Journal of Data-centric Machine Learning Research (2026) Submitted 11/25; Revised 08/26; Published 10/26

QuAnTS: Question Answering on Time Series

Felix Divo[1,*] FELIX.DIVO@CS.TU-DARMSTADT.DE
Maurice Kraus[1,*] MAURICE.KRAUS@CS.TU-DARMSTADT.DE
Anh Q. Nguyen[2] ANHQ.NGUYEN23@GMAIL.COM
Hao Xue[3] HAOXUE@HKUST-GZ.EDU.CN
Imran Razzak[4] IMRAN.RAZZAK@MBZUAI.AC.AE
Flora D. Salim[2] FLORA.SALIM@UNSW.EDU.AU
Kristian Kersting[1,5,6,7] KERSTING@CS.TU-DARMSTADT.DE
Devendra Singh Dhami[8] D.S.DHAMI@TUE.NL

[1]AIML Lab, TU Darmstadt
[2]University of New South Wales
[3]Hong Kong University of Science and Technology (Guangzhou)
[4]MBZUAI, Abu Dhabi
[5]Hessian Center for AI (hessian.AI)
[6]German Research Center for AI (DFKI), Darmstadt
[7]Centre for Cognitive Science, TU Darmstadt
[8]TU Eindhoven
[*]Authors contributed equally

Reviewed on OpenReview: https://openreview.net/forum?id=bNCDElSOXB

Editor: Hongyang Zhang

Abstract

Text offers intuitive access to information. This can, in particular, complement the density of numerical time series, thereby enabling improved interactions with time series models to enhance accessibility and decision-making. While the creation of question-answering datasets and models has recently seen remarkable growth, most research focuses on question answering (QA) on vision and text, with time series receiving minute attention. To bridge this gap, we propose a challenging novel time series QA (TSQA) dataset, QuAnTS, for **Qu**estion **An**swering on **T**ime **S**eries data. Specifically, we pose a wide variety of questions and answers about human motion in the form of tracked skeleton trajectories. We verify that the large-scale QuAnTS dataset is well-formed and comprehensive through extensive experiments. Thoroughly evaluating existing and newly proposed baselines then lays the groundwork for a deeper exploration of TSQA using QuAnTS. Additionally, we provide human performances as a key reference for gauging the practical usability of such models. We hope to encourage future research on interacting with time series models through text, enabling better decision-making and more transparent systems.

Keywords: Time Series, Multimodal Question Answering, Dataset

©2026 Divo, Kraus, Nguyen, Xue, Razzak, Salim, Kersting, and Dhami.

1 Introduction

Language is a highly intuitive medium of communication (Fedorenko et al., 2024). Humans learn to listen and verbally express themselves early in life, and most kids learn to read and write soon after. For example, we can easily describe perceived objects, their relations, chains of events, theories, uncertainties, and much more in this unifying modality (Walker and Gopnik, 2014; Jiang et al., 2023). However, in the age of automation, we often encounter other kinds of data, including temporal data consisting of numerical values recorded over time. Time series are complex and precise pieces of data that are highly relevant in many applications. In contrast to the natural use of text, however, end users are often not immediately able to understand and interact with them (Eggleton, 1982) and might require specific training and tooling (Harold et al., 2015; Zhao et al., 2023). Therefore, employing text as a flexible bridge to more intuitively access time series is an exciting opportunity (Imani et al., 2019; Slack et al., 2023). Figure 1 shows how this even allows for back-and-forth discussions contextualized by data.

Practical machine learning applications require solving a wide variety of tasks on time series, including the classics of forecasting, classification, and anomaly detection. While tremendous progress has been made in building models for each individual use, most are only effective at one, and changes necessitate expensive remodeling and retraining. In the era of large language models (LLMs), we can solve many tasks by embedding both our data and the assignment description as text (Brown et al., 2020; Lewis et al., 2020; Fan et al., 2024). This often does not require specific retraining, as providing some examples by in-context learning is often sufficient. It would be highly desirable to tweak outputs of time series models in a similarly flexible way, as some initial works already started exploring (Xie et al., 2025). However, interacting with time series via text is only just emerging as a field of study.

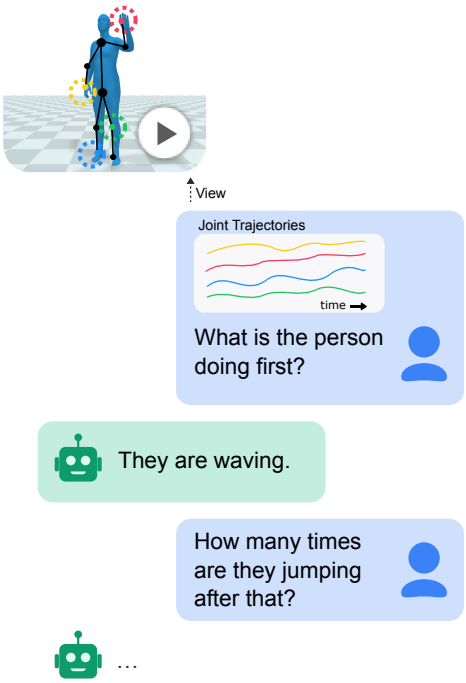


Figure 1: **Text offers a highly intuitive medium of communication to interact with otherwise opaque multivariate time series.**

This combination is further motivated by the common need to enrich data with further context and task descriptions, which is eased by the possibility of encoding such information in text (Han et al., 2023). One example is multimodal forecasting, where textual data enriches numerical time series to provide crucial context (Jia et al., 2024; Williams et al., 2025). Interactive feedback promises even better results by allowing a human-in-the-loop to offer richer information (Imani et al., 2019; Wu et al., 2022; Kraus et al., 2025b). Generally, time series are used to describe and subsequently analyze weather dynamics (Hewage et al., 2020), plant growth (Yasrab et al., 2021), electrocardiograms (Sahoo et al., 2020), electrical brain activity (Andrzejak et al., 2001), human activities (Lara and Labrador, 2013), manufacturing processes (Farahani et al., 2023),

traffic volumes (Jiang and Luo, 2022), financial markets (Nazareth and Ramana Reddy, 2023; Divo et al., 2025), and much more. In all those fields, combining time series with text could help domain experts better leverage data to drive more informed decisions. For instance, one could supplement traffic volume predictions with information about local festivities, or human activity recognition with details on a person's disabilities. In other tasks, the textual domain not only serves as context but is an integral part of it, such as answering questions about time series or their textual summarization (Novák, 2016).

Yet, we find that very little work has been performed in that direction. We attribute this mainly to insufficient high-quality data. This limits both the training data to learn from and the opportunities to evaluate new systems. Therefore, we present **QuAnTS**: A challenging multimodal dataset for question answering on numerical time series data. We use human motion, in the form of tracked skeleton positions, as the signal to reason about. This has several key advantages: (1) It is natural to visualize the numerical values as human figures, making them interpretable for humans while still challenging for machines. (2) This field of application is central to building intelligent devices for, e.g., applications in health and sports. (3) There is a very long line of research on and collection of human motion data, highlighting its significance and providing tools and insights we can leverage in this work, such as the synthetic motion generation and visualization tools.

To move toward rich interactions with complex time series signals via textual conversations, we **contribute** the following:

(i) We investigate the design space of time series QA datasets to compile a comprehensive and challenging set of question and answer types.

(ii) We propose a procedure for generating the diverse and large-scale QuAnTS dataset. The implementation is readily extensible and freely available alongside the dataset.[1]

(iii) Extensive dataset validation and benchmarking of existing and novel baseline models provide a fertile ground for further research into this very young field.

(iv) We provide human performance to contextualize current and future machine capabilities.

Outline. We continue by reviewing related work in Section 2 as a foundation for the subsequent description of QuAnTS's construction in Section 3. Extensive empirical findings on solving QuAnTS are then presented in Section 4. We conclude the work with limitations and an outlook in Section 5 and provide an overview of its broader impact.

2 Related Work

Integrating data sources and models from multiple modalities, such as images, videos, text, time series, tables, graphs, point clouds, etc., has experienced tremendous interest and progress in the last few years. For instance, vision-language models have since arrived in mass usage as new commodity services (Rombach et al., 2022; OpenAI, 2024; Zong et al., 2024; Wadekar et al., 2024). Initial attempts have since worked toward any-to-any multimodal learning, i.e., aligning representations across more than two modalities (Han et al., 2024; Zhang et al., 2023; Wu et al., 2024). However, this field is still emerging and far from mature.

1. We provide the code for generating QuAnTS at `https://github.com/mauricekraus/quants-generate` and for evaluation at `https://github.com/felixdivo/quants-eval`. The dataset is available from `https://huggingface.co/datasets/dasyd/quants` with DOI 10.57967/hf/6663.

2.1 Multimodal Time Series Models and Datasets

Fusing time series with other modalities remains largely unexplored, with a significant portion of multimodal research focusing on language and vision (Baltrusaitis et al., 2019; Gao et al., 2020). This is reflected and possibly caused by a shortage of truly multimodal time series datasets (Liang et al., 2023; Zhou et al., 2025), which is changing only very recently.

Some works already combined time series and text, particularly for augmenting tasks such as forecasting, classification, and anomaly detection with supplemental text data (Sano et al., 2019; Dang et al., 2020; Jia et al., 2024; Liu et al., 2024; Chan et al., 2024; Jin et al., 2024a; Kohaut et al., 2025, 2026). However, in such tasks, the text often serves only as additional context, rather than being integral to the problem. Thus, newer works focus on the text being essential for solving the tasks (Williams et al., 2025; Xu et al., 2025).

These tasks, however, are solved by predicting time series snippets or labels, but not textual responses, such as in time series captioning (Spreafico and Carenini, 2020; Imran et al., 2024; Han et al., 2024; Trabelsi et al., 2025; Strem et al., 2026). These methods, however, do not adapt their responses based on questions or other textual input. This is the case in QA over Temporal Knowledge Graphs (TKGs) (Saxena et al., 2021; Su et al., 2024), where events and their semantic relations have already been identified, or temporal QA, where questions with a specific focus on temporal reasoning based on purely textual context are to be answered (Jia et al., 2018; Jin et al., 2021). While reasoning over TKGs has since been extended to QA on forecasting (Ding et al., 2023), it is not tied to the general context of a numerical time series and thus different from the approach of QuAnTS.

Some initial works have already begun exploring true time series QA (TSQA) and provided datasets, as summarized in Table 1. Firstly, Xing et al. (2021) propose the closed-ended OppQA dataset based on existing human activity recordings and the DeepSQA set of baselines. Similarly, Oh et al. (2023) annotated electrocardiogram recordings to obtain the EGC-QA dataset. However, responses are again framed as a classification problem of one or multiple predefined answers. The Time Series Evol-Instruct dataset accompanying the ChatTS model (Xie et al., 2025) provides open-ended QA data across various fields, time series lengths, and numbers of variates. Lastly, Imran et al. (2024) propose (Tune-)OpenSQA, a human activity TSQA dataset again featuring exclusively open-ended questions. It was obtained by prompting an LLM to generate temporal questions. They focus on a single activity, thereby lacking the completeness of downstream model training and evaluation on the entirety of the scene. Moreover, the diversity of (Tune-)OpenSQA is bounded by the diversity that emerges from a single prompt and an LLM. QuAnTS expands on these works by simultaneously featuring binary/multiple-choice and open-ended responses on a substantial scale and with assured diversity. Furthermore, we ask questions about sequences of consecutive actions and their composition, necessitating models that can both perceive and reason over time series. In addition, the human motion data in QuAnTS challenges models by offering substantially more variates than any other dataset mentioned above (apart from the substantially less diverse OppQA). To contextualize the progress of the emerging field of TSQA, we provide human reference performances on QuAnTS for the first time.

Table 1: **The QuAnTS dataset is large and varied.** It guarantees diversity through the balanced distribution of 45 question and three answer types: *b*inary, *m*ultiple-choice, and *o*pen. The shape of the time series is specified in terms of the number of time steps T and variates V. QuAnTS provides human performances for reference.

			QA Pairs		Questions			Answ.			Time Series		Human
Dataset	Scope	Level	#Total	#Unique	#Unique	Types	Dist.	b	m	o	#Unique	T×V	Eval.
OppQA-1800[1]	scene	high	112,749	89,735	84,389	12	balanced	✓	✓	✗	1,362	1,800×77	✗
ECG-QA	scene	high	414,348	112,455	34,767	7	balanced	✓	✓	✗	16,052	5,000×12	✗
TS Evol-Instr.	mixed	low	44,802	44,516	43,973	10	skewed	✗	✗	✓	4,127	varying	✗
(Tune-)OpenSQA	local	low	183,304	172,379	183,042	n/a	skewed	✗	✗	✓	35,683	varying	✗
QuAnTS (ours)	scene	high	150,000	125,726	100,568	45	balanced	✓	✓	✓	30,000	320×72	✓

[1] OppQA-1800's "counting" answer type is multiple-choice since one picks from a fixed set of four answers.

2.2 Human Motion Data and LLMs

Human motion data, typically captured by wearable sensors such as accelerometers and gyroscopes, presents unique challenges due to its temporal complexity. Since the rise of leveraging LLMs for general time series tasks, researchers have begun exploring their potential in processing and interpreting human motion data (Xue and Salim, 2023; Jin et al., 2024a; Jia et al., 2024; Zhang et al., 2024), despite some more recent questioning (Jin et al., 2024b; Tan et al., 2025; Zhang et al., 2025; Merrill et al., 2024; Jiang et al., 2024). This shift highlights the growing interest in adapting LLMs to handle motion datasets' complex temporal patterns and cross-variate characteristics, such as in zero-shot activity recognition based on inertial measurement units (Ji et al., 2024).

The mainstream of current studies (such as Zhou et al. (2023); Moon et al. (2023); Li et al. (2025)) focuses on aligning sensor data embeddings with textual description embeddings extracted by LLMs. However, a significant challenge in this approach arises from the limitations of existing human motion datasets, as most lack detailed textual descriptions of the activities performed. In recent studies on LLMs for sensor data, the text component is often generated post-hoc through labor-intensive annotation processes, resulting in a lack of standardized criteria for them. This inconsistency complicates the alignment of TS and text embeddings, particularly between different systems, making fair comparisons challenging.

This work presents the readily extensible, large-scale QuAnTS dataset and generation pipeline comprising synthetic human motion data, corresponding textual descriptions, and challenging question-answer pairs. It provides a new data source to support research on time series and textual embedding alignment while simultaneously posing a novel QA challenge.

3 QuAnTS: A Dataset for Question Answering on Time Series

We created the diverse, large-scale QuAnTS dataset as a foundation for training future time series QA models and evaluating them. This section describes the design considerations taken into account during its development. Firstly, we present how the textual descriptions with questions and answers were generated (Section 3.1), followed by the text-conditioned generation of human motion trajectories (Section 3.2). Figure 2 shows the whole pipeline for generating QuAnTS.

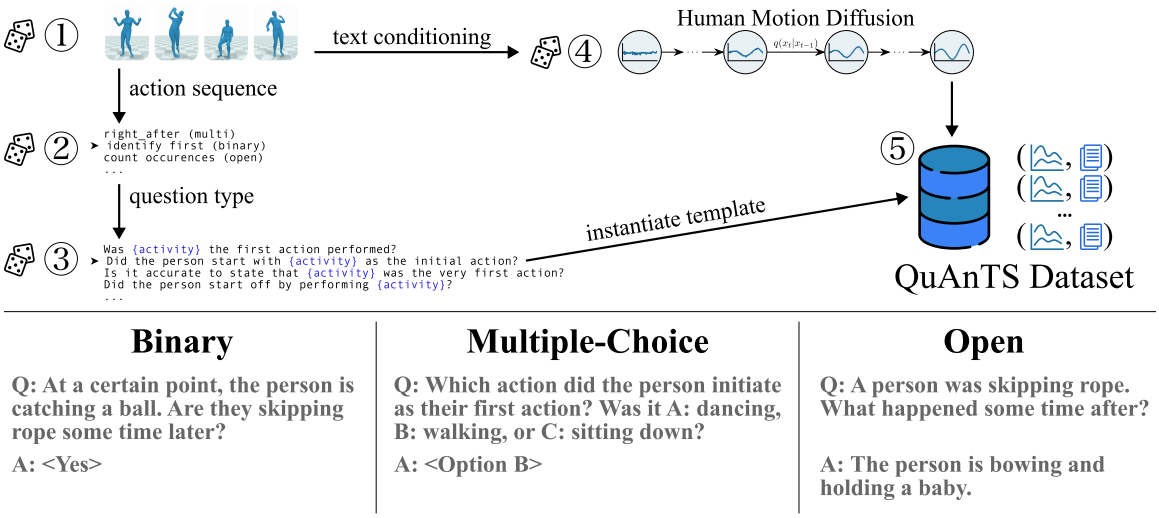


Figure 2: QuAnTS is generated in several steps: An action sequence is sampled ①, where for each we sample five question and answer types ②. For diversity, each of them is then instantiated from a sampled template ③. The time series from the human motion diffusion ④ is then combined with the QA-pair and auxiliary data ⑤. Example QA pairs are shown below. Dice (🎲) indicate randomized operations to enhance dataset diversity.

3.1 Question & Answer Generation Pipeline

Determining suitable questions and corresponding answers is crucial to obtaining a challenging dataset. Question types can either be not categorized at all, specified beforehand, or assigned after manual dataset creation. Since we generate questions synthetically by instantiating templates, it is natural to define them a priori to break down the data generation process into specific parts. This has the added benefit of allowing for the inspection of a model's performance on specific subtasks and ensuring a balanced distribution of question types. Following a substantial review of QA datasets and pipelines for time series and videos, we identified these five main question categories (as shown in Figure 3):

Descriptive questions ask about the name of a performed action, how often it occurred, where it is happening, or insight based on metadata, such as who is performing the action. We exclude the latter two since we abstract from the concrete location, and metadata is often only available in lab settings.

Temporal Ordering questions ask about the temporal ordering of events or if specific actions were performed in the very beginning or end.

Comparison questions ask whether some actions are the same/distinct or if they occur the same/different number of times.

Reasoning questions ask for higher-level concepts in the actions, like predicting the next one, counterfactual reflections, or providing reasons for a person's behavior. Once basic time series QA matures, we propose this as exciting future work.

Overaching questions are a future type that either combines multiple aspects, like counting how often the last action was performed in total, or absurd questions, like asking about what happened after an action that never took place. We omit it for now, as QuAnTS is already very challenging without it.

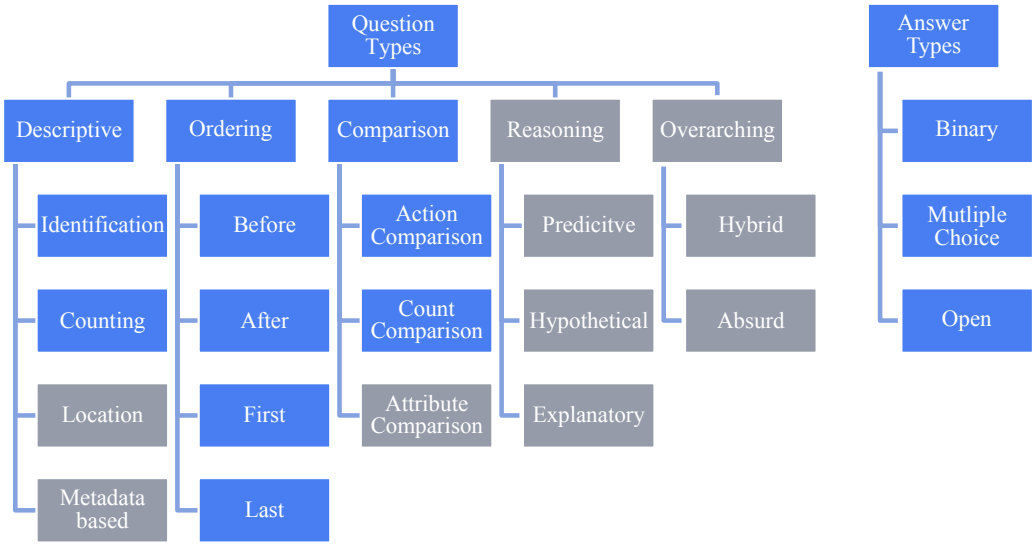


Figure 3: **The hierarchy of question and answer types we identified.** The **blue** types are the most fundamental to time series QA and, therefore, included in QuAnTS. These tasks alone already make for a very challenging dataset, as we will see in Section 4. The **gray** ones are left for future extensions.

As in video QA, the simplest possible answer is a binary "Yes" or "No". QuAnTS goes beyond such simple formats by offering multiple-choice and open-ended questions generated from templates, where we do not expect a label but a textual response. Differentiating between these three answer formats, QuAnTS comprises a balanced distribution of a total of 45 question types. To ensure high diversity in QuAnTS, we sample from about 10 to 20 different question and answer templates, depending on the QA type. To increase data availability for training for open answers, we additionally generate textual answers for binary and multiple-choice questions via templating. We leveraged LLM-assisted paraphrasing, followed by a comprehensive manual review, to enhance the diversity of all templates.

3.2 Leveraging Motion Diffusion Models for Time Series Context

Having generated textual descriptions of the desired sequence of actions and corresponding QA pairs, we next require matching time series. We opted to use synthetic trajectories from a generative human motion model. In contrast to directly using an existing dataset, this approach enables us to generate action sequences tailored to the specific questions we want answered. Furthermore, the human motion generator can sample multiple variations of the same action, reflecting the variability of organic movements. We require a method that allows us to generate motions smoothly following a sequence of activities, and not just a single one, to enable us to ask interesting questions about them. Numerous generative human motion models can be conditioned on natural language prompts, much like diffusion-based vision models work. Due to perceptually well-recognizable motions and very fast generation, we choose to use *Spatio-Temporal Motion Collage* (STMC) by Petrovich et al. (2024).

To generate well-recognizable human motions, we identified 19 particularly distinct action prompts commonly found in STMC's training data (see Appendix A) to build action sequences. We evaluate their identifiability by humans in Appendix C.2. Each context time series in our dataset was obtained by concatenating four action prompts of four seconds each from unique random seeds. Five question-and-answer pairs share one context time series, ensuring high diversity at reasonable generation costs. STMC first assigns textual prompts to one or multiple body parts to create a multi-track timeline. To obtain the individual components, text-guided iterative denoising is performed for each track and time segment. An additional refinement step is required to smoothly blend the transitions between consecutive sections with different prompts. Specifically, we add back some unconditional noise to the transition region before a final denoising procedure, which is then conditioned by the already generated actions before and after the change point. This results in final poses comprising three spatial coefficients for each of the 24 joints in the SMPL scheme (Loper et al., 2015). The final trajectory of 16 seconds at 20 frames per second thus has dimension $320 \times 24 \times 3$. Appendix A provides the full details of the generation procedure.

3.3 Dataset Statistics and Diversity Analysis

We took great care to ensure high diversity of the time series context and textual parts though sampling and diverse templating, while creating a sufficiently large dataset. The samples are independent and identically distributed (i.i.d.), which is a common assumption for data needed in machine learning. We generated 30.000 time series contexts of four consecutive actions, each, totaling ∼133.3 hours. We pose five QA pairs for each of those contexts. For best comparability between approaches adopting QuAnTS, we split the dataset into predefined training, validation, and testing splits of $120,000$ (80%), $15,000$ (10%), and $15,000$ (10%) samples, respectively. To prevent any leakage across splits, we ensure all five QA pairs per context time series are sorted into the same split.

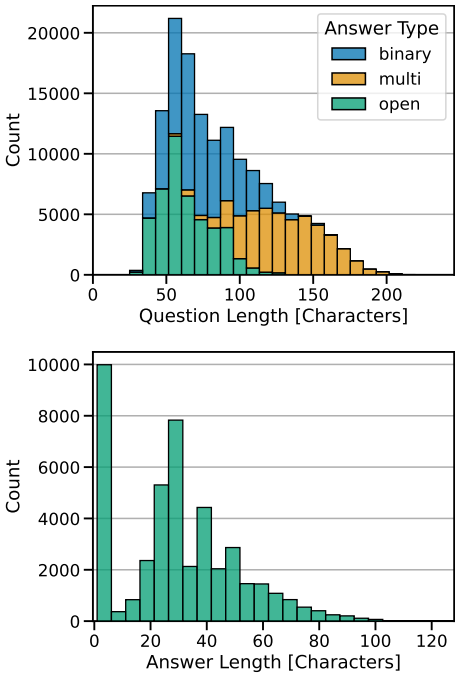


Figure 4: **QuAnTS' question and open answer lengths are diverse.**

For each unique sample, in addition to the question, answer, and motion trajectory, we also provide metadata, such as the question and answer types. This equips QuAnTS for diagnostics of TSQA models, e.g., aiding in understanding which settings prove the most difficult. An optional fully textual version of the answers, as opposed to a binary Yes/No or multiple-choice A/B/C response, allows casting all tasks in an open format and provides additional training material. For debugging and verifying models, we provide the ground-truth action sequence that was used to condition the time series generation. Finally, we provide a textual description of the entire action sequence, again generated from diverse, manually reviewed templates.

When synthesizing the dataset, we ensure high diversity to obtain a challenging, varied dataset for training and evaluating systems. Figure 4 (to the right) shows the wide range of question and open answer lengths. Note that the high number of short answers stems from the counting tasks, where a good response is just the correct action count. Figure 5 in Appendix A visualizes the final balanced distribution of both the question and answer types. Similar to the textual components, diversity in the generated human motions is achieved through the STMC diffusion process, which produces organic, varied trajectories even for a fixed prompt. Appendix A further investigates duplicates in QuAnTS and shows that the presented generation procedure yields an effective sample size of >99%.

3.4 Metrics for Evaluating Answers

We grade the answers to the binary and multiple-choice questions using accuracy and the macro-averaged precision, recall, and F1 scores to account for any remaining class imbalances. Answers that cannot be parsed as a valid answer to a binary or multiple-choice question, e.g., if not unambiguously recognizable as either yes or no, are counted as wrong.

Evaluation of the open questions necessitates comparing the textual reference answer to the answer given by the machine or human. We thus employ the classic metrics ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) commonly used to evaluate QA systems. We omitted BERTScore (Zhang et al., 2020) because, since all questions are on similar topics, both good and bad responses yield uninformative high scores around 0.96. All these metrics, however, capture mostly syntactic similarity (Imran et al., 2024). We thus augment them with a semantic comparison using LLMs as a judge comparing two textual answers (Zheng et al., 2023; Li et al., 2024; Gu et al., 2025). We chose the following scores from 1 to 3 to obtain a nuanced assessment anchored in a well-defined scale:

1: The answer is terrible: The answer is factually incorrect or corrupt in some way (e.g., empty or nonsensical).

2: The answer is bad: Somewhat correctly answers the question like the reference answer, yet contains additional useless/ambiguous/wrong information or is incomplete.

3: The answer is good: Correctly answers the question and is helpful in the context of the scene.

For a more intuitive interpretation, we follow common quality metrics, such as accuracy or F1 scores, by normalizing the range to $[0, 1]$, where 0 represents the poorest and 1 the best. We aggregate the scores across multiple samples using the arithmetic mean.

As is good practice, we evaluate the quality of the LLM Judge by assessing its alignment with human judgments on a representative subset of 200 samples (Bavaresco et al., 2025). Appendix B provides the detailed validation methodology, corresponding results, the full prompt, and detailed configuration. We finally chose Qwen3 (8B-AWQ with Reasoning, Yang et al., 2025) due to its sufficient inference speed at excellent alignment with human judgment, specifically with a Pearson correlation coefficient 0.912. As an intuitive scale, it deviated on average by 7.45%, permitting its use as a reliable measure of open-ended answer quality.

4 On Solving QuAnTS

To validate the dataset and provide stable baselines, we conduct extensive experiments following the protocol defined in the previous Section 3.4. Recall that QuAnTS pairs motion time series with textual questions and target answers in binary, multiple-choice, and open formats. We begin in Section 4.1 with dataset verification and ablation studies to rule out trivial shortcuts (Steinmann et al., 2024) and confirm that each component of the input (i.e., the time series context and question text) is indeed necessary. In Section 4.2, we measure human performance on the same questions using auxiliary video renderings similar to Figure 1, and also test how reliably humans recognize the underlying actions. Finally, in Section 4.3, we evaluate three baselines of increasing capabilities: A naive LLM that sees the serialized numerical time series values and the question, the adapted SOTA time series LLM ChatTS (Xie et al., 2025), and a simple neuro-symbolic pipeline (xQA) built on top of purposefully structured information. xQA is designed to demonstrate that the dataset can be answered by TSQA models, but that current TS-LLMs still require explicit structure.

4.1 Dataset Verification and Ablations

We first ensure the logical consistency of QuAnTS by answering its questions based on ground-truth descriptions of the action sequences. Second, we investigate whether the questions alone are sufficient to answer them, which should not be possible in a carefully constructed dataset. Lastly, we attempt to determine whether LLMs can solve QuAnTS using only the time series, without the questions, again ruling out any shortcuts.

We aim to answer the following set of questions:

(Q1): Does the action sequence provide sufficient information to answer the questions?

(Q2): Are there shortcuts between the questions and answers?

(Q3): Are there shortcuts between the time series and answers?

We base our results on the widely used Llama 3.1 8B model (Llama-Team and MetaAI, 2024) *finetuned* on the whole train split. This model size offers a good balance of reasoning capabilities and the possibility to fine-tune it. See Appendix C.1 for implementation details, and Appendix C.1.3 for how this deteriorates general language modeling capabilities.

Ground-Truth Scene Descriptions and Questions *(Q1).* In this experiment, we evaluate the model performance using textual ground-truth descriptions of the action sequence, which are only available for model debugging, together with the corresponding question as input. This setup acts as a check for soundness, verifying that the dataset is logically consistent and that the questions are answerable given the provided context.

The accuracy and F1 score on the binary variant were 99.10% and 99.12%, and for the multi variant 99.98% and 99.98%, respectively. Furthermore, the very good results for the open-ended questions serve as an aspirational upper bound for subsequent experiments, with an LLMJudge score of 0.9434. The ROUGE and METEOR scores were 0.3380 and 0.5099, respectively, indicating a decent syntactic alignment with the ground truth. The excellent performance for all three variants confirms the well-formedness of QuAnTS and affirms *(Q1)*. Moreover, this validates that Llama 3.1 8B has sufficient capacity for these experiments.

Only the Questions *(Q2).* Next, we verify that the questions cannot be answered independently of the time series context, which we expect to be impossible if no shortcuts are present. In other words, the accuracy should be reasonably close to random guessing.

The results for the *Only Question* ablation in Table 2 show that responses across all three answer types are close to random guessing. We thus reject *(Q2)* as we did not find any spurious correlations.

Only the Time Series Encoded as Text *(Q3).* We assess how much useful information can be extracted from the motion trajectory when the textual question is withheld. The trajectory is serialized as text following Gruver et al. (2023); preprocessing and prompt details are provided in Appendix C.1.

The time-series-only model reaches a binary accuracy of 49.21% and multiple-choice accuracy of 33.29%, both close to uniform guessing, with macro F1 scores of 33.62% and 20.39%, respectively (Table 2). No closed-answer prediction is invalid under the parser used for these configurations. Open-answer performance is low, with an LLMJudge score of 0.1082. Because several different questions share the same trajectory, the missing question makes the requested answer ambiguous. In summary, we can rule out any substantial confounders between the time series and answers in QuAnTS *(Q3)*.

We proceed by evaluating human responses to the questions before baseline models.

4.2 Human Performance as Reference

In the preceding Section 4.1, we first demonstrated that QuAnTS contains sufficient information to be solved, provided that actions can be identified correctly, and second, ensured that handling both TS and text modalities is indeed required. Next, we will investigate whether humans can solve the tasks, given that most can rather reliably identify actions in day-to-day life. Furthermore, this permits judging model performance relative to humans, thereby providing an intuitive scale of progress for the emerging field of TSQA.

Setup. We performed a separate study for each of the three splits (binary, multi, and open), with matching answer formats and the respective evaluation metrics from Section 3.4. We hired a diverse cohort of participants through *Prolific Academic*[2], an established crowdsourcing platform for behavioral research. Participants were laypeople, but selected to be native speakers of English to rule out language barriers as potential sources of error. In the open variant, we additionally verify if humans can correctly identify the individual actions. See Appendix C.2 for details on the study and demographics of its participants.

Results. The results in Table 2 generally confirm that humans can solve the dataset in the provided format, especially for the binary and multi splits. Specifically, the F1 scores—while not quite reaching the upper ceiling from Section 4.1 (Q1)—are overall good. The scores for the binary variant are slightly lower than those for the multiple-choice variant, possibly because participants ponder less when fewer options and guidance are presented. For the open variant, we measured a moderate LLMJudge score of 0.6148. We suspect that the lower performance compared to other answer types partly stems from online annotators being less ambitious in their textual answers due to the increased effort relative to fixed answer types.

2. `https://www.prolific.com`

Table 2: **Dataset variants and ablations show that QuAnTS is well-formed. Furthermore, humans can successfully solve the binary and multi splits. The Naive Baseline (except for open) and ChatTS struggle to solve most splits of the QuAnTS dataset, whereas xQA successfully leverages the additional information.** Human, ChatTS, and xQA rows provide reference scores under their respective evaluation setups; direct comparison is only possible within the sections marked by the gray dividers (—). n indicates the total number of human responses. GT denotes ground-truth action labels and AE a supervised action encoder using pre-cut segments. More details on humans and xQA are provided in Appendix D.2.

	System	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 (↑)
Binary	Ablation: Only Question	64.53%	65.32%	64.70%	64.22%
	Ablation: Only TS	49.21%	50.15%	50.00%	33.62%
	Humans ($n = 840$)	77.50%	81.08%	74.66%	77.74%
	Naive: TS + Question	62.94%	64.54%	63.19%	62.16%
	ChatTS	51.69%	55.38%	25.36%	34.79%
	xQA-Llama on GT	79.44%	84.91%	72.42%	78.17%
	xQA-Qwen on GT	80.92%	92.28%	68.14%	78.39%
	xQA-Llama on AE	79.46%	84.60%	72.84%	78.28%
	xQA-Qwen on AE	81.00%	92.52%	68.10%	78.46%
Multi	Ablation: Only Question	33.32%	30.38%	33.00%	26.25%
	Ablation: Only TS	33.29%	21.94%	33.08%	20.39%
	Humans ($n = 820$)	86.59%	86.55%	86.61%	86.54%
	Naive: TS + Question	34.01%	22.74%	33.68%	26.62%
	ChatTS	30.40%	30.13%	30.58%	29.12%
	xQA-Llama on GT	81.50%	81.64%	81.51%	81.50%
	xQA-Qwen on GT	88.01%	88.04%	88.03%	88.01%
	xQA-Llama on AE	81.18%	81.30%	81.19%	81.18%
	xQA-Qwen on AE	87.97%	87.99%	87.98%	87.97%

	System	ROUGE (↑)	METEOR (↑)	LLMJudge (↑)
Open	Ablation: Only Question	0.4233	0.2838	0.2092
	Ablation: Only TS	0.0956	0.0491	0.1082
	Humans ($n = 440$)	0.0953	0.2635	0.6148
	Naive: TS + Question	0.8199	0.6947	0.8759
	ChatTS	0.0455	0.1060	0.0702
	xQA-Llama on GT	0.0913	0.3677	0.7729
	xQA-Qwen on GT	0.1789	0.3697	0.8279
	xQA-Llama on AE	0.0898	0.3679	0.7753
	xQA-Qwen on AE	0.1782	0.3692	0.8176

Overall, the gap between the results for LLMs trained on ground-truth scene descriptions from Section 4.1 and these human results likely stems from the added task of identifying actions. For instance, `catching a ball` is commonly misidentified as `throwing a ball`. The details provided in Appendix C.2, however, overall confirm that humans can effortlessly discern most actions and solve QuAnTS. Further analysis is provided in Appendix D.1.

4.3 Baselines

To illustrate what our dataset reveals about current TSQA methods, we evaluate three complementary baselines:

(1) A general-purpose fine-tuned text model receives the serialized numerical time series alongside the question, without a dedicated time-series encoder.
(2) ChatTS (Xie et al., 2025) represents purpose-built TS-LLMs and is designed for interactions contextualized by multivariate numerical data.
(3) xQA is a neuro-symbolic pipeline that augments a general-purpose LLM with structured action information extracted from the time series. Its variants help separate action recognition from question answering.

Together, these systems characterize different approaches to solving QuAnTS' TSQA tasks.

(1) Naive Baseline: Time Series Encoded as Text and Questions. Following the setup of the verification experiments in Section 4.1, we finetune Llama 3.1 8B on QuAnTS. The time series is serialized to text as in the "Only TS" ablation and then concatenated with the question and an introductory instruction. The full prompt is shown in Appendix C.1.

(2) Adapting ChatTS. We include ChatTS because it is specifically designed to tokenize and condition on time-series inputs, making it an important state-of-the-art baseline for our task. ChatTS first encodes time series into tokens with attached metadata, such as the number of time steps and normalization coefficients. It then injects these into the prompt, which is fed to a specially finetuned language model to perform standard auto-regressive next-token prediction.

However, the released model is limited to 30 variates and therefore cannot directly process the 72 ones in QuAnTS. To still facilitate a comparison, we only pass 10 particularly distinctive joints (with three coordinates each) to ChatTS, as detailed in Appendix C.3. This restriction of the data channels is simultaneously a potentially advantageous human prior.

(3) xLSTM-based QA Systems: xQA-Qwen & xQA-Llama. End-to-end multimodal systems for TSQA typically require finetuning on substantial amounts of data. Instead, we separate perception (neural) from reasoning (symbolic) by passing only a short sequence of textual task descriptions and discrete action labels to an LLM without retraining. To facilitate this, we train a segmentation model on ground-truth action labels of the action time sequence, which are typically *not* accessible to models solving the dataset.

Let the input to xQA be a multivariate time series $\boldsymbol{x}_{1:T} \in \mathbb{R}^{T \times V}$ partitioned into $M = 4$ equal segments for each action label, with boundaries (s_i, e_i) for $i \in \{1, \ldots, M\}$. We denote the finite action vocabulary by $\mathcal{A}$ (with $|\mathcal{A}| = 19$, see Appendix A). We use xLSTM-Mixer (Kraus et al., 2025a) as an efficient time-series embedding backbone, followed by a lightweight supervised classification head predicting the action label per segment. Specifically, for i, the model predicts $a_i = f_\theta\left(\boldsymbol{x}_{s_i:e_i}\right) \in \mathcal{A}$, yielding an action sequence $\boldsymbol{a} = (a_1, \ldots, a_M)$.

Reasoning is then delegated to an instruction-tuned LLM g consuming an instruction I, the action sequence $\boldsymbol{a}$, the question q, and two few-shot demonstrations D_1 and D_2, yielding a response $r = g(I, \boldsymbol{a}, q; D_1, D_2)$. The instruction I requires a concise, numbered reasoning chain followed by a final answer generated through constrained generation.

To disentangle the evaluation of perception from reasoning, we report the performance of xQA on both the action encoder (AE) and the ground-truth (GT) action labels. xQA-AE serves as a supervised diagnostic baseline jointly performing both perception and reasoning, while xQA-GT provides an oracle-style reasoning ceiling for a given LLM. Both leverage privileged action-sequence information not available to standard methods for solving QuAnTS.

Because the interface is text-only, we instantiate g with the LLMs Llama 3.1 8B and Qwen3 8B *without any finetuning*. Implementation details are provided in Appendix C.4.

Results. The Naive Baseline shows substantially different behavior across answer formats (Table 2). For binary answers, it achieves an F1 score of 62.16%, which is below the performance of the question-only ablation of 64.22%. For multiple-choice answers, accuracy is close to uniform guessing with an F1 of 26.62%. All binary and multiple-choice answers were valid when parsing. For open answers, however, the Naive Baseline attains an LLMJudge score of 0.8759, indicating that a QLoRA-adapted text model can produce accurate open answers from serialized trajectories and questions in this setting.

However, this does not demonstrate general time series understanding, as this lightweight model adaptation causes other model capabilities to collapse entirely. Specifically, in Appendix C.1.3, we measure the change in general language modeling abilities resulting from low-rank model adaptation. We find that perplexity deteriorates drastically: on WikiText-2 (Merity et al., 2016), it increases by seven orders of magnitude, where lower would be better (Table 6). This mirrors the findings for the ablations "Only Question" and "Only TS" from Section 4.1 (see Appendix C.1.3 for details). In contrast, all other baselines retain their language modeling capabilities.

ChatTS does not reliably answer the questions of QuAnTS in the evaluated setup. Its accuracy is close to uniform guessing for the binary and multiple-choice variants, and its open-answer LLMJudge score is 0.0702. Note that ChatTS had access to only 10 of 24 joints due to model limitations (see Appendix C.3), which both limited the information available and assisted through an expert-informed focus.

xQA achieves strong scores using explicit action-sequence information, with Qwen3 8B consistently ahead of Llama 3.1 8B in the reported results. The similar performance with action-encoder and ground-truth labels indicates that the supervised action encoder can provide useful inputs to the downstream question-answering model (see also Figure 11 in Appendix C.4).

Recall that xQA uses privileged action supervision unavailable to the Naive and input-ablation models. Its results, therefore, characterize the potential of hybrid models rather than being directly comparable to the other baselines. A comparison to human performance must be made cautiously because of a mismatch in input modalities due to the videos shown to the human participants.

5 Conclusion and Outlook

There is very limited research on handling complex combinations of modalities, such as time series question answering. To foster research into models with new capabilities, we propose QuAnTS, a novel dataset featuring diverse questions and answers on human motion trajectories in binary, multiple-choice, and open-ended variants. We provide open access to the flexible generation pipeline and dataset and verify its well-formedness and challenges through extensive experiments. Additionally, the baseline performance of three different models provides key insights into the models' capabilities and limitations, highlighting the need for dedicated models. Lastly, we contextualize the state of learned systems by reporting human performance on QuAnTS.

5.1 Limitations

The main limitation of this work stems from its focus on a single, controlled domain: QA on human motion data. While this provides an easily debuggable diagnostic setting, it does not yet provide a benchmark for generic TSQA. By extension, the xQA baseline serves as a diagnostic tool for QuAnTS, but might be challenging to transfer to arbitrary domains.

The time series features do not capture the full range of all human motion and are specific to the signatures generated by STMC. Generally, the structure is also restricted to 19 actions, 4 fixed-length slots each, a fixed set of question/answer templates, a fixed overall sequence duration, regular sampling, and a lack of real-world sensor issues.

We employed LLM judges to assess and define the quality metrics for QuAnTS. While it was evaluated thoroughly in Section 3.4 and appendix B, some biases may persist, especially given that some baselines and the judge share the same foundation model.

As identified by Section 4.2 and appendix C.2, QuAnTS is a challenging task even for humans, with them falling far short of the respective perfect scores. Extra care is thus needed when interpreting results. Reaching human levels in both language modeling and simultaneous time series understanding would already constitute a substantial step forward.

5.2 Future Work

Moving forward, we identify two primary avenues for building on this work. First, regarding the dataset itself, the presented scope of TSQA could be broadened to include time series from other application domains and more complex, compositional question types. In practice, one will likely combine data about generic time series knowledge (such as identifying peaks and seasonal patterns) with domain-specific data (such as the specifics of human motion). Investigating how to curate optimal mixtures of them remains an open question.

This leads to the second key path forward: building truly multimodal models for time series and text. Our results highlight the urgent need for improved architectures specifically designed to address the challenges of TSQA. Firstly, substantially more work is needed to equip pre-trained models with the capabilities to reason about time series in zero-shot settings. Furthermore, moving from single-turn QA to longer conversational systems would open up many new fields of application. It remains unclear to what extent this requires dedicated multi-turn multimodal datasets, or whether conversational abilities can be fully retained in post-trained models.

Broader Impact Statement

Time series data occurs in many real-world applications—often in tight relation to our daily lives. Datasets like QuAnTS work towards models that provide better access to this data modality through more intuitive textual interfaces. This enables many applications with a positive impact, including personalized health recommendations during exercise, insights into sleep quality, personal finance recommendations, and doctors more effectively summarizing complex measurements from patients. On the flip side, some of these fields touch upon very intimate parts of our lives, necessitating a thorough assessment of their potential invasions of privacy. As with any machine learning system deployed to individuals, great care must be taken to adequately represent minorities to ensure a fair model and to clarify the capabilities and limitations of such systems. If applied to less personalized data, such as industrial manufacturing or computer networking logs, we anticipate the potential gains in accessibility and integration to far outweigh the limited potential harm.

Thanks to the entirely synthetic dataset, we do not foresee any major adverse consequences regarding QuAnTS itself. While the STMC human motion generation was trained on data obtained from human subjects, these respective datasets are anonymized and widely used, and, to the best of our knowledge, no major concerns have surfaced so far. Since it might not represent people with different cultural backgrounds, certain (motoric) limitations, or body morphologies, real-world applications should carefully assess their representation of the intended target audiences.

Finally, we stress that QuAnTS is intended mainly as a data source and benchmark of TSQA capabilities, not as a deployment-ready system. We attach a comprehensive datasheet as Appendix E.

Acknowledgments and Disclosure of Funding

We cordially thank Clara Moos and Patrick Goldbeck for their help in drafting the initial iterations of the text generation process.

This work received funding from the German Federal Ministry of Research, Technology and Space (BMFTR) project "KompAKI" within the "The Future of Value Creation – Research on Production, Services and Work" program (funding number 02L19C150), managed by the Project Management Agency Karlsruhe (PTKA), the ACATIS Investment KVG mbH project "Temporal Machine Learning for Long-Term Value Investing", and the EU project EXPLAIN, under the BMFTR (grant 01-S22030D). The project also benefited from the early stage of the German Federal Ministry for Economic Affairs and Energy project "Souveräne KI für Europa" (13IPC040G) as part of the EU funding program IPCEI-CIS. Devendra Singh Dhami received support from the Department of Mathematics and Computer Science at Eindhoven University of Technology and the Eindhoven Artificial Intelligence Systems Institute. Flora Salim would like to acknowledge the support of Cisco's National Industry Innovation Network (NIIN) Research Chair Program. The authors are responsible for the content of this publication.

References

R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, Nov. 2001. doi: 10.1103/PhysRevE.64.061907.

T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, Feb. 2019. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2018.2798607.

S. Banerjee and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fernández, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, A. Martins, P. Mondorf, V. Neplenbroek, S. Pezzelle, B. Plank, D. Schlangen, A. Suglia, A. K. Surikuchi, E. Takmaz, and A. Testoni. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

N. Chan, F. Parker, W. Bennett, T. Wu, M. Y. Jia, J. Fackler, and K. Ghobadi. MedTsLLM: Leveraging LLMs for Multimodal Medical Time Series Analysis, Aug. 2024.

X.-H. Dang, S. Y. Shah, and P. Zerfos. "The Squawk Bot": Joint Learning of Time Series and Text Data Modalities for Automated Financial Information Filtering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4597–4603, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/634.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

Z. Ding, Z. Li, R. Qi, J. Wu, B. He, Y. Ma, Z. Meng, S. Chen, R. Liao, Z. Han, and V. Tresp. ForecastTKGQuestions: A Benchmark for Temporal Question Answering and Forecasting

over Temporal Knowledge Graphs. In T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, and J. Li, editors, *The Semantic Web – ISWC 2023*, volume 14265 of *Lecture Notes in Computer Science*, pages 541–560, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47239-8 978-3-031-47240-4. doi: 10.1007/978-3-031-47240-4_29.

F. Divo, E. Endress, K. Endler, K. Kersting, and D. S. Dhami. Forecasting Company Fundamentals. *Transactions on Machine Learning Research*, May 2025. ISSN 2835-8856.

I. R. C. Eggleton. Intuitive Time-Series Extrapolation. *Journal of Accounting Research*, 20 (1):68–102, 1982. ISSN 0021-8456. doi: 10.2307/2490763. Accounting Research Center, Booth School of Business, University of Chicago, Wiley.

W. Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019. URL `https://github.com/Lightning-AI/lightning`.

W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models, June 2024.

M. A. Farahani, M. R. McCormick, R. Gianinny, F. Hudacheck, R. Harik, Z. Liu, and T. Wuest. Time-series pattern recognition in Smart Manufacturing Systems: A literature review and ontology. *Journal of Manufacturing Systems*, 69:208–241, Aug. 2023. ISSN 0278-6125. doi: 10.1016/j.jmsy.2023.05.025.

E. Fedorenko, S. T. Piantadosi, and E. A. F. Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07522-w.

J. Gao, P. Li, Z. Chen, and J. Zhang. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864, May 2020. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01273.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, Nov. 2021. ISSN 0001-0782. doi: 10.1145/3458723.

N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson. Large Language Models Are Zero-Shot Time Series Forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635, Dec. 2023.

J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo. A Survey on LLM-as-a-Judge, Mar. 2025.

J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, and X. Yue. OneLLM: One Framework to Align All Modalities with Language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26574–26585, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02510.

Z. Han, R. Liao, J. Gu, Y. Zhang, Z. Ding, Y. Gu, H. Koeppl, H. Schütze, and V. Tresp. ECOLA: Enhancing Temporal Knowledge Embeddings with Contextualized Language Representations. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5433–5447, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.335.

J. Harold, K. R. Coventry, I. Lorenzoni, and T. F. Shipley. Making Sense of Time-Series Data: How Language Can Help Identify Long-Term Trends. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 37(0), 2015.

P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and Y. Liu. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing*, 24(21):16453–16482, Nov. 2020. ISSN 1433-7479. doi: 10.1007/s00500-020-04954-0.

S. Imani, S. Alaee, and E. Keogh. Putting the Human in the Time Series Analytics Loop. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 635–644, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317308.

S. A. Imran, M. N. H. Khan, S. Biswas, and B. Islam. LLaSA: Large Multimodal Agent for Human Activity Analysis Through Wearable Sensors, June 2024. DOI 10.48550/arXiv.2406.14498.

S. Ji, X. Zheng, and C. Wu. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers?, Mar. 2024.

F. Jia, K. Wang, Y. Zheng, D. Cao, and Y. Liu. GPT4MTS: Prompt-based Large Language Model for Multimodal Time-series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23343–23351, Mar. 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i21.30383.

Z. Jia, A. Abujabal, R. Saha Roy, J. Strötgen, and G. Weikum. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 1057–1062, Lyon, France, 2018. ACM Press. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191536.

G. Jiang, M. Xu, S. Xin, W. Liang, Y. Peng, C. Zhang, and Y. Zhu. MEWL: few-shot multimodal word learning with referential uncertainty. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 15144–15169, Honolulu, Hawaii, USA, July 2023. JMLR.org.

W. Jiang and J. Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, Nov. 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.117921.

Y. Jiang, View Profile, Z. Pan, View Profile, X. Zhang, View Profile, S. Garg, View Profile, A. Schneider, View Profile, Y. Nevmyvaka, View Profile, D. Song, and View Profile.

Empowering time series analysis with large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, Guide Proceedings, pages 8095–8103, Aug. 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/895.

M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*, Jan. 2024a.

M. Jin, Y. Zhang, W. Chen, K. Zhang, Y. Liang, B. Yang, J. Wang, S. Pan, and Q. Wen. Position: What Can Large Language Models Tell Us about Time Series Analysis. In *Proceedings of the 41st International Conference on Machine Learning, PMLR 235.*, volume 235, pages 22260–22276, Vienna, Austria, 2024b.

W. Jin, R. Khanna, S. Kim, D.-H. Lee, F. Morstatter, A. Galstyan, and X. Ren. ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, pages 4636–4650, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.357.

S. Kohaut, F. Divo, B. Flade, D. S. Dhami, J. Eggert, and K. Kersting. The Constitutional Filter: Bayesian Estimation of Compliant Agents. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3092–3099, Oct. 2025. doi: 10.1109/IROS60139.2025.11246215. ISSN: 2153-0866.

S. Kohaut, F. Divo, J. Hahnewald, B. Flade, J. Eggert, K. Kersting, and D. S. Dhami. Driving, Fast or Slow? Neuro-Symbolic Guidance for Motion Prediction in Multi-Modal Ground Mobility, June 2026. DOI 10.48550/arXiv.2606.15251.

M. Kraus, F. Divo, D. S. Dhami, and K. Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025a.

M. Kraus, D. Steinmann, A. Wüst, A. Kokozinski, and K. Kersting. Right on time: Revising time series models by constraining their explanations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2025b.

O. D. Lara and M. A. Labrador. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013. ISSN 1553-877X. doi: 10.1109/SURV.2012.110112.00192.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *34th Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020.

H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, Dec. 2024. DOI 10.48550/arXiv.2412.05579.

Z. Li, S. Deldari, L. Chen, H. Xue, and F. D. Salim. SensorLLM: Aligning Large Language Models with Motion Sensors for Human Activity Recognition. In *The 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, Nov. 2025.

P. P. Liang, Y. Lyu, X. Fan, A. Agarwal, Y. Cheng, L.-P. Morency, and R. Salakhutdinov. MultiZoo and MultiBench: A Standardized Toolkit for Multimodal Deep Learning. *Journal of Machine Learning Research*, 24(234):1–7, 2023. ISSN 1533-7928.

C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

J. Lin, J. Tang, H. Tang, S. Yang, G. Xiao, and S. Han. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *GetMobile: Mobile Comp. and Comm.*, 28(4):12–17, Jan. 2025. ISSN 2375-0529. doi: 10.1145/3714983.3714987.

H. Liu, S. Xu, Z. Zhao, L. Kong, H. Kamarthi, A. B. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang, and B. A. Prakash. Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, Dec. 2024.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Llama-Team and MetaAI. The Llama 3 Herd of Models, Nov. 2024.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, Sept. 2018.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models, 2016. arXiv:1609.07843.

M. A. Merrill, M. Tan, V. Gupta, T. Hartvigsen, and T. Althoff. Language Models Still Struggle to Zero-shot Reason about Time Series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3512–3533, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.201.

S. Moon, A. Madotto, Z. Lin, A. Saraf, A. Bearman, and B. Damavandi. Imu2clip: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253, 2023.

N. Nazareth and Y. V. Ramana Reddy. Financial applications of machine learning: A literature review. *Expert Systems with Applications*, 219:119640, June 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.119640.

V. Novák. Linguistic characterization of time series. *Fuzzy Sets and Systems*, 285:52–72, Feb. 2016. ISSN 0165-0114. doi: 10.1016/j.fss.2015.07.017.

J. Oh, G. Lee, S. Bae, J.-m. Kwon, and E. Choi. ECG-QA: a comprehensive question answering dataset combined with electrocardiogram. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, pages 66277–66288, Red Hook, NY, USA, Dec. 2023. Curran Associates Inc.

OpenAI. GPT-4 Technical Report, Mar. 2024.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

M. Petrovich, O. Litany, U. Iqbal, M. J. Black, G. Varol, X. B. Peng, and D. Rempe. Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation. In *First Workshop on Human Motion Generation at The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024*, 2024.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01042.

J. Romero, D. Tzionas, and M. J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, Nov. 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130883.

S. Sahoo, M. Dash, S. Behera, and S. Sabut. Machine Learning Approach to Detect Cardiac Arrhythmias in ECG Signals: A Survey. *IRBM*, 41(4):185–194, Aug. 2020. ISSN 1959-0318. doi: 10.1016/j.irbm.2019.12.001.

A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1607–1617, July 2019. ISSN 2168-2208. doi: 10.1109/JBHI.2018.2867619.

A. Saxena, S. Chakrabarti, and P. Talukdar. Question Answering Over Temporal Knowledge Graphs. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, pages 6663–6676, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.520.

Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations*, Jan. 2024.

D. Slack, S. Krishna, H. Lakkaraju, and S. Singh. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8):873–883, Aug. 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00692-8.

A. Spreafico and G. Carenini. Neural Data-Driven Captioning of Time-Series Line Charts. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*, AVI '20, pages 1–5, New York, NY, USA, Oct. 2020. Association for Computing Machinery. ISBN 978-1-4503-7535-1. doi: 10.1145/3399715.3399829.

D. Steinmann, F. Divo, M. Kraus, A. Wüst, L. Struppek, F. Friedrich, and K. Kersting. Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation, Dec. 2024. arXiv:2412.05152.

N. Strem, N. Joshy, D. S. Dhami, and K. Kersting. CATS: Cross-Modal Autoencoding for Time Series Summarization. In *2026 IEEE Conference on Artificial Intelligence (CAI)*, pages 309–316, May 2026. doi: 10.1109/CAI68641.2026.11536584.

M. Su, Z. Li, Z. Chen, L. Bai, X. Jin, and J. Guo. Temporal Knowledge Graph Question Answering: A Survey, July 2024.

M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen. Are language models actually useful for time series forecasting? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NeurIPS '24*, pages 60162–60191, Red Hook, NY, USA, June 2025. Curran Associates Inc. ISBN 979-8-3313-1438-5.

G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*, 2023.

M. Trabelsi, A. Boyd, J. Cao, and H. Uzunalioglu. Time Series Language Model for Descriptive Caption Generation, Jan. 2025. arXiv: 2501.01832.

S. N. Wadekar, A. Chaurasia, A. Chadha, and E. Culurciello. The Evolution of Multimodal Model Architectures, May 2024.

C. M. Walker and A. Gopnik. Toddlers Infer Higher-Order Relational Principles in Causal Learning. *Psychological Science*, 25(1):161–169, 2014. ISSN 0956-7976.

A. R. Williams, A. Ashok, E. Marcotte, V. Zantedeschi, J. Subramanian, R. Riachi, J. Requeima, A. Lacoste, I. Rish, N. Chapados, and A. Drouin. Context is Key: A Benchmark for Forecasting with Essential Textual Information. In *Forty-second International Conference on Machine Learning*, June 2025.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. NExT-GPT: any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 53366–53397, Vienna, Austria, July 2024. JMLR.org.

X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, Oct. 2022. ISSN 0167-739X. doi: 10.1016/j.future.2022.05.014.

Z. Xie, Z. Li, X. He, L. Xu, X. Wen, T. Zhang, J. Chen, R. Shi, and D. Pei. ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning. *Proceedings of the VLDB Endowment*, 18(8):2385–2398, Apr. 2025. ISSN 2150-8097. doi: 10.14778/3742728.3742735.

T. Xing, L. Garcia, F. Cerutti, L. Kaplan, A. Preece, and M. Srivastava. DeepSQA: Understanding Sensor Data via Question Answering. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 106–118, Charlottesvle VA USA, May 2021. ACM. ISBN 978-1-4503-8354-7. doi: 10.1145/3450268.3453529.

Z. Xu, H. Wang, and Q. Xu. Intervention-Aware Forecasting: Breaking Historical Limits from a System Perspective, May 2025.

H. Xue and F. D. Salim. PromptCast: A New Prompt-Based Learning Paradigm for Time Series Forecasting. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2023. ISSN 1558-2191. doi: 10.1109/TKDE.2023.3342137.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 Technical Report, May 2025.

R. Yasrab, J. Zhang, P. Smyth, and M. P. Pound. Predicting Plant Growth from Time-Series Data Using Deep Learning. *Remote Sensing*, 13(3):331, Jan. 2021. ISSN 2072-4292. doi: 10.3390/rs13030331.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *Eighth International Conference on Learning Representations*, Apr. 2020.

X. Zhang, R. R. Chowdhury, R. K. Gupta, and J. Shang. Large Language Models for Time Series: A Survey. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 8335–8343, Jeju, South Korea, Aug. 2024. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/921.

X. Zhang, S. Feng, and X. Li. From Text to Time? Rethinking the Effectiveness of the Large Language Model for Time Series Forecasting, Apr. 2025. DOI 10.48550/arXiv.2504.08818.

Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue. Meta-Transformer: A Unified Framework for Multimodal Learning, July 2023. DOI 10.48550/arXiv.2307.10802.

Z. Zhao, Y. Shi, S. Wu, F. Yang, W. Song, and N. Liu. Interpretation of Time-Series Deep Models: A Survey, May 2023. DOI 10.48550/arXiv.2305.14582.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA, Dec. 2023. Curran Associates Inc.

L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng. SGLang: Efficient Execution of Structured Language Model Programs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, Nov. 2024.

X. Zhou, W. Wang, F. J. Baldán, W. Buntine, and C. Bergmeir. MoTime: A Dataset Suite for Multimodal Time Series Forecasting, May 2025. DOI 10.48550/arXiv.2505.15072.

Y. Zhou, J. Yang, H. Zou, and L. Xie. Tent: Connect language models with iot sensors for zero-shot activity recognition. *arXiv preprint arXiv:2311.08245*, 2023.

Y. Zong, O. M. Aodha, and T. Hospedales. Self-Supervised Multimodal Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299 – 5318, Aug. 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3429301.

Appendix A. Details on Dataset Generation

Action Sequences. To ensure actions are repeated sufficiently often to permit asking, for instance, interesting counting questions, we double the chance of duplicating actions in each activity sequence. We chose the following 19 clearly discernible actions to generate the prompts: `playing guitar`, `punching`, `running`, `jumping once`, `drinking with the left hand`, `bowing`, `throwing a ball`, `shaking hands`, `skipping rope`, `picking something up with both hands`, `dancing`, `eating with the right hand`, `waving`, `catching a ball`, `kicking a ball`, `sitting down`, `golfing (swinging a club)`, `T-posing`, and `holding a baby`. We targeted an equal distribution of question and answer types. Furthermore, we ensured that the correct answer (in case of binary and multi) sits at a randomized position.

Configuration of Human Motion Generation. We experimented with several different STMC configurations to discover the most natural movement. We found that prompting all body parts jointly with a guidance scale of 3.0 and an overlap of $2l = 1.0$ seconds was most effective. As quantitatively evaluated later, this produced a diverse set of unique time series.

Alternative Human Motion Generation Systems. We also experimented with *Motion Diffusion Model* (MDM) (Tevet et al., 2023) using autoregressive stitching by conditioning on previously generated subsequences, and investigated using its successor *PriorMDM* (Shafir et al., 2024). However, neither reached the perceptual quality of STMC.

Diverseness. The question and answer types distribution is visualized in Figure 5. To ensure that almost all instances in the dataset are unique, we also searched for any duplicates. Table 3 shows that even if there exist multiple instances with the same textual question and answer, they refer to different action labels. Moreover, there is not a single time series that is exactly duplicated in the dataset (checked with SHA-256). However, we also want to avoid overly similar motions, not just exact matches.

Each QuAnTS sample carries a context trajectory $x \in \mathbb{R}^{T\times V}$ with $T = 320$ time steps and $V = 72$ variates. Each context is shared across five QA pairs, leaving $n = 30{,}000$ distinct series. We first define the scale at which to compare the variate v of two time series x, y as

$$s_v = \max\left(10^{-8}, \sqrt{\tfrac{1}{2}\left(\mathrm{Var}_t[x_v] + \mathrm{Var}_t[y_v]\right)}\right),$$

using the unbiased variance over the time axis. Pooling s_v per channel makes the score invariant to scales of each joint axis. The distance of x and y is then defined as

$$d(x, y) = \max_{v\in 1,\ldots,V} \mathrm{Q}^t_{0.95}\left(\frac{|x_{t,v} - y_{t,v}|}{s_v}\right).$$

This distance is intuitively thought of in units of channel standard deviations σ: $d(x, y) > \tau$ means that in at least one channel, at least 5% of time steps differ by more than τ pooled standard deviations. Because the outer reduction is a maximum over channels, two time series that differ on a subset of channels differ overall. We therefore first rule out duplicates on a subset of 8 channels, then on the complete set.

We finally compute groups of redundant time series using conservative single-linkage clustering, i.e., by transitively grouping all pairs of similar series. Retaining only a single member per such cluster again yields an effective sample size. Figure 6 shows that under the defined soft equality, almost no duplicates exist.

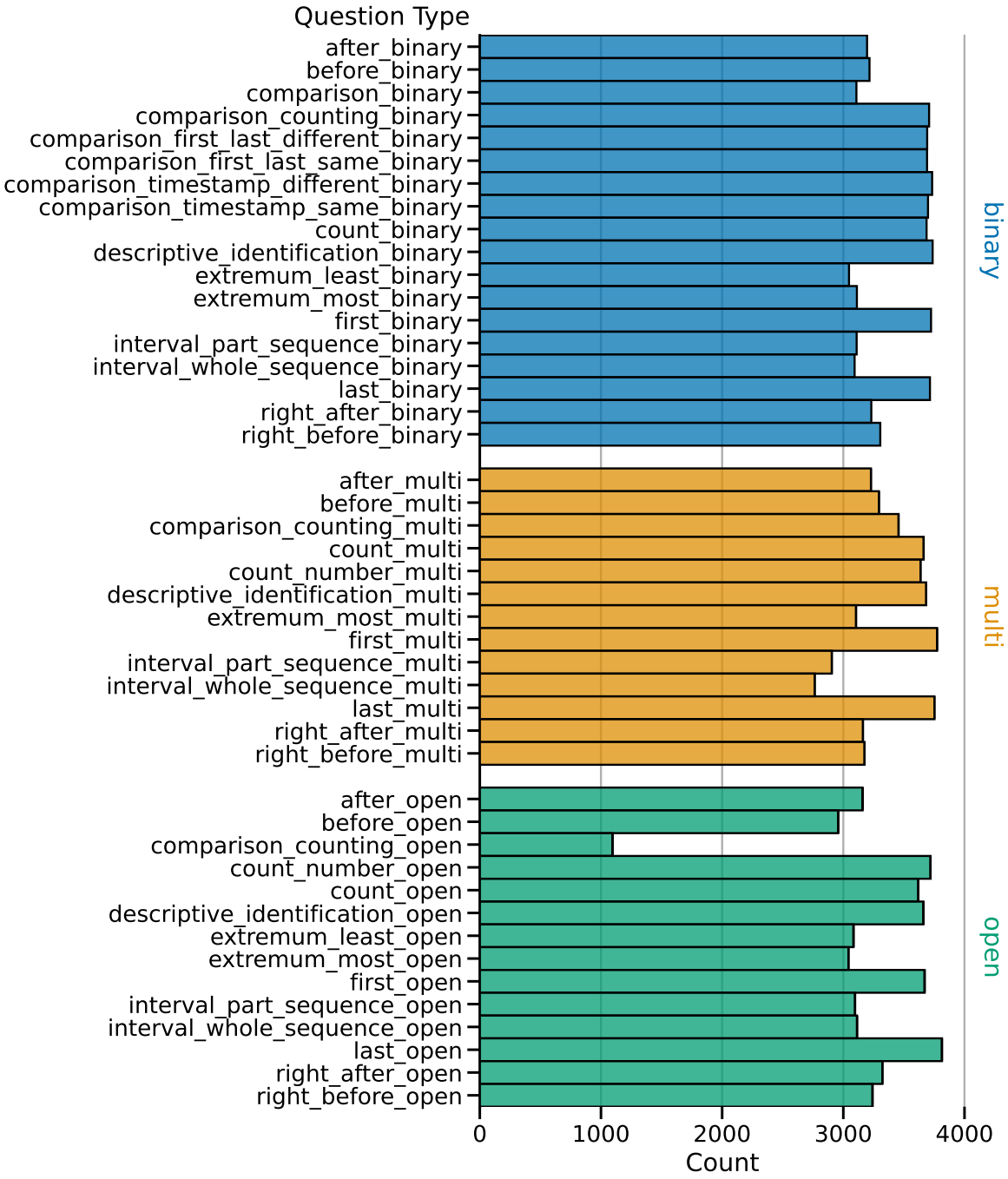


Figure 5: **QuAnTS features a diverse and balanced question and answer types distribution.** Some categories, such as comparison_counting_open, occur less often because of the particular circumstances that allow them to be posed without trivial responses.

Table 3: **Effectively all combinations of question-answer pairs and action sequence labels are distinct.** Furthermore, the probabilistic motion generation did not produce a single time series duplicate. All checks were performed using exact matching.

Data	Answer Type	# Samples	Eff. # Samples	Eff. Samples
Question + Answer	Binary	61,804	45,253	73.220%
	Multi	43,599	41,021	94.087%
	Open	44,597	33,194	74.431%
Question + Answer + Action Labels	Binary	61,804	61,791	99.979%
	Multi	43,599	43,598	99.998%
	Open	44,597	44,540	99.872%
Time Series	—	30,000	30,000	100.00%

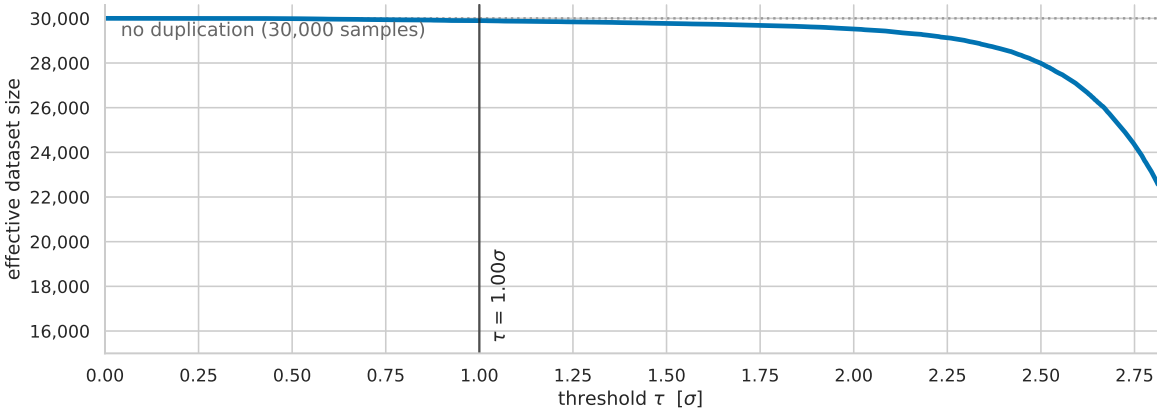


Figure 6: **Effectively all generated numerical motion time series are unique.** At $\tau = 1 \cdot \sigma$, the effective dataset size under soft equality is 29,895 (99.65% of the full 30,000).

Appendix B. Evaluating the LLM Judge

Section 3.4 described the metrics used to evaluate the answers of the machine learning models and humans. This section contains the full details for evaluating the LLM Judge.

Choice and Configuration of the LLM. We evaluated several moderately sized language models to limit the energy consumption and inference time required for the evaluation, while also considering that this evaluation scheme might be replicated by other works adopting QuAnTS. Namely, we tried three models of the Qwen3 family (8B-AWQ with Reasoning, 8B-AWQ without reasoning, and 4B-AWQ without reasoning) (Yang et al., 2025) and Llama 3.2 (3B, instruction tuned) (Llama-Team and MetaAI, 2024), where AWQ stands for Activation-aware Weight Quantization (Lin et al., 2025). The sampling parameters used to generate judge scores from the four LLMs are shown in Table 4. To allocate sufficient space to intermediate "thinking" artifacts, we generate up to 8192 tokens.

Table 4: **Sampling parameters used across the four LLM judge candidates.**

Model	Temperature	Top-p	Top-k	Repetition Penalty
Qwen3 4B/8B (Reasoning)	0.6	0.95	20	1.5
Qwen3 4B	0.7	0.8	20	1.5
Llama 3.2 3B (Instruct)	0.8	0.8	20	1.5

Prompt and Constrained Generation. To ensure a reliably parsable output, we explain the format in the prompt and constrain generated tokens to adhere strictly to the described JSON format using SGLang (Zheng et al., 2024). The rationale was allowed to be up to 500 characters long. We used the following prompt:

User Prompt for the LLM Judge

You will be given a scene_description with timestamps and a scene_question regarding that scene. You will also be given a reference_answer and system_answer couple. Your task is to provide a TotalRating scoring how well the system_answer answers the user concerns expressed in the scene_question. The reference_answer is provided as reference for a very good answer. Give your answer on a scale of 1 to 3, where 1 means that the system_answer is not helpful at all, and 3 means that the system_answer completely and helpfully addresses the scene_question. You will also provide a brief rationale for your rating.

Here is the scale you should use to build your answer:
1: The system_answer is terrible: the answer is factually incorrect or corrupt in some way (e.g., empty or nonsensical)
2: The system_answer is bad: somewhat correctly answers the scene_question like the reference_answer, yet contains additional useless/ambiguous/wrong information or is incomplete
3: The system_answer is good: correctly answers the scene_question and is helpful in the context of the scene

Provide your assessment in the following JSON format:

```
{
    "BriefRationale": "<your rationale, at most 3 sentences>",
    "TotalRating": <your rating>
}
```

You MUST provide values for "BriefRationale" and "TotalRating" in your answer. Do NOT respond with any other text in your answer. Immediately, start with the JSON object, without any additional text or explanation. Make sure to output complete and valid JSON.

Now here are the question and answer:
- scene_description: {scene_description}
- scene_question: {scene_question}
- reference_answer: {reference_answer}
- system_answer: {system_answer}

Provide your assessment.

Verifying Alignment with Human Judgment. To ensure that the LLM Judge aligns well with human assessment, we collected 200 random output samples across the experiments from Section 4.1. Three human annotators were presented with the question, a reference answer, and the model answer. Their agreement with each other was very high, with 88.5% perfect alignment at an average variance of ∼0.04 across annotators. We can thus employ the mean of the human ordinal scale as a baseline to evaluate the LLM Judges against. All four language models are evaluated in Table 5. Due to the high quality of responses at reasonable inference speed, we used Qwen3 8B (AWQ, Reasoning) for the LLM Judge.

Table 5: **Qwen3 8B (AWQ, Reasoning) aligns best with human judgement.** Highlighted **bold** is the best value per column, with the arrows (↓/↑) indicating which direction is desired. Average over 10 seeds.

Model	Avg. Δ (↓)	Pearson (↑)	Spearman (↑)	Kendall (↑)
Qwen3 8B (AWQ, Reasoning)	**0.149**	**0.912**	**0.927**	**0.878**
Qwen3 8B (AWQ)	0.254	0.843	0.857	0.794
Qwen3 4B (AWQ)	0.155	0.883	0.891	0.853
Llama 3.2 3B (Instruct)	0.430	0.640	0.654	0.603

Appendix C. Details on Evaluating QuAnTS

C.1 Verifying and Benchmarking using Language Models

In Section 4, we finetune Llama 3.1 8B independently for the question-only, time series-only, and Naive configurations and each answer format, yielding nine QLoRA adapters. Each adapter uses its corresponding training split for five epochs, with an effective batch size of 16 and a fixed seed of 42. We use AdamW (Loshchilov and Hutter, 2018) with $\beta = (0.9, 0.999)$, weight decay 0.01, learning rate 2×10^{-4} with linear decay, no warmup, and gradient clipping at 1.0. QLoRA (Dettmers et al., 2023) uses rank 16, $\alpha = 32$, dropout 0.1, and no trainable biases. Adapters target the attention and MLP projections; base weights remain frozen. We use NF4 double quantization with BF16 computation and gradient checkpointing.

The time series undergo per-trajectory min-max scaling over the complete tensor to $[-999, +999]$, rounding to integers, and serialization as compact nested JSON preserving the $320 \times 24 \times 3$ rank-3 structure.

Training supervises only answer tokens and the final end-of-sequence token, without a chat template or sequence packing. The final fifth-epoch adapters generate greedily without output constraints, allowing up to 32 new tokens for closed answers and 128 for open answers. Invalid closed answers count as incorrect; the main results use the held-out test split.

Finetuning and evaluating the nine adapters, including prediction generation, open-answer judging, and perplexity evaluation, required approximately 5,492 Nvidia H100 GPU-hours due to the very long context length.

C.1.1 Prompts for Section 4.1.

The following prompts were used to answer the questions in Section 4. Orange denotes the target answer, which is supplied only during training and omitted from the inference prompt. For the question-only, time series-only, and Naive adapters, the answer and end-of-sequence tokens are supervised, conditioned on the preceding prompt.

Prompt For Ground Truth Scene Descriptions and Questions (Q1)

Context: {textual description of the action sequence}

Question: {question}

Answer: {answer}

Prompt For Only the Questions (Q2)

Question: {question}

Answer: {answer}

Prompt For Only the Time Series Encoded as Text (Q3)

Instruction: You are provided with a 3-dimensional dataset of shape [320, 24, 3], representing 3-dimensional spatial locations of 24 human-body joints over 320 time steps. Using this data, please analyze the movements and provide the answer associated with this human activity sequence.

Data: {scaled time series data}

Answer: {answer}

C.1.2 Prompts for Section 4.3.

The experiments "Ablation: Only Question" and "Ablation: Only TS" use the prompts in (Q2) and (Q3), respectively. The Naive Baseline uses the following:

Prompt For Naive Baseline (Time Series Encoded as Text and Questions)

Instruction: You are provided with a 3-dimensional dataset of shape [320, 24, 3], representing 3-dimensional spatial locations of 24 human-body joints over 320 time steps. Using this data, please analyze the movements and respond to the following question related to human activity recognition.

Data: {scaled time series data}

Question: {question}

Answer: {answer}

C.1.3 Measuring Language Modeling Deterioration

For Table 6, we evaluate the unchanged Llama 3.1 8B base model and all nine final adapters on the raw WikiText-2 test corpus from `Salesforce/wikitext` (Merity et al., 2016). We preserve the released row order, including empty rows, and concatenate rows with two newline characters. All models use the same base tokenizer, without a task prompt, chat template, or additional special tokens.

We use overlapping windows of 2,048 tokens with a stride of 1,024 over the complete token stream. Tokens already scored in an earlier window supply context only; every target token is scored exactly once, apart from the first corpus token, which has no preceding context. This gives $N = 289{,}075$ scored targets for every model and precision condition. Models run in evaluation mode without gradients, with token cross-entropy computed in FP32. We accumulate the summed negative log-likelihood before calculating

$$\bar{\ell} = \frac{1}{N} \sum_{i \in \mathcal{S}} -\log p_\theta(w_i \mid c_i), \qquad \text{PPL} = \exp(\bar{\ell}),$$

where $\mathcal{S}$ is the set of scored target positions and c_i is the preceding context available within the corresponding window. The ratio to the precision-matched base is $\exp(\bar{\ell}_{\text{adapter}} - \bar{\ell}_{\text{base}})$; it is a multiplicative factor, not a percentage.

The primary comparison loads the base weights in BF16 and attaches the adapters without merging them. The additional NF4 comparison uses double quantization and BF16 computation, matching the training quantization configuration.

We also perform a small qualitative continuation experiment on five fixed 128-token Wikipedia prefixes for each tested model, generating greedily at most 24 new tokens. The tested models are the NF4 base, question-only binary, time-series-only multiple-choice, Naive binary, and Naive open adapters. The selected closed-answer adapters produce task-label repetitions or immediate termination; the base continues prose, and the Naive open adapter produces short sentence-like continuations that terminate early.

Table 6: **Language modeling deteriorates drastically after fine-tuning for QuAnTS.**

System	Answer Type	BF16		NF4	
		PPL (↓)	PPL/base	PPL (↓)	PPL/base
Base	—	5.61	1.00	5.97	1.00
Ablation: Only Question	Binary	3.17×10^8	5.65×10^7	3.71×10^8	6.21×10^7
	Multi	9870.98	1758.13	6.46×10^4	1.08×10^4
	Open	19.24	3.43	25.65	4.30
Ablation: Only TS	Binary	2.10×10^7	3.74×10^6	1.74×10^7	2.92×10^6
	Multi	6.17×10^8	1.10×10^8	7.32×10^8	1.23×10^8
	Open	10.39	1.85	12.60	2.11
Naive: TS + Question	Binary	1.14×10^8	2.03×10^7	1.48×10^8	2.48×10^7
	Multi	1.35×10^7	2.40×10^6	2.90×10^7	4.86×10^6
	Open	10.61	1.89	12.99	2.18

C.2 Measuring Human Performance

Setup. To ensure clear stimuli, each page of the online study form showed the single relevant time series context as a rendered video clip using the gender-neutral SMPL-H model (Loper et al., 2015; Romero et al., 2017), and below the question with matching answer modality. Participants were able to watch the context video clip as many times as they wished, even after seeing the question (and possibly answer options). The excerpts of the survey introduction screen in Figure 9 and questions in Figure 10 illustrate this.

Demographics. We paid a fair median hourly wage of at least 12 Euros. The diverse self-reported demographics are shown in Figure 7.

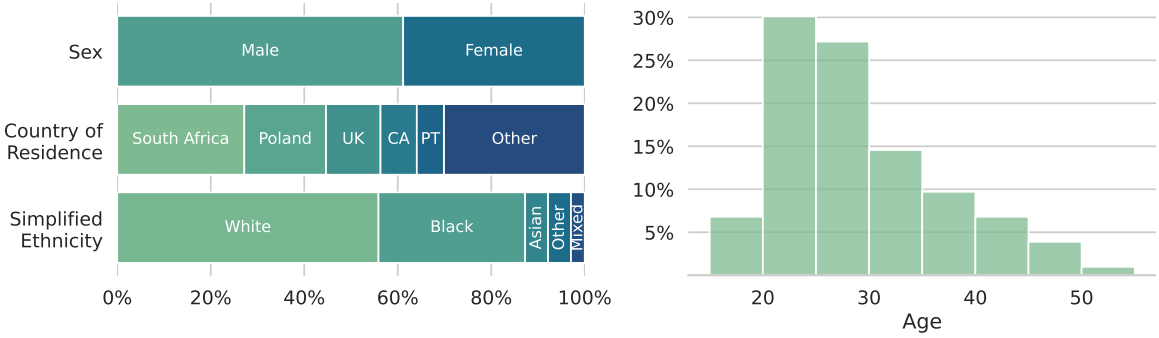


Figure 7: **The study was performed on a diverse set of 105 participants**, where each provided 20 answers spread over the three answer types.

Identification of Actions. In addition to the findings presented in the main body of the work, we evaluated how well humans can recognize individual actions within the sequences. The data was obtained via the questions shown in Subfigure 10c. For each of the $n = 440$ results from Table 2, four actions had to be identified. For only 136 (7.73%) cases out of those $440 \times 4 = 1,760$, the option `<not recognized>` was selected. Figure 8 shows the binary F1 score, where unrecognized actions were counted as wrong.

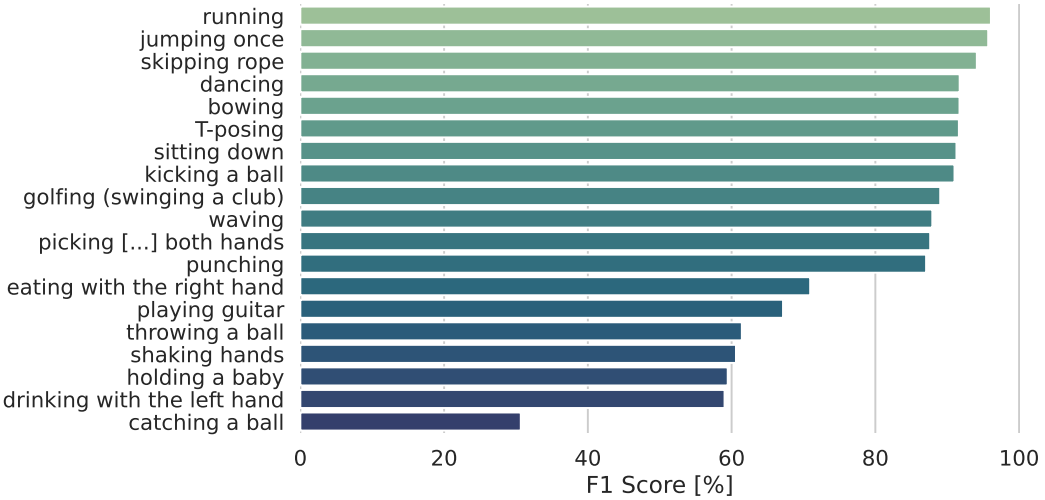


Figure 8: **Most actions are easily recognizable from the animated video clips.**

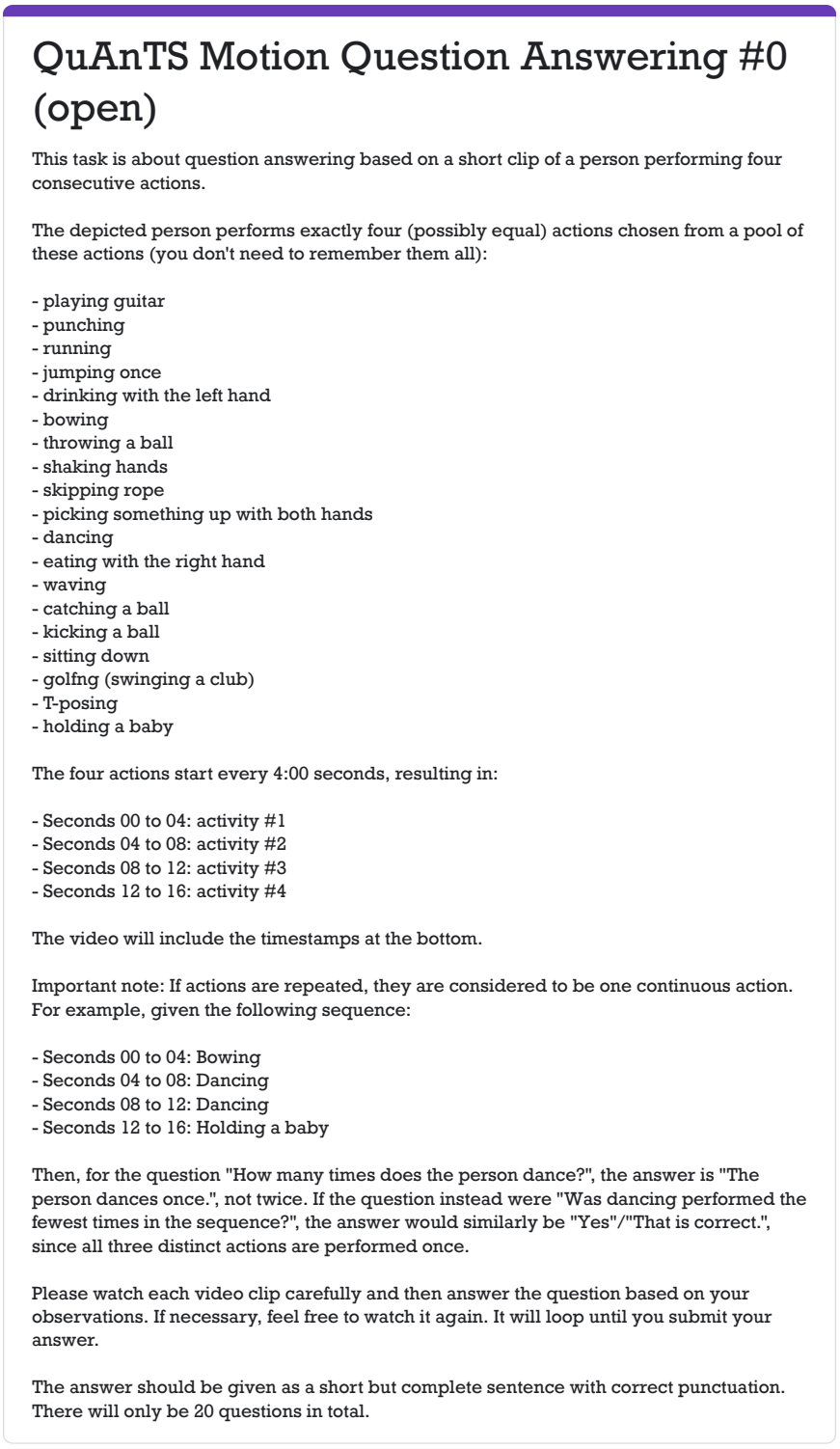
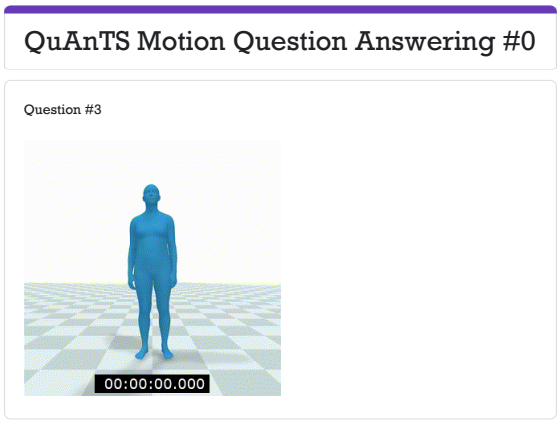
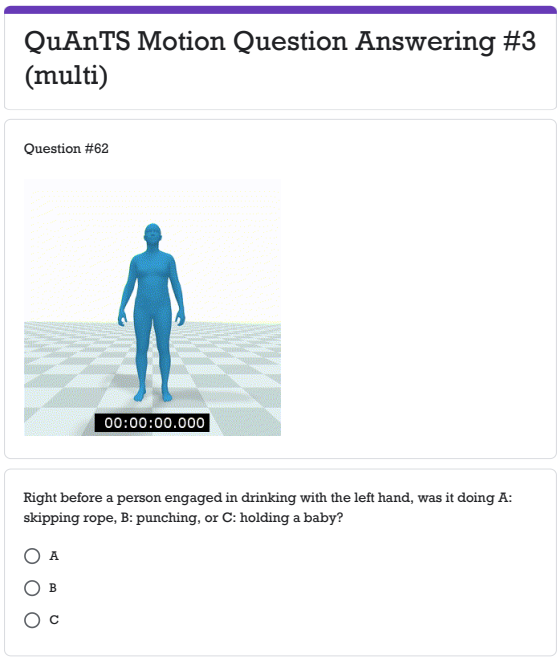


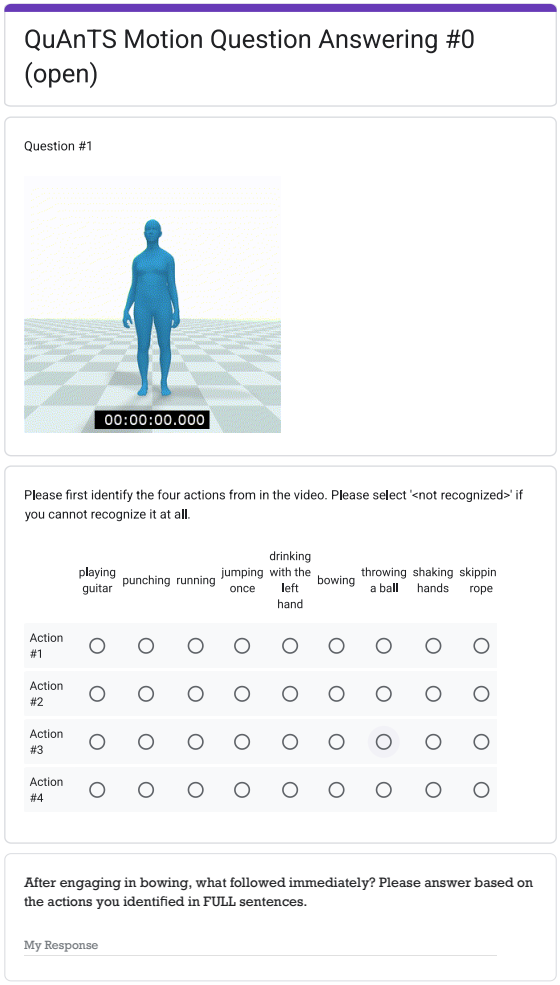
Figure 9: **Study participants were first presented with this introduction screen.** The text for the binary and multi variants only differed in the description of the answer modalities in the last three paragraphs.



(a) Binary



(b) Multi



(c) Open

Figure 10: **Study participants were presented with screens matching the respective answer modality.** For the open variant, we also collected information on the identifiability of the action sequences.

C.3 Adapting ChatTS to QuAnTS

Most importantly, to stay within the upper limit of the 30 variates that ChatTS can handle, we reduce the joints to only `pelvis`, `left_foot`, `right_foot`, `head`, `left_shoulder`, `right_shoulder`, `left_elbow`, `right_elbow`, `left_wrist`, and `right_wrist`, which we deemed the most discriminative for the actions in QuAnTS.

Due to implementation limitations, we were unable to strictly constrain the tokens generated by ChatTS to adhere to the described format, resulting in slightly elevated rates of non-parsable responses.

We employed the following prompt:

Prompt For ChatTS

<|im_start|>system
You are a helpful assistant.<|im_end|><|im_start|>user
I have 30 time series representing human motion data from skeleton tracking. All time series are of length 320.
The time series:

- pelvis x axis: <ts><ts/>
[...]
- right_wrist z axis: <ts><ts/>

The recorded time series tracks four subsequent actions. Each takes 80 steps, reflecting four seconds of wall-clock time.

The possible action categories being performed are:
- playing guitar
[...]
- holding a baby

First, identify the sequence of actions. If the same action occurs multiple times directly after one another, it is counted as a single continuous one. Only merge actions that have identical labels. Recall that there are four actions in total.

Finally, answer the following question in a full sentence: "{question}".

Only output the following:
1. "Actions: <4 identified action names before merging>",
2. "Merged Actions: <1 to 4 potentially merged actions>", and
3. then the final response as "Answer: <your response>".<|im_end|><|im_start|>assistant
Actions:

C.4 Details on xQA

To evaluate the difficulty of QuAnTS, we evaluate the xQA baseline family based on xLSTM-Mixer in Section 4.3. It factorizes the problem into (i) action extraction from the time series and (ii) question answering with an LLM. We report results on two LLM backends, xQA–Qwen and xQA–LLaMA, with an otherwise identical pipeline. To separate LLM reasoning quality from perception errors, we provide results on ground-truth action sequences (GT) from the debugging metadata of QuAnTS and the actual sequences predicted by our action encoder (AE).

Action Encoder. We utilize a segmentation model based on xLSTM-Mixer (Kraus et al., 2025a). It is trained on the QuAnTS action labels, which are not part of the regular data available to TSQA models. The model predicts one of 19 actions per segment, i.e., using a human motion time series of length 80 with 72 variates. The hidden size in the xLSTM blocks was set to 256, and training was performed using RAdam (Liu et al., 2019) with a learning rate of 10^{-4}. We built the implementation on PyTorch 2.8 (Paszke et al., 2019) with Lightning 2.4 (Falcon and The PyTorch Lightning team, 2019). For the xLSTM-Mixer implementation, we rely on a custom fork[3] of the transformers library (Wolf et al., 2020). The action encoder reached high accuracies above 99% as detailed by Figure 11.

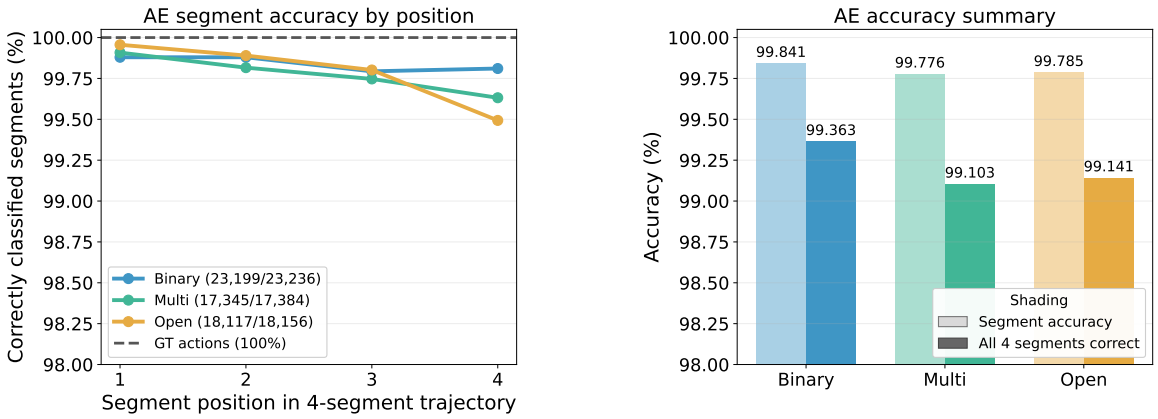


Figure 11: **The action encoder recognizes the 19 generated actions almost perfectly.** Visualized are the accuracies on the QuAnTS test set.

Prompt and Constrained Generation. Similar to the design of the LLMJudge laid out in Appendix B, we explain the format in the prompt and constrain generated tokens to adhere strictly to the described JSON format using SGLang (Zheng et al., 2024). Both are prompted to return a single JSON object with fields `actions`, `steps`, and `answer`. The answer field is task-specific: for binary it must be `"true"`/`"false"`, for multi-choice it must be `"A"`/`"B"`/`"C"`, and for open-ended it is free-form. Invalid generations are retried a small number of times. We used the following prompts for the three different answer types:

3. `https://github.com/mauricekraus/transformers#xlstm-mixer`

Binary xQA Prompt

<|im_start|>system
You are a helpful timeseries question answering model.
You are given a timeseries (TS) of **action names** and a question about them.
If an action appears multiple times in a row in the TS, it counts as **one** occurrence (consecutive repeats collapse conceptually).
Translate the timeseries into reasoning steps and then answer the question. Provide your thought process in "steps: [...]".

Example 1
TS: ["running", "bowing", "jumping once", "playing guitar"], QS: Do the person's actions stay the same before and after they are jumping once?

```
{
    "actions": ["running", "bowing", "jumping once", "playing guitar"],
    "steps": [
        "1. Identify the index of \"jumping once\" which is 2.",
        "2. Now we need to look before and after index 2, thus index 1 and
            index 3.",
        "3. At actions[1] the action is bowing and actions[3] the action is
            playing guitar.",
        "4. Because these two actions are different, the answer is false."
    ],
    "answer": "false"
}
```

Example 2
TS: ["catching a ball", "running", "catching a ball", "running"], QS: Is the person conducting catching a ball the same amount as running?

```
{
    "actions": ["catching a ball", "running", "catching a ball", "running" ],
    "steps": [
        "1. Let's count catching a ball. Catching a ball occurs at index 0
            and 2, thus 2 times.",
        "2. Let's count running. Running occurs at index 1 and 3, thus 2
            times.",
        "3. Therefore, the answer is true."
    ],
    "answer": "true"
}
```

<|im_end|><|im_start|>user
TS: ["{action_1}", "{action_2}", ..., "{action_N}"], QS: {question}
Respond with exactly one JSON object with the keys "actions", "steps", and "answer".
<|im_end|><|im_start|>assistant

Multiple-Choice xQA Prompt

<|im_start|>system
Initial Instruction as in the Binary xQA Prompt

Example 1
TS: ["running", "jumping once", "shaking hands", "jumping once"], QS: Which activity does the person carry out 1 time? A: jumping once, B: shaking hands, C: golfing (swinging a club)

```
{
    "actions": ["running", "jumping once", "shaking hands", "jumping once"],
    "steps": [
        "1. Identify the actions that are performed only once. From the
            timeseries we see that running and shaking hands occur only once
            .",
        "2. Among A,B,C, only shaking hands matches this (so B).",
        "3. Therefore, the answer is **B**."
    ],
    "answer": "B"
}
```

Example 2
TS: ["golfing (swinging a club)", "golfing (swinging a club)", "shaking hands", "running"], QS: Which activity does the person carry out 0 times? A: golfing (swinging a club), B: shaking hands, C: punching

```
{
    "actions": ["golfing (swinging a club)", "golfing (swinging a club)", "
        shaking hands", "running"],
    "steps": [
        "1. Rule out actions performed > 0 times.",
        "2. golfing (swinging a club) appears (consecutive repeats count as
            1), so rule out A.",
        "3. shaking hands appears once, so rule out B.",
        "4. punching does not appear, so answer is **C**."
    ],
    "answer": "C"
}
```

Respond with exactly one JSON object and no additional explanation or narrative. If none of the options seem to be valid choose a random one.
<|im_end|><|im_start|>user
TS: ["{action_1}", "{action_2}", ..., "{action_N}"], QS: {question with options A/B/C}
Respond with exactly one JSON object with the keys "actions", "steps", and "answer", where "answer" is one of "A", "B", or "C".
<|im_end|><|im_start|>assistant

Open-Ended xQA Prompt

<|im_start|>system
Initial Instruction as in the Binary xQA Prompt

Example 1
TS: ["running", "punching", "punching", "playing guitar"], QS: Describe in plain English what the person is doing.

```
{
    "actions": ["running", "punching", "punching", "playing guitar"],
    "steps": [
        "1. Notice two consecutive occurrences of the same action count as
            one.",
        "2. Unique action sequence = [running, punching, playing guitar].",
        "3. Summarize in plain English."
    ],
    "answer": "The person runs, then punches, then plays guitar."
}
```

Example 2
TS: ["jumping once", "jumping once", "shaking hands", "golfing (swinging a club)", "golfing (swinging a club)", "golfing (swinging a club)", "holding a baby"], QS: Explain step by step how many times each action occurs, and then summarize.

```
{
    "actions": ["jumping once", "jumping once", "shaking hands", "golfing (
        swinging a club)", "golfing (swinging a club)", "golfing (swinging a
        club)", "holding a baby"],
    "steps": [
        "1. Count unique consecutive:",
        " - jumping once appears consecutively ⇒ count=1",
        " - shaking hands appears once ⇒ count=1",
        " - golfing (swinging a club) appears consecutively ⇒ count=1",
        " - holding a baby appears once ⇒ count=1",
        "3. Totals: jumping once=1, shaking hands=1, golfing (swinging a club
            )=1, holding a baby=1.",
        "4. Summarize in one sentence."
    ],
    "answer": "The person jumps once, shakes hands, plays golf with a club,
        and then holds a baby."
}
```

Respond with exactly one JSON object and no additional explanation or narrative.
<|im_end|><|im_start|>user
TS: ["{action_1}", "{action_2}", ..., "{action_N}"], QS: {free-form question}
Respond with exactly one JSON object with the keys "actions", "steps", and "answer".
<|im_end|><|im_start|>assistant

Appendix D. Extended Results

D.1 Action Recognizability and Human Performance

We also assess whether human performance depends on the humans' ability to recognize actions in the provided video clips, or on linguistic and reasoning abilities. Figure 12 shows that the presence of certain action labels in the question or answer has an influence on the respective task performance. This positive correlation is explored quantitatively in Figure 13, which shows that the relationship is particularly pronounced for the open split. This partly explains the shortcomings of human responses to certain actions (see Section 4.2).

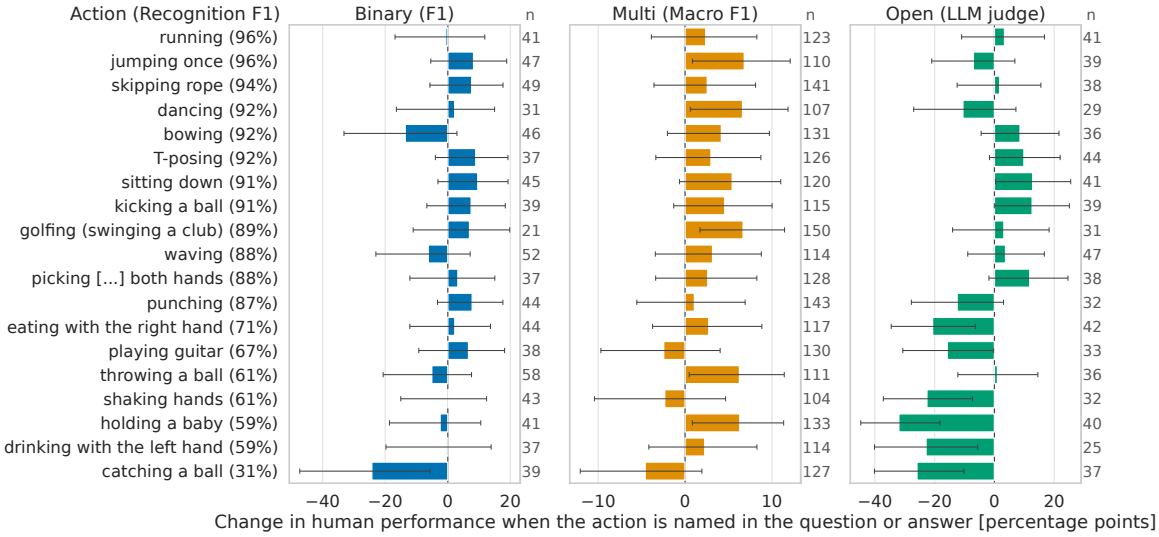


Figure 12: **The presence of different actions in the question or answer drives changes in QA performance.** For open, the relation is less obvious, since in 2 out of 3 cases, a hard-to-detect action is in a distractor rather than the correct answer.

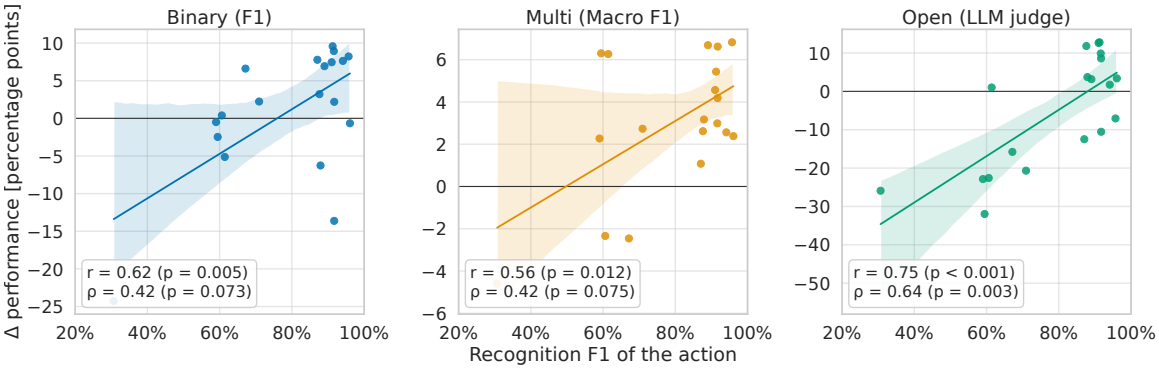


Figure 13: **Overall, action recognizability is positively correlated with performance.** We also report the respective Person's correlation coefficient r and Spearman's rank correlation coefficient ρ, along with their p-values. The relationship is significant only for the open split.

D.2 Results by Answer and Question Type

A more detailed breakdown of selected results is provided in Figure 14.

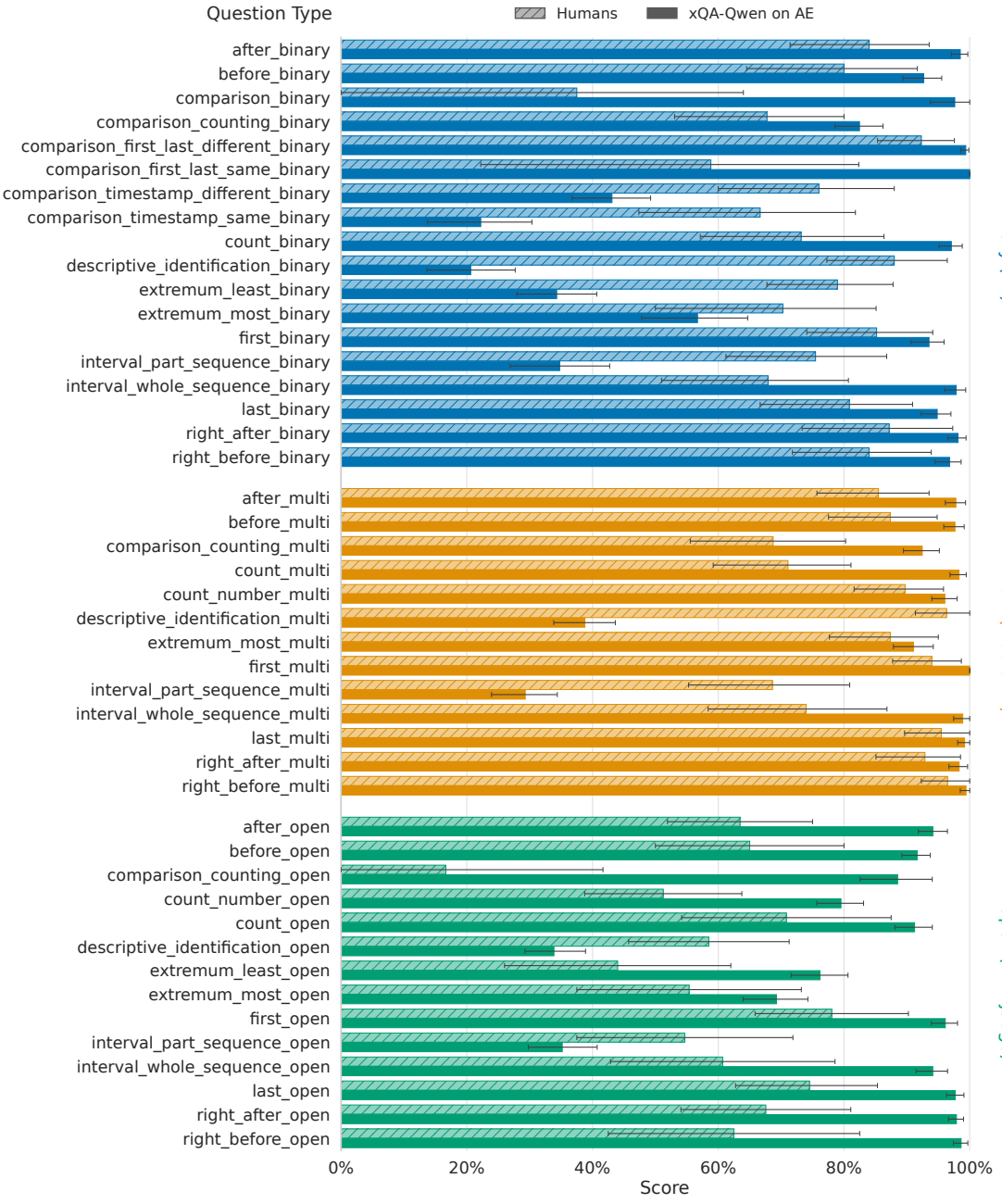


Figure 14: **xQA-Qwen on AE is ahead of humans on 35 of the 45 question types, especially on the multi and open splits.** The clearest exception to that is descriptive_identification, where humans lead decisively across splits. The performance of humans ($n = 2{,}100$ responses) and xQA (all $n = 15{,}000$ pairs) is shown with 95-percentile bootstrap intervals from 2,000 resamples, with wider ones for humans due to fewer samples.

Appendix E. Datasheet for QuAnTS

In compliance with the DMLR Submission Guidelines[4], this section provides a datasheet following Gebru et al. (2021). Specifically, it contains documentation and intended uses (E.5), links for viewing and obtaining the dataset (E.6), as well as a hosting, licensing, and maintenance plan (E.6 & E.7).

E.1 Motivation

For what purpose was the dataset created? ▷ *QuAnTS provides both training and evaluation data for time series question answering (TSQA). It poses textual questions regarding numerical multivariate time series and provides reference answers alongside additional metadata.*

Who created the dataset and on behalf of which entity? ▷ *The dataset was created by Felix Divo and Maurice Kraus at the AIML Group, Department of Computer Science, Technische Universität Darmstadt in Darmstadt, Germany.*

Who funded the creation of the dataset? ▷ *The work was funded by multiple entities as listed in the Section "Acknowledgments and Disclosure of Funding" following Section 5.*

E.2 Composition

What do the instances that comprise the dataset represent? ▷ *The dataset comprises triplets of (1) a numerical time series of synthetic human motion tracking data of four actions, (2) a textual question, and (3) an answer. The latter can be binary (Yes/No), multiple-choice (one of A/B/C), or open (free text). In addition, QuAnTS provides supplemental textual versions for all answer types, the underlying action sequences (for debugging only), ground-truth scene descriptions (for debugging only), and the question/answer type for statistical purposes.*

How many instances are there in total (of each type, if appropriate)? ▷ *We generated 30.000 time series contexts, each with five question-answer pairs, resulting in a total of 150.000 instances.*

Does the dataset contain all possible instances, or is it a sample of instances from a larger set? ▷ *QuAnTS was created synthetically from scratch to reach a set target size. We put in great effort to ensure diversity of questions and answers (see, in particular, Section 3.3) and a balanced distribution of question and answer types (see Appendix A). The probabilistic motion diffusion process ensures sufficient diversity of the numerical time series.*

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? ▷ *The components of each instance are described above. Each time series is a motion trajectory of 16 seconds at 20 frames per second (320 steps) of 24 joints with a 3-dimensional position, and thus a tensor of rank* $320 \times 24 \times 3$.

Is there a label or target associated with each instance? ▷ *The target value to predict is the answer, which is in either binary, multiple-choice, or open format.*

4. See `https://data.mlr.press/submissions`.

Is any information missing from individual instances? ▷ *No.*

Are relationships between individual instances made explicit? ▷ *The samples are independent and identically distributed (i.i.d).*

Are there recommended data splits (e.g., training, development/validation, testing)? ▷ *For best comparability, we split the dataset into pre-defined training, validation, and testing splits of 120.000, 15.000, and 15.000 samples, respectively.*

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. ▷ *The synthetic generation process limits potential sources of error. Beyond manual inspection, we provide extensive validation and ablation experiments in Section 4.1. The vast majority of question-answer pairs are unique, as shown in Table 1. Possible errors can arise if the human motion generation fails to faithfully represent the conditioning activity description.*

Is the dataset self-contained, or does it link to or otherwise rely on external resources? ▷ *It is fully self-contained.*

Does the dataset contain data that might be considered confidential? ▷ *No.*

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or otherwise cause anxiety? ▷ *No.*

E.3 Collection Process

How was the data associated with each instance acquired? What mechanisms or procedures were used to collect the data? ▷ *The dataset was created synthetically, as detailed in Section 3. We sampled random actions and generated randomized matching human motions. For each instance, a question-answer type was chosen randomly, for which we instantiate a randomly selected text template.*

If the dataset is a sample from a larger set, what was the sampling strategy? ▷ *n/a*

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated? ▷ *No third parties were involved in the dataset's creation. Crowdworkers assisted in evaluating the dataset. The details, including their fair compensation, are provided in Section 4.2 and appendix C.2.*

Over what timeframe was the data collected? ▷ *See above for context. The evaluation data was collected in Q3 and Q4 of 2024.*

Were any ethical review processes conducted (e.g., by an institutional review board)? ▷ *No such review was deemed necessary as the study posed negligible risk: adult participants viewed synthetic, non-sensitive animations and answered comprehension questions, posing no risks beyond those of daily life. Voluntary withdrawal was possible at any point and a fair compensation of at least 12 Euros per hour median was provided. We report demographics only in aggregate (see Appendix C.2, specifically Figure 7).*

E.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done? ▷ *No, the desired target format was generated directly.*

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? ▷ *n/a*

Is the software that was used to preprocess/clean/label the data available? ▷ *The software used to generate the entire dataset is available at* `https://github.com/mauricekraus/quants-generate`.

E.5 Uses

Has the dataset been used for any tasks already? ▷ *We provide initial experiments using the brand new dataset in Section 4.*

Is there a repository that links to any or all papers or systems that use the dataset? ▷ *n/a*

What (other) tasks could the dataset be used for? ▷ *We anticipate usage to focus on TSQA.*

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? Is there anything a dataset consumer could do to mitigate these risks or harms? ▷ *We anticipate very few of such issues. Potential misrepresentation of actions could be inherited from STMC, e.g., as it does not represent people with certain (motoric) disabilities. This is a very common limitation in human motion data. Consumers of the dataset should be aware of these limitations when deploying systems in the real world.*

Are there tasks for which the dataset should not be used? ▷ *No (see previous).*

E.6 Distribution

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created? ▷ *Yes.*

How will the dataset be distributed? Does the dataset have a digital object identifier (DOI)? ▷ *The dataset is freely available at* `https://huggingface.co/datasets/dasyd/quants`, *which provides various common data formats. The DOI of the dataset is 10.57967/hf/6663.*

Will the dataset be distributed under a copyright or other intellectual property license, and/or under applicable terms of use? ▷ *The dataset is available under the permissive CC BY 4.0 license, see* `https://creativecommons.org/licenses/by/4.0`.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? ▷ *We are aware of none.*

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? ▷ *We are aware of none.*

E.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? ▷ *QuAnTS is distributed by Huggingface (DOI 10.57967/hf/6663). We would seek other means of archival and distribution, such as university repositories, if long-term accessibility became uncertain at some point in the future.*

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? ▷ *Email adresses of all authors are provided on p. 1.*

Is there an erratum? ▷ *We will publish any important changes on the respective repositories for generation (`https://github.com/mauricekraus/quants-generate`), evaluation (`https://github.com/mauricekraus/quants-generate`), or QuAnTS itself (`https://huggingface.co/datasets/dasyd/quants`).*

Will the dataset be updated (e.g., to correct labeling errors, add new instances, or delete instances)? ▷ *We currently do not plan any updates.*

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? ▷ *n/a*

Will older versions of the dataset continue to be supported/hosted/maintained? ▷ *Yes.*

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? ▷ *The license permits freely adapting the dataset. However, we currently do not plan to collect any community contributions.*