Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: www.tandfonline.com/journals/tadr20

Anticipating daily human actions: comparing pipelines for long-term skeleton-based prediction in real-world scenarios

Junhan Wen, Xucong Zhang & Jouh Yeong Chew

To cite this article: Junhan Wen, Xucong Zhang & Jouh Yeong Chew (24 Feb 2026): Anticipating daily human actions: comparing pipelines for long-term skeleton-based prediction in real-world scenarios, Advanced Robotics, DOI: 10.1080/01691864.2026.2626369

To link to this article: https://doi.org/10.1080/01691864.2026.2626369

Published online: 24 Feb 2026.

ADVANCED ROBOTICS
https://doi.org/10.1080/01691864.2026.2626369

SHORT PAPER

OPEN ACCESS Check for updates

Anticipating daily human actions: comparing pipelines for long-term skeleton-based prediction in real-world scenarios

Junhan Wen[a,b], Xucong Zhang[b] and Jouh Yeong Chew[a]

[a]Honda Research Institute Japan, Wako, Saitama, Japan; [b]Delft University of Technology, Delft, Netherlands

ABSTRACT

Human action anticipation remains a key challenge to achieve efficient human-robot interaction due to the difficulties to learn the higher level of abstraction. This work explores three action anticipation pipelines as a guideline for future work. Specifically, two pipelines adopt a top-down approach: they recognize current actions and then anticipate future actions using either traditional machine learning models or Large Language Models (LLMs). The third pipeline follows a bottom-up strategy by first forecasting future motions and then inferring actions. Our results show that top-down pipelines achieve higher accuracy and robustness, demonstrating the advantage of abstract reasoning over direct motion-based inference.

ARTICLE HISTORY
Received 6 July 2025
Revised 25 September 2025
and 2 December 2025
Accepted 22 December 2025

KEYWORDS
Action anticipation; action recognition; motion forecasting; large language models; daily action analysis

1. Introduction

Action anticipation is critical for applications like human-robot collaboration [1] and robot motion generation [2]. However, understanding and forecasting human behavior from visual data is a long-standing challenge. These tasks become increasingly difficult as the prediction horizon extends further into the future, due to greater uncertainty and variability in human behavior [3–5]. In this work, we distinguish between *'motions'* as the low-level physical trajectories of body joints over time, and *'actions'* as higher-level, semantically meaningful descriptions of activities (e.g. cutting, reaching). Following this distinction and in line with prior works, we use *(action) 'anticipation'* to denote high-level action predictions and *(motion) 'forecasting'* to denote low-level, coordinate-based motion predictions [6,7]. We use *prediction* only as a general term for these computer vision tasks in introductory and related discussions. In addition, we use the expression *predicted actions* to refer to labels produced by predictive models, distinguishing them from the direct outputs of language models, which we call *anticipated actions*.

In realistic settings, people interact with other individuals and objects in shared environments. Modeling these interactions is important for accurate forecasting, as they can serve as contextual cues [8–10]. However, such modeling remains difficult due to the complexity of social dynamics and the challenges of curating representative and accurate interaction datasets. As a result, multi-person scenarios are underexplored despite being common in real-world applications [11].

Many prior approaches rely on RGB or multimodal inputs, which can be sensitive to visual variations such as background and appearance [12,13]. Skeleton data, by contrast, offers a compact and generalizable representation of motion that abstracts away low-level visual cues [14,15]. While often used for low-level motion forecasting, skeleton data has high potential for appearance-agnostic action forecasting [15,16]. Nonetheless, how to robustly perform action anticipation in a social setting remains an open question.

In this paper, we investigate three pipelines for action anticipation using the *'Humans in Kitchens'* (HIK) dataset [7], which represents realistic, multi-person interactions through 3D skeletonlines investigated in this. As shown in Figure 1, two of the pipelines follow a top-down strategy: they first recognize ongoing actions and then forecast future ones, using either classical machine learning (ML) models or large language models (LLMs) for the anticipation task. The third pipeline adopts a bottom-up approach, first forecasting future motion sequences and then inferring actions from the predicted movements. This setup enables a systematic comparison of different anticipation strategies in a unified and realistic input-output setting.

CONTACT Jouh Yeong Chew jouhyeong.chew@jp.honda-ri.com

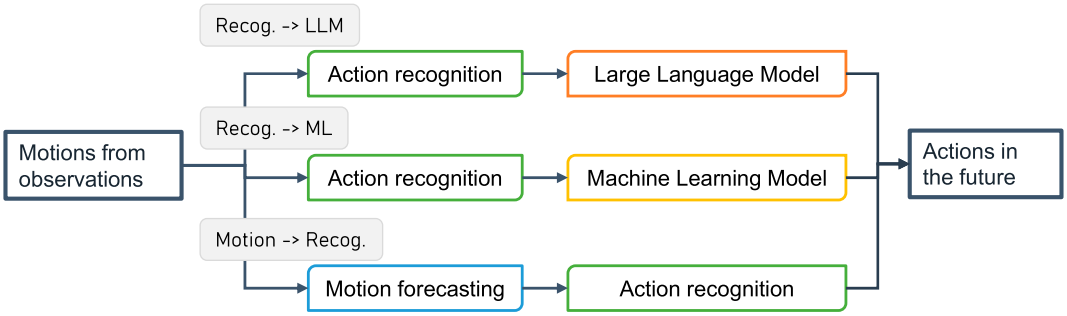

Figure 1. The three pipelines investigated in this paper. The first pipeline *Recog.- > LLM* uses a machine learning model to recognize current actions, which are then fed into an LLM to forecast future actions. The second pipeline *Recog.- > ML* also recognize current actions while using a predictive model for the action anticipation. The third pipeline *Motion- > Recog.* first forecast the future motions, and then uses these predicted motions to recognize the actions.

The contributions of this paper are therefore threefold:

- We formulate a challenging action anticipation task, featuring complex observer-view scenarios with multitasking subjects, and support varied label formats.
- We design and implement three distinct model pipelines to solve this task, including two top-down (action-first) and one bottom-up (motion-first) approaches.
- We benchmark these pipelines on a naturalistic, multi-person dataset of solely skeleton data, providing a systematic comparison of performance baselines.

2. Related work

The task of action prediction is compelling in computer vision, spanning tasks from recognizing ongoing actions to anticipating future ones. The task of anticipation is particularly challenging, as uncertainty and potential errors grow with the prediction horizon [3]. While action recognition is a well-established field with large-scale benchmarks such as NTU RGB+D [17] and AVA [9], action anticipation remains less explored despite its clear real-world utility.

To model the complex temporal dynamics of human action, a variety of architectures have been explored, including CNNs, RNNs, Transformers, and transitional models [5,18–21]. Recently, Large Language Models (LLMs) have emerged as a powerful tool for this task, leveraging textual cues or recognized action labels for high-level sequential reasoning [22–24]. Since LLMs do not operate directly on video features, this necessitates a two-stage, top-down approach: first recognizing actions, then prompting the LLM for future anticipations [25,26]. The alternative is a bottom-up strategy, which first forecasts low-level motion and then applies recognition to those predicted movements [27,28]. While both top-down and bottom-up strategies are cognitively inspired, they have not yet been systematically compared, particularly in a realistic, multi-person setting.

A significant gap in this field stems from the limitations of datasets. Many benchmarks for observer-view action recognition, for instance, rely on scripted behaviors or provide only a single action label per clip [17,29–31], which simplify the task at the cost of real-world applicability. While newer, LLM-based studies leverage more natural datasets, they are almost exclusively benchmarked from an egocentric perspective [25,26,32], which limits the analysis to a single actor's primary task and overlooks the complexity of natural multitasking. Meanwhile, the related field of motion forecasting has demonstrated that incorporating contextual and social cues significantly enhances anticipation performance of models [16,33]. These parallel efforts highlight a gap in benchmarks that combine realism, multi-person interaction, and observer-view settings. To address this, we turn our attention to the recently-released *'Humans in Kitchens'* (HIK) dataset [7], which was previously benchmarked only for motion forecasting but also includes high-level action annotations, making it well-suited for our study.

3. Method and materials

3.1. Modeling approaches

Our work systematically compares three distinct pipelines for action anticipation using skeleton data, as illustrated in Figure 2. Specifically, we focus on high-level action anticipation using low-level skeleton data, which provides rich spatio-temporal information suitable for modeling human motion while avoiding biases present in ego-vision or RGB-based approaches. We investigate two *top-down* (action-first) pipelines that first recognize ongoing actions and then use these high-level labels to anticipate

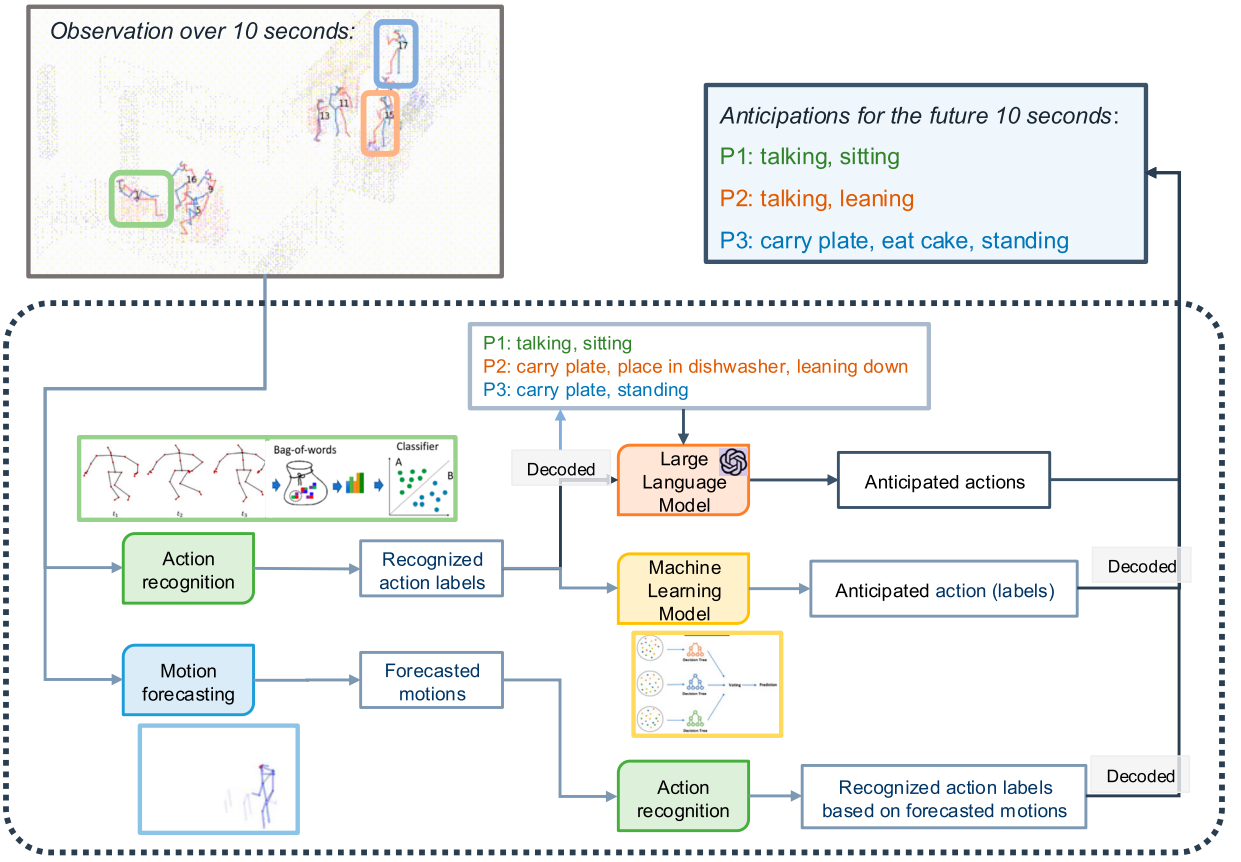

Figure 2. Detailed pipelines investigated in this paper, with the functionality blocks color-coded. The upper pipeline represents the *top-down* methodology, which branches into two variants due to different forecasting strategies used in the second stage. The lower pipeline represents the *bottom-up* methodology, where the second stage employs the same models as the first stage of the top-down approach.

future actions. These are contrasted with a *bottom-up* (motion-first) pipeline that first forecasts future motions and then infers actions from the predicted motions. This experimental design allows us to directly compare the effectiveness of abstract, label-based reasoning against low-level temporal modeling within a unified framework. Unlike prior benchmarks on skeleton-based action recognition that typically focus on scripted or independent actions, our setup leverages a realistic, interaction-rich dataset, extending recognition to inter-dependent, real-world actions and connecting functional blocks to achieve generalizable anticipation.

All pipelines are built upon a shared action recognition backbone, a tailored model built upon *Hyperformer* [34], a high-performing transformer that demonstrated strong action recognition performance on the NTU RGB+D dataset [17], with reconfigured and fine-tuned to ensure it is fully adapted to the setting of the multi-person and multi-label dataset. To ensure compatibility with both upstream and downstream modules, we train the action recognition models to produce two distinct output formats. We use *hard labels* (one-hot vectors) to identify the dominant action, as their clear class boundaries are beneficial for tasks like class re-balancing. In parallel, we use *soft labels* (float vectors to show the proportion of each action's presence in a period) because they provide a detailed representation of the multiple, overlapping behaviors common in our setting. The two top-down pipelines use these labeled recognitions to feed either a classical ML model or an LLM, which learn to anticipate future actions. The bottom-up pipeline, in contrast, employs a fine-tuned implementation of SAST [16] to first future skeleton motions before they are interpreted by the same recognition module.

3.2. Dataset and pre-processing

Aligning with our motivation, we select skeleton data as our primary input modality to promote model generalizability and ensure robustness against variations in appearance, lighting, and background. Our experiments are grounded in the *'Humans in Kitchens'* (HIK) dataset [7], which collected 3D skeleton sequences from realistic, unscripted multi-person interactions. HIK is particularly suitable as a benchmark for action anticipation because it provides both low-level motion

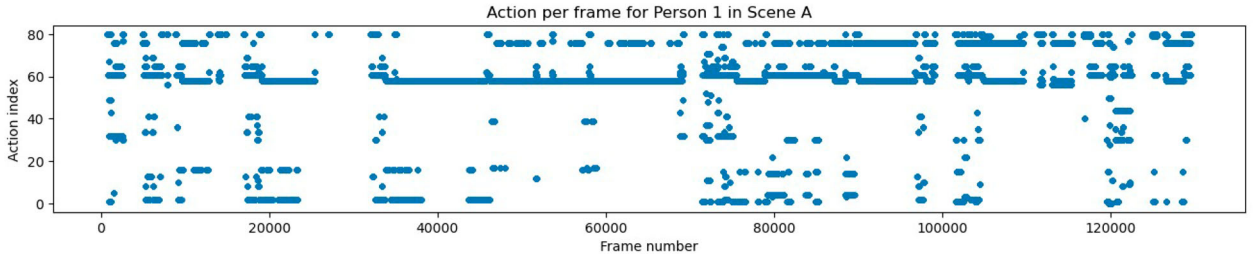

Figure 3. Temporal presence of action labels for Person 1 in Scene A. The x-axis represents video frames, and the y-axis denotes action label indices. Each dot indicates that a specific action (y-axis) is annotated as occurring at a given frame (x-axis).

(skeletons) and high-level action annotations of real-world activities at a per-frame resolution, as presented in Figure 3, thereby enabling direct comparisons across levels of abstraction. Meanwhile, to maintain clarity while preserving meaningful distinctions, the annotation scheme in HIK [7] merges actions that differ semantically but not functionally, e.g. all eating or drinking activities are grouped together, while keeping distinguishable actions, e.g. *carrying a plate* versus *carrying a whiteboard marker*. This ensures that the filtered label set aligns with our skeleton-only representation and avoids unsupported object- or tool-specific distinctions. Prior benchmarks of it, however, have focused almost exclusively on motion forecasting [16], leaving the task of action anticipation underexplored in this modality.

A key distinction of our setting is the need for a multi-label formulation. In realistic scenarios, individuals often perform several actions concurrently (e.g. *standing* while *talking* and *carrying a plate*). Rather than collapsing these into a single class through one-hot encoding, we represent each time window with a *soft label*, which is a vector and every entry is a proportion between 0 and 1 indicating the fraction of frames annotated with that action. This provides a faithful description of overlapping behaviors and better reflects the richness of real-world activity.

While the realism of the HIK dataset makes it an excellent setting for understanding human activities in multi-person scenarios, it also introduces challenges that are common in real-world data. First, the social environment is highly dynamic, meaning the number of individuals present is not constant. People naturally appear and disappear from view, which leads to dense periods of co-occurrence, with up to nine individuals present simultaneously in the test scene (Scene A) and even up to 14 in others. Second, this complexity extends to individual behavior, as people frequently perform multiple actions concurrently. This results in a single person having up to six simultaneous action labels, as visualized in Figure 3. Finally, the unscripted nature of the actions leads to a highly imbalanced data distribution, where common activities like *sitting* are far more noted than events like *starting a dishwasher*. Taken together, it is precisely these challenges that make HIK a difficult yet well-suited benchmark for developing anticipation models robust enough for real-world applications.

To establish a standardized framework, we first define the overall pipeline configuration. The core anticipation task is set as *observing a 10-second (250-frame) window to predict the subsequent 10 seconds*. We refined the original 82 action categories of HIK into a more tractable set of 34 labels. This reduction follows two clear criteria: (1) the action must appear for at least 250 frames in more than one scene; and (2) it must have at least one continuous instance of more than 300 frames. Actions that failed both conditions were not taken into account, as they are insufficient for training anticipation models. For example, highly fragmented or single-scene activities were excluded from the dataset that we used, while ambiguous actions that cannot be resolved from skeleton data alone, e.g. *talking*, were also removed. This yields the benchmark to be less affected by label scarcity and better aligned with what skeleton-based models can realistically capture.

In addition, HIK has an inherent significant imbalance due to its setting that applying a 250-frame window with a uniform stride would contain up to 80% of a dominant class, leaving all others below 10%. In such cases, a trivial model that always outputs the dominant action could already achieve deceptively strong scores. Since augmentation of skeleton data might include more noise, we apply a dynamic sliding window strategy during training to specifically counter this. Specifically, the stride is adjusted between 1 and 250 based on frequency and continuity of each action, ensuring that the resulting training samples are distributed more evenly across the filtered 34 classes. This reduces the risk of bias in dominant categories and allows less frequent but meaningful actions to contribute more effectively to model learning.

3.3. Metrics and evaluations

We use a 10-second observation window followed by a 10-second anticipation window for the core anticipation task, providing a consistent temporal context for comparison across all pipelines. Two types of inputs are considered: (i-1) one hard label per second and (i-2) one soft label over the period; and three types of outputs: (o-1) one hard label per second, (o-2) one hard label over the period, and (o-3) one soft label over the period as previously described.

Pipeline performance is primarily evaluated using Top-1 and Top-5 accuracy on the predicted actions, expressed as percentages. While some shared metrics, such as Top-k accuracy, are reported, these cannot be directly compared across different output types. For example, an 'any-hit' criterion is applied for soft-label anticipation, favoring models that better capture probabilities or action presence. For LLM-based pipelines, accuracy is assessed after a lemmatization process, where if the processed label cannot be matched to any action in the predefined pool, it is treated as an anticipation that no relevant action occurred. In addition, the first generated response is taken as the effective prediction for calculating the Top-1 accuracy because their predictions are not ranked by probabilities. Additionally, edit distance is used to compare LLM outputs, providing insight into the models' ability to anticipate verbs and nouns correctly. Component-level evaluations employ task-specific metrics that, at the action-label level, hard-label anticipations use mainly cross-entropy, while soft-label anticipations are assessed additionally with mean squared error (MSE); at the motion level, forecasting in the bottom-up pipeline is evaluated using both MSE and Normalized Directional Motion Similarity (NDMS) [35]. To enhance readability, we use ↑/↓ indicators in the result tables in Section 4, but include them only at their first occurrence to avoid unnecessarily size expansion of them.

To ensure consistency across pipelines and enable fair comparisons, we create a fixed test set from Scene A of HIK, based on the aforementioned filtering criteria. For more detailed, component-level analyses, we include additional configurations and input–output tasks, covering train–test splits by person versus by scene, comparisons across ML methods, and evaluations of zero-shot versus fine-tuned LLMs. Because the accuracies are mostly computed once over a single fixed test set, we report it without standard deviation; in contrast, per-sample metrics include standard deviation (std.) when they serve as primary criteria, e.g. MSE or NDMS for motion forecasting. In addition, we additionally report the std. of Top-1 accuracy of LLMs, since the responses are unranked. These metrics and evaluation procedures provide a standardized framework to quantify both end-to-end pipeline performance and individual module contributions, highlighting strengths and limitations across diverse input–output configurations.

4. Experimental results

In this section, we present the experimental results. We first compare the performance of the three complete action anticipation pipelines, followed by a detailed analysis of the individual modules, i.e. skeleton-based action recognition, skeleton-based motion forecasting, and label-level action anticipation. This two-level evaluation allows us to assess both the end-to-end effectiveness of each pipeline and the specific contribution of each component.

4.1. Overall performance comparison of pipelines

As shown in Table 1, our evaluation reveals clear trade-offs between the pipelines, with the preferable method depending on the specific anticipation task and desired output.

As shown in the table, top-down approaches exhibit a distinct performance pattern depending on the granularity and output type of the task. The *Recog.- > LLM* pipeline performs best for fine-grained and probabilistic reasoning, achieving the highest Top-1 accuracy for both per-second hard label and period-level soft label anticipations. Its strength comes from leveraging knowledge and reasoning capabilities of LLMs acquired during their pre-training phase to model temporal dependencies, making it well-suited for applications that require either the most likely instantaneous action or a comprehensive anticipation over time.

On the other hand, the *Recog.- > ML* pipeline is more effective at producing a final probabilistic output for an aggregated future period. This suggests it excels at consolidating complex observations into a reliable probability forecast. These results emphasize the complementary expertise of the two pipelines: the *Recog.- > ML* pipeline is preferable for a single aggregated probabilistic anticipation, whereas the *Recog.- > LLM* pipeline is better for sequential action anticipation or generating a diverse set of possibilities.

In contrast, the bottom-up pipeline suffers from error propagation: inaccuracies in motion forecasting degrade action recognition performance, particularly when recognition is based on individual (separate) and short observation windows (i.e. conducting the recognition on motions per second). Nevertheless, this approach still yields reasonable Top-5 accuracy, suggesting the

Table 1. Performance comparison of the complete action anticipation pipelines, evaluated by Top-1 and Top-5 accuracy.

Anticipation Task	Method	Specified Configuration	Top1 Acc.[%] ↑	Top5 Acc.[%] ↑
Hard label per second	Recog. - > LLM	*recog. - > original GPT-4*	14.02	17.91
		Using hard label per second	**24.24**	35.89
	Recog. - > ML	Using hard label per second	17.54	34.56
	Motion - > Recog.	*random forecasts - > recog.*	0.82	9.81
		Forecast with full observations	18.15	**68.83**
		Forecast with last 25 frames∗	8.44	58.5
Hard label over 10 second	Recog. - > LLM	Using hard label per sec.	21.30	27.27
		Using soft label over 10 sec.	20.61	24.60
	Recog. - > ML	Using hard label per sec.	15.49	29.47
		Using soft label over 10 sec.	**44.08**	48.16
	Motion - > Recog.	Forecast with full observations	15.03	**62.06**
		Forecast with last 25 frames	15.78	**58.13**
Soft label over 10 second	Recog. - > LLM	Using hard label per sec.	**23.15**	**72.89**
		Using soft label over 10 sec.	20.01	**76.09**
	Recog. - > ML	Using hard label per sec.	18.16	36.16
		Using soft label over 10 sec.	**27.43**	40.06
	Motion - > Recog.	Forecast with full observations	13.36	50.52
		Forecast with last 25 frames	12.92	49.95

Note: Two baselines, shown in italics, are benchmarked on the hard-label-per-second task: using a zero-shot learning for the *Recog.- > LLM* pipeline ('recog.- > original GPT-4') and a random forecaster for the *Motion- > Recog.* pipeline ('random forecasts→recog'.). Configurations marked with an asterisk (∗) use a different input to match the original setting as in [16], as discussed further in Section 4.3.

Table 2. Action recognition performance on its independent test set.

Recognition Task	Training Loss	Data Split	Top1 Acc. ↑	Top5 Acc. ↑	CE ↓	MSE ↓
Hard label for 1 sec.	Cross-Entropy	person	35.25	72.90	3.29	–
	Cross-Entropy	scene	7.94	24.20	6.28	–
Hard label for 10 sec.	Cross-Entropy	person	62.21	93.03	1.52	–
	Cross-Entropy	scene	20.52	71.00	3.69	–
Soft label for 10 sec.	Cross-Entropy	person	61.64	92.68	8.44	178.6
	Cross-Entropy	scene	27.86	72.00	10.93	265.4
	MSE	person	60.41	91.59	7.74	235.2
	MSE	scene	28.17	71.95	11.04	267.4

Note: Metrics include Top-1/Top-5 accuracy, Cross-Entropy (CE), and MSE for soft-label tasks.

action recognition model is robust to suggest proper sets of candidate anticipations against noisy forecasts, yet to a limited extent. Overall, the results confirm that the *Motion- > Recog.* pipeline is a viable method, but top-down approaches are relatively more coherent and robust against error accumulations.

4.2. Action recognition performance

The action recognition models are evaluated independently, with results presented in Table 2. The models achieve higher accuracy with a longer, 10-second observation window. Interestingly, while cross-entropy for soft-label recognitions is higher due to their complexity, the 'any-hit' accuracy remained comparable to that of hard labels. We also observed that the choice of training loss (MSE or cross-entropy) does not significantly impact the final accuracy on soft-label anticipation tasks. Nevertheless, Table 2 shows a shared and noticeable performance drop over all configurations when the train-test split is based on the scene rather than the person. This suggests the model generalizes effectively to new action performers but struggles with unfamiliar scene layouts, likely because variations in object placement and orientation are difficult to normalize.

4.3. Motion forecasting performance

For motion forecasting, the first stage of our bottom-up pipeline, we replicate the SAST model from [16] but adapt its configuration to align with the pipelines' shared configurations by extending the input observation window from the original 1 second to 10 seconds. Our benchmark results, presented in Table 3, show a slight performance improvement over the original SAST publication. We attribute this to methodological differences, including the use of a comparatively easier test set (Scene A vs. D) and an action-label-based frame filtering strategy in place of uniform sampling. More critically, we find that forecast quality measured by MSE does not always correlate with the success of the downstream recognition task. Instead, higher NDMS scores prove to be a more reliable forecaster of improved recognition performance, a relationship we summarize in Table 1.

4.4. Action-level anticipation performance

4.4.1. Machine learning on categorical action labels

Table 4 summarizes the performance of three classical ML methods on the test set of label-level anticipation. Among them, k-Nearest Neighbors (k-NN) significantly outperforms in terms of Top-1 accuracy, showing its

Table 3. Motion forecasting performance on its independent test set. Metrics include MSE and NDMS [35]. The first row shows the published benchmarks [16] for comparison.

Forecasting Task	Model	MSE	NDMS ↑
Observe 25 frames, forecast next 250 frames	*Benchmarked in [16]*	–	0.17
	Replicated	0.330 ± 0.11	0.19 ± 0.11
Observe 250 frames, forecast next 250 frames	Replicated	0.339 ± 0.38	0.23 ± 0.05

Table 4. Action anticipation performance on its independent test set.

Anticipations	Observations	Model Type	Top1 Acc.	Top5 Acc.	CE ↓	MSE ↓
Hard label per sec.	Hard label per sec.	k-Nearest Neighbors	59.44	63.03	3.97	–
		Random forest	18.62	76.97	2.42	–
		Logistic Regression	12.06	59.16	12.6	–
Hard label over 10 sec.	Hard label per sec.	k-Nearest Neighbors	58.22	61.55	4.22	–
		Random Forest	4.62	67.67	2.62	–
		Logistic Regression	4.02	31.73	13.60	–
	Soft label over 10 sec.	k-Nearest Neighbors	57.15	61.39	1.59	–
		Random Forest	50.52	64.56	1.42	–
		Logistic Regression	4.46	65.98	3.13	–
Soft label over 10 sec.	Soft label over 10 sec.	k-Nearest Neighbors	98.68	99.97	4.56	0.015
		Random Forest	98.93	100.0	13.3	0.015
		Logistic Regression	98.21	99.91	36.0	0.015

Note: Metrics include Top-1/Top-5 accuracy, Cross-Entropy (CE), and MSE for soft-label tasks.

capability to more effectively identify the most likely action by directly comparing observations in the three forms to similar patterns in the training data. Random forests slightly outperform in Top-5 accuracy by capturing a broader distribution of plausible actions, which could also lead to the lower cross-entropy that it achieves. Logistic regression generally underperforms due to its limited capacity to model the complex relationships inherent in this task. This illustrates a trade-off between anticipating the single most probable action versus modeling the overall distribution.

Taken together, while random forests offer some advantages in Top-5 accuracy and cross-entropy, these benefits are marginal compared to the substantial deficit in Top-1 accuracy relative to k-NN, motivating its selection as the preferred model for label-level anticipation, where accurately identifying the next action is critical.

4.4.2. LLMs on descriptive action labels

Our evaluation of LLM-based anticipation, summarized in Table 5, demonstrates clear gains from integrating LLMs into the action anticipation task. Even without adaptation, the zero-shot model performs reasonably well in settings where inputs and outputs share the same representational form (e.g. per-second hard label to per-second hard label, or over-period soft label to over-period soft label). After fine-tuning, the models show substantial improvements over zero-shot baselines, producing more structured and accurate outputs across conditions. They not only become more reliable in these point-to-point tasks but also succeed in resolving harder input–output mismatches, leading to greater robustness and overall stronger performance as summarized in Table 1. Taken together, these results indicate that the LLM stage is not the primary limiting factor of the *Recog.- > LLM* pipeline. Furthermore, when errors from the upstream task, i.e. action recognition, inevitably propagate downstream and cap the achievable performance, the LLM stage demonstrates higher robustness than the *Recog.- > ML* pipeline.

As shown in the two examples below, fine-tuned LLMs can produce outputs that are properly formatted and consistent with task-specific instructions. In many cases, the model correctly highlights the relevant action keywords once its vocabulary has been adapted through training prompts. This makes the anticipations readily usable for further qualitative or quantitative analyses.

However, drawbacks remain that require additional processing to fully exploit these outputs. A central issue is semantic inaccuracy, where the model generates plausible but incorrect labels, often in the form of synonyms or over-detailed descriptions. For instance, the process of 'making coffee' may be expanded into 'stir coffee, taste coffee, adjust taste, sip coffee'. While the format is correct, such verbosity complicates mapping anticipations back to the predefined action categories. To address this, an additional language model would be needed to post-process the outputs (e.g. in this case, we used spaCy for lemmatization), so as to align with the target action dictionary.

Another challenge arises when the task desires soft labels. Although the anticipations remain reasonable and achieve strong Top-k accuracy under the 'one-hit' evaluation policy, their MSE remains higher than that of classical ML models, even after fine-tuning. This likely stems from the LLM's tendency to distribute probability mass across a broader set of plausible actions, which

Table 5. Performance of LLM models for action anticipation on the test set, including fine-tuned results and zero-shot performance using the original OpenAI GPT-4 model.

Antici-pations	Observ-ations	Model	Top1 Acc.	Top5 Acc.	CE	CE′	ED verb	ED noun	MSE	MSE′
Hard label per sec.	Hard label per sec.	zeroshot	82.4 ± 0.3	86.8	6.33	4.74	0.96	1.01	–	–
		finetuned	87.3 ± 0.4	94.9	4.57	1.84	0.49	0.51	–	–
Hard label over 10 sec.	Soft label over 10 sec.	zeroshot	12.1 ± 0.1	73.0	36.0	35.8	0.99	0.76	–	–
		finetuned	92.1 ± 0.4	95.1	2.85	1.77	0.04	0.05	–	–
Soft label over 10 sec.	Soft label over 10 sec.	zeroshot	62.2 ± 0.3	95.7	15.0	5.12	–	–	0.316	0.071
		finetuned	83.3 ± 0.1	100	1.50	0.32	–	–	0.125	0.019

successfully broadens the 'hit space' but has limited capability in achieving numerical precision.

As illustrated in the examples below, formatting inconsistencies can still occur even with properly crafted prompts. For fine-tuned models, outputs occasionally deviate from the required structure, producing verbose or conversational text that complicates automated parsing. These cases typically require additional post-processing to convert the anticipations into structured action labels usable by downstream modules. The issue is more pronounced in zero-shot deployments, where the model frequently generates unstructured outputs that cannot be directly mapped to the expected labels. Such outputs necessitate alternative evaluation strategies and highlight the practical challenges of using LLMs without adaptation, consistent with observations in prior work [25].

These persistent challenges with both semantic and structural reliability underscore a practical reality: leveraging LLMs within an automated pipeline requires a robust post-processing setting. This fact hints that, rather than being used purely as direct prediction engines, LLMs may be most effective in a complementary role alongside structured models. A promising future direction could be to reposition the LLM as a supervisory layer that validates, refines, or critiques the outputs of more structured models, a concept explored in recent works such as [26].

5. Conclusion

In this work, we present and systematically evaluate three distinct pipelines for skeleton-based action anticipation. Our experiments are grounded in the realistic multi-person *Humans in Kitchens* dataset, where social interactions emerge naturally rather than following scripted patterns. This setting poses challenges for action representation, as behaviors are continuous, overlapping, and observed from multiple perspectives. Under this situation, soft labels, which take proportionate of action presence time instead of binary values that form one-hot encoding, can more faithfully capture concurrent actions than simply hard labeling the dominant action.

Our results show that top-down strategies, which first recognize current actions to anticipate future ones, are generally more reliable than a bottom-up approach that suffers from error accumulation. Specifically, among the top-down methods, classical machine learning models demonstrate robust and efficient performance on simplified, temporally abstracted inputs. LLM-based forecasters achieve higher accuracy and can suggest a broader range of plausible future actions; however, they entail greater computational cost, require additional post-processing, and cannot fully exploit the numerical information encoded in soft labels. Together, these results highlight the complementary strengths of the two approaches and suggest opportunities for hybrid pipelines or further refinements in model design.

While it is demonstrated that explicit contextual information, such as other individuals or environmental features, benefits low-level motion forecasting [16], it can also introduce scene-specific biases. In our study, shared performance drops are noted when the train-test split is based on scenes rather than individuals in the action recognition section indicates that the skeleton-based models are able to generalize to new performers but struggle with unfamiliar layouts. These results highlight a trade-off between robustness and informativeness: avoiding environmental biases improves generalization, whereas leveraging context can enhance high-level recognition. Future work could pursue complementary directions, including mitigating scene-specific biases, incorporating richer contextual cues into high-level modules, and integrating semantic knowledge of actions to focus on practically relevant distinctions, such as differentiating 'sit and eat' from merely 'eat' or 'drink'.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Junhan Wen was a Ph.D. student at Delft University of Technology (TU Delft) from 2021 to 2025. Her research focused on machine learning and optimization, with specific applications in deep learning and computer vision for precision agriculture. During her doctoral candidacy, she also conducted research at Honda Research Institute Japan (HRI-JP), focusing on motion

analysis and action forecasting. Her research interests include computer vision, predict-then-optimize paradigms, and data mining. To date, she has authored four journal publications and three peer-reviewed conference and workshop contributions.

Xucong Zhang is an Assistant Professor at Delft University of Technology. He was a postdoctoral researcher at ETH Zurich from 2018 to 2021, and prior to that completed his PhD at the Max Planck Institute for Informatics in Germany from 2013 to 2018. His research focuses on human-centered artificial intelligence, computer vision, and embodied intelligence, with particular interests in human behavior modeling, gaze estimation, and human-robot interaction. He has published extensively in leading journals and conferences in computer vision and artificial intelligence, and his research has been supported by competitive funding from both public agencies and industry partners.

Jouh Yeong Chew is a Senior Scientist and Project Leader at Honda Research Institute Japan, where he works on the analysis and modeling of nonverbal cues for cooperative intelligence. He is interested in developing embodied AI to realize a hybrid society in which ubiquitous embodied AI agents and humans coexist in the real world. His research interests include human-robot interaction, machine learning, generative AI, and robotics. He has conducted collaborative research with Toyota Industries Co. and Tadano Ltd. on the development of adaptive support systems for industrial machines, such as forklifts and mobile cranes, by incorporating behavioral intelligence, including gaze saliency. He serves as a Senior/Guest Editor for Advanced Robotics and has organized workshops at IEEE/ACM conferences such as IROS, ICRA, and HAI. He was also a member of the Advisory Committee of the Innovative Manufacturing, Mechatronics and Materials Forum (iM3F) from 2020 to 2023.

References

[1] Huang CM, Mutlu B. Anticipatory robot control for efficient human-robot collaboration. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Christchurch: IEEE; 2016. p. 83–90.

[2] Koppula HS, Saxena A. Anticipating human activities using object affordances for reactive robotic response. IEEE Trans Pattern Anal Mach Intell. 2015;38(1):14–29. doi: 10.1109/TPAMI.2015.2430335

[3] Kong Y, Tao Z, Fu Y. Deep sequential context networks for action prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii; 2017. p. 1473–1481.

[4] Kong Y, Fu Y. Human action recognition and prediction: a survey. Int J Comput Vis. 2022;130(5):1366–1401. doi: 10.1007/s11263-022-01594-9

[5] Girdhar R, Grauman K. Anticipative video transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual conference; 2021. p. 13505–13515.

[6] Zhong Z, Martin M, Voit M, et al. A survey on deep learning techniques for action anticipation. arXiv preprint arXiv:230917257; 2023.

[7] Tanke J, Kwon OH, Mueller FB, et al. Humans in kitchens: a dataset for multi-person human motion forecasting with scene context. Adv Neural Inf Process Syst. 2023;36:10184–10196.

[8] Zhou Y, Ni B, Hong R, et al. Interaction part mining: A mid-level approach for fine-grained action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA; 2015. p. 3323–3331.

[9] Gu C, Sun C, Ross DA, et al. AVA: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah; 2018. p. 6047–6056.

[10] Vesper C. How to support action prediction: evidence from human coordination tasks. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Edinburgh: IEEE; 2014. p. 655–659.

[11] Zhang HB, Zhang YX, Zhong B, et al. A comprehensive survey of vision-based human action recognition methods. Sensors. 2019;19(5):1005. doi: 10.3390/s19051005

[12] Sun Z, Ke Q, Rahmani H, et al. Human action recognition from various data modalities: a review. IEEE Trans Pattern Anal Mach Intell. 2022;45(3):3200–3225.

[13] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision. Amsterdam: Springer; 2016. p. 20–36.

[14] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana; Vol. 32; 2018. p. 1–9.

[15] Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana; 2022. p. 2969–2978.

[16] Mueller FB, Tanke J, Gall J. Massively multi-person 3D human motion forecasting with scene context. In: European Conference on Computer Vision. Milano: Springer; 2025. p. 130–147.

[17] Liu J, Shahroudy A, Perez M, et al. NTU RGB+ D 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell. 2019;42(10):2684–2701. doi: 10.1109/TPAMI.34

[18] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile; 2015. p. 4489–4497.

[19] Schydlo P, Rakovic M, Jamone L, et al. Anticipation in human-robot cooperation: a recurrent neural network approach for multiple action sequences prediction. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane: IEEE; 2018. p. 5909–5914.

[20] Gong D, Lee J, Kim M, et al. Future transformer for long-term action anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana; 2022. p. 3052–3061.

[21] Zhao H, Wildes RP. On diverse asynchronous activity anticipation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16; Springer; 2020. p. 781–799.

[22] Liu J, Chen C, Liu M. Multi-modality co-learning for efficient skeleton-based action recognition. In: Proceedings

of the 32nd ACM International Conference on Multimedia, Melbourne, Australia; 2024. p. 4909–4918.
[23] Liu R, Li C, Tang H, et al. ST-LLM: large language models are effective temporal learners. In: European Conference on Computer Vision. Milano: Springer; 2024. p. 1–18.
[24] Wang X, Feng M, Qiu J, et al. From news to forecast: integrating event analysis in llm-based time series forecasting with reflection. Adv Neural Inf Process Syst. 2024;37:58118–58153.
[25] Zhao Q, Wang S, Zhang C, et al. AntGPT: can large language models help long-term action anticipation from videos? arXiv preprint arXiv:230716368; 2023.
[26] Beedu A, Haresamudram H, Samel K, et al. On the efficacy of text-based input modalities for action anticipation. arXiv preprint arXiv:240112972; 2024.
[27] Rodriguez C, Fernando B, Li H. Action anticipation by predicting future dynamic images. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany; 2018.
[28] Gammulle H, Denman S, Sridharan S, et al. Predicting the future: a jointly learnt model for action anticipation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea; 2019. p. 5562–5571.
[29] Soomro K, Zamir AR, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402; 2012.
[30] Denina G, Bhanu B, Nguyen HT, et al. Videoweb dataset for multi-camera activities and non-verbal communication. In: Bhanu B, Ravishankar C, Roy-Chowdhury A, et al., editors. Distributed Video Sensor Networks. London: Springer; 2011. https://doi.org/10.1007/978-0-85729-127-1_23
[31] Joo H, Liu H, Tan L, et al. Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile; 2015. p. 3334–3342.
[32] Wang B, Tian Y, Wang S, et al. Multimodal large models are effective action anticipators. arXiv preprint arXiv:250100795. 2025.
[33] Furnari A, Battiato S, Grauman K, et al. Next-active-object prediction from egocentric videos. J Vis Commun Image Represent. 2017;49:401–411. doi: 10.1016/j.jvcir.2017.10.004
[34] Ding K, Liang AJ, Perozzi B, et al. Hyperformer: learning expressive sparse feature representations via hypergraph transformer. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan; 2023. p. 2062–2066.
[35] Tanke J, Zaveri C, Gall J. Intention-based long-term human motion anticipation. In: 2021 International Conference on 3D Vision (3DV). London: IEEE; 2021. p. 596–605.