Multiple Human Motion Understanding

Lei Li[1,3*†], **Sen Jia**[1,2*], **Jenq-Neng Hwang**[1]

[1]University of Washington
[2]VitaSight
[3]AI Thinktank

Abstract

We introduce **LLaMMo** (**L**arge **La**nguage and **M**ulti-Person **Mo**tion Assistant), the **first** instruction-tuning multimodal framework tailored for multi-human motion analysis. LLaMMo incorporates a novel human-centric and social-temporal learner that models and fuses both intra-person dynamics and inter-person dependencies, yielding robust, context-aware representations of complex group behaviors while maintaining low computational overhead. To support LLaMMo, we construct **LLaVerse**, a large-scale dataset with fine-grained manual annotations covering diverse multi-person activities spanning daily social interaction and professional team sports. Built on top of LLaVerse, we also propose **LLaMI-Bench**, a dedicated benchmark for evaluating multi-human behavior understanding across motion and video modalities. Extensive experiments demonstrate that LLaMMo consistently outperforms baselines in understanding multi-person interactions under low-latency settings, with notable gains in both social and sport-specific contexts.

Introduction

Human motion understanding is a pivotal research area in multimodal learning, with recent advances in motion understanding greatly impacting applications such as digital avatars, human-computer interaction, sports and personalized healthcare (Deng et al. 2025; Ghosh et al. 2024; Zhou et al. 2025; Wu et al. 2024b; Deng et al. 2024; Lai et al. 2025; He et al. 2025a; Jin et al. 2024). Despite these advancements, existing research primarily focuses on individual, struggling to capture the interaction of human activities (Li et al. 2024a; Jiang et al. 2023; Yao et al. 2025; He et al. 2025b; Cai et al. 2025). However, as most real-world scenarios, from collaborative tasks to team sports such as soccer, involve complex multi-agent interactions (Wu et al. 2024b; Jia and Li 2024; Jin et al. 2024; Lan et al. 2025; Lu et al. 2025) and multiple humans. Consequently, accurately analyzing these realistic contexts necessitates a shift toward multi-person motion understanding (Yu et al. 2025; Xu, Wang, and Gui 2023; Liu et al. 2025; Li et al. 2025; Li 2024).

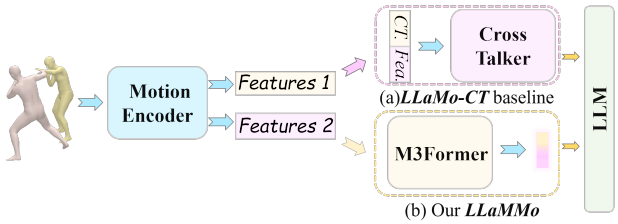

Figure 1: (a) Extension of LLaMo, a SOTA single-human motion understanding model, for multi-person input as a baseline. (b) LLaMMo introduces fusion-token-based interaction modeling for multi-human motion understanding.

Recent work in human motion understanding focusing on a single person, such as MotionGPT (Jiang et al. 2023) and MotionLLM (Chen et al. 2024), which encode motion sequences into discrete tokens through vector quantization. Building upon this, LLaMo tried to encode motion without vector quantization, better preserving spatial-temporal detail, and demonstrating strong performance in both social and professional contexts(Li et al. 2024a). However, these methods remain predominantly single-person oriented, limiting their applicability in real-world multi-human scenarios (Peng, Mao, and Wu 2023; Jeong, Park, and Yoon 2024; Xu et al. 2023; Wang et al. 2025; Jin et al. 2025).

Extending human motion understanding to multi-human contexts faces two primary challenges. Firstly, accurately modeling multi-person interactions requires handling intricate causal and temporal dependencies, a complexity inadequately addressed by merely concatenating individually encoded motions, which also introduces computational inefficiencies (Guo et al. 2025; Wang et al. 2021; Tanke et al. 2023; Gu et al. 2025). Secondly, current datasets like HumanML3D (Guo et al. 2022) and MotionX (Lin et al. 2023) primarily focus on single-person motion, while interaction-focused datasets (e.g., InterX (Xu et al. 2024), InterHuman (Liang et al. 2024)) typically involve only pairs of individuals, leaving comprehensive multi-person scenarios underrepresented.

To address these limitations, we propose **LLaMMo** (**L**arge **La**nguage and **M**ulti-Person **Mo**tion Assistant), the **first** framework for general multi-human motion under-

*These authors contributed equally.
†Corresponding author, lenny.lilei.cs@gmail.com

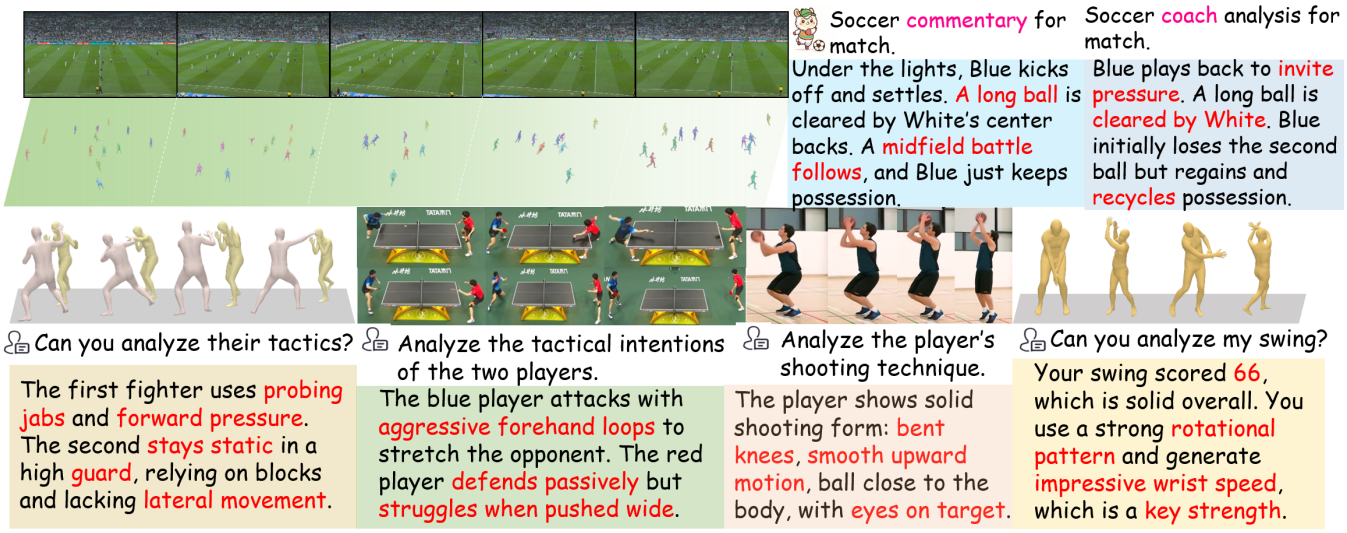

Figure 2: Qualitative results of LLaMMo on complex multi-person sports scenarios, including soccer, boxing, table tennis, golf, and basketball. For soccer, we assess LLaMMo's performance in both commentator-style narration and coach-like strategic analysis. These highly dynamic activities pose greater challenges than typical social interactions for LLaMMo.

standing, spanning both social scenarios and sports settings. LLaMMo incorporates a Human Encoder that facilitates motion or/and video input, accommodating arbitrary keypoint representations. Additionally, LLaMMo introduces M3Former for multi-person modeling, which jointly reasons over intra- and inter-person dynamics across frames. This design generates robust, unified representations for multi-persons without additional computation overhead.

For training and fair comparison, we introduce **LLaVerse**, the first large-scale multihuman motion dataset. LLaVerse comprises over 200,000 annotated sequences spanning diverse contexts, from intricate social interactions to professional sports, offering a richly labeled foundation for future studies. Based on LLaVerse, we propose **LLaMI-Bench**, a benchmark designed for multi-person understanding evaluation, scoring social understanding from five aspects and professional insight with coach-style metrics. Evaluations demonstrate that LLaMMo delivers state-of-the-art performance and lower latency than strong baselines, establishing a scalable foundation for future multi-human motion analysis and real-world applications. Our contributions can be summarized as follows:

- We present **LLaMMo**, the **first** framework for multi-human motion understanding, delivering interaction-aware representations across social and sports scenes without introducing additional computation overhead.
- We propose **LLaVerse**, a comprehensive datasets involving 600k motion-text pairs spanning social and sports interactions, establishing a broad, richly-labeled foundation for multi-person motion models.
- We introduce **LLaMI-Bench**, LLaMI-Bench aims to serve as a standardized benchmark for assessing multi-human motion understanding capabilities for future studies and applications.

Related Work

Human Representation

Recent studies in motion representation have explored various modalities for encoding dynamic content, including joint sequences, 3D body parameters such as SMPL (Loper et al. 2023), and discrete latent codes. Joint-based datasets like HumanML3D (Guo et al. 2022) capture skeletal trajectories, while parametric models abstract detailed body motion. Methods like MotionGPT (Jiang et al. 2023) and MotionLLM (Wu et al. 2024a) further introduce vector-quantized latent spaces, facilitating diverse and controllable motion synthesis across different tasks. Motion representations inherently offer stronger privacy protection compared to raw video data, as they abstract appearance while retaining essential dynamic information.

Human Activities Understanding

Early approaches to human motion understanding relied on handcrafted features and sequence models (e.g., DTW (Müller and Röder 2006), HMMs (Yamato, Ohya, and Ishii 1992)), but were limited to simple, single-person actions. The emergence of deep learning brought substantial gains via RNNs (Du, Wang, and Wang 2015), spatio-temporal CNNs (Tran et al. 2015), and GNNs (Yan, Xiong, and Lin 2018), enabling richer motion representations. More recently, unified language-conditioned frameworks have gained traction, treating motion as a modality aligned with text. Models such as MotionLLM (Wu et al. 2024a) and LLaMo (Li et al. 2024b) leverage LLMs to support diverse human understanding. However, these methods remain restricted to single-person settings, limited to real-world multi-human scenarios. A big challenge for these models is to expand understanding to include interactions between humans, a crucial aspect of real-world behaviors.

Human Motion Datasets

Existing high-quality motion datasets predominantly focus on single-person activities. HumanML3D and MotionX provide approximately 70.6K motion clips with textual annotations (Guo et al. 2022; Lin et al. 2023). In professional sports, LLaMo's Golf-swing dataset offers 30K coach-annotated golf motions (Li et al. 2024b). For multi-person datasets, Inter-X and InterHuman deliver rich motion and text descriptions focusing on interactions between two persons (Xu et al. 2024; Liang et al. 2024). MuPoTS-3D and the CMU Panoptic provide multi-view MoCap data capturing everyday activities (Mehta et al. 2018; Joo et al. 2015). In the sports domain, SoccerNet (500 matches, 764 hours) (Giancola et al. 2018; Gautam et al. 2024), WorldPose, comprising 2.5 million 3D poses from the FIFA World Cup (Jiang et al. 2024), OpenTTGames (Voeikov, Falaleev, and Baikulov 2020), and an Olympic boxing dataset annotated by referees expand domain-specific resources (Stefański, Kozak, and Jach 2024). Collectively, these corpora form the foundation for LLaVerse.

Method

As is shown in Figure 3. LLaMMo consists of four core components designed for multi-human behavior understanding: a Human Encoder for multimodal input processing, an Interaction-Aware Selector (IA-Selector) for salient frame selection, a Social-Temporal Transformer (M3Former) for modeling inter/intra-person dynamics, and a Crosser for aligning motion features with language.

Human Encoder

The Human Encoder extracts compact multimodal representations from motion and video inputs to support downstream interaction modeling. It includes a motion estimator, a motion encoder, and a feature enhancer. Given a video sequence $V = \{v_1, \ldots, v_T\}$, and optionally a motion sequence M_o, we directly use M_o if available; otherwise, we estimate motion from video using $f_e(V)$.

The resulting motion sequence M and video V are processed by dedicated encoders $f_m(\cdot)$ and $f_v(\cdot)$, respectively, and then fused through a cross-attention module $f_h(\cdot)$:

$$\tilde{F}_M = f_h(f_m(M), f_v(V)).$$

The output multimodal feature captures joint visual-motion cues and serves as the input to subsequent modules.

Interaction-Aware Selector

Based on $\tilde{F}_M$, the IA-Selector identifies interaction-critical frames. Temporal pooling is applied to obtain global features. A cross-attention mechanism computes the relevance between F_P and text embeddings T_L, yielding language-aware importance weights that are injected into $\tilde{F}_M$:

$$\hat{F}_M = \tilde{F}_M \odot \mathcal{B}\Big[\mathrm{softmax}_P\Big(\frac{(\mathcal{P}(\tilde{F}_M)W_Q)\,(T_L W_K)^\top}{\sqrt{H}}\Big)\Big]$$

where $\mathcal{P}(\cdot)$ denotes temporal pooling and $\mathcal{B}[\cdot]$ broadcasts the computed weights to all frames and channels, producing language-aware features $\hat{F}_M$. Then, $\hat{F}_M$ are pooled across persons yields global representations $F_T \in \mathbb{R}^{T\times H}$ and a self-attention matrix is computed to derive frame importance scores. Finally, the top-K most critical frames are selected, resulting in a compressed representation $\tilde{F}'_M \in \mathbb{R}^{P\times T'\times H}$, which can be summarized as:

$$\tilde{F}'_M = \mathrm{TopK}_{T\to T'}\Big(\tilde{F}_M,\ \mathrm{softmax}_T\big[\phi\big(\mathcal{P}_p(\hat{F}_M^\top)\big)\big]\Big)$$

where $\phi(\cdot)$ is the scoring function. By jointly aligning person-level dynamics with language and adaptively filtering salient temporal cues, the IA-Selector yields a compact, interaction-informative representation while improving computational efficiency.

M3Former

Unlike naive token concatenation approaches, where attention complexity grows quadratically with the number of individuals, M3Former leverages a compact fusion token mechanism to maintain stable sequence length and efficient reasoning even in large-group scenarios. M3Former processes $\tilde{F}'_M$ from IA-Selector through TempFormer and SocFormer, capturing intra-person temporal dynamics and inter-person social dependencies, respectively. The outputs are fused via the TSTalker to produce unified multi-human motion representations, enabling coherent downstream behavior understanding.

TempFormer TempFormer is motivated by human social cognition, where individuals dynamically attend to socially relevant counterparts (Bertenthal and Boyer 2015; Kuang 2016). It models fine-grained temporal dependencies and evolving multi-person relations via a **tri-branch** relational attention mechanism, which captures **personalized dynamics**, **focused social cues**, and **global context**. Given the interaction-aware features $\tilde{F}'_M$ from the IA-Selector, we first pool over each individual's frames to obtain a global representation $F_P \in \mathbb{R}^{P\times H}$. The sequence $\tilde{F}'_M$ is then flattened, and for each query q_i, we compute attention weights over F_P to identify the most socially relevant counterpart, denoted as:

$$p^\star(i) = \arg\max_p \left(q_i F_P^{(p)\top}\right),$$

Then, through the focus attention branch (**F**), each query attends to the full set of frames $\mathcal{T}_{p^\star(i)}$ corresponding to the selected individual, enabling the model to capture interactions over time.

While the focus attention captures social relevance, modeling the internal temporal consistency within each person's own remains essential. As such, we incorporate a personal attention (**P**) where the query attends to its own frame sequence $\mathcal{T}_{p(i)}$. Moreover, to capture the broader scene-level dynamics, the query performs a global attention (**G**) over the global contexts F_P, allowing the model to integrate global semantic context across all individuals. Finally, the outputs from the three attention branches are adaptively fused through a gated mixture:

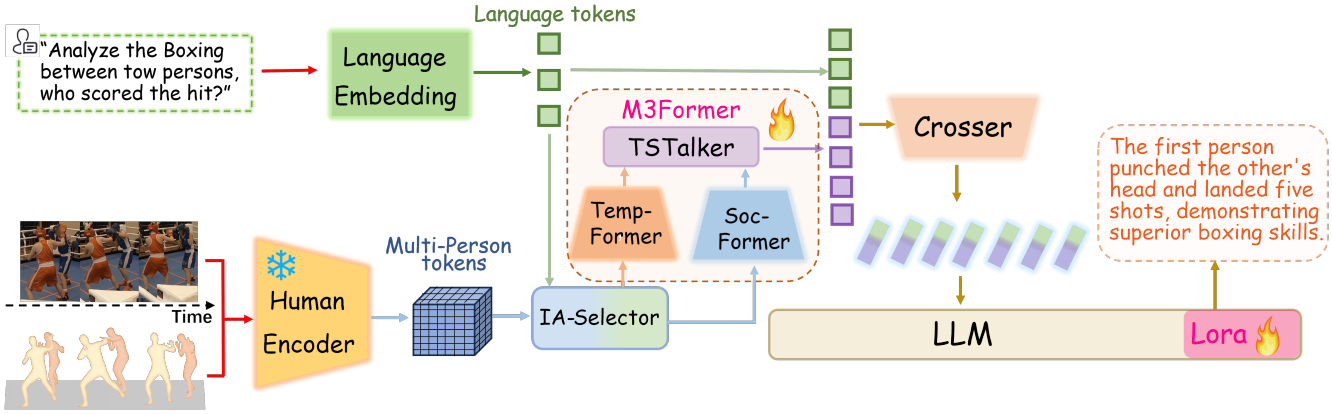

Figure 3: Overview of LLaMMo. The model takes video and/or motion input, encodes multi-person features, and selects salient frames via an IA-Selector. It then models the interactions of persons to generate fused multi-person representations, which are aligned with language embeddings and fed into an LLM for text generation.

$$\hat{I}_i = \sum_{b\in\{F,P,G\}} \Gamma_i^b \left(\sum_{x_j \in \mathcal{S}_i^b} \alpha_{ij}^b x_j \right),$$

$$\Gamma_i^b = \text{softmax MLP} \left(\sum_{x_j \in \mathcal{S}_i^b} \alpha_{ij}^b x_j \right),$$

where α denotes the attention weights. $\mathcal{S}_i^b$ denotes the target set for branches, corresponding to the focus sets $\mathcal{T}_{p^\star(i)}$, the personal frames $\mathcal{T}_{p(i)}$, and the global contexts $\{F_P^{(p)}\}$. This tri-branch design allows TempFormer to integrate personalized dynamics, interaction cues, and scene-level semantics into interaction-aware temporal representations.

SocFormer SocFormer models interpersonal dependencies within each frame. Given the interaction-aware features $\tilde{F}'_M$ from the IA-Selector, we transpose the temporal and person dimensions and apply self-attention across the P individuals for each frame t, yielding social-interaction representations $\hat{I}_S \in \mathbb{R}^{T'\times P\times H}$. This frame-wise inter-person modeling equips LLaMMo with an explicit understanding of social interactions, enabling socially aware motion representations beyond isolated temporal patterns.

TSTalker Given the temporal features $\hat{\mathbf{I}}_T$ from TempFormer and social features $\hat{\mathbf{I}}_S$ from SocFormer, TSTalker first performs bidirectional cross-attention to produce temporal and social outputs, and then fuses the transposed temporal output back into the social stream to refine social context, formally denoted as:

$$[\mathbf{Z}_T, \mathbf{Z}_S],\ \mathbf{Z}_{TS} = \text{BiCrossAttn}(\hat{\mathbf{I}}_T,\ \hat{\mathbf{I}}_S),\ \mathbf{Z}_S + f(\mathbf{Z}_T^\top).$$

Finally, a person-aware aggregation function is applied over $\mathbf{Z}_{TS}$ at each time step to produce interaction-fused representations. This fusion token sequence serves as a unified embedding that jointly models intra-person temporal coherence and inter-person social interactions, facilitating socially aware motion generation in LLaMMo.

Crosser

Given the interaction-fused representations from M3Former, we introduce Crosser, a lightweight two-layer bidirectional cross-attention module that aligns aggregated motion features with language embeddings before LLM decoding. This early-stage fusion enhances motion-language alignment by bridging semantics across modalities.

Training Objective

We finetune the language model to generate socially-aware descriptions conditioned on multi-person motion and video features, while keeping the feature encoders frozen. Given training samples $(V^{(i)}, M^{(i)}, \hat{Y}^{(i)})$, the objective minimizes the negative log-likelihood:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L^{(i)}} \log p(\hat{y}_t^{(i)} \mid \hat{y}_{1:t-1}^{(i)}, F_{\text{fusion}}^{(i)}).$$

This encourages the model to align motion inputs with coherent multi-human behavioral descriptions.

LLaVerse with LLaMI-Bench

LLaVerse Construction

LLaVerse is constructed from ten high-quality corpora, encompassing both daily activities and professional sports. For daily activities, we leverage HumanML3D and MotionX for rich single-person motion-text pairs, Inter-X and InterHuman for diverse two-person interactions, and MuPoTS-3D and CMU Panoptic for multi-person scenes. In particular, we enhance them with GPT-4o-mini assisted annotation and manual refinement to ensure semantic quality.

For professional sports, we aggregate data from five high-quality sources: the LLaMo Golf-Swing dataset, WorldPose for soccer motion and video, SoccerNet with paired broadcast videos and commentary, OpenTTGames for table tennis, and an Olympic boxing dataset with referee-annotated recordings. SoccerNet, OpenTTGames, and boxing videos are processed via large-scale motion estimation to obtain 3D

Track / Model	TA		AR		IU		SA		NC		All		BertScore	FPS
	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score		
(A) LLaMI-Sports														
GT	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	–	–
LLaMo-CT	38.76	2.66	33.78	2.28	35.12	2.57	40.49	2.82	55.39	3.23	42.69	2.61	0.34	185
LLaMMo	**48.04**	**2.91**	**39.23**	**2.60**	**43.79**	**3.22**	**46.78**	**3.31**	**62.00**	**4.05**	**50.16**	**2.91**	**0.42**	**279**
(B) Challenging WorldPose														
GT	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	–	–
LLaMo-CT	23.51	1.77	20.08	2.10	18.34	1.57	19.22	2.10	37.90	2.68	21.29	1.95	0.32	80
LLaMMo	**31.67**	**2.38**	**26.67**	**2.27**	**28.33**	**2.33**	**26.67**	**2.55**	**41.67**	**2.97**	**28.00**	**2.48**	**0.42**	**264**

Table 1: Performance of LLaMMo on LLaMI-Bench across two tracks: (A) LLaMI-Sports and (B) Challenging WorldPose. Both accuracy and scores are reported.

Model	IN		SC		CA		RM		CO		All		BertScore	FPS
	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score		
GT	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	100.00	5.00	–	–
LLaMo-CT	34.16	2.20	60.05	2.88	35.44	2.08	47.39	2.39	55.20	2.62	46.41	2.40	0.39	189
LLaMMo	**41.32**	**2.32**	**76.98**	**3.43**	**40.05**	**2.35**	**56.25**	**2.63**	**67.23**	**2.99**	**56.37**	**2.89**	**0.45**	**288**

Table 2: Performance of LLaMMo on LLaMI-Bench-Life.

Model	KE	TD	TA	EE	SE	All
GT	5.00	5.00	5.00	5.00	5.00	5.00
LLaMo-CT	2.40	2.10	2.00	2.20	2.83	2.31
LLaMMo	**3.40**	**2.95**	**3.05**	**3.50**	**3.76**	**3.33**

Table 3: Evaluation of LLaMMo's commentator-style generation in soccer scenarios. Scores range from 1 to 5, with higher values indicating better commentary quality.

motion data. In WorldPose, soccer coaches provide expert-annotated frame-level interactions. Implementation details are provided in the supplementary material.

LLaMI-Bench

To benchmark multi-human motion understanding, we propose **LLaMI-Bench**, spanning both social and sports scenarios. Unlike BLEU and similar text metrics that focus on surface-level matching, our benchmark evaluates interaction reasoning through social plausibility, causal coherence, and semantic clarity.

LLaMI-Bench-Life This track focuses on daily group interactions and assesses a model's capacity to reason about relational structure, semantic alignment, temporal causality, role inference, and behavioral coordination. We formulate five core dimensions: Interactivity (IN), Semantic Consistency (SC), Causality Awareness (CA), Role Modeling (RM), and Coordination (CO). These dimensions are motivated by key principles in cognitive science. For example, interactivity enhances comprehension (Chi 2008), semantic consistency supports discourse coherence (Zwaan and Radvansky 1998), and social role inference is essential for understanding group behavior (Li et al. 2019). Together, these five axes provide a principled basis for evaluating multi-human interaction understanding.

LLaMI-Bench-Sports Targeting structured, high-tempo domains such as soccer, table tennis, and boxing, this track evaluates models across five core dimensions: Temporal Accuracy (TA), Action Recognition (AR), Interaction Understanding (IU), Spatial Awareness (SA), and Narrative Coherence (NC), each reflecting key priorities in real-world tactical and analytical evaluation. These dimensions were defined in consultation with professional coaches from multiple team sports. For commentary-style generation, we additionally assess Key Events (KE), Technical Detail (TD), Tactical Analysis (TA), Emotional Expression (EE), and Style (SE), reporting only **Score** due to their inherently subjective nature.

Together, LLaMI-Bench-Life and LLaMI-Bench-Sports form a holistic protocol for evaluating models' abilities to interpret, describe, and reason about complex multi-person behavior across domains.

Experiment

Implementation Details

Training Setup: We train LLaMMo on LLaVerse training split described in Section , covering both daily and sports scenarios with diverse interaction complexities. During training, we freeze the human encoder and apply LoRA with rank 64 to the language model. LLaMMo is trained on 8× AMD MI250X GPUs over 8 hours for 5 epochs and optimized with a learning rate of $4e{-}4$.

Evaluation Details: We evaluate LLaMMo across both single- and multi-person settings. For single-person understanding, we follow LLaMo's protocol on Golf-Swing, using accuracy and quality scores aligned with coach annotations. For multi-person scenarios, we adopt LLaMI-Bench, which covers daily interactions (Life) and structured sports domains (Sports). Evaluation follows the defined dimensions in each track and is conducted via GPT-4o-based automatic

Model	Reasonableness		Coherence		Pertinence		Adaptability		All	
	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score
GT	100	5.00	100	5.00	100	5.00	100	5.00	100	5.00
MotionGPT (Jiang et al. 2023)	10.53	1.25	15.42	1.66	12.64	1.10	14.79	1.35	14.35	1.40
MotionLLM (Chen et al. 2024)	12.33	1.51	19.20	1.87	17.98	1.79	10.20	1.22	16.53	1.57
LLaMo	21.10	2.11	27.10	2.71	**31.81**	**3.12**	20.22	**1.98**	24.80	2.48
LLaMMo	**22.87**	**2.35**	**29.21**	**2.89**	31.08	3.10	**20.30**	1.88	**25.60**	**2.63**

Table 4: Performance on the golf-swing dataset across four indicators. Higher accuracy and score values indicate better performance.

IA	Temp	Soc	TS	IN		SC		CA		RM		CO		All	
✗	✗	✗	✗	34.16	2.20	60.05	2.88	35.44	2.08	47.39	2.39	55.20	2.62	46.41	2.40
✓	✗	✗	✗	35.07	2.17	64.22	3.05	36.02	2.14	48.50	2.44	57.21	2.69	47.89	2.52
✓	✓	✗	✗	37.33	2.24	66.50	3.19	38.54	2.25	50.12	2.50	59.50	2.75	50.07	2.65
✓	✓	✓	✗	39.11	2.29	71.21	3.32	39.86	2.31	53.75	2.58	62.06	2.83	53.58	2.79
✓	✓	✓	✓	**41.32**	**2.32**	**76.98**	**3.43**	**40.05**	**2.35**	**56.25**	**2.63**	**67.23**	**2.99**	**56.37**	**2.89**

Table 5: Ablation study of IA-Selector, TempFormer, SocFormer, and TSTalker on LLaMI-Bench-Life.

scoring, complemented by expert review for semantic reliability.

Baselines: **LLaMo-CT** is implemented following the LLaMo (Li et al. 2024a) architecture. Each person's motion is encoded via a Human Encoder and filtered through a language-guided selector. Then, for each person, the representation undergoes self-attention and cross-modal alignment with language tokens, followed by augmentation with a learnable identity token. All embeddings are flattened, concatenated with the language tokens into a unified sequence, and fed into an LLM for multimodal reasoning.

Results

We evaluate LLaMMo on single- and multi-person scenarios, including social interactions and sports, covering both coach-level analysis and commentary-style generation for soccer. Results show that LLaMMo consistently outperforms baselines in multi-human understanding, achieving stronger interaction modeling and notably faster inference, making it well-suited for real-time applications.

Evaluations on LLaMI-Bench-Life We evaluate LLaMMo on the LLaMI-Bench-Life, which covers diverse real-world social interactions. As shown in Table 2, LLaMMo outperforms the baseline LLaMo-CT by **9.9%** in overall Accuracy and by **0.49** in Score, with large gains in Interactivity and Role Modeling. These improvements arise from IA-Selector's key-frame pruning and M3Former's interaction modeling, which capture relational cues more effectively than simple concatenation. To measure how naturally generated descriptions mirror human narratives, we also report BertScore. The higher BertScore of LLaMMo confirms its effectiveness in both accuracy and linguistic coherence.

Beyond performance, LLaMMo processes **288 FPS**, representing a **52% speedup** over LLaMo-CT's 189 FPS and demonstrating real-time inference as group size grows. This efficiency gain results from M3Former's fusion-token strategy, which maintains a constant sequence length regardless of the number of people. In contrast, the baseline's flattening approach causes tokens length to grow quadratically. Together, these results show that LLaMMo's advanced interaction modeling not only strengthens understanding across all metrics but enables accelerated, high-fidelity reasoning in practical, low-latency scenarios.

Evaluations on LLaMI-Bench-Sports As reported in Table 1(A), LLaMMo improves overall Accuracy by **7.5%** and average Score by **0.30** over LLaMo-CT, showcasing deeper domain-specific insights. Notably, the largest gains appear in TA and IU, underscoring LLaMMo's ability to capture precise temporal localization and inter-player dependencies in challenging, fine-grained sports scenarios. These results confirm that our interaction modeling framework effectively addresses the complexity of sports analysis.

We further assess LLaMMo on the commentary-generation track, which demands precise capture of fine-grained motion and visual cues. As Table 3 shows, LLaMMo achieves nearly **+1.0** improvement in commentary metrics (TD, TA, KE), highlighting its capability in fine-grained motion analysis. Additionally, it boosts narrative quality, with gains of **1.3 and 0.9 points** in Emotional Expression and Style, highlighting LLaMMo's strength in delivering vivid, multimodal commentary. In sports, LLaMMo excels at both coach-level analysis and live commentary, showcasing strong multimodal generalization and the ability to deliver fine-grained motion analysis and vivid, context-aware commentary.

Evaluations on Worldpose WorldPose dataset poses an extreme challenge with up to **22** simultaneous players. Modeling such large groups demands scalable attention mechanisms and robust relational reasoning, making this benchmark a critical test of LLaMMo's real-world team-sport understanding. As Table 1(B) shows, LLaMMo raises overall

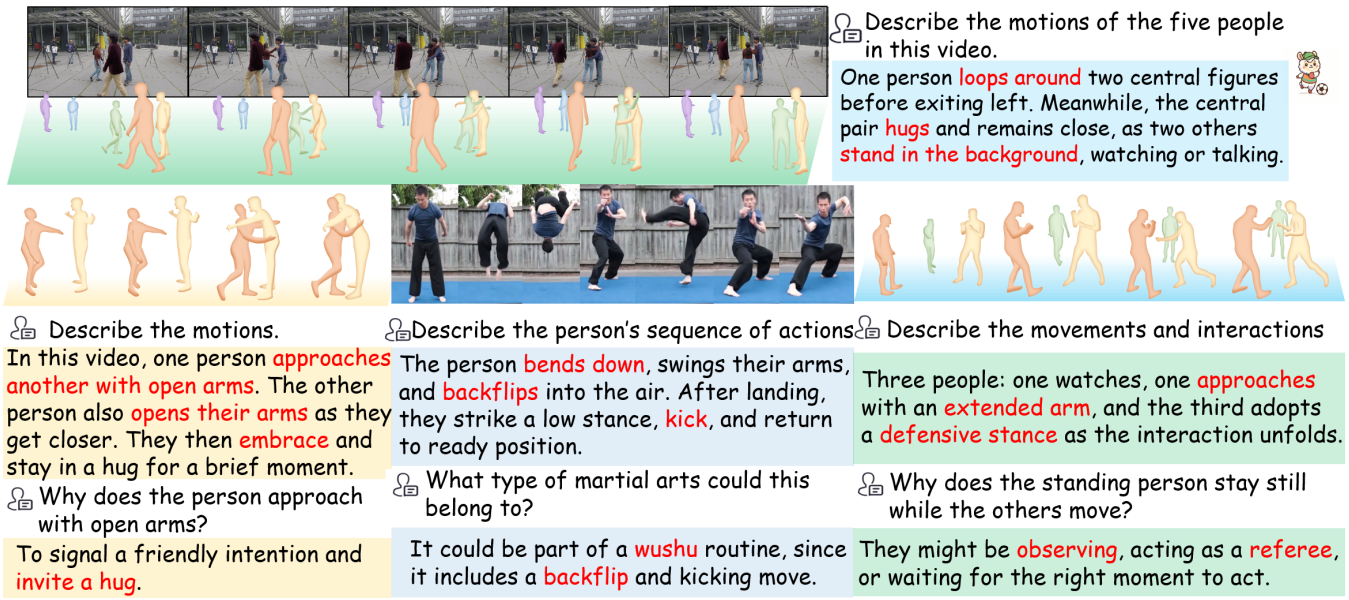

Figure 4: Qualitative results of LLaMMo on LLaMI-Bench-Life, demonstrating its strong generalization to diverse social interaction scenarios.

Accuracy from 21.29% to 28.00% and average Score from 1.95 to 2.48. On the LLaMI-Bench-Life, LLaMMo's inference speed is **288 FPS**, and it only decreases to **264 FPS** on WorldPose; by contrast, LLaMo-CT falls from **189 FPS** to **80 FPS** under the same conditions. This efficiency gap underscores the scalability of LLaMMo's design, where structured token fusion ensures consistent inference speed, unlike baselines that scale poorly with group size due to flat token representations. Coupled with consistent gains across all evaluation dimensions and improved BertScore, **these results confirm LLaMMo's scalability and fine-grained motion understanding in full-team scenarios**.

Evaluations on Golf-swing To assess LLaMMo's adaptability to fine-grained single-person tasks, we evaluate it on the Golf-Swing dataset, which requires precise temporal reasoning and domain-specific structure beyond everyday interactions. Multi-person pretraining improves motion understanding and response quality, increasing Overall Accuracy from 24.8% to 26.0% and average Score from 2.48 to 2.60. Minor declines in Pertinence and Adaptability suggest that golf-specific training still offers an advantage for domain-specialized feedback. These results confirm that interaction-driven pretraining transfers positively to single-player tasks, while highlighting the value of targeted fine-tuning in high-precision domains.

Ablations on IA-selector Compared to the baseline, introducing the IA-Selector yields a +1.48% gain in accuracy and a +0.12 improvement in average score. Semantic Consistency also increases from 2.88 to 3.05, suggesting that language-guided keyframe selection enables the model to focus on more semantically aligned interactions and delivers refined representations to downstream modules.

Ablations on M3Former We ablate M3Former to quantify each module's impact. TempFormer improves consistency and accuracy, SocFormer boosts relational metrics, and TSTalker further lifts accuracy and coherence, confirming effective integration of temporal and social cues. Overall, the results validate the M3Former design.

Qualitative Study

To evaluate LLaMMo in real-world multi-human settings, we qualitatively test it on LLaMI-Bench-Life and LLaMI-Bench-Sports (Figure 4, 2). LLaMMo captures fine-grained social dynamics in daily interactions and scales to large sports scenes (up to 22 players), generating tactically aware, persona-adaptive commentary. In soccer, it models team structure and maintains coherent narratives, demonstrating strong competence in complex group behavior reasoning.

Conclusion

We present **LLaMMo**, the first instruction-tuned multimodal framework for multi-human motion understanding, capable of reasoning about interactions across both daily and professional domains. Built on a novel social-temporal architecture and trained with **LLaVerse**, the first large-scale dataset for multi-human motion understanding, LLaMMo produces interaction-aware, scalable, and context-adaptive representations. Extensive evaluations on **LLaMI-Bench**, a comprehensive benchmark for multi-person interaction reasoning, confirm its effectiveness in generating coherent, role-sensitive descriptions, even in large-group, real-time scenarios. We envision LLaMMo and LLaVerse as foundational tools for future multimodal agents that interpret human interactions in dynamic real-world environments and support a wide range of interaction-centric applications.

References

Bertenthal, B. I.; and Boyer, T. W. 2015. The development of social attention in human infants. *The many faces of social attention: Behavioral and neural measures*, 21–65.

Cai, C.; Zhao, X.; Liu, H.; Jiang, Z.; Zhang, T.; Wu, Z.; Hwang, J.-N.; and Li, L. 2025. The Role of Deductive and Inductive Reasoning in Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, L.-H.; Lu, S.; Zeng, A.; Zhang, H.; Wang, B.; Zhang, R.; and Zhang, L. 2024. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *arXiv preprint arXiv:2405.20340*.

Chi, M. T. 2008. Interactive learning environments: A literature review and a meta-analysis. In *Proceedings of the 8th International Conference on Learning Sciences*.

Deng, T.; Shen, G.; Qin, T.; Wang, J.; Zhao, W.; Wang, J.; Wang, D.; and Chen, W. 2024. PLGSLAM: Progressive Neural Scene Represenation with Local to Global Bundle Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19657–19666.

Deng, T.; Shen, G.; Xun, C.; Yuan, S.; Jin, T.; Shen, H.; Wang, Y.; Wang, J.; Wang, H.; Wang, D.; et al. 2025. Mneslam: Multi-agent neural slam for mobile robots. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1485–1494.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Gautam, S.; Sarkhoosh, M. H.; Held, J.; Midoglu, C.; Cioppa, A.; Giancola, S.; Thambawita, V.; Riegler, M. A.; Halvorsen, P.; and Shah, M. 2024. SoccerNet-Echoes: A Soccer game audio commentary dataset. In *2024 International Symposium on Multimedia (ISM)*, 71–78. IEEE.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2024. ReMoS: 3D Motion-Conditioned Reaction Synthesis for Two-Person Interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Giancola, S.; Amine, M.; Dghaily, T.; and Ghanem, B. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1711–1721.

Gu, R.; Jia, S.; Ma, Y.; Zhong, J.; Hwang, J.-N.; and Li, L. 2025. MoCount: Motion-Based Repetitive Action Counting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9026–9034.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5152–5161.

Guo, T.; Lu, B.; Wang, F.; and Lu, Z. 2025. Depth-aware super-resolution via distance-adaptive variational formulation. *Journal of Electronic Imaging*, 34(5): 053018–053018.

He, Y.; Li, S.; Li, K.; Wang, J.; Li, B.; Shi, T.; Xin, Y.; Li, K.; Yin, J.; Zhang, M.; et al. 2025a. GE-Adapter: A General and Efficient Adapter for Enhanced Video Editing with Pretrained Text-to-Image Diffusion Models. *Expert Systems with Applications*, 129649.

He, Y.; Li, S.; Wang, J.; Li, K.; Song, X.; Yuan, X.; Li, K.; Lu, K.; Huo, M.; Tang, J.; et al. 2025b. Enhancing low-cost video editing with lightweight adaptors and temporal-aware inversion. *arXiv preprint arXiv:2501.04606*.

Jeong, J.; Park, D.; and Yoon, K.-J. 2024. Multi-Agent Long-Term 3D Human Pose Forecasting via Interaction-Aware Trajectory Conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16975–16984.

Jia, S.; and Li, L. 2024. Adaptive Masking Enhances Visual Grounding. *arXiv preprint arXiv:2410.03161*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795*.

Jiang, T.; Billingham, J.; Müksch, S.; Zarate, J.; Evans, N.; Oswald, M. R.; Polleyfeys, M.; Hilliges, O.; Kaufmann, M.; and Song, J. 2024. WorldPose: A world cup dataset for global 3D human pose estimation. In *European Conference on Computer Vision*, 343–362. Springer.

Jin, C.; Che, T.; Peng, H.; Li, Y.; Metaxas, D.; and Pavone, M. 2024. Learning from teaching regularization: Generalizable correlations should be easy to imitate. *Advances in Neural Information Processing Systems*, 37: 966–994.

Jin, C.; Peng, H.; Zhang, Q.; Tang, Y.; Metaxas, D. N.; and Che, T. 2025. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*.

Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; and Sheikh, Y. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, 3334–3342.

Kuang, S. 2016. Two polarities of attention in social contexts: From attending-to-others to attending-to-self.

Lai, Z.; Yang, J.; Xia, S.; Lin, L.; Sun, L.; Wang, R.; Liu, J.; Wu, Q.; and Pei, L. 2025. RadarLLM: Empowering Large Language Models to Understand Human Motion from Millimeter-wave Point Cloud Sequence. *arXiv preprint arXiv:2504.09862*.

Lan, T.; Xu, J.; He, X.; Hwang, J.-N.; and Li, L. 2025. Attention Consistency for LLMs Explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1736–1750. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Li, L. 2024. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 513–522.

Li, L.; Jia, S.; Wang, J.; An, Z.; Li, J.; Hwang, J.-N.; and Belongie, S. 2025. Chatmotion: A multimodal multi-agent for human motion analysis. *arXiv preprint arXiv:2502.18180*.

Li, L.; Jia, S.; Wang, J.; Jiang, Z.; Zhou, F.; Dai, J.; Zhang, T.; Wu, Z.; and Hwang, J.-N. 2024a. Human Motion Instruction Tuning. *arXiv preprint arXiv:2411.16805*. Accepted to CVPR 2025.

Li, L.; Jia, S.; Wang, J.; Jiang, Z.; Zhou, F.; Dai, J.; Zhang, T.; Wu, Z.; and Hwang, J.-N. 2024b. Human motion instruction tuning. *arXiv preprint arXiv:2411.16805*.

Li, Y.; Zhu, Q.; Zhou, L.; et al. 2019. Social role-aware emotion recognition in conversations. In *ACL*.

Liang, H.; Zhang, W.; Li, W.; Yu, J.; and Xu, L. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9): 3463–3483.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280.

Liu, P.; Liu, H.; Liu, X.; Li, Y.; Chen, J.; He, Y.; and Ma, J. 2025. Scene-Aware Explainable Multimodal Trajectory Prediction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10786–10792. IEEE.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Lu, B.; Lu, Z.; Qi, Y.; Guo, H.; Sun, T.; and Zhao, Z. 2025. Predicting asphalt pavement friction by using a texture-based image indicator. *Lubricants*, 13(8): 341.

Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; and Theobalt, C. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3D vision (3DV)*, 120–130. IEEE.

Müller, M.; and Röder, T. 2006. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 137–146.

Peng, X.; Mao, S.; and Wu, Z. 2023. Trajectory-Aware Body Interaction Transformer for Multi-Person Pose Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stefański, P.; Kozak, J.; and Jach, T. 2024. Boxing Punch Detection with Single Static Camera. *Entropy*, 26(8): 617.

Tanke, J.; Zhang, L.; Zhao, A.; Tang, C.; Cai, Y.; Wang, L.; Wu, P.-C.; Gall, J.; and Keskin, C. 2023. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9601–9611.

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

Voeikov, R.; Falaleev, N.; and Baikulov, R. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 884–885.

Wang, J.; He, Y.; Li, K.; Li, S.; Zhao, L.; Yin, J.; Zhang, M.; Shi, T.; and Wang, X. 2025. MDANet: A multi-stage domain adaptation framework for generalizable low-light image enhancement. *Neurocomputing*, 627: 129572.

Wang, J.; Xu, H.; Narasimhan, M.; and Wang, X. 2021. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34: 6036–6049.

Wu, Q.; Zhao, Y.; Wang, Y.; Tai, Y.-W.; and Tang, C.-K. 2024a. Motionllm: Multimodal motion-language learning with large language models. *arXiv e-prints*, arXiv–2405.

Wu, T.; He, R.; Wu, G.; and Wang, L. 2024b. SportsHHI: A Dataset for Human-Human Interaction Detection in Sports Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18537–18546.

Xu, C.; Tan, R. T.; Tan, Y.; Chen, S.; Wang, Y. G.; Wang, X.; and Wang, Y. 2023. EqMotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1410–1420.

Xu, L.; Lv, X.; Yan, Y.; Jin, X.; Wu, S.; Xu, C.; Liu, Y.; Zhou, Y.; Rao, F.; Sheng, X.; et al. 2024. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22260–22271.

Xu, S.; Wang, Y.-X.; and Gui, L.-Y. 2023. Stochastic Multi-Person 3D Motion Forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yamato, J.; Ohya, J.; and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. In *CVPR*, volume 92, 379–385.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yao, Z.; Cheng, X.; Huang, Z.; and Li, L. 2025. Countllm: Towards generalizable repetitive action counting via large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19143–19153.

Yu, H.; Zhang, J.; Chen, C.; Xiang, T.; Fang, Y.; Niebles, J. C.; and Adeli, E. 2025. SocialGen: Modeling Multi-Human Social Interaction with Language Models. *arXiv preprint arXiv:2503.22906*.

Zhou, Y.; He, Y.; Su, Y.; Han, S.; Jang, J.; Bertasius, G.; Bansal, M.; and Yao, H. 2025. ReAgent-V: A Reward-Driven Multi-Agent Framework for Video Understanding. *arXiv preprint arXiv:2506.01300*.

Zwaan, R. A.; and Radvansky, G. A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2): 162–185.