Kinaema: a recurrent sequence model for memory and pose in motion

December 5th, 2025


Mert Bulent Sariyildiz


Philippe Weinzaepfel


Guillaume Bono

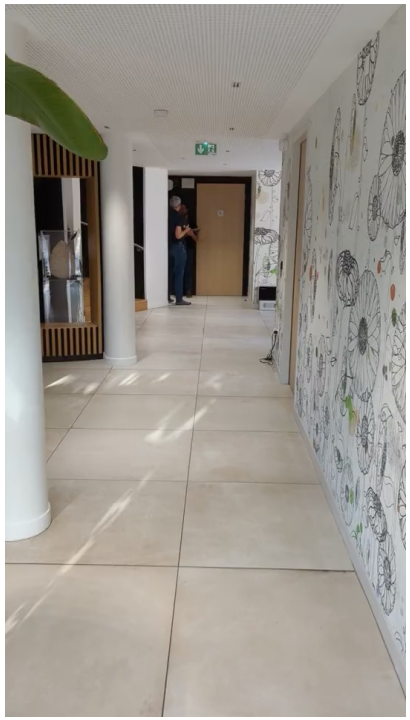

Gianluca Monaci


Christian Wolf

NEURAL INFORMATION PROCESSING SYSTEMS

NAVER LABS Europe

Goal: exploiting continuous operation of robots



Now, bring me to the coffee machine!

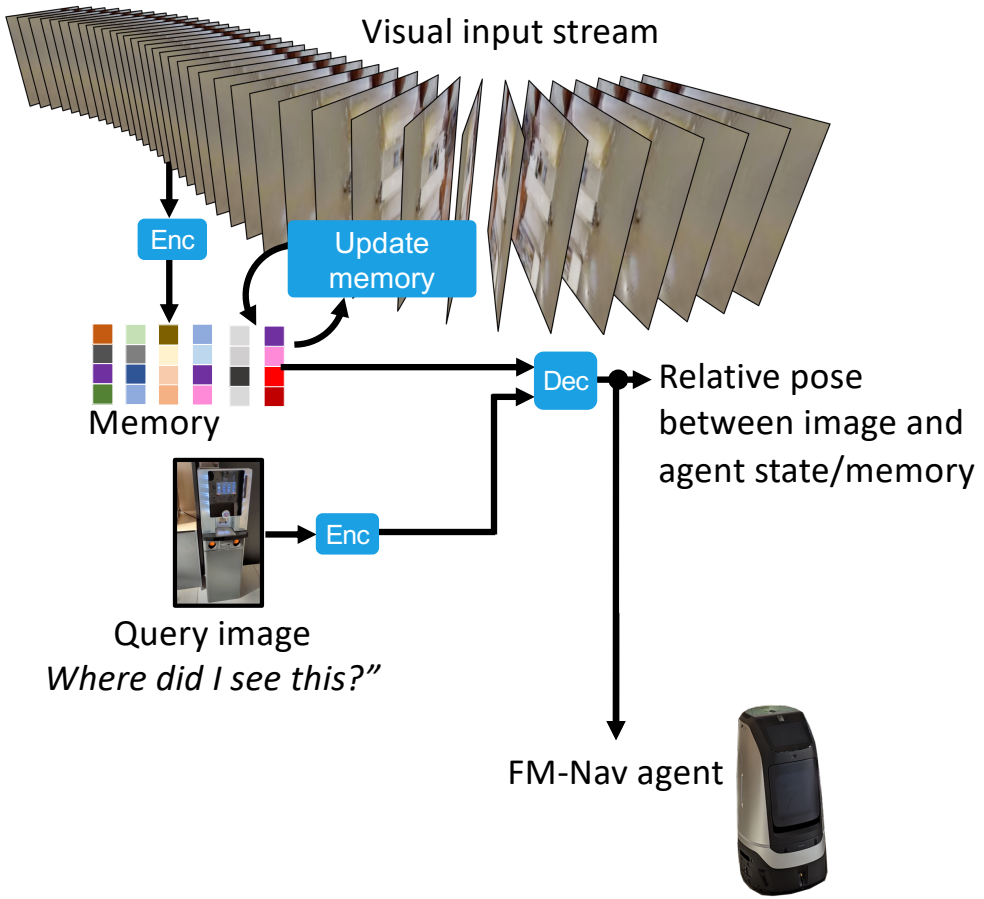
2 new tasks:

Mem-Nav:

Navigate to a place which was potentially observed before the nav episode start.

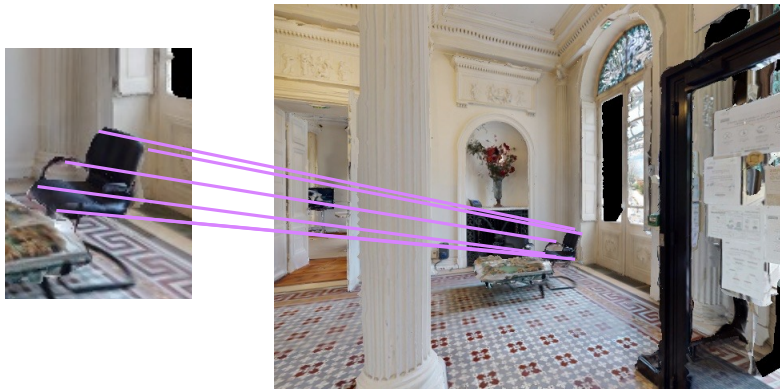
Mem-RPE:

“Situate” a place which has been potentially observed before the nav episode start.



How do we (humans or models) estimate relative poses?

Local feature matching



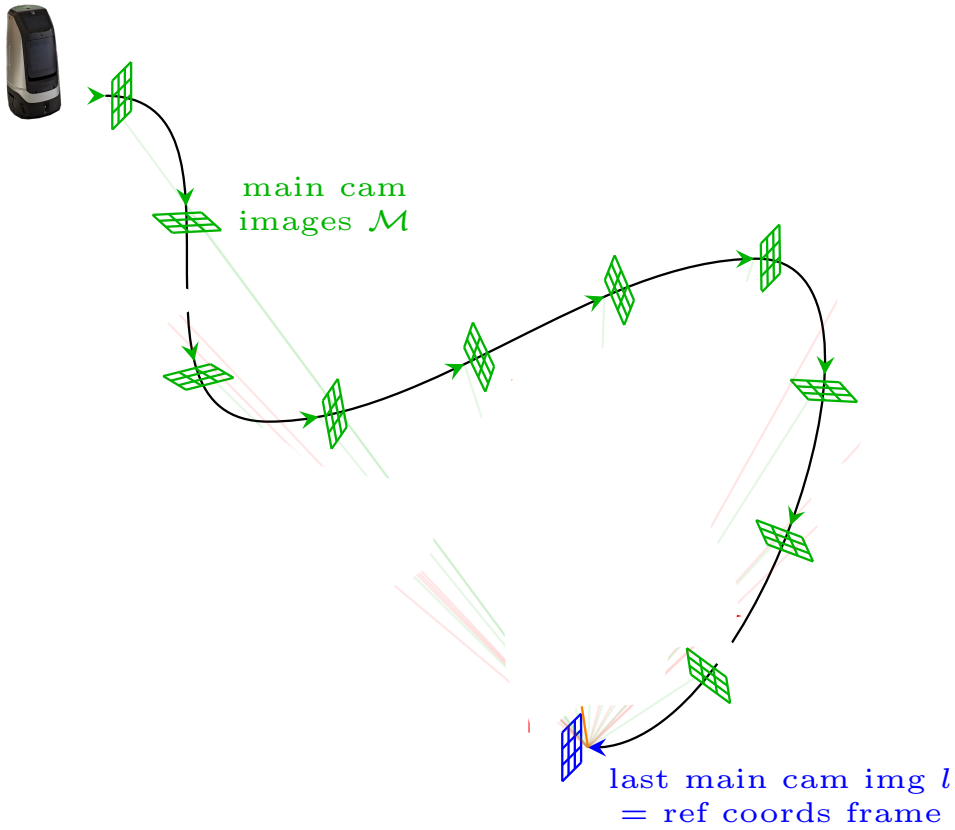
Regularities of “impossible matching” (DUSt3R, MASt3R)

Wang et al, DUSt3R: Geometric 3D Vision Made Easy, CVPR 2024.

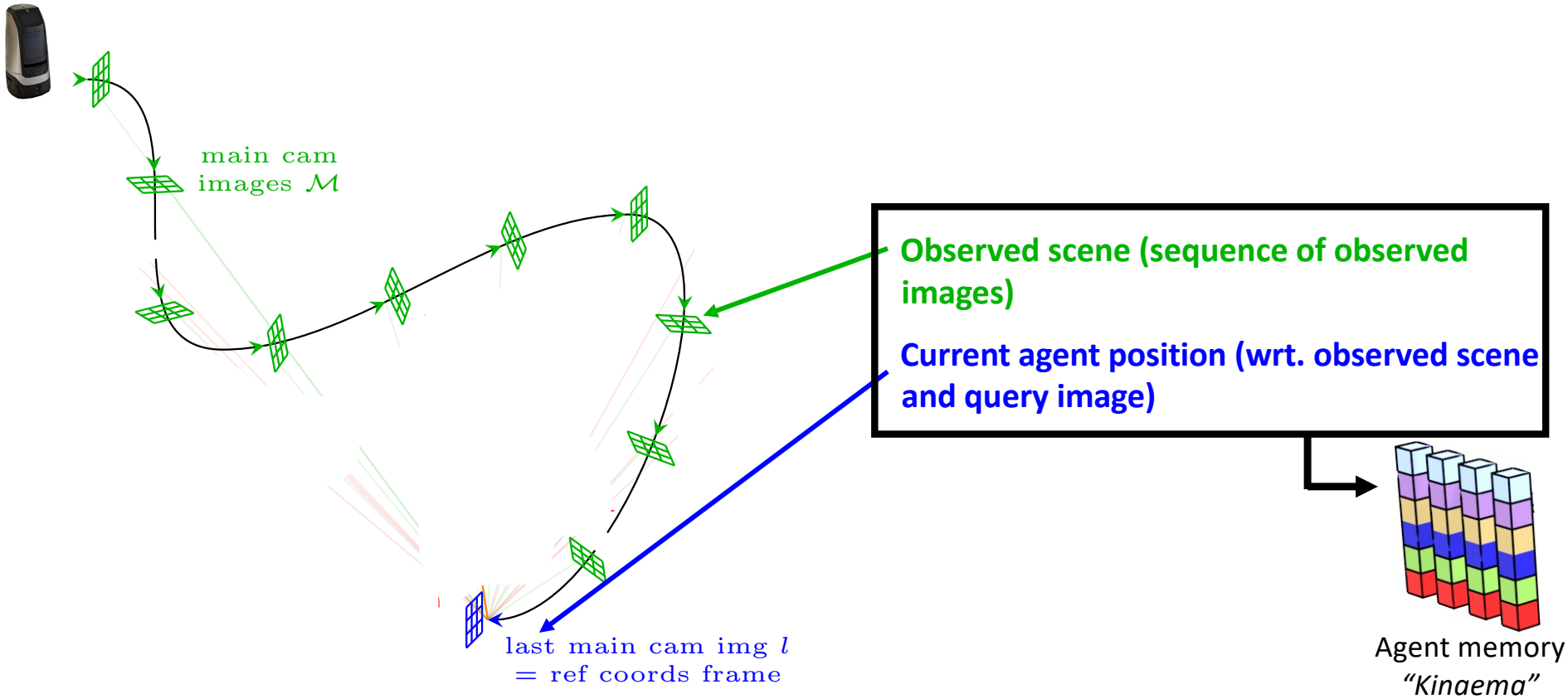
Leroy et al., Grounding Image Matching in 3D with MASt3R, ECCV 2024.

+ agent motion

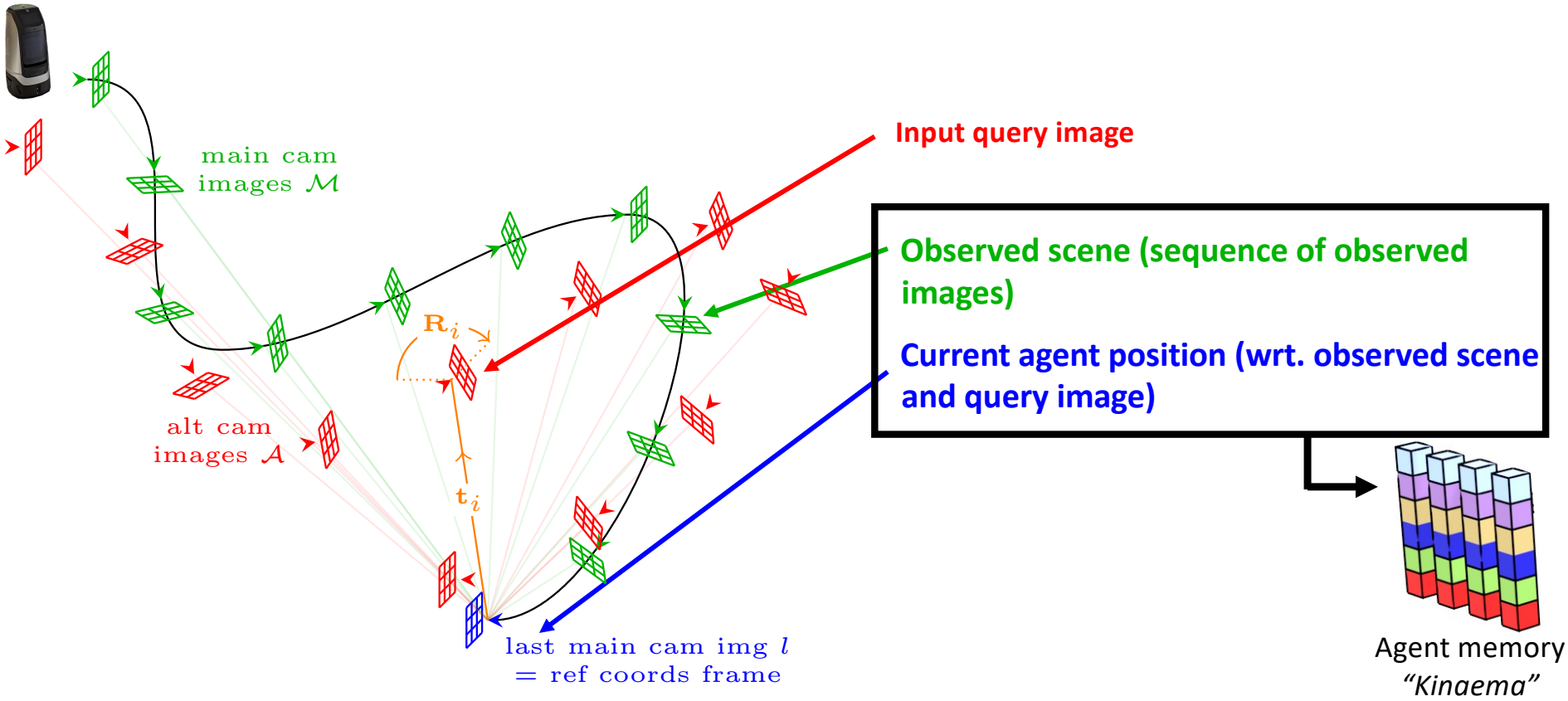
Mem-RPE: exact task definition



Mem-RPE: exact task definition

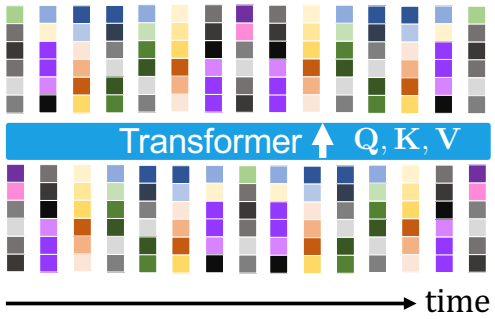


What is involved in Mem-RPE?



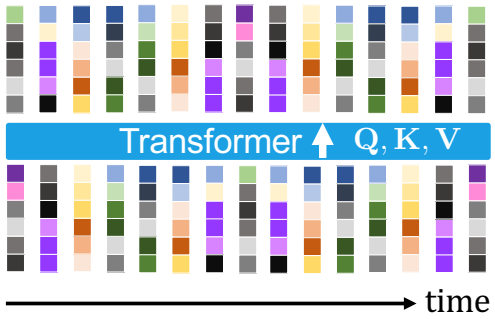
Types of memory

- **Transformer on obs history**
  - Used for LLMs, nav-world models (eg. GAIA-1,2)
  - Needs truncation (limit context length)
  - High complexity: $O(N^2)$ for each step



Types of memory

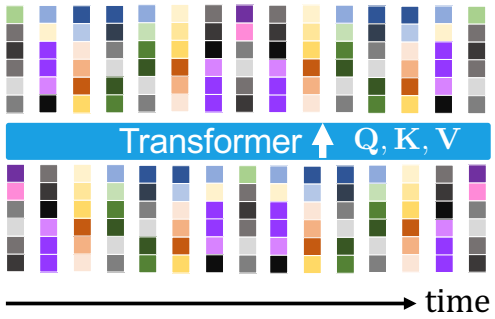
- **Transformer on obs history**
  - Used for LLMs, nav-world models (eg. GAIA)
  - Needs truncation (limit context length)
  - High complexity: $O(N^2)$ for each step



- **Recurrent/classical** (GRU, LSTM)
  - O(1) for each step
  - Does not scale (not enough mem capacity)

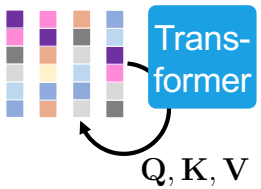
Types of memory

- **Transformer on obs history**
  - Used for LLMs, nav-world models (eg. GAIA)
  - Needs truncation (limit context length)
  - High complexity: $O(N^2)$ for each step

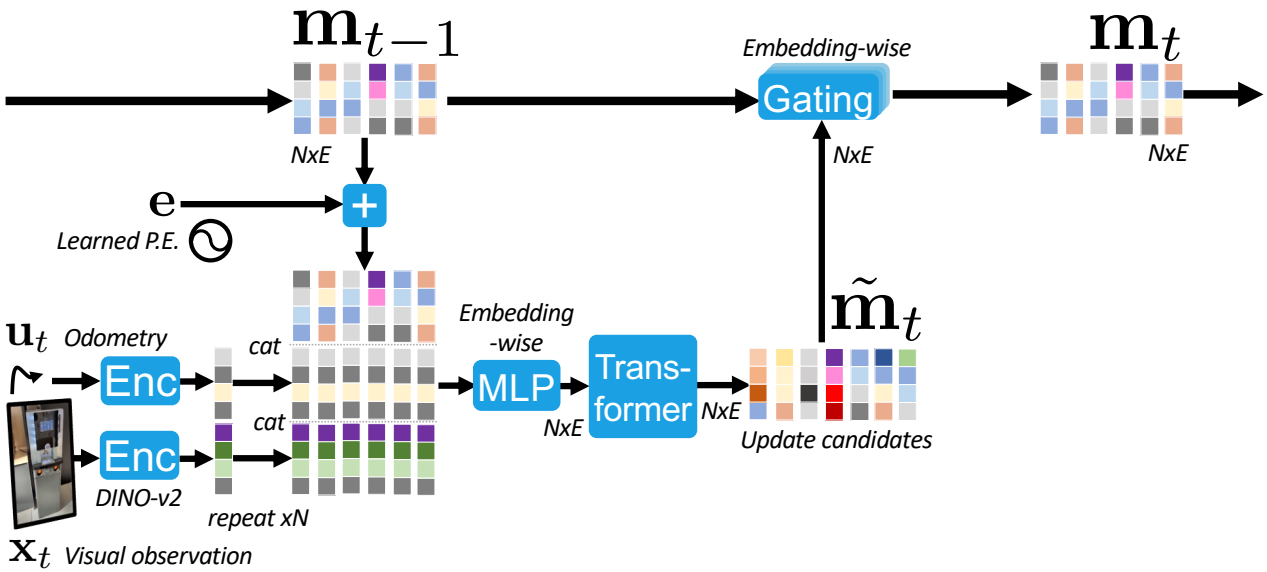


- **Recurrent/classical** (GRU, LSTM)
  - O(1) for each step
  - Does not scale (not enough mem capacity)

- **Recurrent transformer** **(Ours, "Kinaema")**
  - O(1) for each step
  - High mem capacity



The Kinaema model



Trade-offs in recurrent memory

Less (no?) latent dynamics
More inductive bias

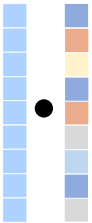
Highly expressive
latent dynamics

EMA
(mem decay only)

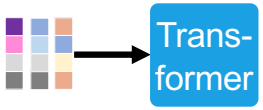
Kinaema

RNN/GRU

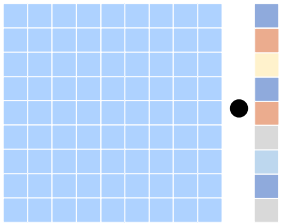
$$\mathbf{m}_{t+1} = \lambda \mathbf{m}_t + \mathbf{V}\mathbf{x}_t$$



$$\mathbf{m}_{t+1} = SelfAttn(\mathbf{m}_t \cup \mathbf{x}_t)$$



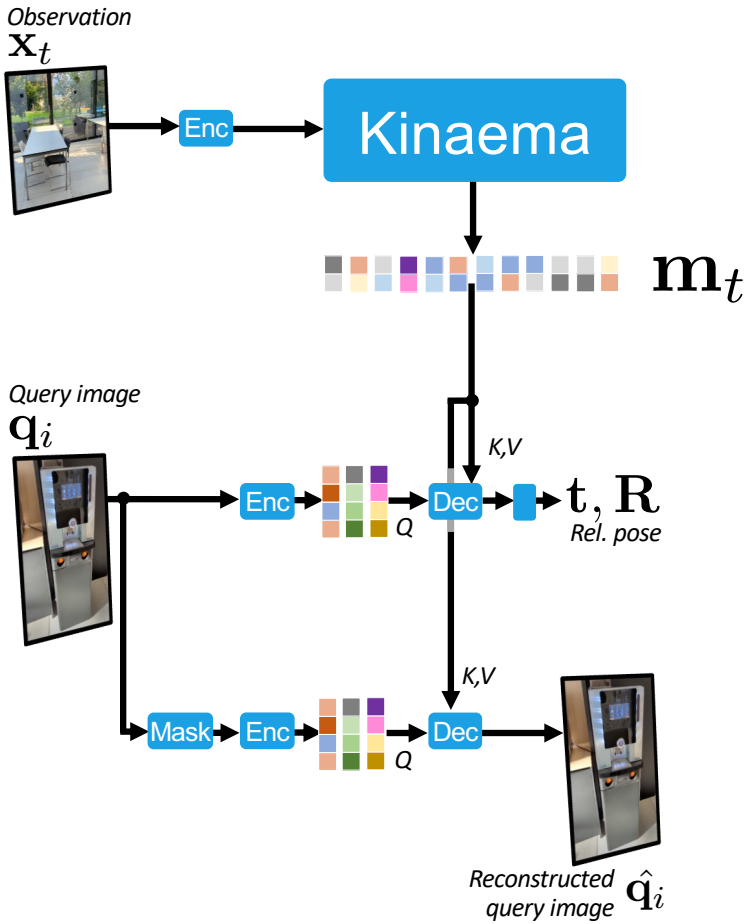
$$\mathbf{m}_{t+1} = \mathbf{W}\mathbf{m}_t + \mathbf{V}\mathbf{x}_t$$



Eberhard et al., Partially observable reinforcement learning with memory traces, ICML 2025

S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computing, 1997.

Training on sequences



1. Mem-RPE loss

2. Masked image modelling

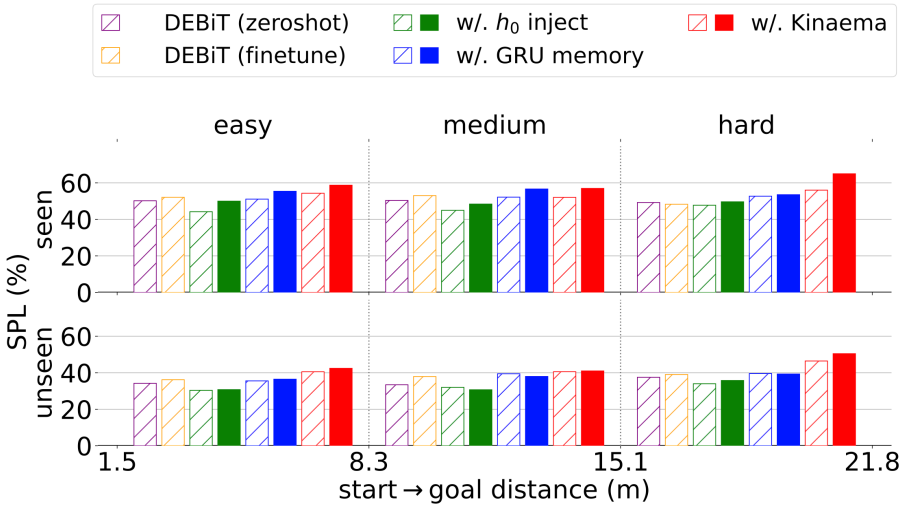
(task requires memory usage!)

Experimental Results

Mem-RPE
Pose estimation accuracy

Model	Mem size	Obs hist	Seq len 200			Seq len 800		
			$1m$ 10^o	$1m$ 90^o	$2m$ 90^o	$1m$ 10^o	$1m$ 90^o	$2m$ 90^o
Trunc.Hist.	41.6k	✓	2	11	28	1	6	16
MooG [58]	524.3k	✗	0	5	14	0	3	9
LRU [45]	3.1k	✗	4	18	34	2	9	20
EMA [21]	153.6k	✗	6	18	34	3	11	24
xLSTM [5]	2,359.3k	✗	8	23	47	5	13	29
GRU [16]	3.1k	✗	12	32	56	4	14	31
Kinaema	61.4k	✗	21	41	63	10	21	37

Mem-Nav
Navigation efficancy (SPL)



Conclusion

- A new transformer-based recurrent sequence model
- Trained to predict poses of query images wrt. previously observed content
- Memory is latent and of constant size
- Outperforms other sequence models on memory based pose estimation and navigation

Mert Bulent Sariyildiz
Philippe Weinzaepfel
Guillaume Bono
Gianluca Monaci
Christian Wolf

NAVER LABS Europe