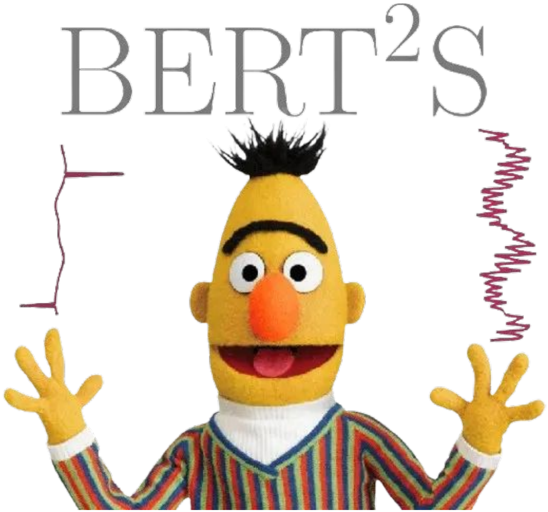
Do We Really Need Another Time-Series Forecasting Model?

Maurice Kraus



TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

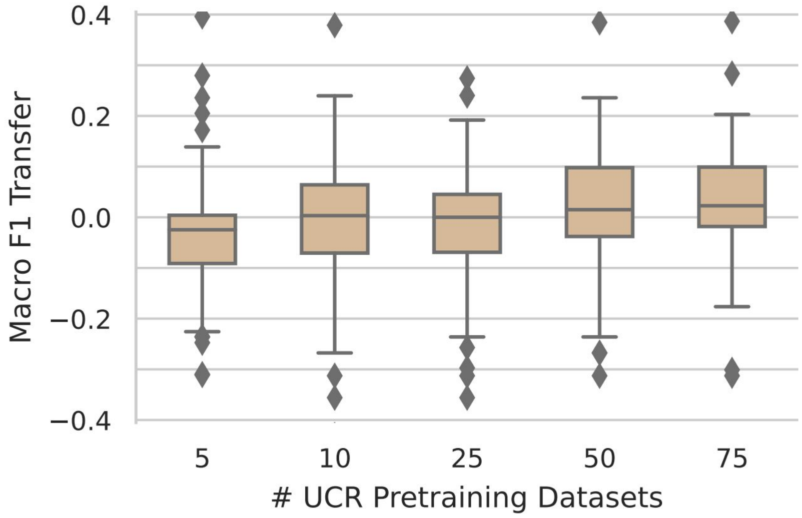
NEURAL INFORMATION PROCESSING SYSTEMS

About AIML Lab

- Who we are: A research group focused on various different fields in machine learning
- Based in Germany, TU Darmstadt
- Prof. Kristian Kersting

Rise & Challenges of TSFM

- NLP & Vision FMs inspired universal, zero-shot TS forecasting
- Early multi-dataset pretraining (XIT) proved cross-dataset transfer is possible
- Modern TSFMs scale via huge real+synthetic corpora + LLM influenced architectural choices



(b) Finetuning on a hold-out set of 25 datasets each.

[1] Kraus, Maurice; Divo, Felix et al.
"United We Pretrain, Divided We Fail! Representation Learning for Time Series by Pretraining on 75 Datasets at Once." Preprint, 2023.

Mixed Evidence Baseline

- Benchmark results vary widely.
- Lightweight supervised models often match TSFMs.
- Benchmarks disagree
  - GIFT-Eval vs OpenTS vs FoundTS vs TSLib
  - Challenged by Lorenzo et al. 2025 [2]
- No model dominates across tasks.

[2] Brigato, Lorenzo et al. “Position:There are no Champions in Long-Term Time Series Forecasting”, Preprint, 2025.

Do We Always Need a Foundation Model?

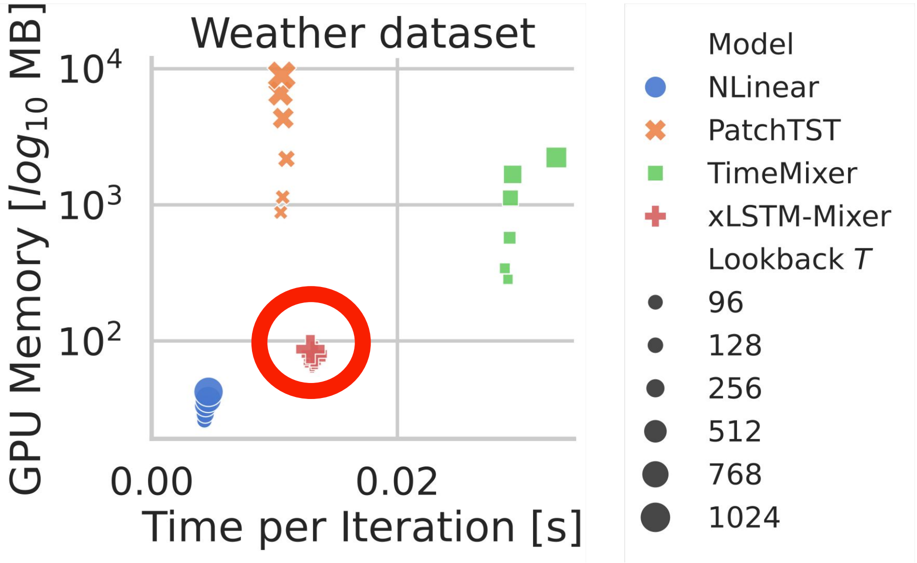
- Universal vs purpose-built trade-off.
- Zero-shot helps when data is scarce.
- Supervised/domain specific wins in data-rich settings (e.g., finance).
- Specialization can exceed generalization.
- Gupta et al. 2024 shows marginal gains of fine tuned over fully supervised in medical data

[3] Gupta, Divij et al. “Beyond LoRA: Exploring Efficient Fine-Tuning Techniques for Time Series Foundational Models”, Preprint, 2024.

Efficiency

Why xLSTMs here?

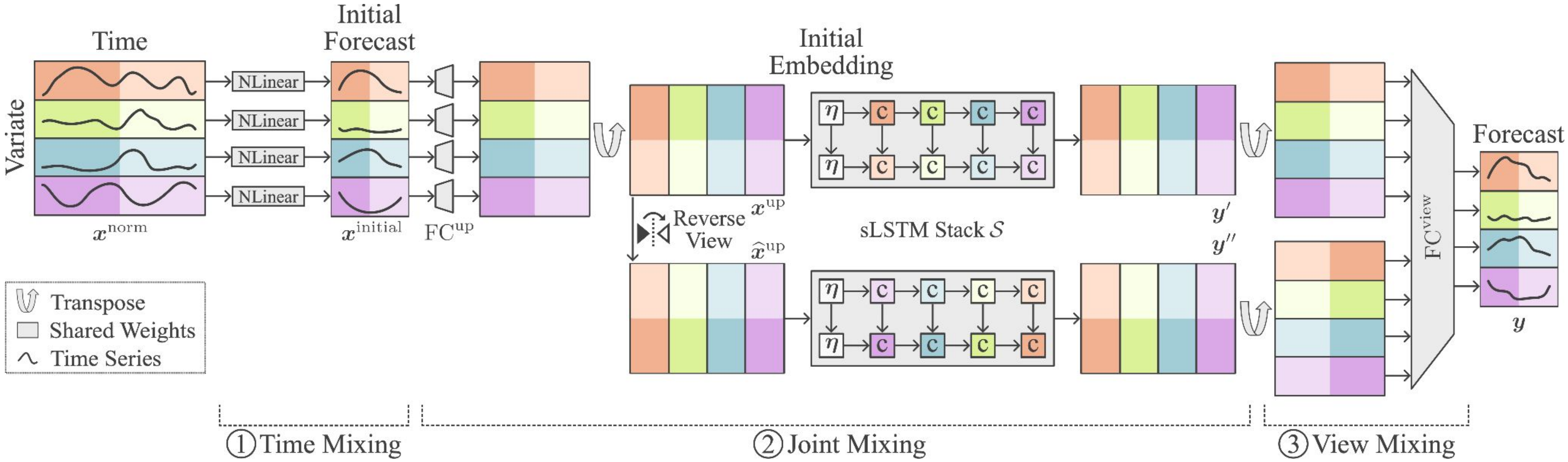
- xLSTMs [4] use scalar memories and gating → strong sequence modeling without quadratic attention.
- Very low GPU memory and competitive iteration time.
- Fits edge / constrained deployments.



[4] Beck, Maximilian, Korbinian Pöppel, Markus Spanring, et al. “Xlstm: Extended Long Short-Term Memory.” NeurIPS, 2024.

xLSTM-Mixer:

Multivariate Time Series Forecasting by Mixing via Scalar Memories [5]

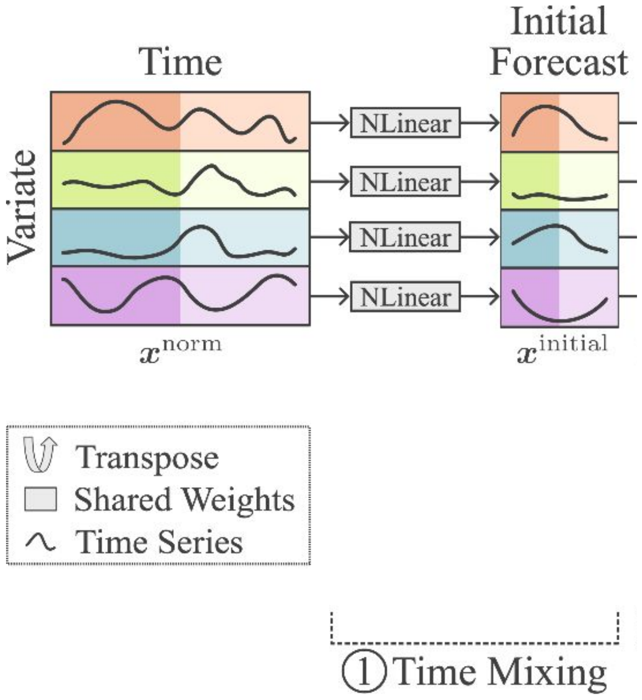


[5] Kraus, Maurice; Divo, Felix; Singh Dhami, Devendra; Kersting, Kristian.
"xLSTM-Mixer: Multivariate Time Series Forecasting by Mixing via Scalar Memories." NeurIPS 2025

The Mixing Process

Time Mixing

- Start with a **shared linear forecast** [3] (cheap, channel-independent).



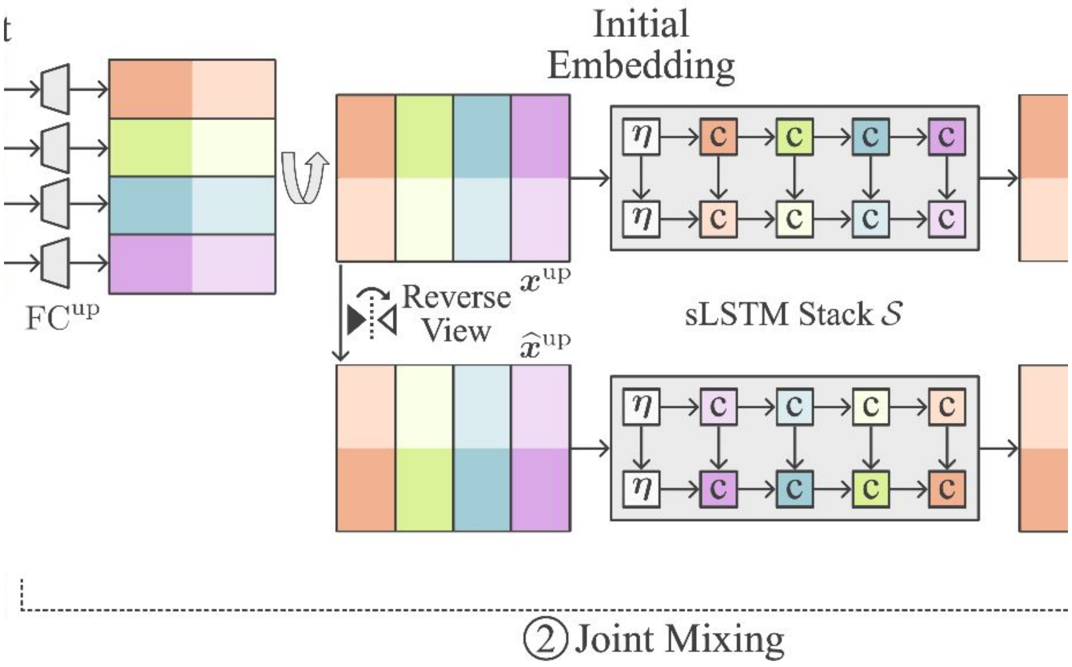
[3] Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu. “Are Transformers Effective for Time Series Forecasting?” *AAAI*, 2023.

The Mixing Process

Joint Mixing

- Start with a **shared linear forecast** [3] (cheap, channel-independent).

- **Refine** it with xLSTM block(s) that mix time + variates.

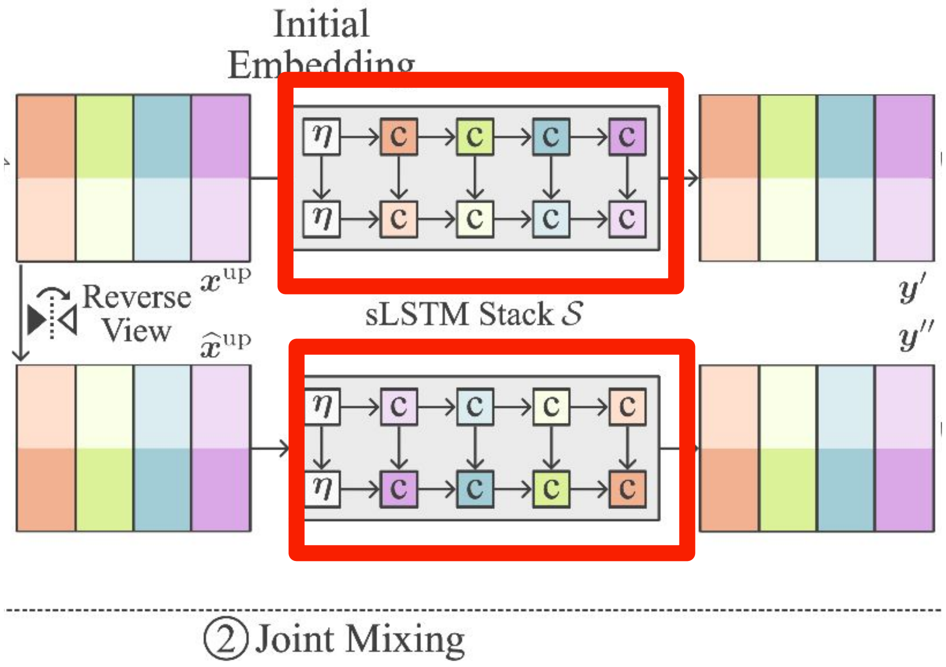


[4] Liu, Yong, Tengge Hu, Haoran Zhang, et al. “Itransformer: Inverted Transformers Are Effective for Time Series Forecasting.” *ICLR*, 2023.

The Mixing Process

Joint Mixing

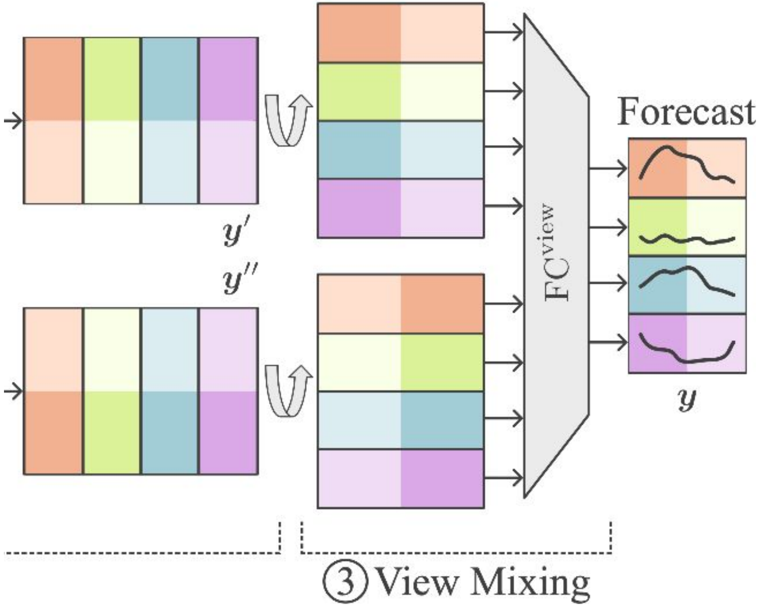
- Start with a **shared linear forecast** [3] (cheap, channel-independent).
- **Refine** it with xLSTM block(s) that mix time + variates.
- **Two views** (forward + reversed) → **view mixing** → final forecast



The Mixing Process

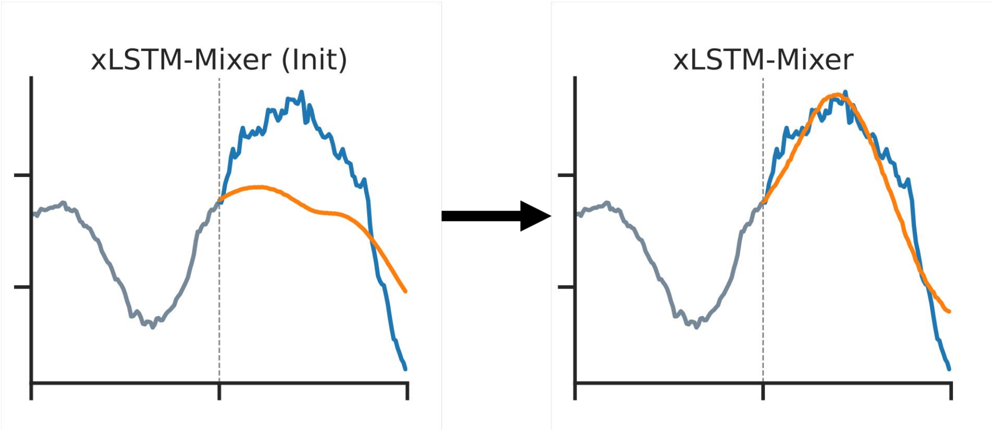
View Mixing

- Start with a **shared linear forecast** [3] (cheap, channel-independent).
- **Refine** it with xLSTM block(s) that mix time + variates.
- **Two views** (forward + reversed) → **view mixing** → final forecast
- **Result**: As expressive as big models yet parameter-frugal like RNNs.



Iterative refinement

- Think: ‘**rough guess** → s**marter correction**’.
- Early stage handles what’s easy - xLSTM stages focus capacity on what’s hard.
- Multi-view mixing regularizes and reduces parameters via shared weights.



Benchmark Performance

- SOTA on standard multivariate benchmarks.
- Strong probabilistic forecasts on GIFT-Eval.
- Can be used in a versatile fashion

Models	Recurrent						Mixer						MLP					
	xLSTM-Mixer		xLSTMTime 2024		LSTM 1997 †		TimeMix.++ 2025a		TimeMix. 2024a		TSMixer 2023c		CycleNet 2024		DLinear 2023		TiDE 2023	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	**0.219**	**0.250**	0.222	0.255	0.444	0.454	0.226	0.262	0.222	0.262	0.225	0.264	0.223	0.264	0.246	0.300	0.236	0.282
Electricity	**0.153**	**0.245**	0.157	0.250	0.559	0.549	0.165	0.253	0.156	0.246	0.160	0.256	0.156	0.251	0.166	0.264	0.159	0.257
Traffic	0.392	**0.253**	0.391	0.261	1.011	0.541	0.416	0.264	0.387	0.262	0.408	0.284	0.403	0.282	0.434	0.295	**0.356**	0.261
ETTh1	**0.397**	0.420	0.408	0.428	1.198	0.821	0.419	0.432	0.411	0.423	0.412	0.428	0.435	0.440	0.423	0.437	0.419	0.430
ETTh2	0.340	0.382	0.346	0.386	3.095	1.352	0.339	0.380	**0.316**	0.384	0.355	0.401	0.367	0.405	0.431	0.447	0.345	0.394
ETTm1	**0.339**	**0.366**	0.347	0.372	1.142	0.782	0.369	0.378	0.348	0.375	0.347	0.375	0.360	0.388	0.357	0.379	0.355	0.378
ETTm2	**0.248**	**0.307**	0.254	0.310	2.395	1.177	0.269	0.320	0.256	0.315	0.267	0.322	0.263	0.324	0.267	0.332	0.249	0.312

Model	**MASE ↓**	**CRPS ↓**	**Rank (CRPS) ↓**
TiRex	0.724	0.498	1
xLSTM-Mixer (ours)	0.780	0.510	2
TEMPO_ensemble	0.862	0.514	3
Toto_Open_Base_1.0	0.750	0.517	4
TabPFN-TS	0.771	0.544	5
YingLong_300m	0.798	0.548	6
timesfm_2_0_500m	0.758	0.550	7
YingLong_110m	0.809	0.557	8
sundial_base_128m	0.750	0.559	9
YingLong_50m	0.822	0.567	10

The Current Landscape of Architectures

Model	From	Architecture	#Parameters
TimesFM	Das et al. (2023)	Transformer (Decoder)	200M
Chronos-1	Ansari et al. (2024)	Transformer (Encoder-Decoder)	20M - 710M
Chronos-2	Ansari et al. (2025)	Transformer (Encoder)	120M
Moirai 1.0	Woo et al. (2025)	Transformer (Encoder)	14M-311M
Moirai 2.0	Liu et al. (2025)	Transformer (Decoder)	11M-?
FlowState	Graf et al. (2025)	SSM + Functional Bases	2.6M-9.1M
TiRex	Auer et al. (2025)	Recurrent (xLSTM)	35M
xLSTM-Mixer	Kraus et al. (2025)	Recurrent (xLSTM) + Mixing	Per Dataset ~(50k-100M)

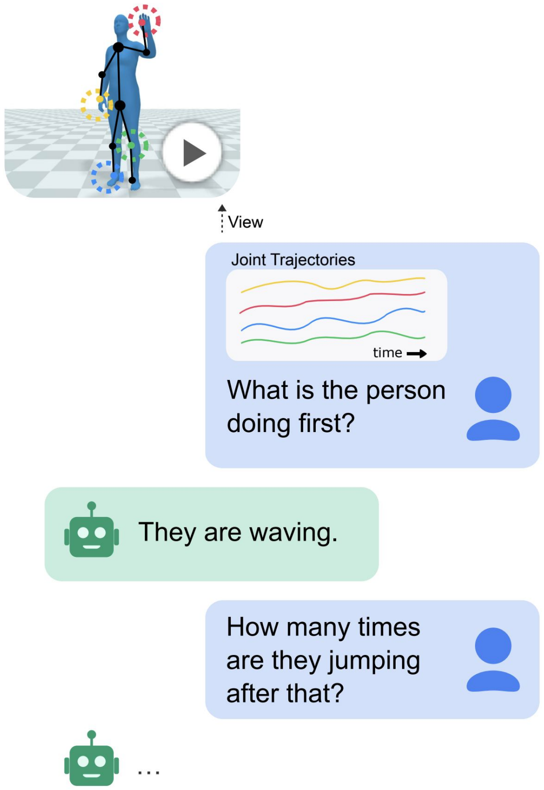
Versatility needs to be shown for true FMs

- Forecasting alone doesn’t prove foundation status
- We need FMs that work across modalities, tasks, and domains
- Models: SensorLM [7], ChatTS [8], LLaSA [9] → TS ↔ language, richer reasoning
- Benchmarks: QuAnTS [10], BEDTime [11] → TS QA + natural-language description

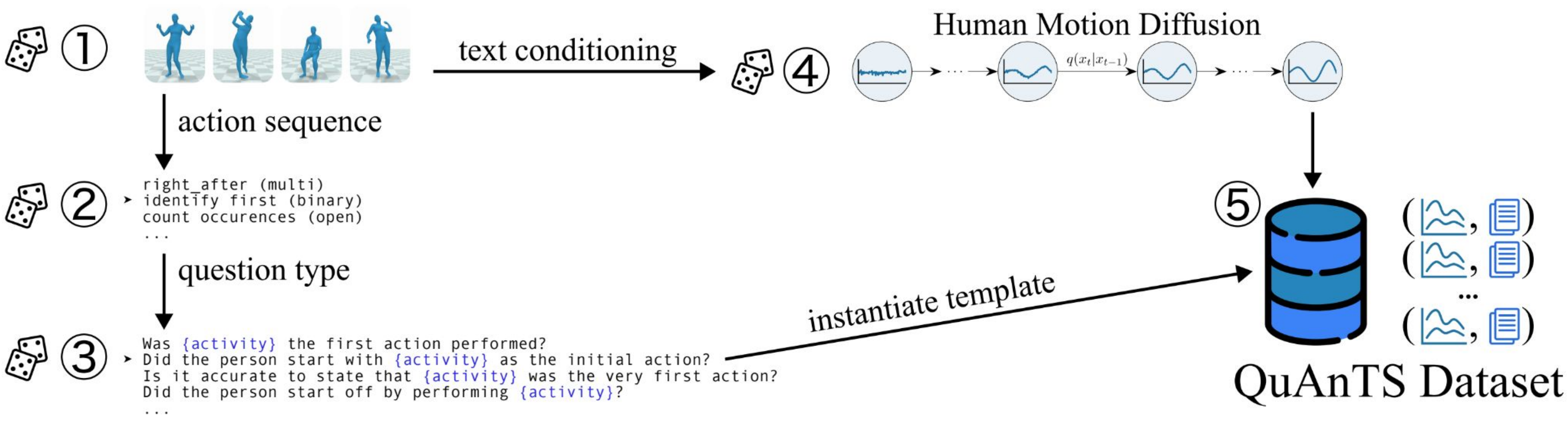
[7] Zhang, Yuwei; Ayush, Kumar et al. "SensorLM: Learning the Language of Wearable Sensors", NeurIPS 2025
[8] Xie, Zhe; Li, Zeyan et al. "ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning", PVLDB 2025
[9] Imran, Asif; Khan, Mohammad Nur Hossain et al. "LLaSA: A Sensor-Aware LLM for Natural Language Reasoning of Human Activity from IMU Data", UbiComp 2025
[10] Divo, Felix; Kraus, Maurice et al. "QuAnTS: Question Answering on Time Series", Preprint, 2025.
[11] Sen, Medhasweta; Gottesman, Zachary et al. "BEDTime: A Unified Benchmark for Automatically Describing Time Series", Preprint, 2025.

QuAnTS: Question Answering on Time Series

- Users want to be able to speak with time series
- "Why did sales drop?" "Find the anomaly."
- The Challenge: LLMs are **suboptimal** at **processing raw** numerical **time series** directly.



QuAnTS: Question Answering on Time Series



QuAnTS: Question Answering on Time Series

- ChatTS suffers big drop in performance
- High agreement of our xQA and humans

	System	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 (↑)
	Ablation: Only Question	64.47%	64.47%	64.47%	64.38%
	Ablation: Only TS	0.28%	0.28%	0.28%	0.28%
	Humans ($n = 820$)	86.59%	86.55%	86.61%	86.54%
Multi	Naive: TS + Question	9.69%	9.69%	9.69%	9.62%
	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%

xLSTM-Mixer

TSFMs are the Future, But...

- No "One Size Fits All" (Yet): We do not have a "BERT" that solves everything perfectly.
- Forecasting models slowly get better and better.
- The Data Scale Argument: If a lot of data is available, efficient supervised training on domain data beats zero-shot generalization.
- Proper ablations are still needed

Do We Really Need Another Time-Series Forecasting Model?

The Hybrid Future:

- High-Volume/Low-Latency: Efficient supervised models for large amounts of data (e.g., xLSTM-Mixer).
- Specific Tasks: Domain-Specific FMs.

AIML Lab

TECHNISCHE UNIVERSITÄT DARMSTADT

Centre for Cognitive Science

dfki ai

With funding from the:

Federal Ministry of Research, Technology and Space

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

hessian.AI

KompAKI