Recent developments in embodied AI

Neurips 2025

Roland Memisevic

Sr. Director, Qualcomm AI Research(*)

In collaboration with Apratim Bhattacharyya, Daniel Dijkman, Reza Ebrahimi, Sanjay Haresh, Litian Liu, Pulkit Madan, Sunny Panchal, Reza Pourreza, ...

memisevr@qti.qualcomm.com

(*) Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc

Agenda

1. What is embodied AI?
2. Training data for end-to-end learning
3. Models for end-to-end learning
4. State tracking as unsolved open problem
5. Discussion

Embodied AI: Training AI models to understand the real world

Visual interaction

Language grounded in real-world understanding

Real-world robotics

Embodied AI = AI with *world model*

"World model" means many things:

- Knowledge about the physical world (objectness, gravity, cause-and-effect)
- Situated reasoning (understand the meaning of "Pick up *that* one")
- Sense of time ("Why is it taking so long?", "Should I check in on them?")
- Agency (planning, concept of self, "Can I do this", "How long will this take me?")
- Social understanding ("When should I respond?", "Are they confused? Interested? Distracted?")

A better term than "world model" may be "human-like common sense"

Embodiment as vital aspect of intelligence

- In the quest to understand intelligence, embodiment may be more than a nice-to-have application
- It may be the foundational core, and understanding it may be prerequisite to truly understand intelligence
- This is a long-held view in **cognitive metaphor** and related areas (Lakoff, Johnson, Hofstadter, Rosch, …)
- Language is full of embodied metaphors:

"did they truly *grasp* the idea?"

"transformers are the *foundation* of current AI"

"don't *waste so much* time on these details"

Embodiment as vital aspect of intelligence

- From the perspective of cognitive metaphor, low-level perceptual processing and high-level thinking rely on one another in a constant feed-back loop

- Much of AI research was concerned with the first in the previous decade and with the latter in the current decade

- Embodied AI is concerned with **combining them**

2010 → 2012 → 2014 • • •

SPEECH TO TEXT

Audio
↓
~~Pipeline~~
Neural network
↓
Text

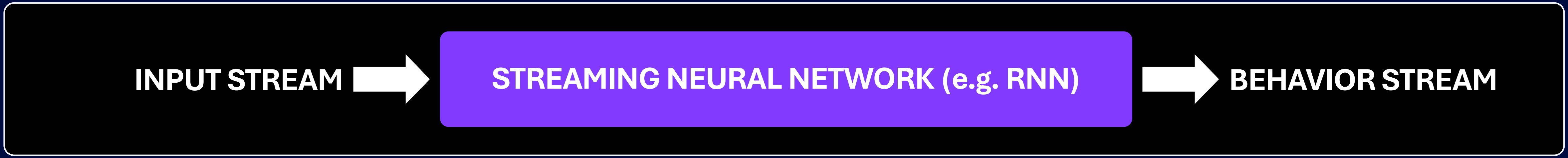
OBJECT RECOGNITION

Pixels
↓
~~Pipeline~~
Neural network
↓
Objects

LANGUAGE TRANSLATION

YES English
↓
~~Pipeline~~
Neural network
↓
YES OUI French

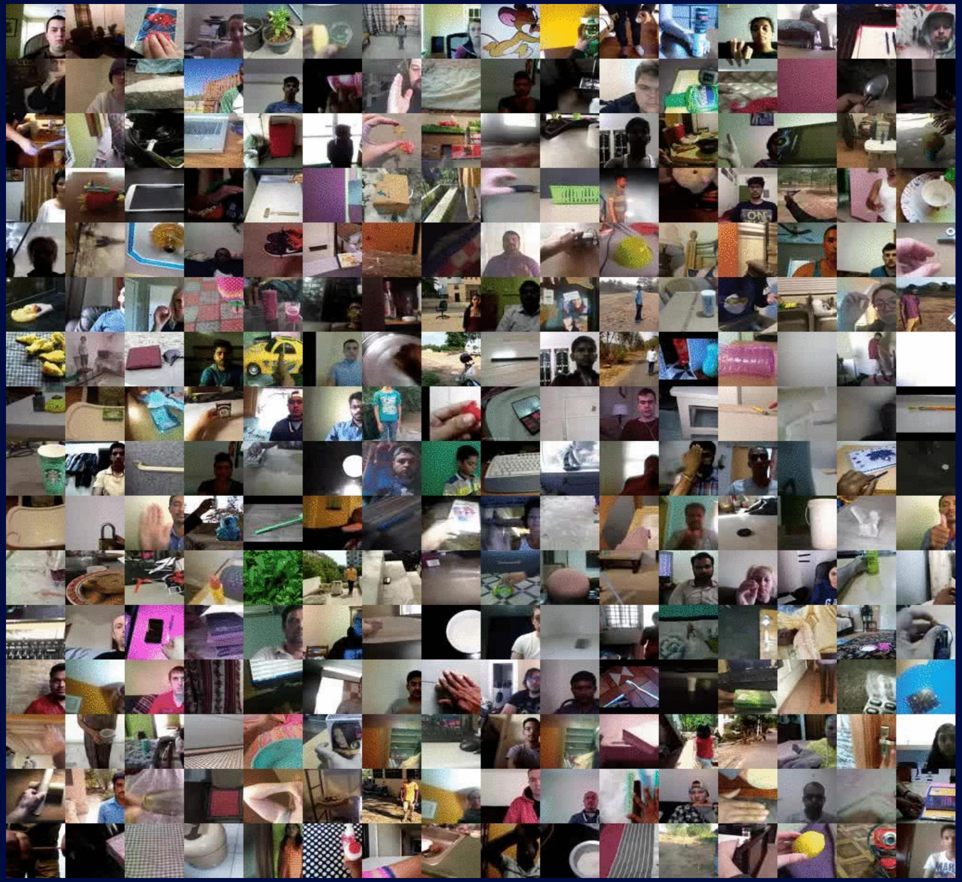
End-to-end trained neural networks have been replacing modular, computational pipelines for decades



“Taking end-to-end learning to the end”:

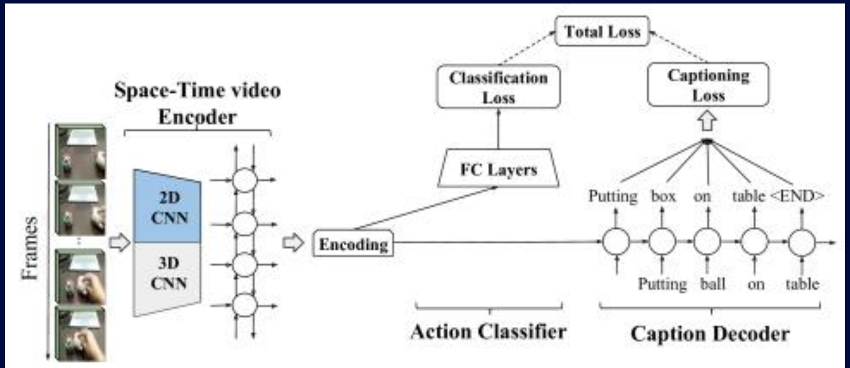
Can aspects of world models and common sense emerge in response to end-to-end training?

Can physical common sense emerge from video-language pre-training?



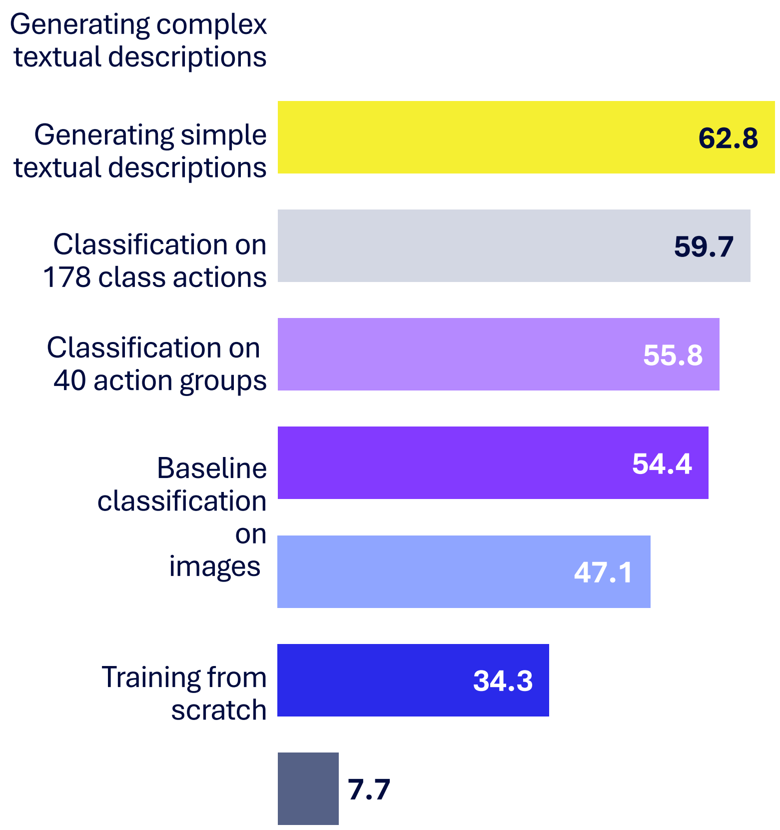
“The Something-something video database for learning and evaluating visual ‘common sense’”, Goyal et al. 2017

Vision-language models pre-trained on video captioning leads to the emergence of general-purpose features

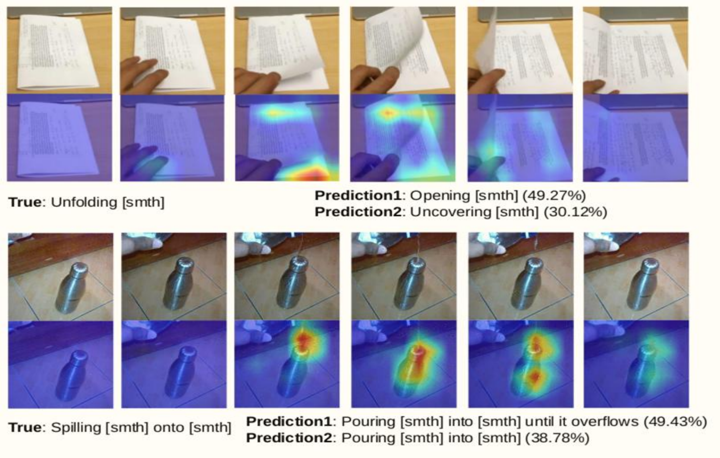


Pre-training a vision-language model on a difficult captioning task (Something-something by Goyal et al. 2017)...

...allows us to improve prediction accuracy on a separate home cooking Task:



"On the effectiveness of task granularity for transfer learning" (Mahdisoltani, et al. 2018)



Grad-cam visualizations show the emergence of highly sensible attention policies

Training data for learning about the real world end-to-end

Fitness coaching as a test-bed for situated interactions

- Fitness coaching (of all things) may be the ideal test-bed for exploring physical common sense
    - It is a streaming scenario (“pixels in -> behaviors out”)
    - Models need to know what to say and when to say it
    - Models need to understand real physical stuff
    - Models need to show fundamental social competencies in the presence of humans
    - However: interactions have strict guard-rails: the interaction is like playing a “real-world game”
    - It is a true real-world use-case with true value

FIT-Coach

benchmark and dataset

A novel interactive visual coaching benchmark and dataset as a test-bed for real-time, real-world situated interaction

Fitness questions dataset

148
exercises

300k
short-clip videos

470+
hours

1900
unique participants

1.1M+
high-level question-answer pairs

400k+
fine-grained question- answer pairs

Fitness feedback dataset

9+
hours of fitness coaching session

148
exercise sessions

~3.5
minutes long sessions with 5 to 6 exercises

21
unique participants

Learning situated live interactions based on fitness coaching

Live Fitness Coaching as a Testbed for Situated Interaction, Panchal, Bhattacharyya, et al. Neurisp 2024

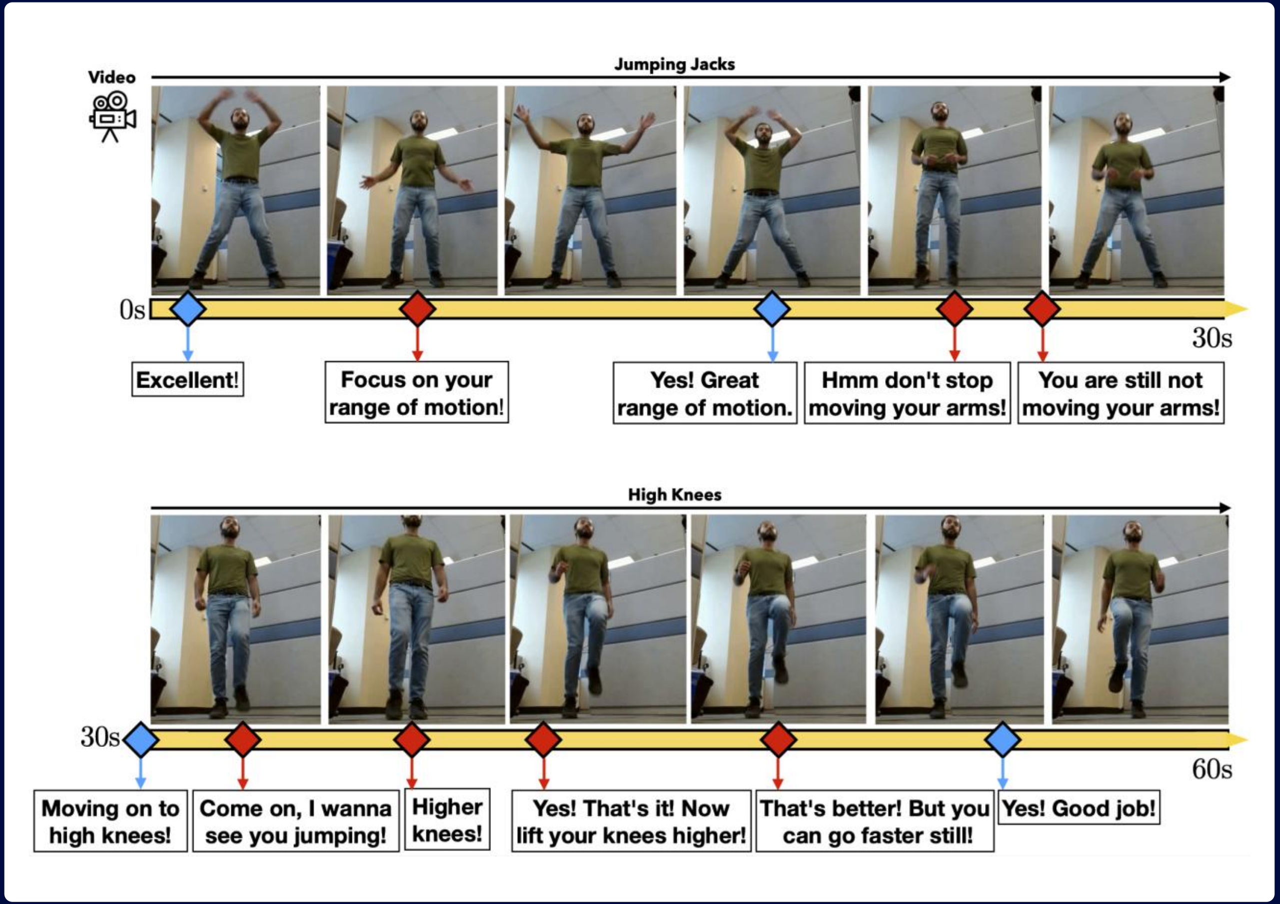
Fitness assistant dataset and benchmark

Short video clips showing the user performing individual exercises, along with labels for performance and common mistakes (~300k clips of duration ~5-10 seconds each)

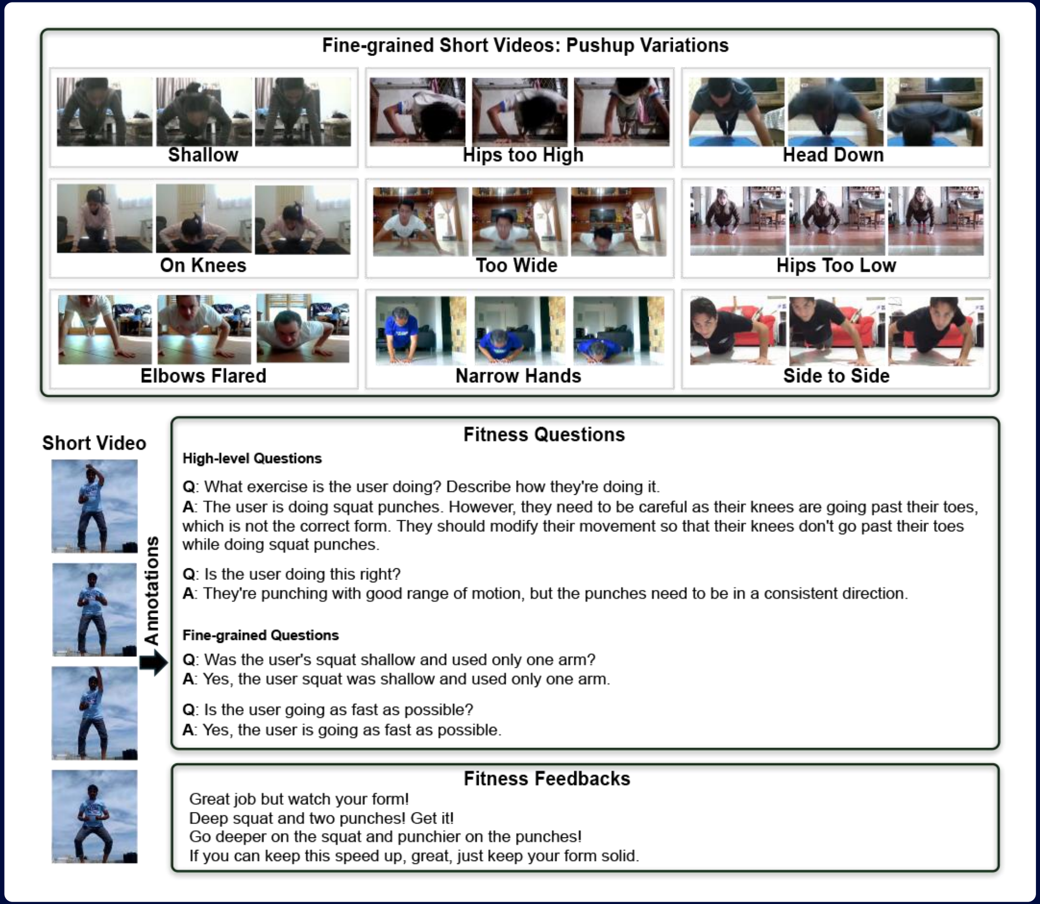
Long-range videos showing the user exercising, along with aligned comments by the coach (~200 sessions across 5-6 exercises each)

	SHORT CLIPS		LONG-RANGE	
	Train	Test	Train	Test†
Number of videos	290,775	16,429	153	69
Unique Participants	1,800+	100	21	7
Average Duration (s)	5.6 ± 1.1	5.6 ± 1.2	213.4 ± 3.1	213.7 ± 3.3
Exercises per Video	1	1	5-6	5-6
Total Number of Exercises	148	148	23	23
Total Classes	1866	1690	—	—
Fitness Questions				
Total High-level Questions	1,193,056	78,390	—	—
Total Fine-grained Questions	404,082	80,694	—	—
Fitness Feedbacks				
Average Feedbacks per Exercise	2.0 ± 10.1	2.4 ± 6.9	5.0 ± 1.3	5.0 ± 1.2
Average Silence Period (s) ††	n/a	n/a	5.2 ± 1.4	5.3 ± 1.2
Average Feedback Length (words)	9.0 ± 6.1	9.1 ± 5.0	6.3 ± 3.8	6.6 ± 4.0

Fitness assistant dataset and benchmark



Long fitness sessions dataset



Short fitness clips dataset

Comparison to other video datasets

DATASET	DOMAIN	HUMAN ACTIONS	INTERACTIVE	MISTAKES	CORRECTIVE FEEDBACKS	DOMAIN EXPERTISE	LENGTH
Action Recognition Datasets							
NTU RGB+D	Fitness	✓	X	X	X	✓	—
FineGym	Fitness	✓	X	X	X	✓	708
Procedural Activity Datasets							
YouCook2	Cooking	X	X	X	X	X	176
Epic-Kitchens	Cooking	X	X	X	X	X	100
HowTo100M	Daily-life	✓	X	X	X	X	134k
Ego-4D	Daily-life	X	X	X	X	X	3670
Ego-Exo4D	Daily-life	X	X	✓	X	X	1422
Assembly-101	Toy assm.	X	X	✓	X	X	513
Interactive AI Assistant Datasets							
WTAG	Cooking	X	X	✓	✓	X	10
HoloAssist	Obj. manip.	X	X	✓	✓	X	166
QEVD (Ours)	Fitness	✓	✓	✓	✓	✓	474

Datasets for end-to-end training of live visual assistants

Key requirement for end-to-end training:
aligned video feed (frames) + assistant’s comments (tokens)

“HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World”
Wang et al. 2024

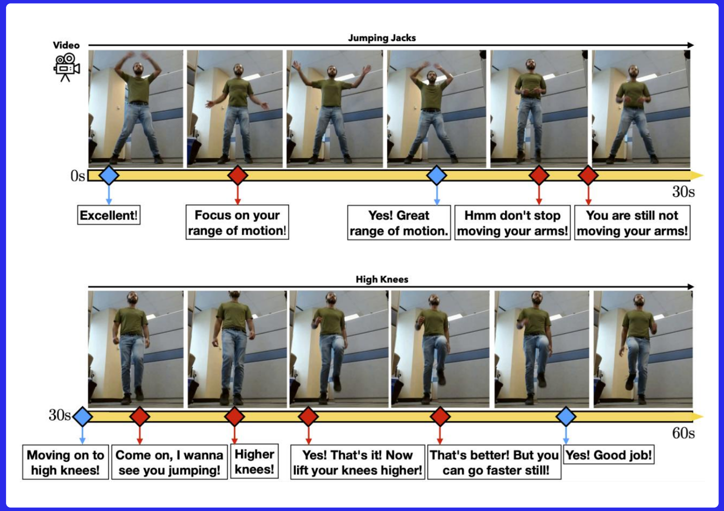
1st person videos showing a variety of tasks (20 tasks across 16 objects)

“Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?”
Bao et al. 2023

1st person videos showing preparation of cupcakes

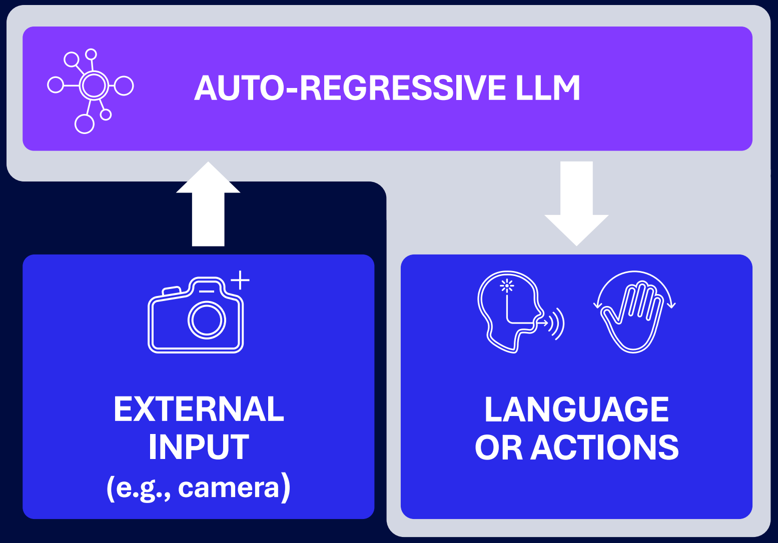
“Live Fitness Coaching as a Testbed for Situated Interactions”
Panchal et al. 2024

3rd person videos showing fitness exercises and their corrections



Real-time (audio-)visual streaming models

End-to-end learning with a multi-modal *streaming* architecture



- Vision-language models combine image features with language, based on various adapter mechanisms, e.g.:
  - Cross-attention (e.g., Flamingo)
  - Dedicated vision tokens (e.g., Llava)

They are very effective in applications like image captioning and visual question answering.

However, ...

... a *streaming* model, that continuously responds to a real-time camera feed, comes with extra challenges:

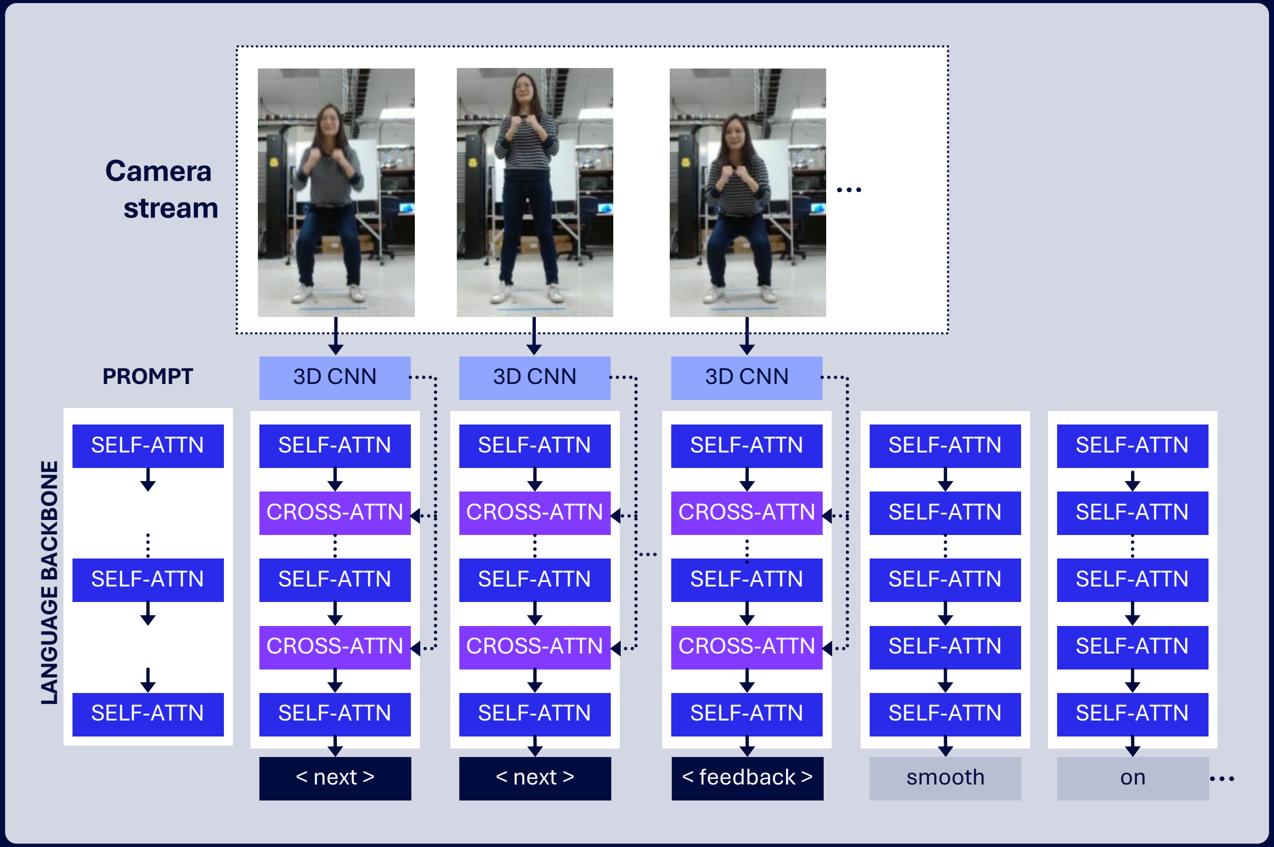
- **Need to freely interleave vision frames and language or action tokens**
- **Need to run in real-time: pay careful attention to vision frame-rate vs. token rate / efficiency / etc.**
- **Need for training data, e.g. allowing a model to learn what to do or say, and when**

- Related recent work: "*VideoLLM-online: Online Video Large Language Model for Streaming Video* ", Chen et al., 2024

Flamingo: a Visual Language Model for Few-Shot Learning", Alayrac et al. 2022

"Visual Instruction Tuning", Liu et al. 2023

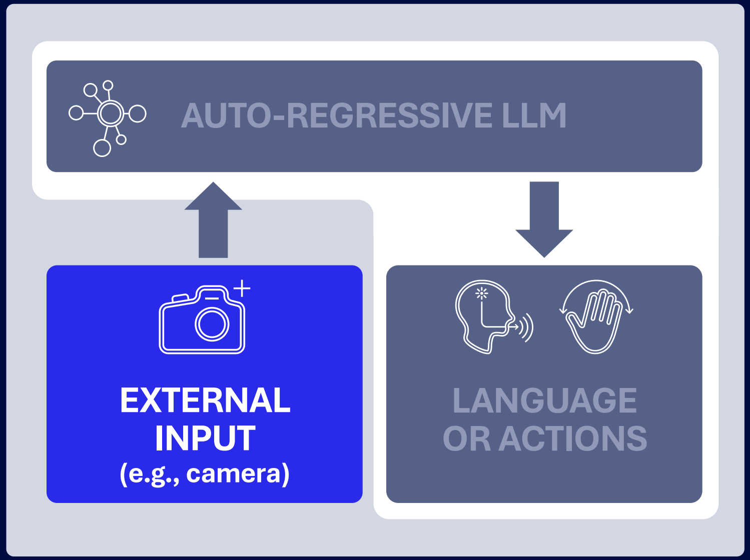
A vision-language model that can learn what to say and when to say it



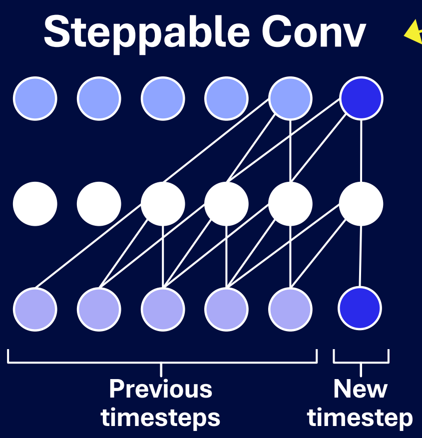
End-to-end training on data + probes that support grounding of key concepts (user behaviors, typical, mistakes, counting, etc.)

3D CNN details

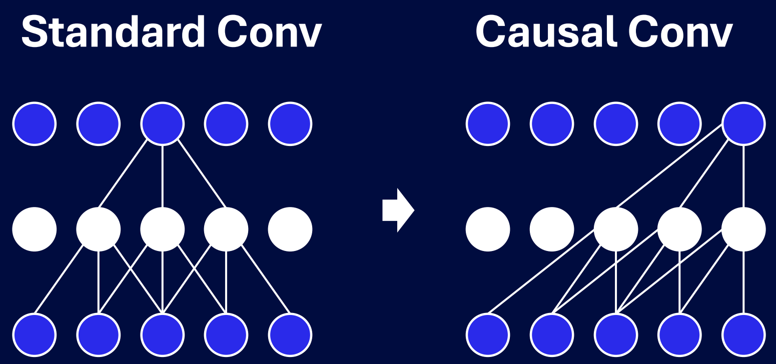
- Existing vision language models use a 2d CNN or vision transformer to represent visual input
- This makes them unsuitable for tasks that require an understanding of motions and human behaviors
- We use a 3d CNN as the feature extractor, which are well-suited to end-to-end learning (e.g. “Is end-to-end learning enough for fitness activity recognition?”, Mercier et al. 2023)



Efficient visual streaming at inference time can be enabled using steppable, causal convolutions:



github.com/quic/sense



“Is end-to-end learning enough for fitness activity recognition?”, Mercier et al. 2023

Qualitative result: end-to-end trained visual dialogue

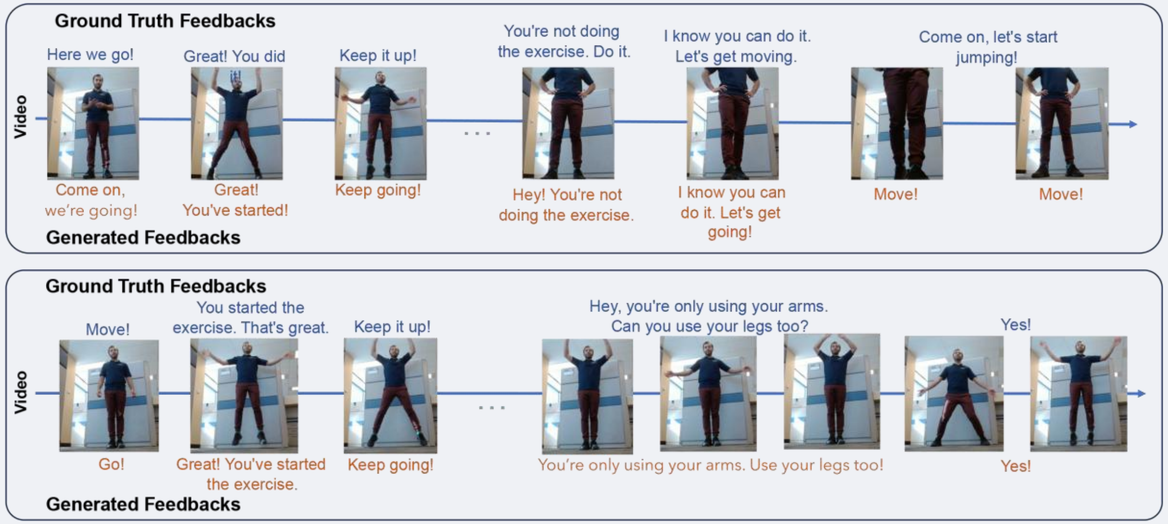
Qualitative results: end-to-end trained visual dialogue

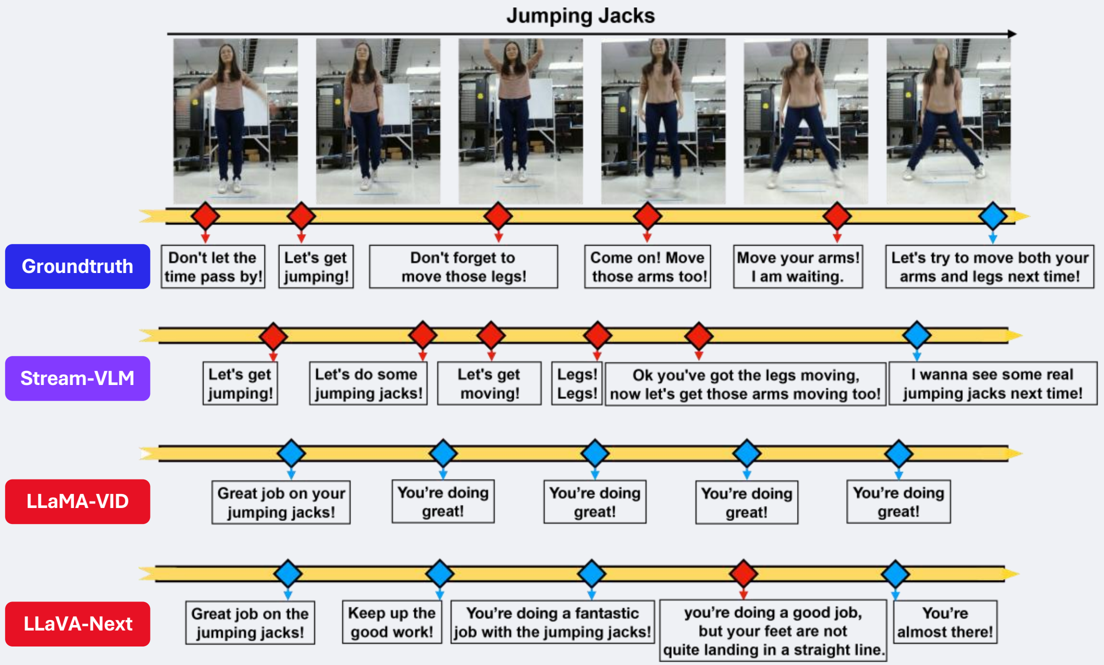
Question: Provide an appropriate feedback for the user

Video-LLaMA: We see a young man standing in a kitchen, wearing a red shirt and white shorts. ✗

Video-ChatGPT: The user has successfully demonstrated the ability to perform a balancing act on a pair of stools. ✗

Coach-LLaMA: This is awesome. Let's keep the intensity high! ✓





Quantitative results: end-to-end learning enables video LLMs to deliver accurate live feedback

Zero-shot prompting results:

METHOD	METEOR↑	ROUGE-L↑	BERT↑	LLM-Acc.↑
InstructBLIP	0.047	0.040	0.839	1.64
Video-LLaVA	0.057	0.025	0.847	1.82
Video-ChatGPT	0.098	0.078	0.850	2.27
Video-LLaMA	0.101	0.077	0.859	2.28
LLaMA-VID	0.100	**0.079**	**0.859**	2.33
LLaVA-Next	**0.104**	0.078	0.858	**2.39**

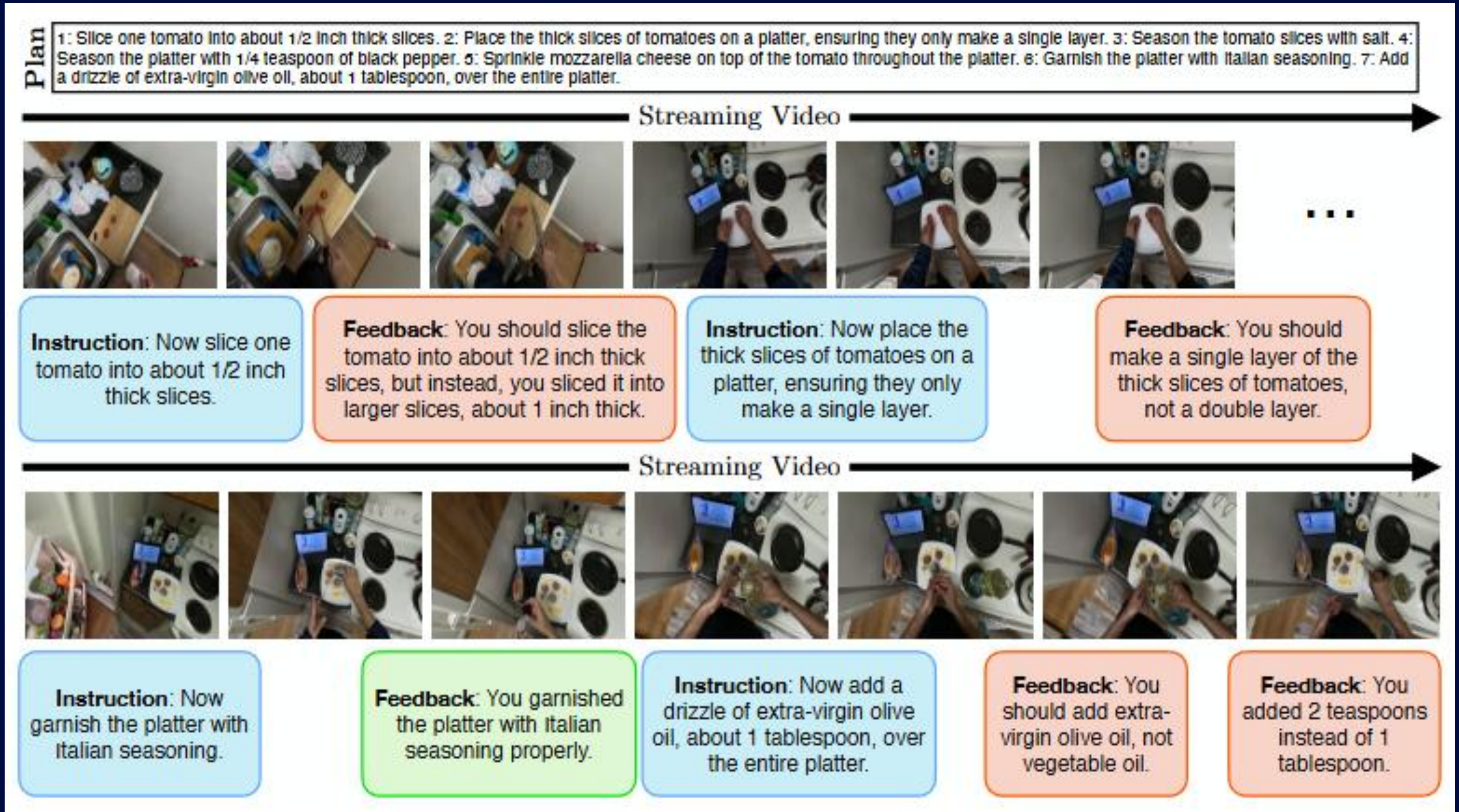
Fine-tuning results:

METHOD	METEOR↑	ROUGE-L↑	BERT↑	LLM-Acc.↑	T-F-Score↑
Socratic-Llama-2-7B	0.094	0.071	0.860	2.39	0.50†
Video-ChatGPT *	0.108	0.093	**0.863**	2.42	0.50†
LLaMA-VID *	0.106	0.090	0.860	2.40	0.50†
STREAM-VLM	**0.125**	**0.116**	**0.863**	**2.56**	**0.59**
STREAM-VLM (w/o 3D CNN)	0.090	0.083	0.857	2.17	0.51
STREAM-VLM (w/o Action-Tokens	**0.125**	0.110	0.861	**2.56**	0.50†

Related task: Responding to questions in the real world

"Can Vision-Language Models Answer Face to Face Questions in the Real-World?", Pourreza et al. 2025

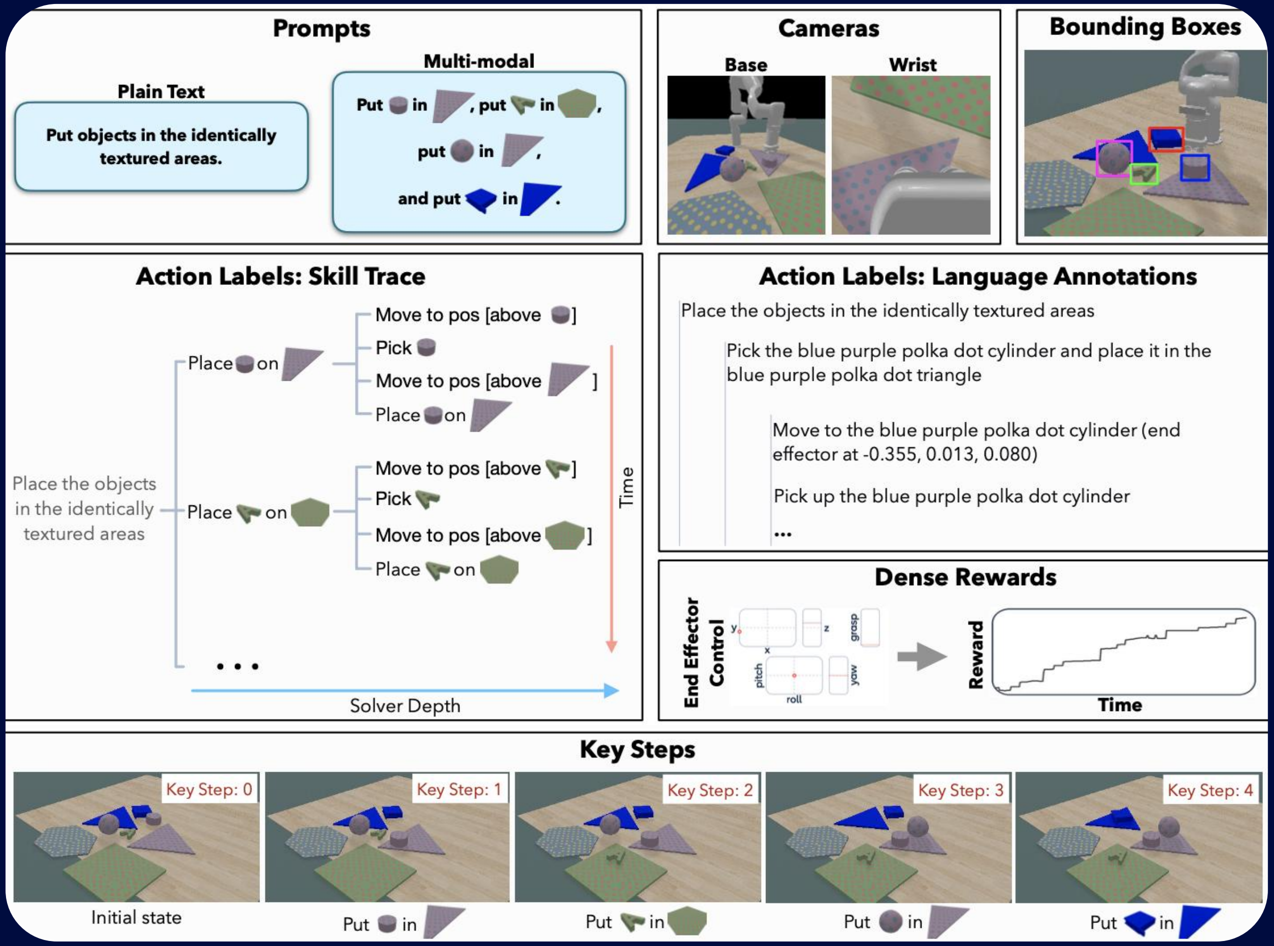
Related task: Cooking instruction



SAN DIEGO POSTER | Wed, Dec 3, 2025 • 11:00 AM – 2:00 PM PST | Exhibit Hall C,D,E #5403

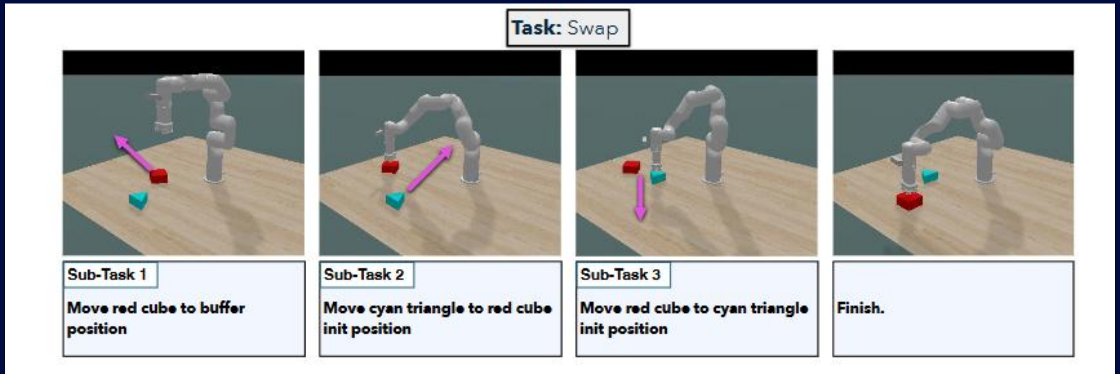
"Can Multi-Modal LLMs Provide Live Step-by-Step Task Guidance?", Bhattacharyya, et al., Neurips 2025

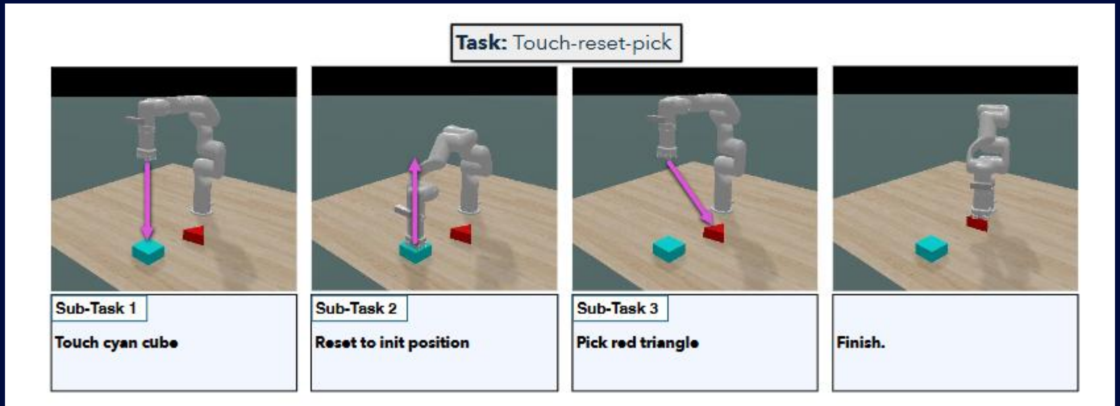
Related task: End-to-end learning for robot control



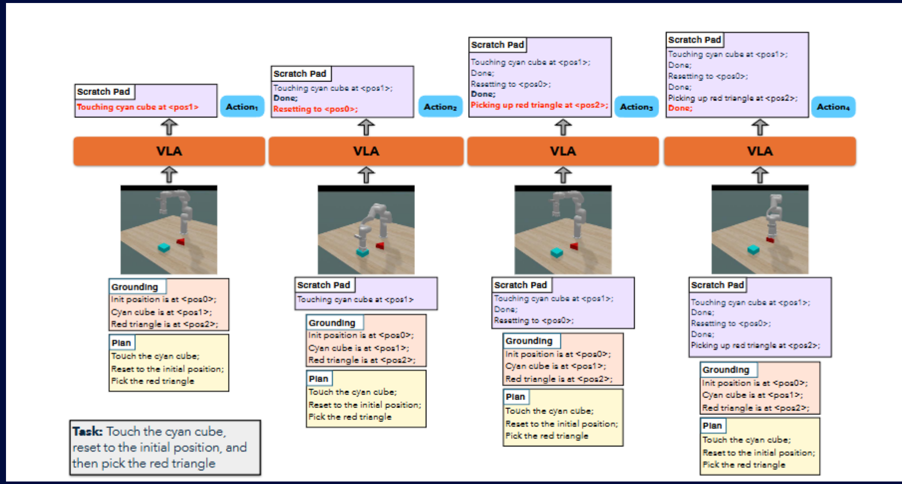
CLEVRSKILLS: “Compositional Language and Visual Reasoning in Robotics” ,Haresh, et al., Neurips 2024

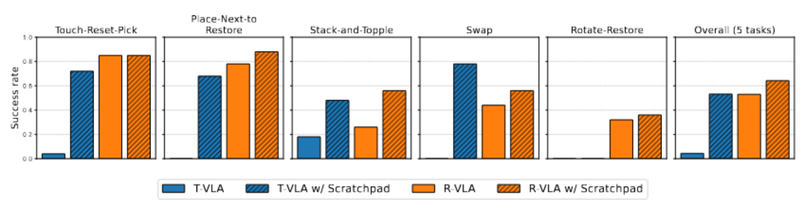
Outlook: Non-Markovian manipulation is hard





Simple memory-dependent tasks





Task-specific scratchpads can improve accuracy:

OpenVLA

OpenVLA w/ scratchpad

Task: "Place tomato in the bowl, then put it back to where it was"

The VLA without memory fails as the start and end of the task are exactly the same. A VLA w/ scratchpad can learn to solve the task.

"Augmenting VLAs with Language Scratchpad for Memory Dependent Manipulation Tasks", Haresh et al. (2025)

State tracking and the importance of a recurrent hidden state

State tracking is transformers' Achilles' heel

- Two simple tasks:

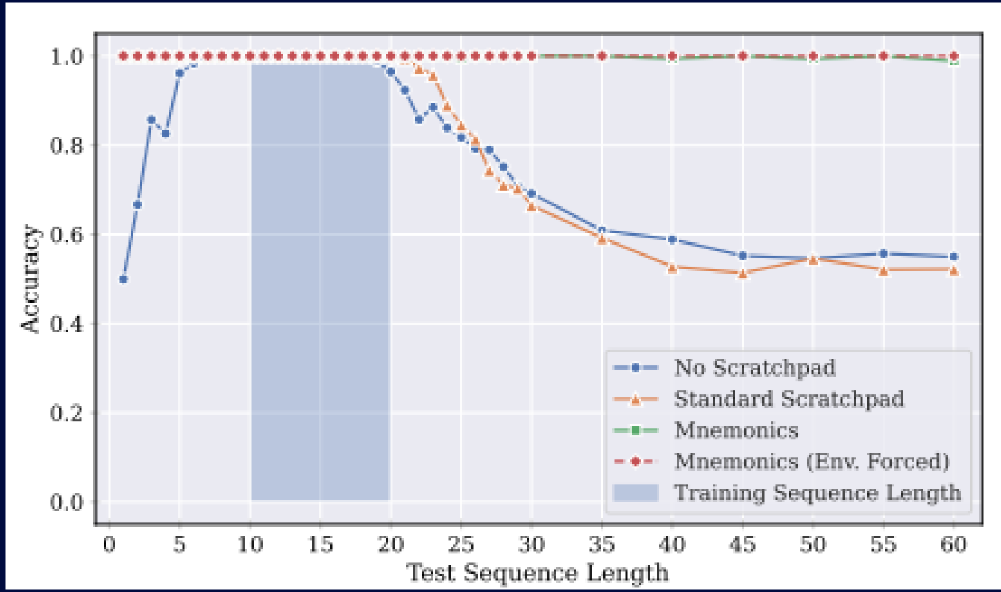
Parity :	1000101 -> 1
Parity with scratchpad/COT :	1000101 -> 1 1 1 1 0 0 1

- Neither are learnable by transformers for arbitrary sequence length!

- Same problem for addition/multiplication and any other algorithmic task

- See, for example:
  - Exploring length generalization in Large Language Models (Anil et al., 2022)
  - Faith And Fate: Limits of Transformers on Compositionality (Dziri et al., 2023)
  - The Illusion of State in State-Space Models (Merrill et al., 2025)

- The problem is non-existent for **non-linear recurrent networks**!



"*Your context is not an array: revealing random access limitations in transformers*" Ebrahimi et al. 2024

Inductive bias: RNNs learn to compress the past to predict the future. Transformers don't.

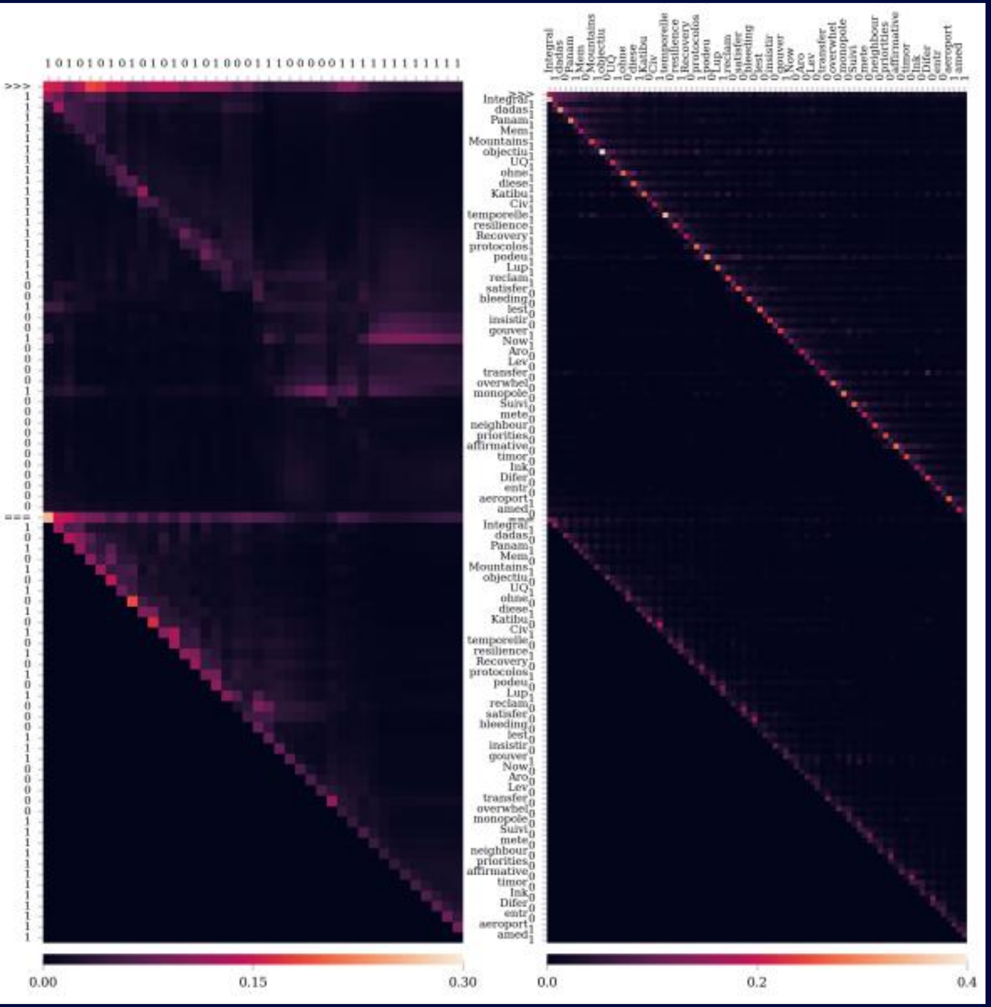
Your context window is not an array

Hypothesis: The key problem is the inability of a transformer to perform "random access" read operations in the context window

Evidence 1:

- Interleaving temporal anchors ("mnemonics") resolves the problem for Parity
- Parity task w/ mnemonics: a 1 b 0 c 0 d 1 -> a 1 b 1 c 1 d 0
- Perfect length generalization!
- Also works for addition / multiplication
- Caveat: This solution is highly *task-specific*

Evidence 2:

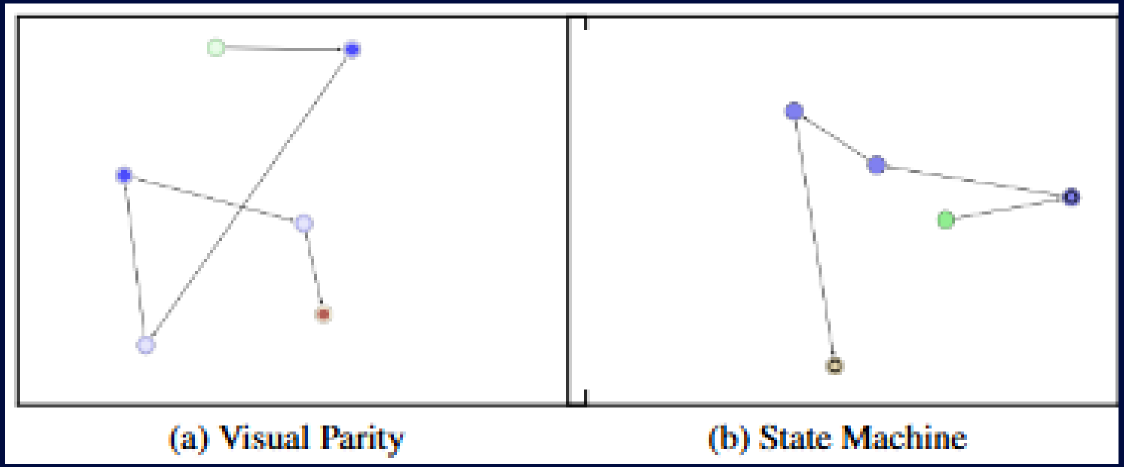


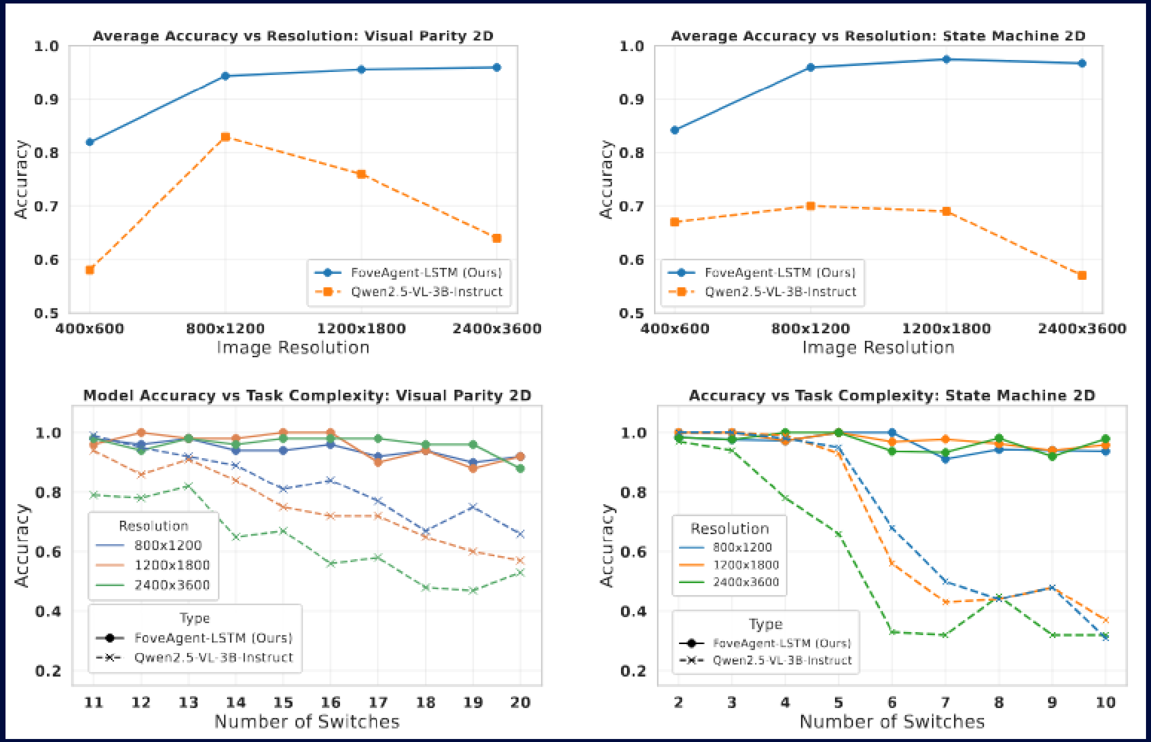
Attention maps with (right) and without (left) mnemonics

"*Your context is not an array: revealing random access limitations in transformers*" Ebrahimi et al. 2024

State tracking in visual reasoning

- Length-generalization is of the same concern in vision as it is in language!
- Vision models based on local, visual attention are much better at generalization out-of-distribution





Generalizing over resolution

Generalizing over task-length

"*On locality and length-generalization in visual reasoning*", Madan et al. 2025

Transformer-RNN hybrids do *not* solve the problem

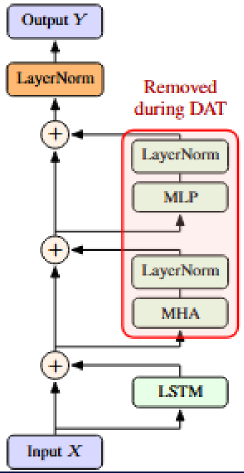
- Transformers generalize in recall tasks, RNNs generalize in state tracking tasks
- Combined task: $\langle \text{bos} \rangle\, k_1 v_1\, k_2 v_2 \cdots k_n v_n\, \langle \text{modulo} \rangle\, m\, \langle \text{recall} \rangle\, k_j v_j$
- In hybrid models, self-attention takes over, preventing the RNN from learning the state tracking part
- Delaying self-attention allows the RNN to learn, but it is highly task-specific

Model	5	10	15	20	25	30	35	40	45	50	75	100
Transformer	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.90	0.80
Mamba	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Memory Networks												
LSTM	1.00	0.98	0.95	0.89	0.83	0.76	0.70	0.64	0.59	0.53	0.37	0.28
LSTM ∘ Attn	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.90	0.76
Attn ∘ LSTM	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.96	0.89
Hybrid Block	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.92
LSTM ∘ Attn ∘ LSTM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.97	0.90
LSTM + Attn	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.92	0.82

Table 1: Results on the Recall-Only setting.

Model	5	10	15	20	25	30	35	40	45	50	75	100
Transformer	1.00	1.00	1.00	0.99	0.98	0.51	0.11	0.10	0.10	0.10	0.10	0.10
LSTM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LSTM ∘ Attn	1.00	1.00	1.00	1.00	1.00	0.20	0.10	0.10	0.10	0.10	0.10	0.10
Attn ∘ LSTM	1.00	1.00	1.00	1.00	1.00	0.63	0.10	0.10	0.10	0.10	0.10	0.10
Hybrid Block	1.00	1.00	1.00	1.00	1.00	0.45	0.22	0.13	0.10	0.10	0.10	0.10
LSTM ∘ Attn ∘ LSTM	1.00	1.00	1.00	1.00	1.00	0.73	0.16	0.08	0.09	0.10	0.10	0.10
LSTM + Attn	1.00	1.00	1.00	1.00	0.99	0.11	0.11	0.10	0.10	0.10	0.10	0.10

Table 2: Results on the Modulo-Only setting.



Model	5	10	15	20	25	30	35	40	45	50	75	100
Transformer	1.00	1.00	0.99	0.99	0.99	0.49	0.10	0.09	0.09	0.09	0.05	0.03
LSTM	1.00	0.73	0.52	0.41	0.35	0.32	0.28	0.25	0.23	0.22	0.17	0.15
LSTM ∘ Attn	1.00	1.00	0.99	0.99	0.99	0.45	0.12	0.10	0.09	0.09	0.05	0.04
(+ DAT)	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.93	0.88
Attn ∘ LSTM	1.00	1.00	0.99	0.99	0.99	0.83	0.36	0.18	0.12	0.09	0.07	0.05
(+ DAT)	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.80	0.56
Hybrid Block	1.00	1.00	1.00	1.00	1.00	0.56	0.21	0.12	0.10	0.10	0.09	0.08
(+ DAT)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.93	0.85
LSTM ∘ Attn ∘ LSTM	1.00	1.00	1.00	1.00	1.00	0.52	0.13	0.05	0.05	0.07	0.06	0.06
(+ DAT)	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.91	0.80
LSTM + Attn	1.00	1.00	1.00	1.00	1.00	0.28	0.20	0.19	0.18	0.16	0.12	0.08
(+ DAT)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.95	0.85

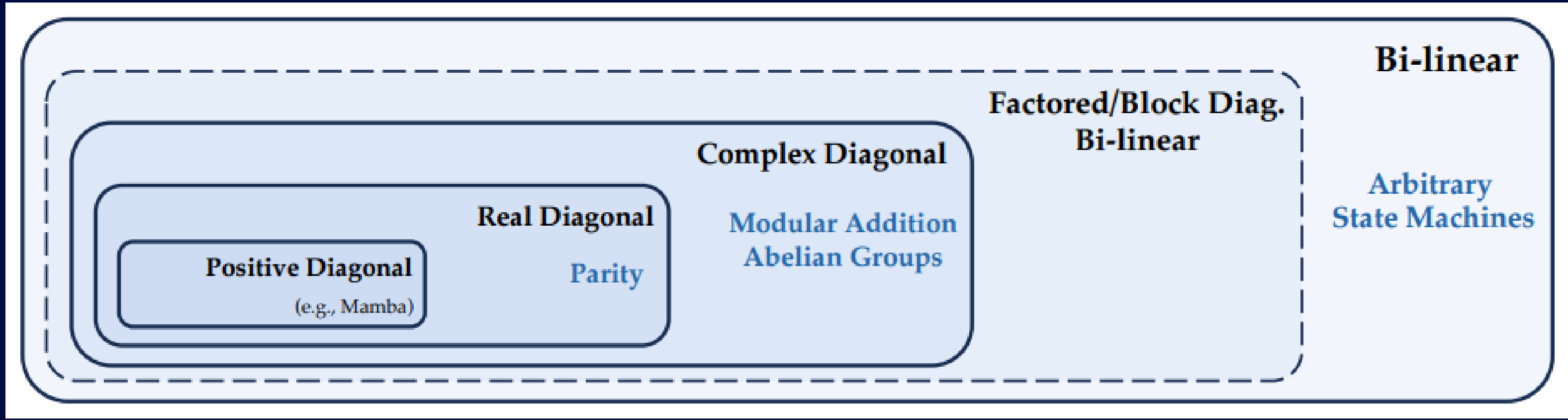
"Delayed Attention Training Improves LengthGeneralization in Transformer–RNN Hybrids", Buu et al. 2025

Bi-linear state transitions are the ideal inductive bias for state tracking

- Vanilla RNN: $h^t = \mathcal{A}h^{t-1} + \mathcal{B}x^t + b$

- Bi-linear RNN: $$h_i^t = (h^{t-1})^\top \mathcal{W}_i\, x^t = \sum_{jk} \mathcal{W}_{ijk} x_k^t h_j^{t-1}$$

Hidden-to-hidden state transitions in an RNN and state tracking tasks it can solve:



Modulus / State Size		Validation Accuracy (Length 2-10) 2	3	5	10	25	50	OOD Accuracy (Length 500) 2	3	5	10	25	50
Modular Addition													
Bilinear		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Factored Bilinear		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95
Block Diag. (block size)	1	1.00	0.96	0.88	0.85	0.45	0.32	1.00	0.00	0.00	0.10	0.00	0.02
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\mathcal{R}_2$ Block Diag.		1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.66	0.37	0.00
LSTM		1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.98	1.00	0.00	0.02
RNN		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.37	0.07
Mamba (layers)	1	0.99	0.92	0.96	0.85	0.74	0.61	0.00	0.01	0.01	0.00	0.00	0.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.02	0.01	0.00	0.00	0.00
	4	1.00	1.00	1.00	1.00	1.00	0.47	0.01	0.01	0.00	0.01	0.00	0.00
Transformer (layers)	1	1.00	1.00	1.00	0.47	0.98	0.19	0.03	0.01	0.01	0.00	0.00	0.00
	2	1.00	1.00	1.00	0.99	0.89	0.00	0.01	0.02	0.00	0.00	0.00	0.00
	4	1.00	1.00	1.00	0.99	0.92	0.02	0.04	0.00	0.00	0.00	0.01	0.00
State Machine													
Bilinear		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Factored Bilinear		1.00	1.00	1.00	1.00	1.00	0.19	1.00	1.00	1.00	1.00	1.00	0.01
Block Diag. (block size)	1	1.00	0.28	0.20	0.14	0.09	0.07	1.00	0.03	0.00	0.00	0.00	0.00
	2	1.00	1.00	0.84	0.49	0.25	0.15	1.00	0.34	0.16	0.06	0.06	0.02
	8	1.00	1.00	1.00	1.00	0.48	0.21	1.00	1.00	1.00	0.41	0.13	0.04
	64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\mathcal{R}_2$ Block Diag.		1.00	0.29	0.19	0.11	0.07	0.02	1.00	0.00	0.00	0.00	0.00	0.01
LSTM		1.00	1.00	1.00	1.00	1.00	0.30	1.00	1.00	1.00	1.00	0.64	0.09
RNN		1.00	1.00	1.00	1.00	0.41	0.18	1.00	1.00	1.00	0.99	0.19	0.07
Mamba (layers)	1	1.00	1.00	0.96	0.55	0.34	0.19	0.00	0.99	0.87	0.31	0.16	0.07
	2	1.00	1.00	1.00	0.79	0.44	0.30	0.00	1.00	0.96	0.42	0.18	0.09
	4	1.00	1.00	1.00	0.99	0.62	0.41	0.03	0.99	0.97	0.47	0.24	0.10
Transformer (layers)	1	1.00	0.94	0.83	0.46	0.27	0.18	0.03	0.01	0.02	0.01	0.00	0.00
	2	1.00	1.00	0.97	0.61	0.39	0.17	0.01	0.01	0.01	0.01	0.00	0.00
	4	1.00	1.00	1.00	0.84	0.49	0.17	0.00	0.02	0.01	0.00	0.00	0.00
Modular Arithmetic													
Bilinear		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Factored Bilinear		1.00	0.34	0.90	0.09	0.03	0.03	1.00	0.24	0.37	0.06	0.04	0.03
Block Diag. (block size)	1	0.60	0.30	0.24	0.27	0.16	0.14	0.19	0.15	0.09	0.11	0.04	0.03
	2	0.99	0.77	0.53	0.52	0.27	0.21	0.37	0.00	0.08	0.12	0.03	0.05
	8	1.00	1.00	1.00	1.00	0.54	0.47	1.00	1.00	0.41	0.24	0.06	0.08
	64	1.00	1.00	1.00	1.00	1.00	0.66	1.00	1.00	1.00	1.00	0.40	0.21
$\mathcal{R}_2$ Block Diag.		0.61	0.34	0.21	0.23	0.03	0.04	0.02	0.00	0.04	0.04	0.02	0.03
LSTM		1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	0.99	0.64
RNN		1.00	1.00	1.00	1.00	0.82	0.22	1.00	1.00	1.00	1.00	0.55	0.16
Mamba (layers)	1	0.99	0.83	0.56	0.59	0.24	0.14	0.74	0.55	0.29	0.32	0.08	0.08
	2	1.00	0.99	0.80	0.93	0.35	0.33	0.85	0.38	0.41	0.29	0.11	0.07
	4	1.00	1.00	0.99	0.99	0.55	0.29	0.92	0.75	0.48	0.51	0.17	0.09
Transformer (layers)	1	0.88	0.63	0.46	0.32	0.10	0.08	0.19	0.02	0.04	0.01	0.03	0.01
	2	1.00	0.97	0.81	0.32	0.11	0.08	0.19	0.15	0.05	0.04	0.02	0.01
	4	1.00	0.99	0.75	0.32	0.07	0.08	0.19	0.12	0.03	0.03	0.02	0.03

“Revisiting bi-linear state transitions in recurrent neural networks”, Ebrahimi et al., Neurips 2025

SAN DIEGO POSTER | Wed, Dec 3, 2025 • 11:00 AM – 2:00 PM PST | Exhibit Hall C,D,E #4210

Discussion

...embodiment as vital aspect of intelligence

- **Cognitive metaphor** (Lakoff, Johnson, Hofstadter, Rosch, ...):
- Language is full of embodied metaphors

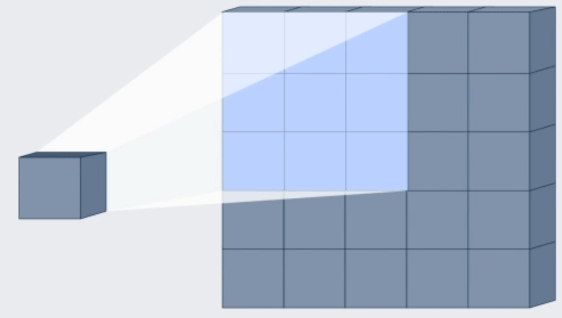
"did they truly *grasp* the idea?"

"transformers are the *foundation* of current AI"

"don't *waste so much* time on these details"

AI = “back-prop + weight sharing”

- The arguably most fundamental driving force in AI, besides back-prop, has been weight sharing
- The use of high-level metaphors and analogies is like “weight sharing in concept space”
- Just like convolutional networks allow us to share low-level image filters (“System-1 weight sharing”), metaphors allow us to share high-level, dynamical circuitry (“System-2 weight sharing”)
- Embodiment creates an extra bottleneck and thus opportunity for sharing circuitry:
  - all processing must go through a single perceptual representation and a single action space
- Visual attention, as discussed above, is a special case of this, where the fovea takes the role of the “System-1” low-level learner, and the attention policy that learns where to look takes the role of the “System-2” system



Convolution is weight sharing across image positions.

“Can neural networks ‘think’ in analogies?”, Memisevic 2022
“Metaphors We Learn By”, Memisevic 2022

Discussion

- Language models <-> world models:
  - Should they really be separate??
  - Instead of learning separate world models that operate on their own, why not combine vision and language, and learn both at the same time
- Human-like common sense may a more appropriate concept to guide research, that the idea of a generalist, “objective” world model

Thank you

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog