion approaches recognize actions by correlating visual cues with complementary data, their reliance on pattern recognition means they don't understand the implicit physical dynamics within an action space. Our approach extends this direction by fusing the predictive capabilities of a high-level world model [46, 16] with explicit, low-level human skeletal information [20]. We introduce this fusion to directly ground actions within their contextual space so that an enriched, holistic understanding of this space is learned.

3 Methodology

Feature extraction and temporal alignment. To generate the visual feature sequence F_V, we first sample 64 frames uniformly from each InHARD and UCF101 video clip using the Temporal Segment Network (TSN) sampling strategy [10]. Each T frame is independently processed by V-JEPA 2's ViT-g_{384} encoder that provides physical world dynamics. The transformer's [CLS] token corresponding to each frame is extracted and used as that frame's representative feature vector. This procedure results in a visual feature sequence $F_V \in \mathbb{R}^{T \times D_V}$ where $T = 64$ is the number of time tokens, and $D_V = 1408$ is the feature dimension. For the skeletal stream, CoMotion processes each frame per video clip to produce a sequence of SMPL [47] parameters (pose θ, translation t, and shape β) for the original clip length T_{clip}. These parameters are then decoded by an SMPL layer to obtain 3D coordinates for all $J = 24$ joints that are then stored in a matrix of dimensions $[J, 3]$. To ensure each representation is invariant to the person's global position, a root-relative normalization is applied by subtracting the coordinates of the root joint (pelvis at index 0) from all other joints. The

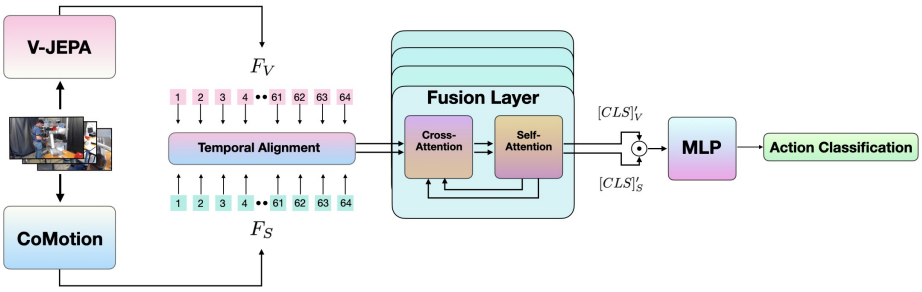

Figure 1: Fusion model architecture showing the pipeline of visual and skeletal feature sequences undergoing cross-attention and refinement to classify actions.

normalized coordinates for each frame are converted into a feature vector by flattening the coordinate matrix into a single dimension of size $D_S = 3J = 72$. This yields the final skeleton feature sequence $F_S \in \mathbb{R}^{T_{clip} \times D_S}$. To temporally align F_S with F_V, we employ the same TSN sampling strategy in the visual feature extraction, enabling the fusion model to capture a holistic view of the action while enforcing a consistent input structure [48, 49]. For a video with T_clip frames, it is divided into $N = 8$ segments of equal duration. From each segment, $k = 8$ frames are sampled to form the final sequence of $T = n \times k = 64$. Though seemingly redundant for the visual feature stream, this process ensures the tokens from each modality correspond to the same frame to enable proper integration.

Modality embedding and positional encoding. To enable parameter sharing and preservation of modality, the temporally aligned visual features F_V and F_S are first projected into a common embedding dimension $D_{model} = 512$, using the separate linear layers: $X_V = F_V W_V + b_V$ and $X_S = F_S W_S + b_S$ where $W_V \in \mathbb{R}^{D_V \times D_{model}}$ and $W_S \in \mathbb{R}^{D_S \times D_{model}}$ are learnable weight matrices with associated bias vectors, b. We prepend a learnable $[CLS]_V$ and $[CLS]_S$ token to X_V and X_S, respectively, to preserve modality classification before aggregation. Finally, a sinusoidal positional encoding ($PE \in \mathbb{R}^{(T+1) \times Dmodel}$) is added to each sequence to provide a continuous mapping of each feature sequence's temporal order. This results in updated visual and skeletal sequences represented by $Z_V^{(0)} = Concat([CLS]_V, X_V) + PE$ and $Z_S^{(0)} = Concat([CLS]_S, X_S) + PE$.

Cross-attention fusion transformer. The core architecture of our model is a stack of identical fusion layers $L = 4$ that perform bidirectional cross-attention followed by self-attention. This sequence enables mutual enrichment of each feature stream through an information exchange [50] followed by refinement through self-attention [51]. To achieve this, we perform a bidirectional cross-attention operation where each modality's representation is updated by attending to the other. This step consists of a multi-head attention sub-layer and an FFN, each with its own residual connection and layer normalization to stabilize training. The visual stream is updated with: $\tilde{Z_V} = LayerNorm(Z_V^{(l-1)} + MultiHeadAttn(Q = Z_V^{(l-1)}, K = Z_S^{(l-1)}, V = Z_S^{(l-1)}$, where $l-1$ represents output from the previous layer. An identical, parallel operation computes the updated skeleton Stream $\tilde{Z_S}$ by swapping the roles of Z_V and Z_S. These newly enriched feature representations are then passed through a standard self-attention block to enforce feature contextualization. This operation contains a multi-head attention sub-layer and an FFN to produce final layer outputs Z'_V and Z'_S, where the updated visual sequence is represented by $Z'_V = LayerNorm(\tilde{Z}_V + MultiHeadAttn(Q = \tilde{Z}_V, K = \tilde{Z}_V, V = \tilde{Z}_V)$, and an identical operation is applied to update the skeletal sequence to Z'_S.

Action classification. After the final fusion layer, the updated [CLS] tokens, $[CLS]'_V$ and $[CLS]'_S$, are extracted from their respective feature streams Z'_V and Z'_S . We concatenate these two tokens into a final feature vector that is passed through a multi-layer perception (MLP) with a softmax activation to generate action class probabilities. The full model architecture is shown in Figure 1.

4 Experiments

We conduct experiments to evaluate our multimodal fusion model by exploring the following questions: (1) Does synthesizing contextual visual data with explicit skeletal data provide a more robust

Table 1: Evaluation of action recognition performance on the InHARD and UCF-19-Y-OCC benchmarks. Best results per benchmark are indicated in **bold**.

Model	InHARD			UCF-19-Y-OCC		
	Top-1 Acc. (%)↑	Macro mAP (%)↑	Macro F1 (%)↑	Top-1 Acc. (%)↑	Macro mAP (%)↑	Macro F1 (%)↑
V-JEPA 2 baseline	80.76 ± 0.21	80.93 ± 0.43	76.24 ± 0.30	31.83 ± 0.76	$\mathbf{58.48 \pm 0.47}$	14.23 ± 0.88
CoMotion baseline	75.92 ± 0.17	74.60 ± 0.41	69.52 ± 0.11	6.20 ± 0.00	8.84 ± 0.25	1.72 ± 0.16
Fusion model (cross-attention)	$\mathbf{83.47 \pm 0.03}$	$\mathbf{84.96 \pm 0.10}$	$\mathbf{80.21 \pm 0.31}$	$\mathbf{38.62 \pm 0.22}$	54.10 ± 0.14	$\mathbf{16.30 \pm 0.41}$
Gated recursive fusion	79.25 ± 1.54	76.90 ± 0.90	73.69 ± 1.58	29.54 ± 1.78	50.07 ± 0.79	11.44 ± 0.64

Table 2: Ablation study on different fusion methods evaluated on InHARD and UCF-19-Y-OCC.

Fusion method	InHARD			UCF-19-Y-OCC		
	Top-1 Acc. (%)↑	Macro mAP (%)↑	Macro F1 (%)↑	Top-1 Acc. (%)↑	Macro mAP (%)↑	Macro F1 (%)↑
Early fusion (concatenation)	79.52 ± 0.83	78.55 ± 1.16	75.91 ± 1.46	33.34 ± 1.10	$\mathbf{56.50 \pm 2.51}$	14.87 ± 0.41
Late fusion (score averaging)	80.24 ± 0.53	83.59 ± 0.73	77.31 ± 0.55	34.42 ± 0.78	53.2 ± 0.25	14.93 ± 0.44
Fusion model (cross-attention)	$\mathbf{83.47 \pm 0.03}$	$\mathbf{84.96 \pm 0.10}$	$\mathbf{80.21 \pm 0.31}$	$\mathbf{38.62 \pm 0.22}$	54.10 ± 0.14	$\mathbf{16.30 \pm 0.41}$

and spatially grounded representation of human actions? (2) Does this fusion architecture have significantly better understanding of human actions in scenes with heavy occlusion compared to the proposed baselines? To answer these questions, we validate the proposed model on action recognition using the InHARD [25] dataset and the UCF-19-Y-OCC high occlusion dataset [27], and we further conducted a fusion mechanism ablation study, detailed in subsection 4.2. Our model is compared to state-of-the art V-JEPA 2, CoMotion, and fusion architecture [52] baselines. Further details on the datasets and experiment configurations are contained in A.1 and A.2, respectively.

4.1 Results

Table 1 shows the results of each model's action recognition performance on each benchmark. We find that our fusion model outperforms the other baseline models in all three metrics, and the V-JEPA 2 baseline only has slightly lower performance compared to our model. These results emphasize the robust and spatially-grounded action recognition capabilities achieved through our modality fusion, yet they also showcase V-JEPA 2's powerful prediction and understanding capabilities.

The right side of Table 1 shows each model's performance on the high occlusion benchmark UCF-19-Y-OCC. Our model significantly outperforms the baseline models, with a 6.79% higher accuracy score than V-JEPA 2's baseline. Notably, CoMotion alone collapses under heavy occlusion. This performance gap shows that the fusion of a contextual vision model with explicit pose data significantly enhances human action understanding in complex scenes. To evaluate the effectiveness of our cross-attention mechanism, an ablation study to compare different fusion techniques is seen in Table 2. Our model is seen to have a slight performance advantage compared to these techniques, reflecting cross-attention's effectiveness in enriching and contextualizing complementary feature streams.

5 Conclusion

In this paper, we propose a multimodal fusion architecture that effectively integrates rich, contextual representations from the V-JEPA 2 video model with precise, geometric skeleton trajectories from CoMotion. Our experiments confirm that this multimodal approach leads to a more robust and spatially grounded understanding of human action. Our model achieved superior performance on both action recognition and high-occlusion action recognition benchmarks. These findings attest to the value of complementing finer data representations within broader contextual data to enhance spatial understanding. However, limitations of our work come from its dependency on feature extraction from V-JEPA 2 and CoMotion, along with its limited testing on action recognition benchmarks and sparse availability of state-of-the-art fusion baseline performances on these benchmarks. While this research has the potential to positively impact society by advancing applications in collaborative robotics and assistive technologies, it's important to consider potential negative implications, such as malicious use of human activity monitoring in surveillance applications. Ultimately, our results advocate for a shift from statistical pattern action recognition toward recognition gained from a holistic understanding of both human interaction's and their spatial context.

References

[1] D. Aarno and D. Kragic, "Motion intention recognition in robot assisted applications," *Robotics and Autonomous Systems*, vol. 56, no. 8, pp. 692–705, 2008.

[2] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 299–306, 2013.

[3] R. Hu, J. Smith, and W. Liu, "Embodied human activity recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[4] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, and J. Malik, "On the benefits of 3d pose and tracking for human action recognition," 2023.

[5] C. Cheng and H. Xu, "A 3d motion image recognition model based on 3d cnn-gru model and attention mechanism," *Image and Vision Computing*, vol. 146, p. 104991, 2024.

[6] S. Win and T. L. L. Thein, "Real-time human motion detection, tracking and activity recognition with skeletal model," in *2020 IEEE Conference on Computer Applications(ICCA)*, pp. 1–5, 2020.

[7] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014. Celebrating the life and work of Maria Petrou.

[8] A. Doering and J. Gall, "A gated attention transformer for multi-person pose tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3189–3198, 2023.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 568–576, 2014.

[10] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision (ECCV)*, pp. 20–36, 2016.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new benchmark and model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4728–4737, 2017.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12018–12027, 2019.

[14] R. Hang, S. Wu, and M. Li, "Spatial-temporal adaptive graph convolutional network for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274, 2022.

[15] H. Zhang, J. Zhang, and Z. Cai, "A comprehensive survey of rgb-based and skeleton-based human action recognition," *Sensors*, vol. 23, no. 12, p. 5433, 2023.

[16] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas, "V-jepa 2: Self-supervised video models enable understanding, prediction and planning," 2025.

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.

[18] S. Al-Shorbaji and B. Al-Abdallah, "A review on computer vision-based methods for human action recognition," *PLoS ONE*, vol. 15, no. 6, p. e0234825, 2020.

[19] Y. Yuan, K. Zhou, and J. Liu, "On occlusions in video action detection: Benchmark datasets and analysis," *arXiv preprint arXiv:2410.19553*, 2023.

[20] A. Newell, P. Hu, L. Lipson, S. R. Richter, and V. Koltun, "Comotion: Concurrent multi-person 3d motion," 2025.

[21] Y. Duan, X. Tian, and H. Li, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] G. Li, C. Wu, and J. Han, "A survey on 3d skeleton-based action recognition using learning methods," *CB Systems*, vol. 1, no. 2, p. 100, 2024.

[23] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *CoRR*, vol. abs/2107.00135, 2021.

[24] H. Yan, S. Xiong, L. Wang, L. Jian, and G. Vivone, "Atfusion: An alternate cross-attention transformer network for infrared and visible image fusion," 2025.

[25] M. DALLEL, V. HAVARD, D. BAUDRY, and X. SAVATIER, "Inhard - industrial human action recognition dataset in the context of industrial collaborative robotics," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6, 2020.

[26] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012.

[27] S. Grover, V. Vineet, and Y. S. Rawat, "Revealing the unseen: Benchmarking video action recognition under occlusion," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[28] D. Tran, H. Wang, L. Torresani, C. Feichtenhofer, M. Bulo, and D. Fejzic, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[29] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *International Conference on Machine Learning*, pp. 1018–1029, 2021.

[30] C. Meo, K. Sycheva, A. Goyal, and J. Dauwels, "Bayesian-lora: Lora based parameter efficient fine-tuning using optimal quantization levels and rank values trough differentiable bayesian gates," *arXiv preprint arXiv:2406.13046*, 2024.

[31] P. Weinzaepfel, V. Leroy, T. Lucas, R. Brégier, Y. Cabon, V. Arora, L. Antsfeld, B. Chidlovskii, G. Csurka, and J. Revaud, "Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion," 2023.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[33] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3," 2025.

[34] Y. Geng, Z. Zhou, J. Lin, Q. Hou, H. Chen, and L. Wang, "Videomae v2: Scaling video masked autoencoders with dual masking," *arXiv preprint arXiv:2303.16727*, 2023.

[35] Y. Xu, X. Li, J. Zhang, J. Zhang, K. Huang, C. Zhao, C. Chen, X. Wang, H. Li, L. Wang, *et al.*, "Internvideo2: Scaling video foundation models for multimodal video understanding," *arXiv preprint arXiv:2403.15377*, 2024.

[36] Y. Qian, Y. Sun, A. Kargarandehkordi, P. Azizian, O. C. Mutlu, S. Surabhi, P. Chen, Z. Jabbar, D. P. Wall, and P. Washington, "Advancing human action recognition with foundation models trained on unlabeled public videos," 2024.

[37] C. Meo, A. Nakano, M. Lică, A. Didolkar, M. Suzuki, A. Goyal, M. Zhang, J. Dauwels, Y. Matsuo, and Y. Bengio, "Object-centric temporal consistency via conditional autoregressive inductive biases," *arXiv preprint arXiv:2410.15728*, 2024.

[38] A. Dave, Y. Tsvetkov, and R. Schwartz, "What's missing from self-supervised representation learning?," in *ICLR 2021 Workshop on Representation Learning on Graphs and Manifolds (RLGM)*, 2021.

[39] Y. Wei, A. Gupta, and P. Morgado, "Towards latent masked image modeling for self-supervised visual representation learning," 2024.

[40] J. Shin, N. Hassan, A. S. M. Miah, and S. Nishimura, "A comprehensive methodological survey of human activity recognition across diverse data modalities," *Sensors*, vol. 24, no. 10, p. 3167, 2024.

[41] D. Xie, X. Zhang, X. Gao, *et al.*, "Maf-net: A multimodal data fusion approach for human action recognition," *PLoS ONE*, vol. 20, no. 4, p. e0319656, 2025.

[42] J. Chen *et al.*, "Grounding multimodal large language models in actions," in *Advances in Neural Information Processing Systems*, 2024.

[43] K. Deng, Y. Wang, and L.-P. Chau, "Egocentric human-object interaction detection: A new benchmark and method," *arXiv preprint arXiv:2506.14189*, 2025.

[44] X. Zhu, Y. Zhu, H. Wang, H. Wen, Y. Yan, and P. Liu, "Skeleton sequence and rgb frame based multi-modality feature fusion network for action recognition," 2022.

[45] N. Zheng and H. Xia, "Snn-driven multimodal human action recognition via event camera and skeleton data fusion," 2025.

[46] C. Meo, M. Lica, Z. Ikram, A. Nakano, V. Shah, A. R. Didolkar, D. Liu, A. Goyal, and J. Dauwels, "Masked generative priors improve world models sequence modelling capabilities," *arXiv preprint arXiv:2410.07836*, 2024.

[47] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, Oct. 2015.

[48] G. Yang, Y. Yang, Z. Lu, J. Yang, D. Liu, C. Zhou, and Z. Fan, "Sta-tsn: Spatial-temporal attention temporal segment network for action recognition in video," *PLOS ONE*, vol. 17, pp. 1–19, 03 2022.

[49] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1895–1904, June 2021.

[50] Y. Lin, J. Lu, Y. Yong, and J. Zhang, "Mv-gmn: State space model for multi-view action recognition," 2025.

[51] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023.

[52] Y. Shihata, "Gated recursive fusion: A stateful approach to scalable multimodal transformers," 2025.

[53] H. Zhang, Y. Li, J. Cao, W. Wang, M. Zhou, and Y. Wang, "Pymaf-x: Towards robust and accurate 3d human mesh recovery from single-view images," *arXiv preprint arXiv:2204.09459*, 2022.

[54] G. Li, C. Wu, J. Han, B. Tang, and Q. Sun, "Spatiotemporal focus for skeleton-based action recognition," *Pattern Recognition*, vol. 136, p. 109231, 2022.

A Appendix

A.1 Dataset details

Our experiments are conducted on the Industrial Human Action Recognition Dataset (InHARD) that contains over two million frames collected from 16 different subjects over 13 different industrial action classes. This benchmark is designed from complex, interaction-heavy human activities that feature frequent occlusions and presents a challenge for models that do not effectively reason about spatial and postural information. The dataset provides 14 meta action labels, which we use as the standard for training and evaluation.

A critical pre-processing step in our methodology is the cropping of the InHARD video clips. The raw videos in the dataset are presented as mosaics of three synchronized camera views:top-left, top-right, and bottom-right, with an empty bottom-left quadrant. Therefore, we cropped each of these views into separate video clips that created singular left, right, and top views for each InHARD video. The cropping ensures actors remain centered so that CoMotion, which heavily relies on visual cues from body joints, can more accurately detect SMPL joints. The cropping presents clear, per-frame skeletons of the actor, so occlusion patterns can be properly detected to eventually create an occlusion-heavy subset of InHARD for model evaluation.

To specifically test our model's performance in complex scenes with occluded interactions, we use the UCF-19-Y-OCC [27] occlusion benchmark, which is a subset of the UCF-101 [26] action recognition benchmark. This benchmark was manually curated for action recognition performance under real-world occlusion. It consists of 1,732 video clips that cover 19 action classes that frequently involve natural occlusions (environmental clutter, challenging camera angles, and limbs blocked by objects). Unlike other synthetic occlusion datasets, UCF-19-Y-OCC provides a realistic benchmark for evaluating a model's ability to generalize and maintain performance when presented with incomplete visual evidence in a complex environment.

A.2 Experiment configuration

For our experiments involving InHARD, we train all models on the cropped dataset and use its official train and validation splits across the 14 meta action labels. For our high occlusion experiment, each model is trained on the first UCF101 training split, and they are evaluated on the UCF-19-Y-OCC high occlusion benchmark. Each model is trained on a RunPod NVIDIA A100 SXM GPU for 30 epochs, employing the AdamW optimizer with a learning rate of 310^{-4} and a weight decay of 0.05 for the attention probes. The learning rate was managed by a cosine decay schedule with a 5% warmup period. The training set was split into batches of size 128. To handle variability in T_{clip}, all models employ TSN with the previously described protocol of being divided into eight clip segments each containing eight feature vectors.

To ensure stable training, we utilize a dropout rate of 0.1, gradient clipping at a max norm of 1.0, and AMP to accelerate computation. The baseline attentive probes are configured with a model dimension D_{model} of 1408 for the V-JEPA 2 baseline and 256 for the CoMotion baseline while both have two transformer layers and eight attention heads. Our fusion model has a $D_{model} = 512$, with four fusion layers and eight attention heads. The best-performing checkpoint for each model is selected based on the highest mean Average Precision (mAP) on the validation set for each run. Three different seeds are tested to ensure experiment variability, and the mean values with standard deviation are presented in the tables.

A.3 Ablation study details

To evaluate our model's cross-attention fusion mechanism, we conduct an ablation study against two other fusion models: early fusion via feature concatenation, and late fusion via averaging prediction scores. For a fair comparison, all other model components and training hyperparameters were held constant across experiments. The performance of each fusion strategy was evaluated on the InHARD benchmark using the same evaluation metrics as the first experiment. The results clearly demonstrate that Cross-Attention fusion consistently outperforms both early and late fusion, highlighting its superior capability at modeling complex relationships between fused features and action recognition.

B Extended Related Works

A complementary line of research focuses on explicit representations of human movement. Recent progress in 3D human pose estimation has enabled the recovery of detailed skeletal and body mesh information from a single camera [53]. These methods can track multiple individuals through time, even when they are partially obscured, by leveraging temporal dependencies and learned priors [20, 8]. Such skeletal data provides a robust, geometry-based understanding of posture and limb configurations, which is often difficult to infer implicitly from raw video pixels alone. Researchers have also explored action recognition solely using these skeletal streams, often employing graph convolutional networks to model the spatial and temporal relationships of the joints [12, 54].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our introduction and abstract contains the summarization of our new model and its contributions. The further prove can be found in the methodology and experiment section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are described in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results, theorems, formulas, or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the novel architecture established in our fusion model and details in order to fully reproduce our results in the Methodology section as well as in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code for all the models and experimentation process will be provided on github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We describe all of the training and test details under the Methodology section and in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports standard deviation and mean from multiple testing seeds for each experiment and metric.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information about the computer resources utilized in our experiments are listed under Appendix A.3 Experiment configuration

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research abides by all of the standards mentioned in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both positive societal impacts and potential malicious uses of our work in the Conclusion section.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not pose any risks since there is no release of models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We perform all the experiments under the license and have cited the original papers behind the InHARD and UCF-19-Y-OCC benchmarks, the V-JEPA 2 model, the CoMotion model, etc. in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiments are involved in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No IRB is involved in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: All the core methods were developed without the involvement of LLMs for any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.