# Cherish every MOMENT:

## Long-Context Time Series Foundation Models

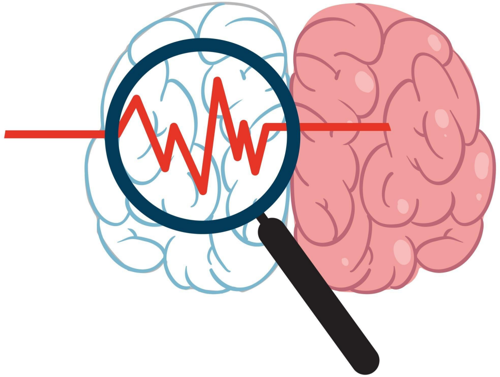
Nina Żukowska, Mononito Goswami, Michał Wiliński, Willa Potosnak, Prof. Artur Dubrawski

7/11/ 2024

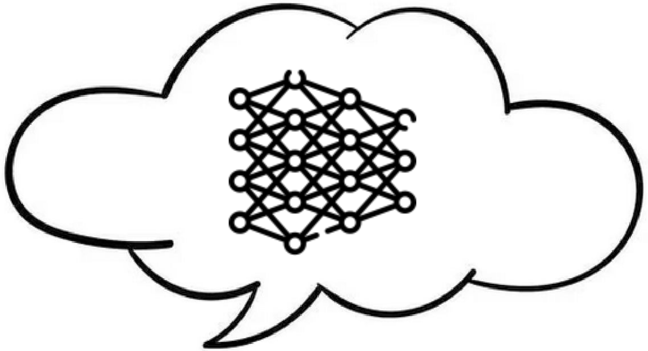
# Contents

1. Time Series and Foundation Models
2. Context extension in Time Series Foundation Models.
3. Our Approach: Infini-Channel Mixer.

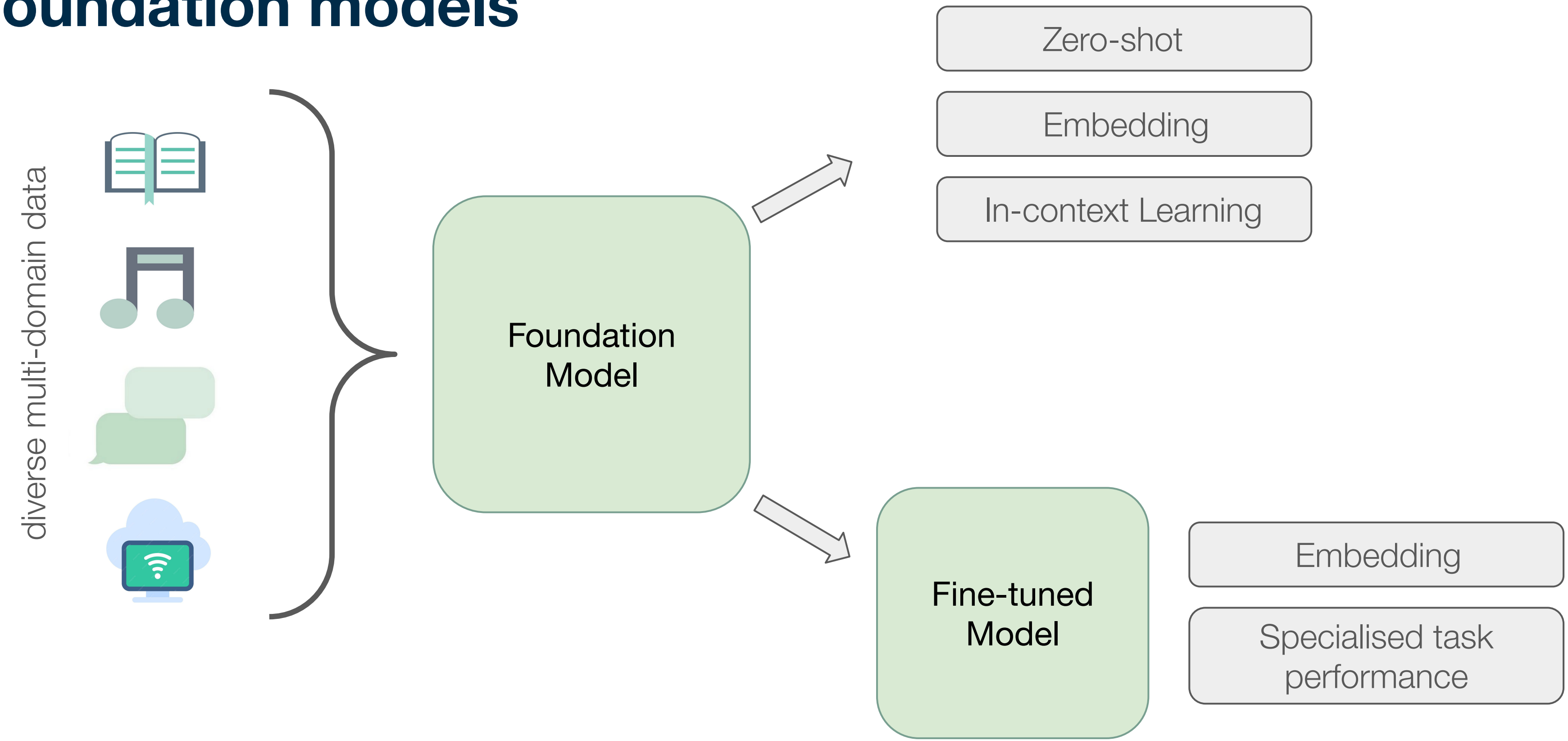
# Time Series are heterogeneous.



# Time Series are heterogeneous.



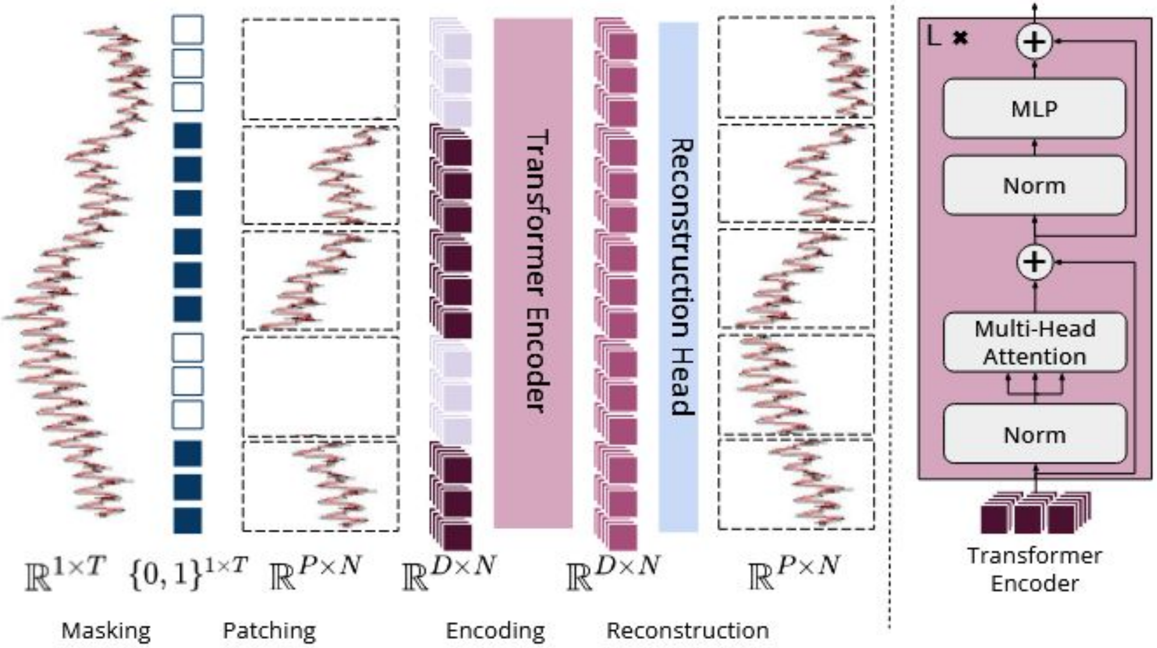
# Foundation models



# MOMENT

Time Series Foundation Models[1,2,3] generally model short **univariate time series.**

- Strong representation learning
- Multiple tasks





[1]Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A Family of Open Time-series Foundation Models. In International Conference on Machine Learning. arXiv preprint arXiv:2402.03885. Retrieved from https://arxiv.org/abs/2402.03885.

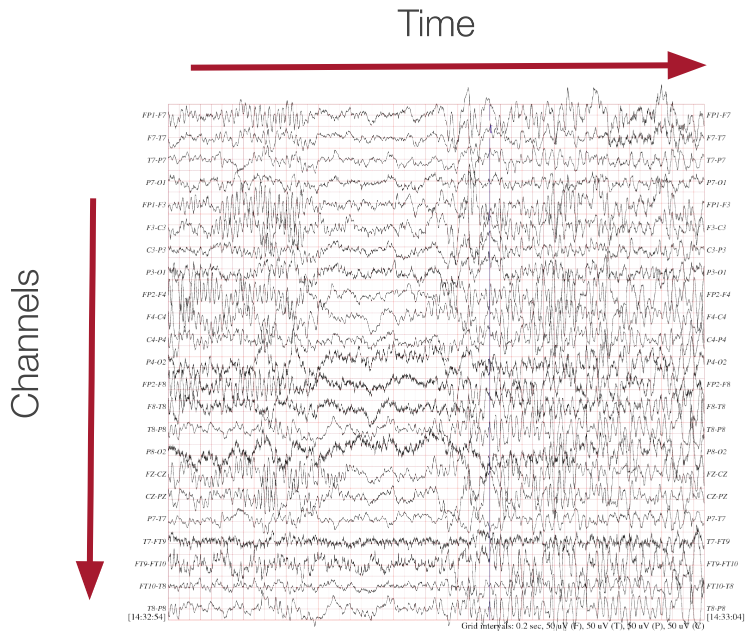
[2]Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2024). Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. arXiv preprint arXiv:2310.08278.

[3]Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified Training of Universal Time Series Forecasting Transformers. arXiv preprint arXiv:2402.02592. Retrieved from https://arxiv.org/abs/2402.02592.

# But what is Context Expansion?

Expansion of context means:

- Capture intricate dependencies **between channels**, e.g. different leads in electrocardiogram
- To capture long-term dependencies in the same channel
- Improve the predictive accuracy of time series foundation models



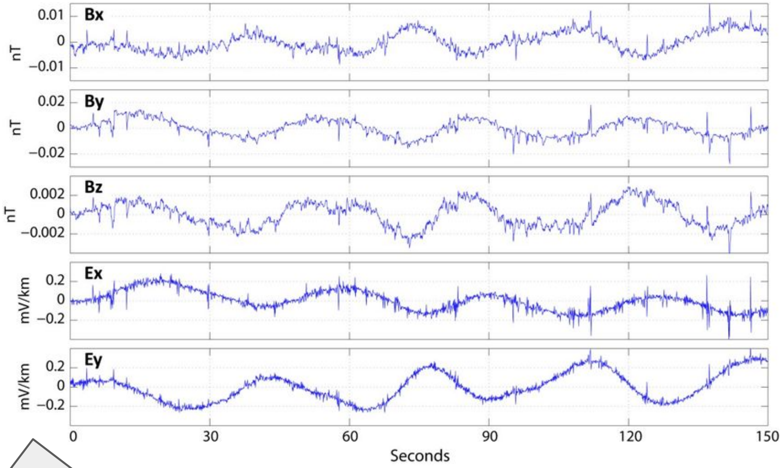

Expand context length along 2 dimensions: time and channels

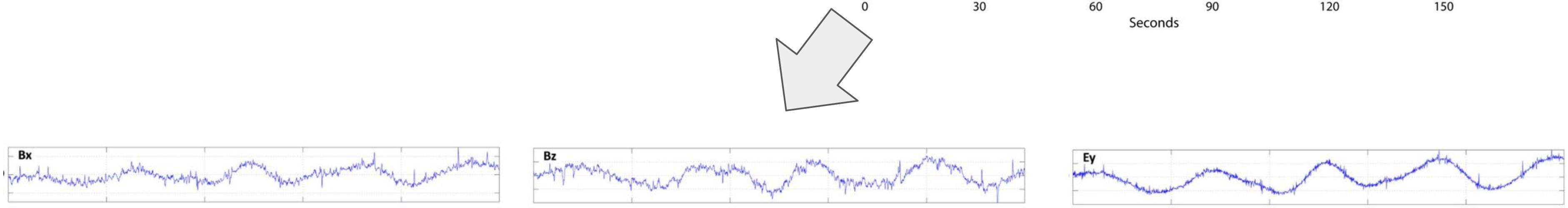
[1]Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A Family of Open Time-series Foundation Models. In International Conference on Machine Learning. arXiv preprint arXiv:2402.03885. Retrieved from https://arxiv.org/abs/2402.03885.
[2]Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2024). Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. arXiv preprint arXiv:2310.08278.
[3]Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified Training of Universal Time Series Forecasting Transformers. arXiv preprint arXiv:2402.02592. Retrieved from https://arxiv.org/abs/2402.02592.

# A possible approach[3] to mixing channels...

**Method:**

- Flattening input sequence
- Relative Channel Encoding
    - 
- High memory requirement 



# Channel Mixing

## Adapters

Uses already established univariate representations

**Graph Transformer layers**

E.g., UP2ME

**Mixing head**

E.g., Tiny Time Mixers

## End-to-End Channel Mixers

### Homogenous

Channel-mixing is embedded in each layer

**Intra-channel Patching**

E.g., iTransformer

**Relative Encodings**

E.g., Moirai

### Non-homogenous

Channel-mixing is not embedded in each layer

**Dedicated Layers**

E.g., Crossformer

**Compressive Memory**

E.g., Ours

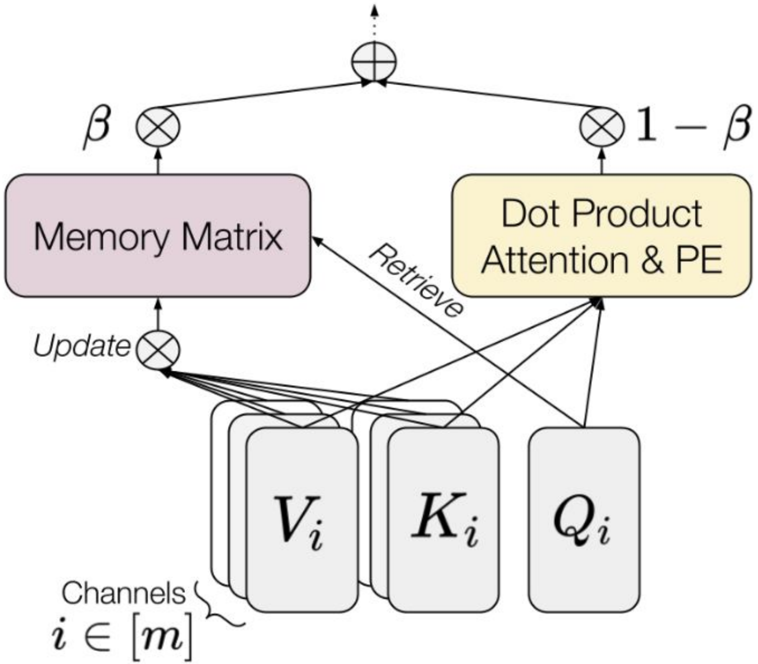
# Our approach …

**Method:**

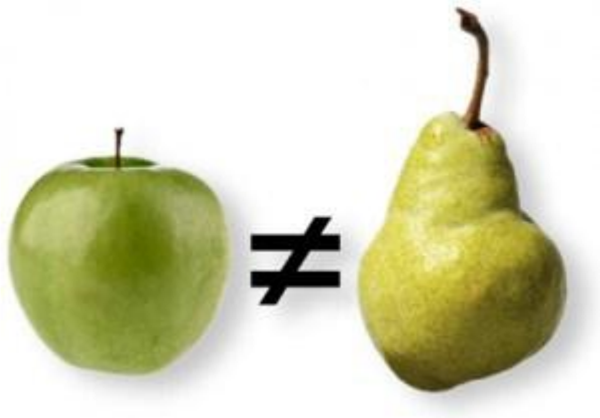
Introduce a compressive memory matrix, adding **one trainable parameter** per attention head

**Step 1: Aggregate Cross-Channel Information**
Initialize compressive memory matrix and normalization term. Aggregate information from all channels.

**Step 2: Retrieve Cross-Channel Data**
Use query matrix to retrieve and combine inter- and intra-channel information, adjusting with a learned gating scalar for balance.



# Experiments and comparing pears and apples...




**We use the same model architecture!**

# Results (Supervised Settings)

| Model / Example | Class | Design | Exchange | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather |
|---|---|---|---|---|---|---|---|---|
| N-BEATS | No Channel Mixing | Channel Independence | 0.524 | 0.461 | 0.410 | 0.346 | 0.278 | 0.211 |
| MOMENT-Tiny | | | 0.249 | 0.418 | 0.359 | 0.339 | **0.234** | 0.206 |
| UP2ME | Adapter | Graph Transformer | 0.240 | 0.435 | 0.367 | 0.340 | 0.237 | **0.204** |
| Crossformer | Non-homogeneous End-to-End Mixer | Dedicated Intra-Channel Attention | 0.559 | 0.571 | 0.654 | 0.390 | 0.515 | 0.227 |
| iTransformer | Homogeneous End-to-End Channel Mixer | Multivariate Patching | 0.245 | 0.429 | 0.380 | 0.353 | 0.251 | 0.212 |
| MOIRAI | | Concatentation + Relative Encoding | 0.243 | 0.426 | 0.357 | 0.340 | 0.249 | 0.216 |
| ICM (Ours) | | Compressive Memory | **0.232** | **0.416** | **0.349** | **0.333** | **0.234** | 0.205 |

# Does Beta Matter? Yes!

| **Model name** | Fine-tune $\beta$ | Exchange | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather |
|---|---|---|---|---|---|---|---|
| MOMENT-Tiny | – | 0.250 | <u>0.437</u> | 0.343 | 0.333 | <u>0.230</u> | 0.222 |
| +Infini-Channel Mixer | × | <u>0.249</u> | 0.439 | **0.336** | <u>0.332</u> | <u>0.230</u> | <u>0.219</u> |
| | ✓ | **0.247** | **0.436** | <u>0.337</u> | **0.330** | **0.228** | **0.214** |

# Summary

1. Taxonomy of Context Extension

2. Compressive Memory Matrix Design for Time Series

3. Experiments and Benchmarking

# Shameless plug…

RISS | ROBOTICS INSTITUTE SUMMER SCHOLARS

# Thank you