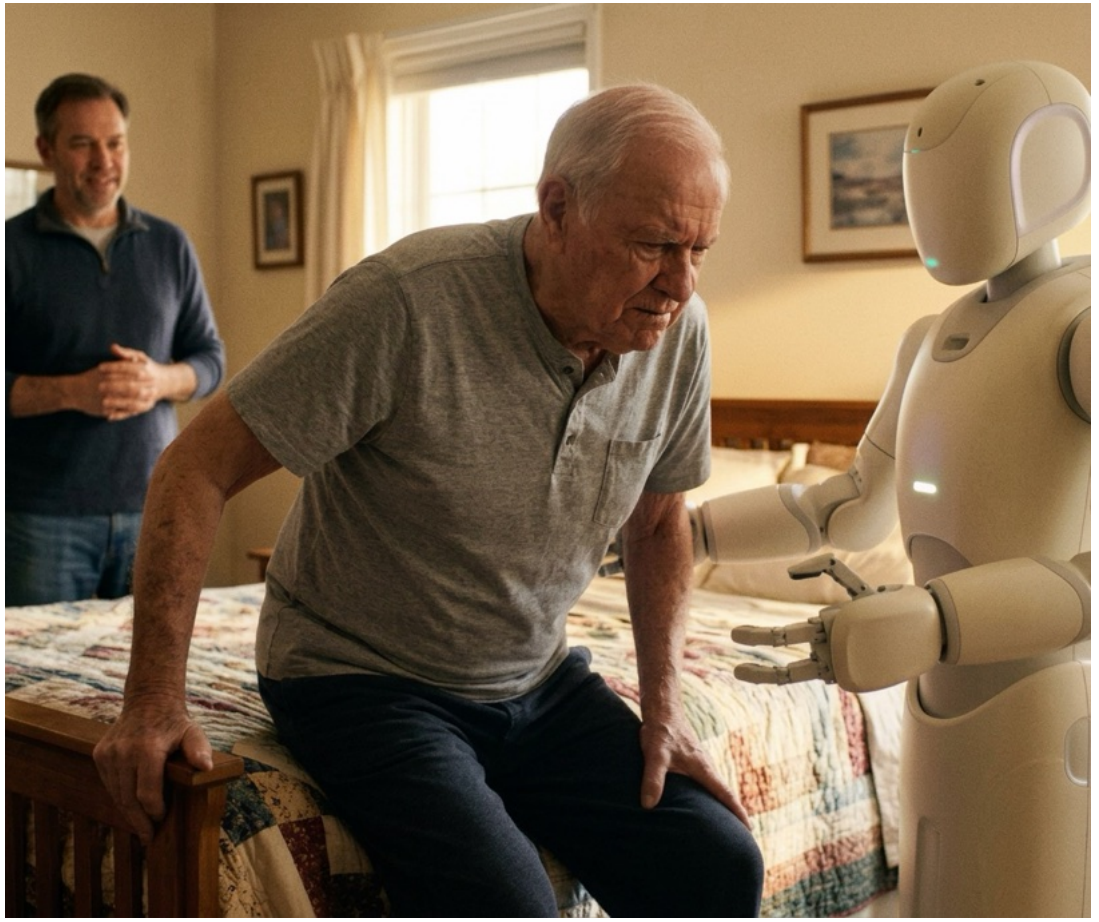
# Perception, Evidence & Efficiency for AI in the Physical World

Juan Carlos Niebles


Samsung Research

# Perception, Evidence & Efficiency for AI in the Physical World



# Traditional vs what we need

**Traditional Video AI:**

- Third-person static cameras
- Short, curated video clips (few seconds)
- Opaque textual outputs ("Yes" or "No")
- Offline batch processing

**AI in the Physical World**

- First-person, egocentric, dynamic camera
- Continuous streaming
- High-stakes decisions require grounding & evidence
- Real-time latency constraints

# Perception, Evidence & Efficiency for AI in the Physical World

Predictive Motion Understanding

Evidence-backed Reasoning

Streaming Efficiency

# Perception, Evidence & Efficiency for AI in the Physical World

Predictive Motion Understanding

Evidence-backed Reasoning

Streaming Efficiency

# Part I: Predictive Motion Understanding

# Predictive Motion Understanding: Three related tasks

## Reconstruction

**input** video + camera trajectory

*What was my motion?*

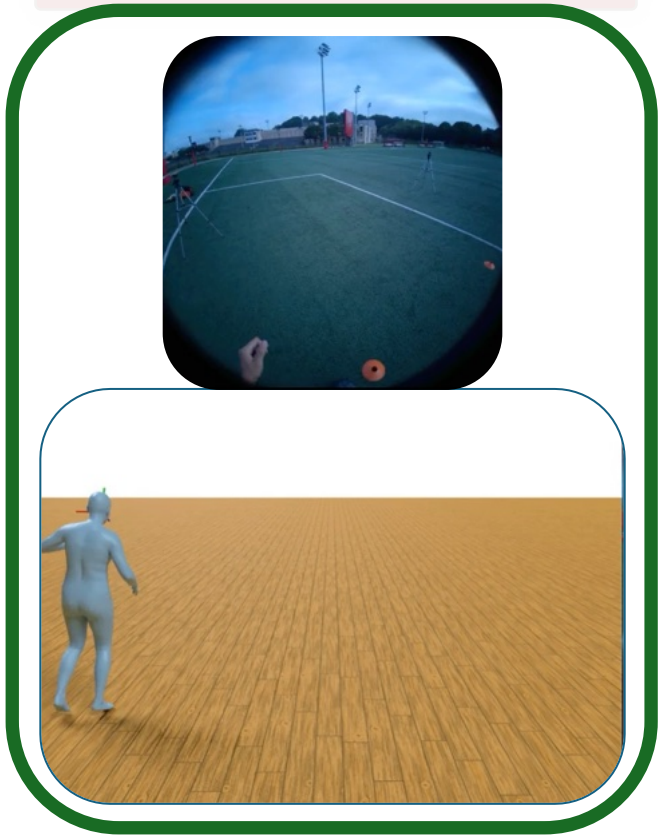
FROM PRIOR WORK

## Forecasting

**input** past video + camera trajectory

*How does this motion continue?*

NEW TASK



## Generation

**input** a single egocentric image

*What motion is plausible here?*

NEW TASK

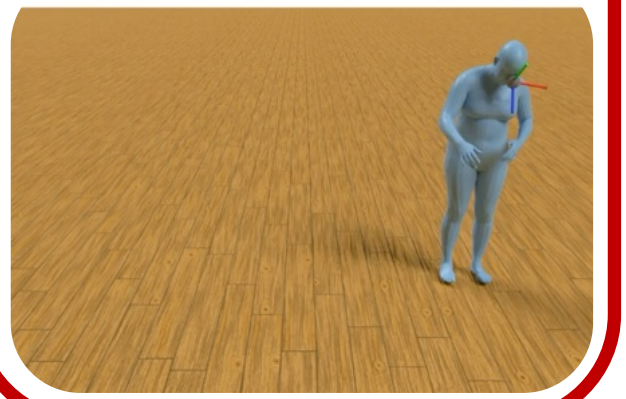
# Predictive Motion Understanding: Three related tasks

## Reconstruction

**input** video + camera trajectory

*What did I just do?*

FROM PRIOR WORK



## Forecasting

**input** past video + camera trajectory

*How does this motion continue?*

NEW TASK





## Generation

**input** a single egocentric image

*What motion is plausible here?*

NEW TASK

# Predictive Motion Understanding: Three related tasks

## Reconstruction

**input** video + camera trajectory

*What did I just do?*

FROM PRIOR WORK



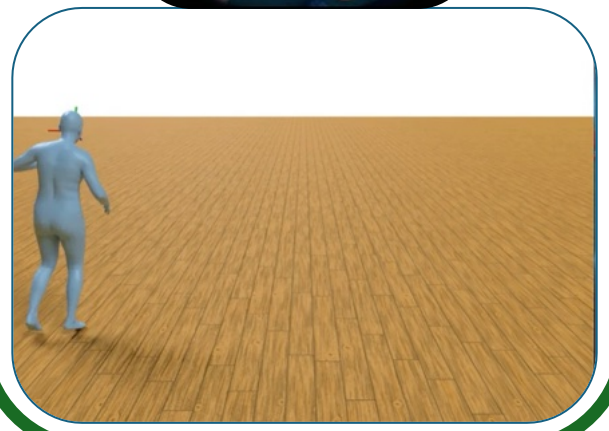


## Forecasting

**input** past video + camera trajectory

*How does this motion continue?*

NEW TASK



## Generation

**input** a single egocentric image

*What motion is plausible here?*

NEW TASK

# Predictive Motion Understanding: Three related tasks

## Reconstruction

**input** video + camera trajectory

*What did I just do?*

FROM PRIOR WORK



## Forecasting

**input** past video + camera trajectory

*How does this motion continue?*

NEW TASK



## Generation

**input** a single egocentric image

*What motion is plausible here?*

NEW TASK

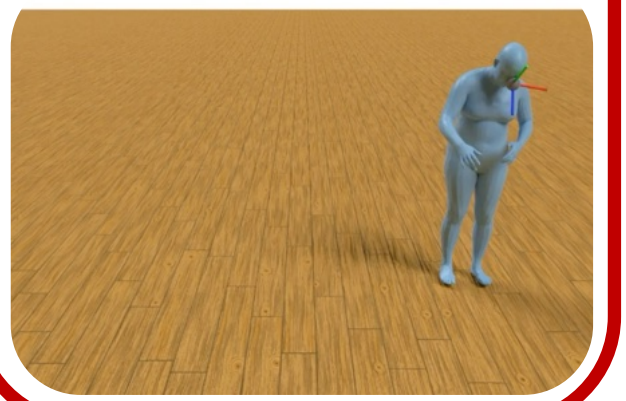
# Predictive Motion Understanding: Three related tasks

## Reconstruction

**input** video + camera trajectory

*What did I just do?*

FROM PRIOR WORK

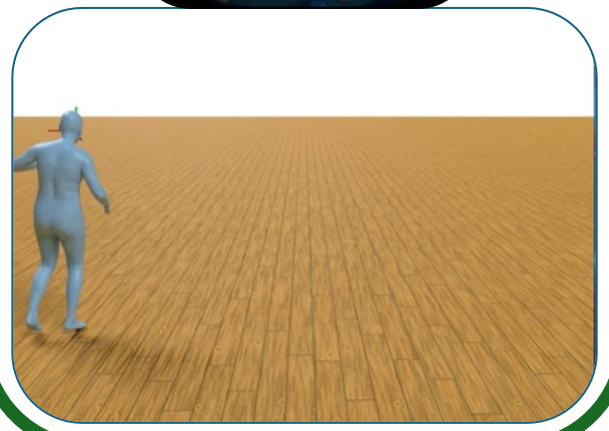


## Forecasting

**input** past video + camera trajectory

*How does this motion continue?*

NEW TASK

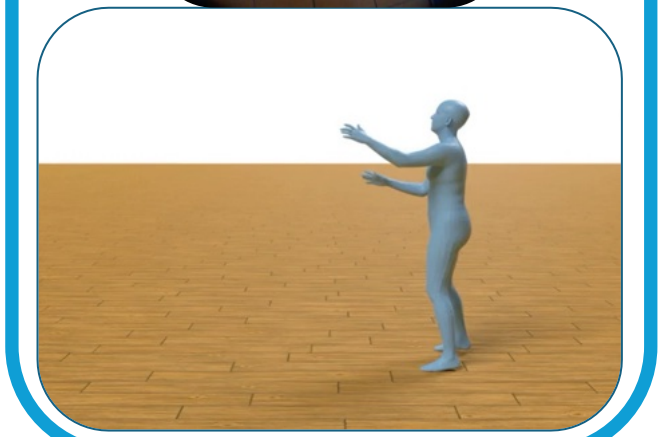


## Generation

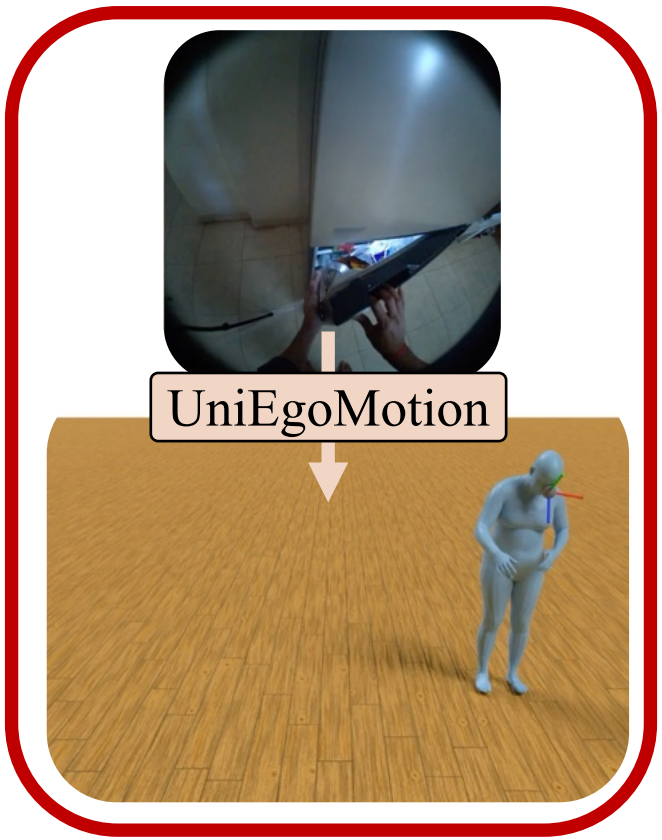
**input** a single egocentric image

*What motion is plausible here?*

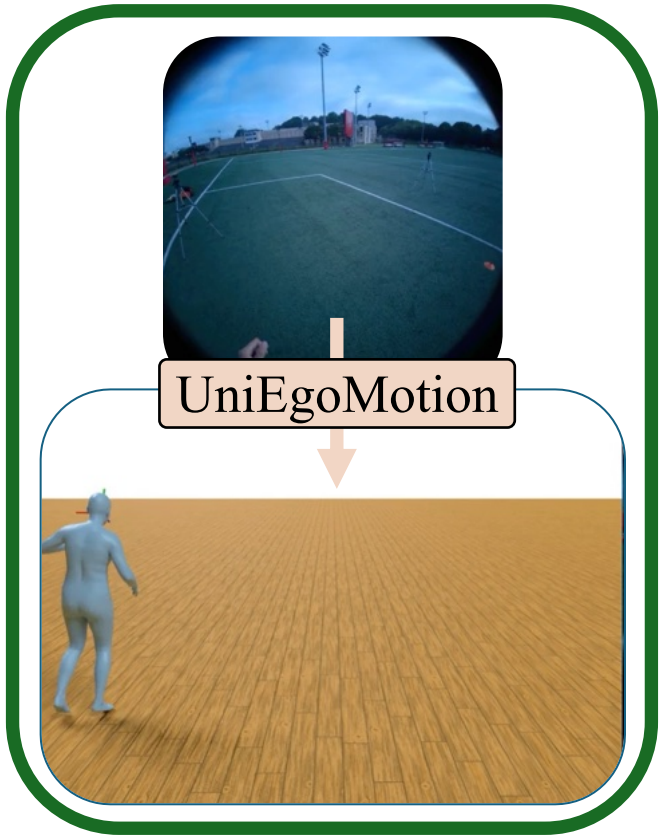
NEW TASK



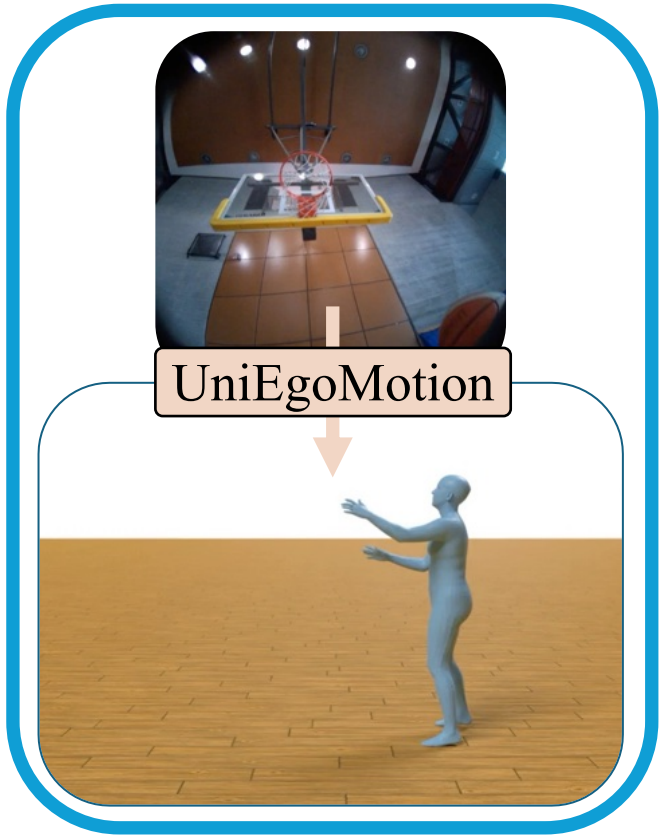
# UniEgoMotion: A Single Diffusion Model for All Three Tasks




Motion Reconstruction



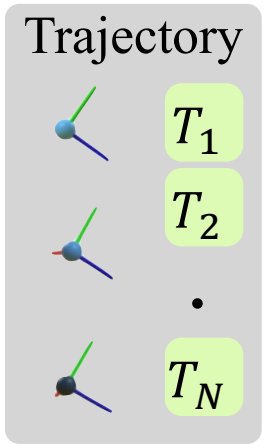

Motion Forecasting




Motion Generation

[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Model Inputs




$\boldsymbol{T}_{1:N}$

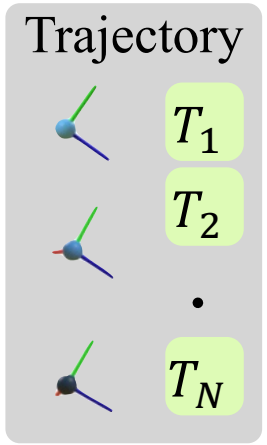
3D trajectory of the egocentric camera


$\boldsymbol{I}_{1:N}$

Visual observations from egocentric camera

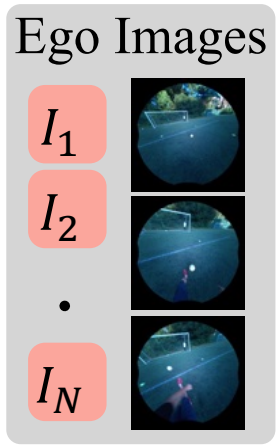
[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Generative Formulation




$\boldsymbol{T}_{1:N}$

3D trajectory of the egocentric camera




$\boldsymbol{I}_{1:N}$

Visual observations from egocentric camera

## Motion Reconstruction

$$\boldsymbol{X}_{1:N} \sim p(\,.\,|\,\boldsymbol{T}_{1:N}\,,\;\boldsymbol{I}_{1:N})$$

## Motion Forecasting

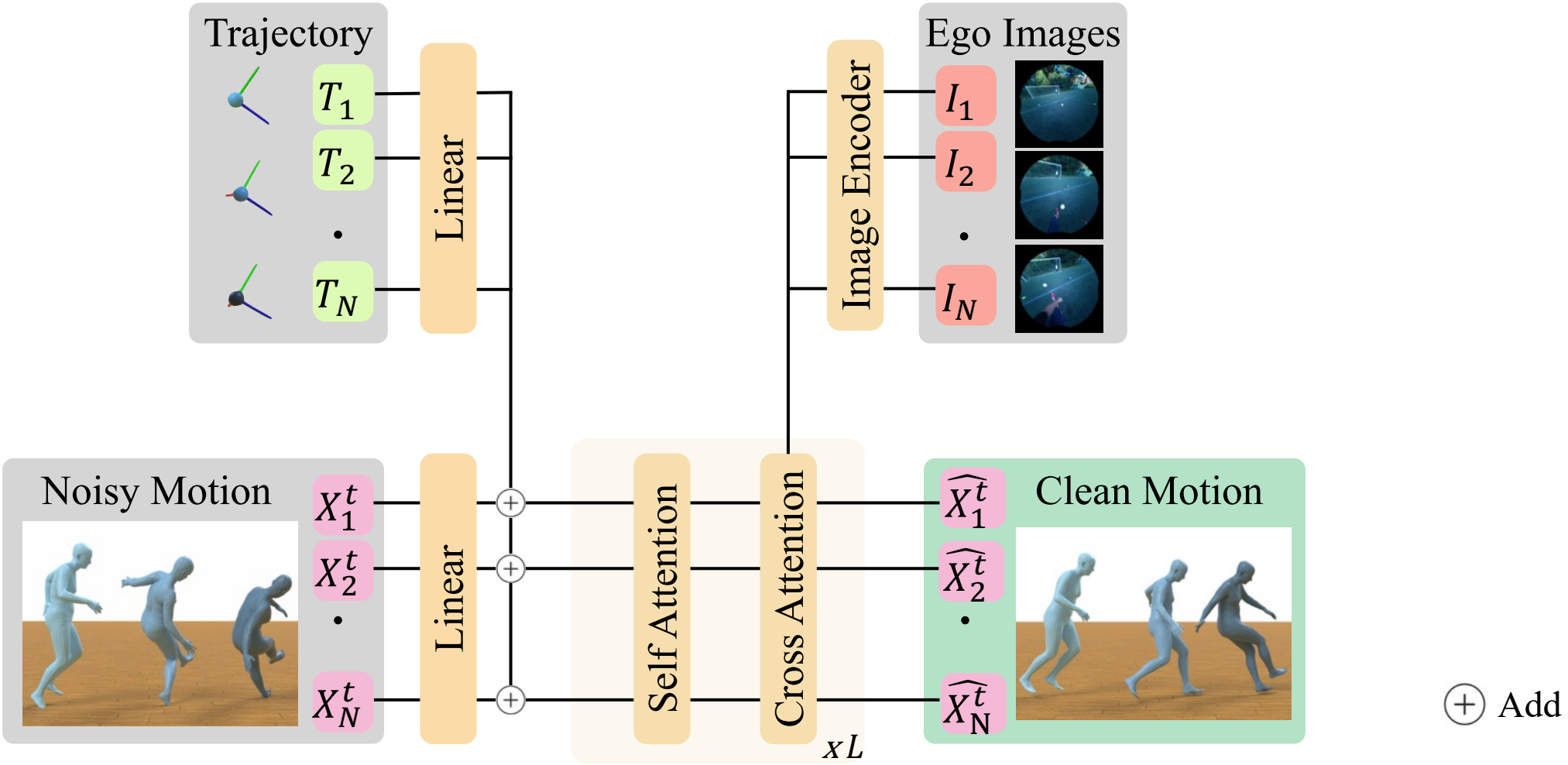
$$\boldsymbol{X}_{\mathrm{n}+1:N} \sim p(\,.\,|\,\boldsymbol{T}_{1:\mathrm{n}}\,,\;\boldsymbol{I}_{1:\mathrm{n}})$$

## Motion Generation

$$\boldsymbol{X}_{1:N} \sim p(\,.\,|\,I_1)$$

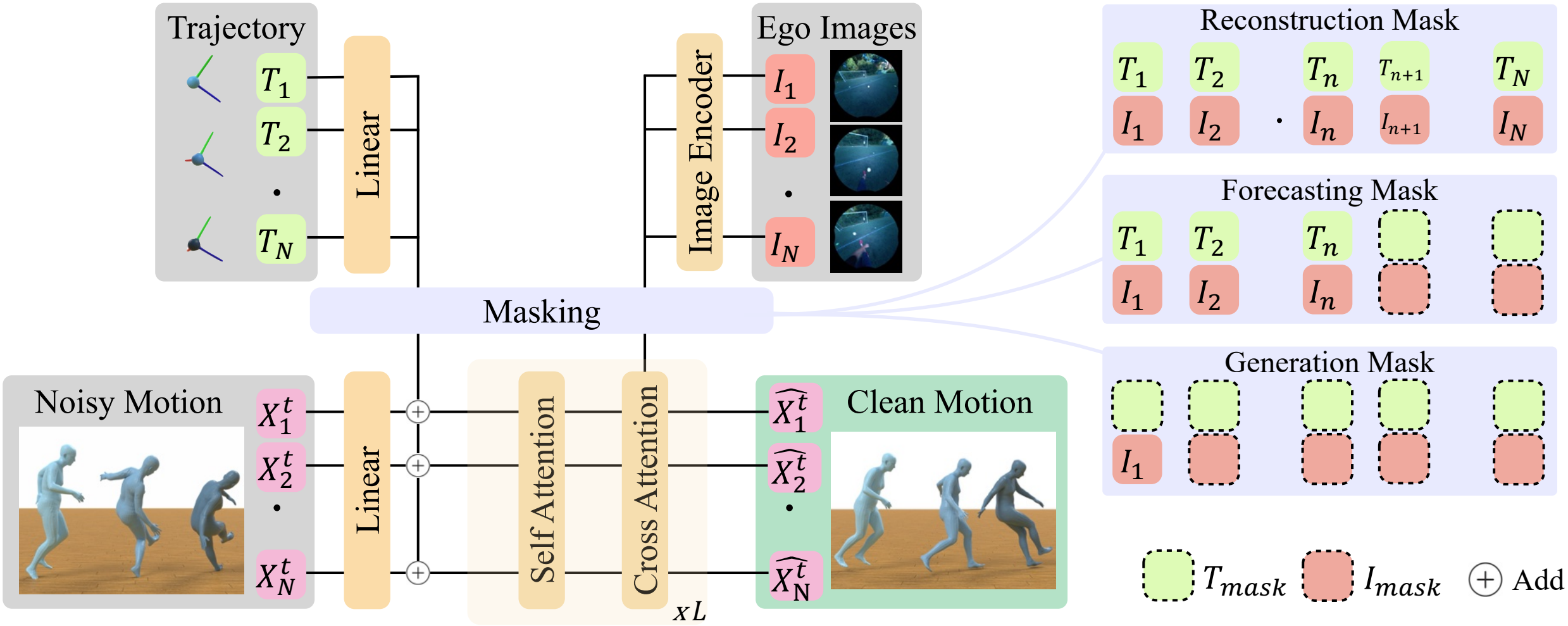
[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# UniEgoMotion Diffusion Architecture



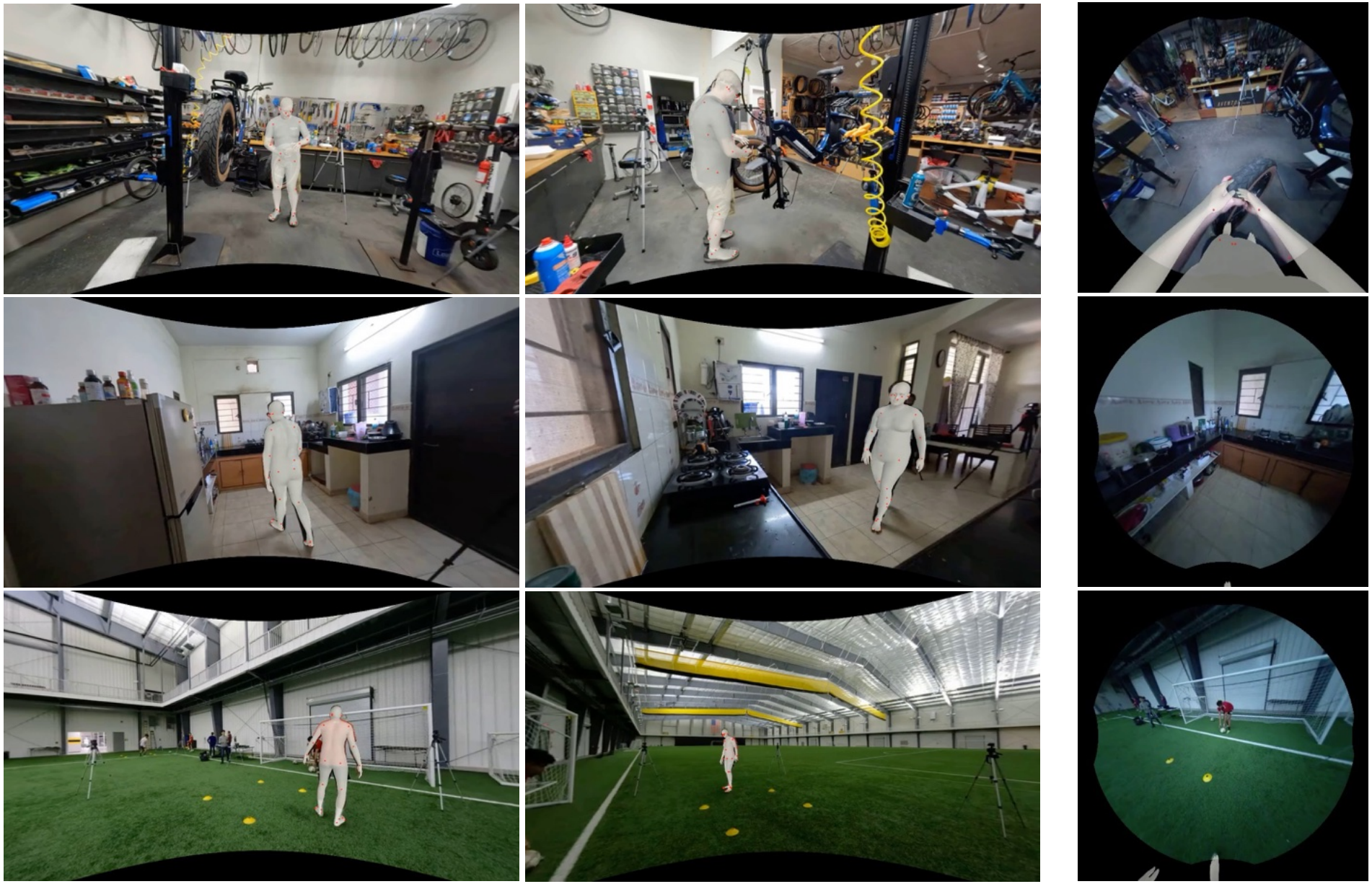

[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Input Masking to Simulate All Three Tasks




[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

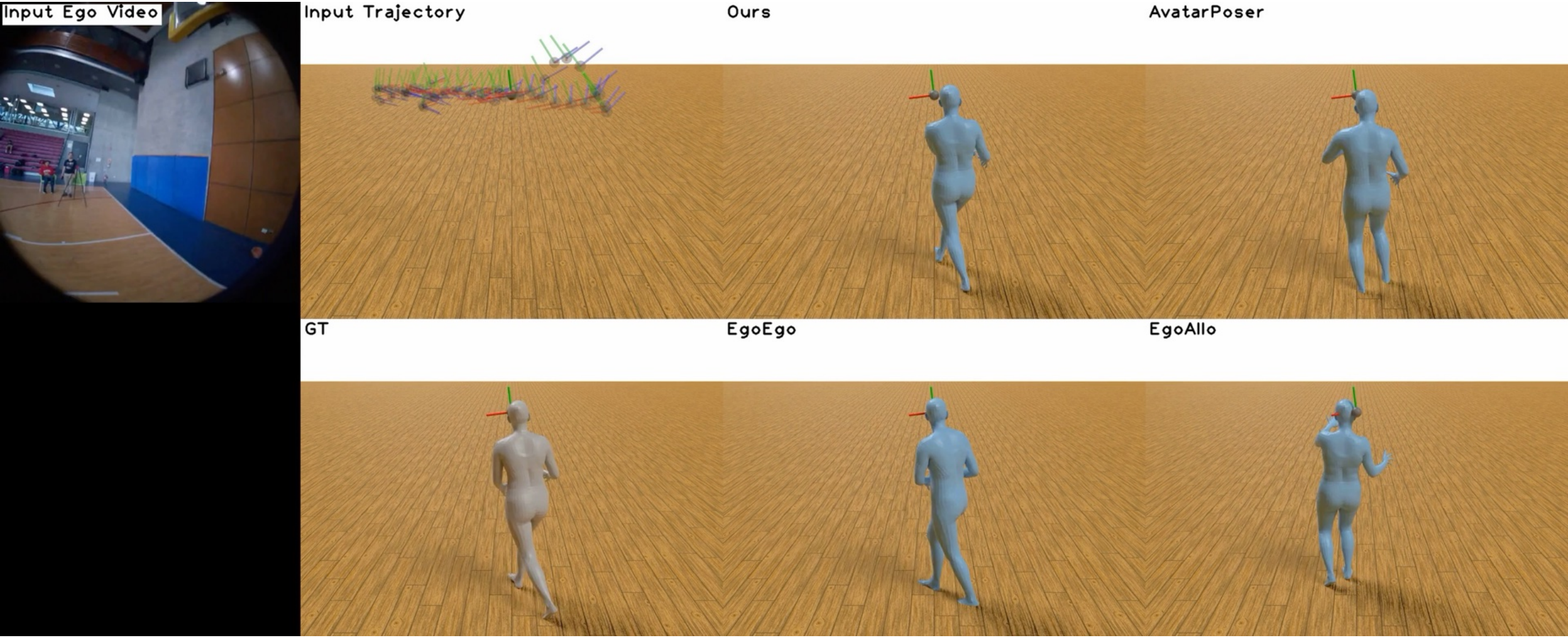
# Dataset: EE4D-Motion



EE4D-Motion dataset:
~208hrs of SMPL-X motion fitting on EgoExo4D dataset.

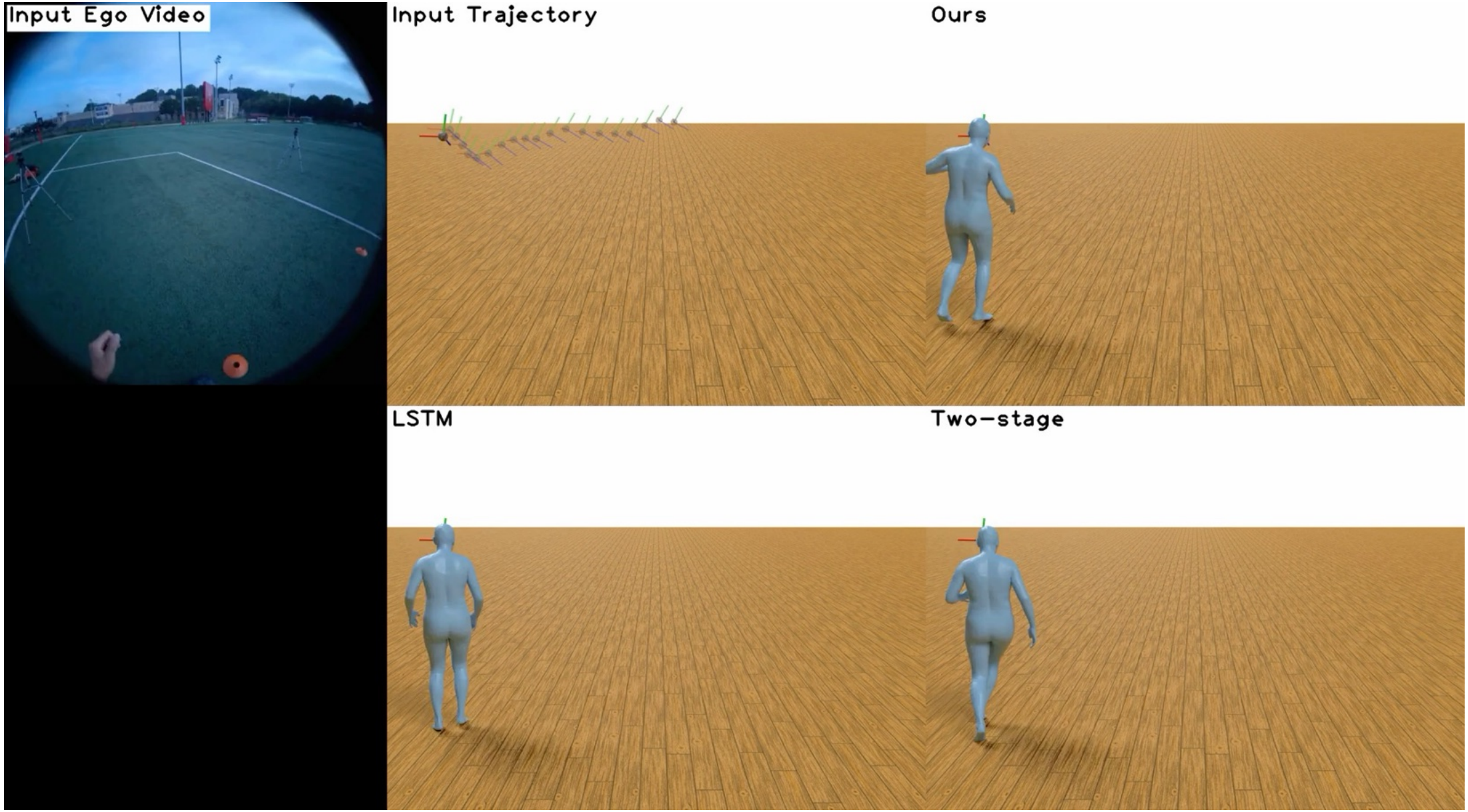
Publicly available for download!

# Results: Egocentric Motion Reconstruction



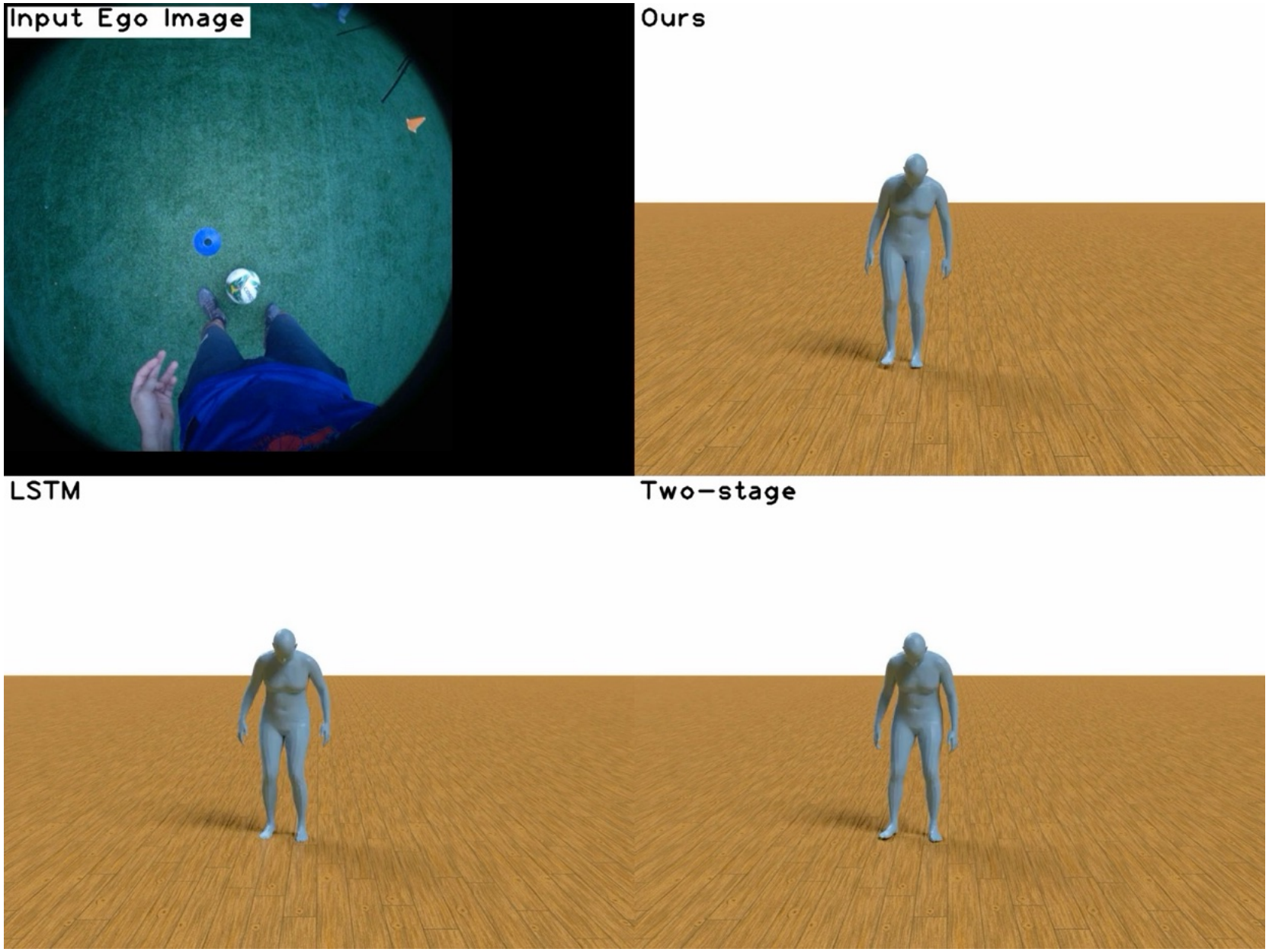

[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Results: Egocentric Motion Forecasting




[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Results: Egocentric Motion Generation




[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation

- **One model, three capabilities**: Reconstruction, forecasting, and generation unified in a single diffusion model
- **EE4D-Motion Dataset**: paired egocentric video and 3D motion from real environments
- **Open:** Code and data, are publicly available

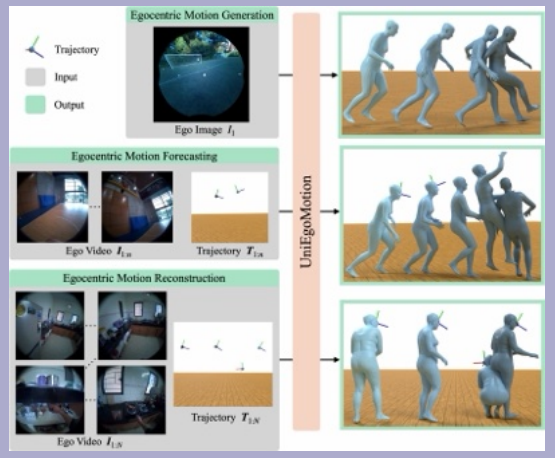
🔗 **code/data/paper**

[C. Patel et al. UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation. ICCV 2025]

# Perception, Evidence & Efficiency for AI in the Physical World

## Predictive Motion Understanding

[UniEgoMotion ICCV'25]




## Evidence-backed Reasoning

## Streaming Efficiency

# Perception, Evidence & Efficiency for AI in the Physical World

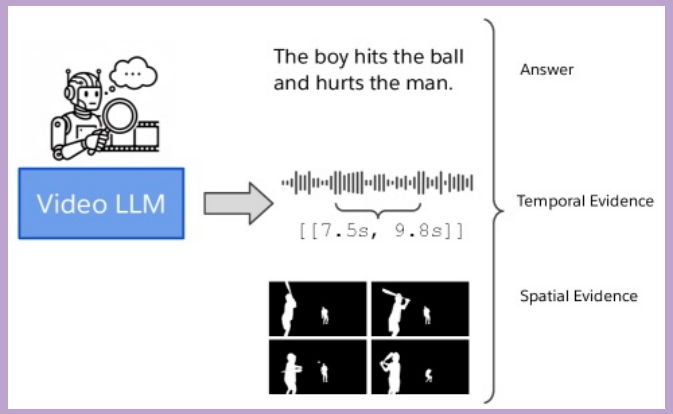
## Predictive Motion Understanding

[UniEgoMotion ICCV'25]



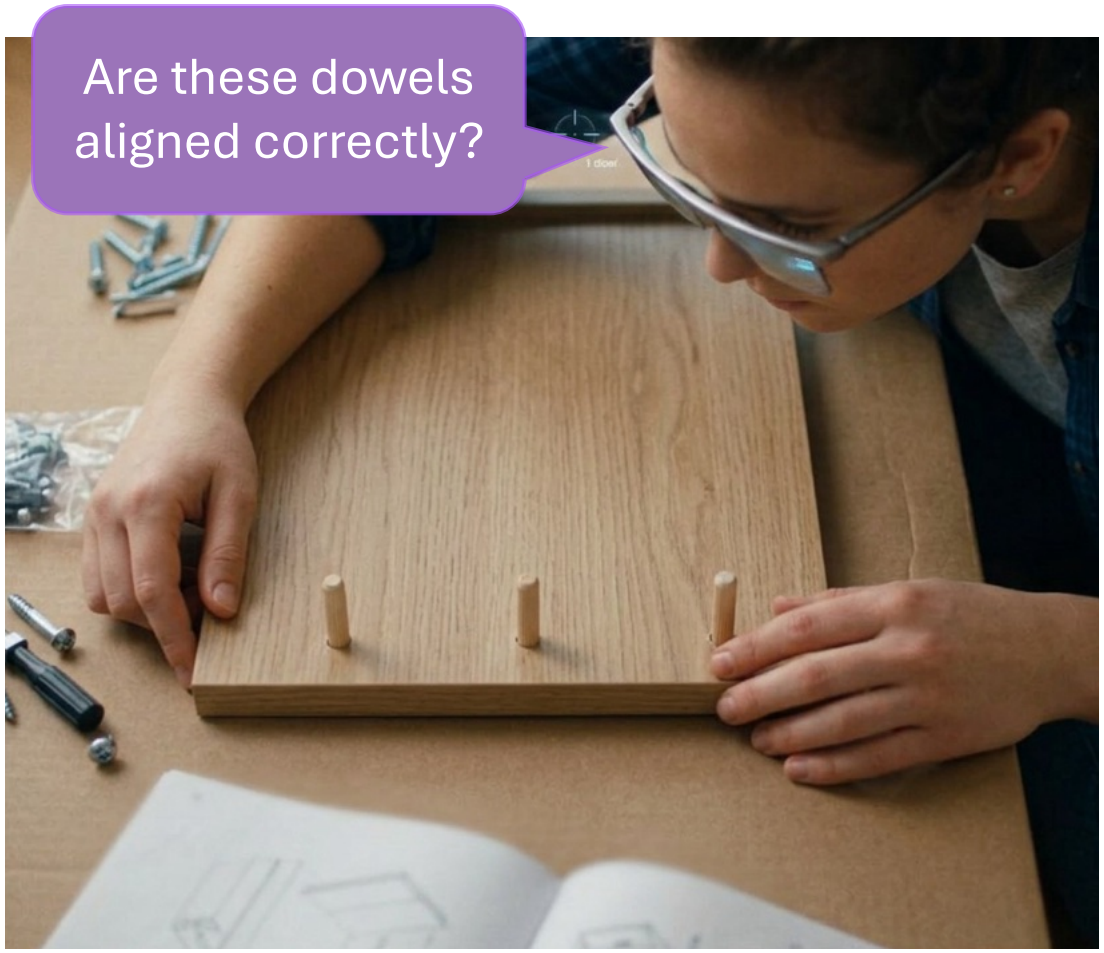

## Evidence-backed Reasoning

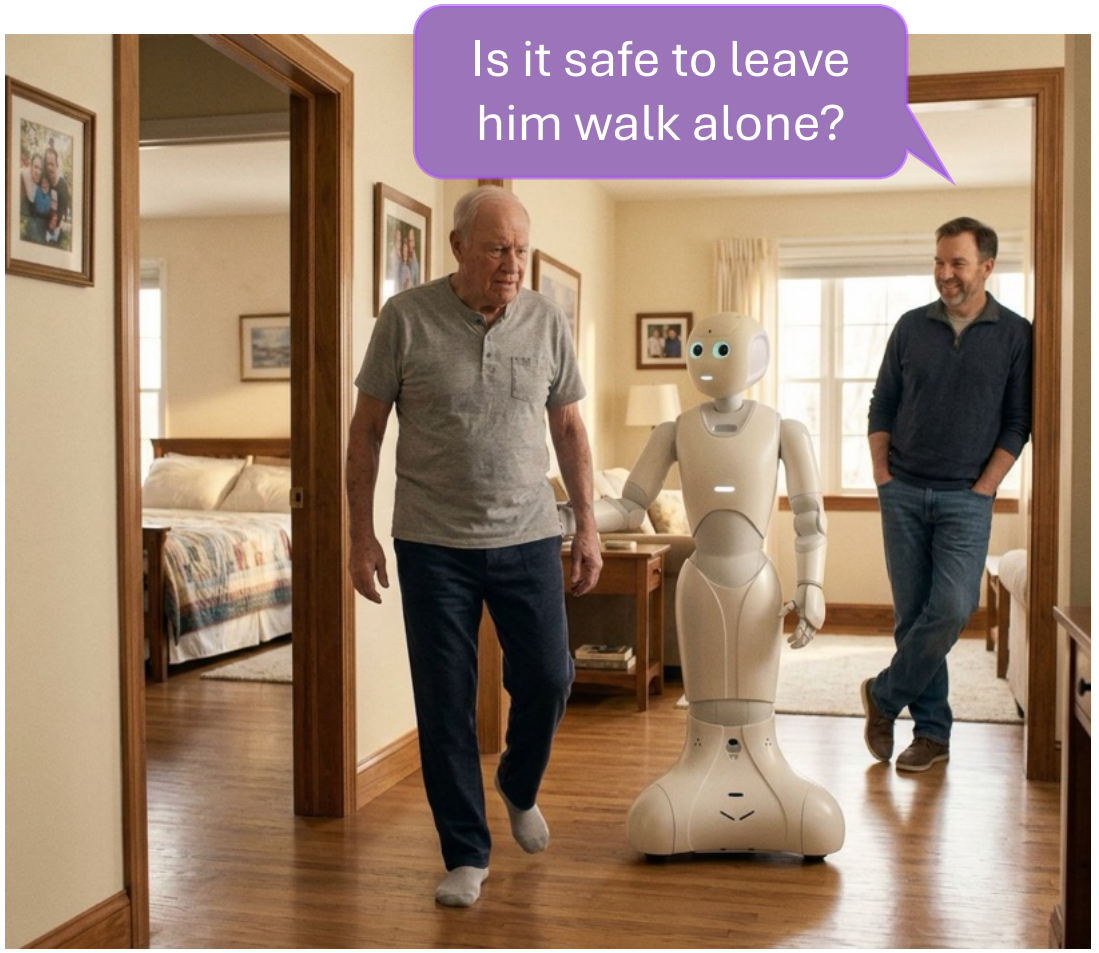
[E-VQA ECCV'26]




## Streaming Efficiency

# Part II: Evidence-backed Reasoning



MOTIVATION

# Today's Video LLMs: fluent answers, invisible reasoning


Why did the man fall down?

→ **Video LLM** →

*“The boy hits the ball and hurts the man.”* ✓

When did it happen? *not shown*

What did it look at? *not shown*

**Right answer — but did the model actually see it, or just guess from language priors?**

MOTIVATION

# Existing forms of “evidence” cannot verify perception

## Textual chain-of-thought

*“The boy swings the bat, the ball flies toward the man, so the man gets hurt and falls…”*

✗ words, not pixels — unverifiable

## Boxes on sparse keyframes


✗ too sparse for dynamics

## Dense tracked masklets

✓ pixel-level, frame-by-frame, verifiable

**Only a dense, tracked evidence trail can support that the model perceived the right thing at the right time.**

TASK SETTING

# E-VQA: answer + when + where, in one output


Why did the man fall down?

Video LLM

The boy hits the ball and hurts the man.

[[7.5s, 9.8s]]

Answer

Temporal Evidence

Spatial Evidence

**A** Semantic answer to the question

**$E_t$** Temporal segments {[$start_i$, $end_i$]}

**$E_s$** Dense tracked masklets of evidence objects

*The triplet couples reasoning with pixel-level grounding to defeat language-prior shortcuts.*

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

TASK SETTING

# What's new compared to existing video tasks?

| Task | Semantic answer | Temporal evidence | Pixel-level masks | Dense tracking | Joint output |
|---|---|---|---|---|---|
| Video QA | ✓ | — | — | — | — |
| Temporal grounding | — | ✓ | — | — | — |
| Referring video segmentation | — | — | ✓ | ✓ | — |
| Answer grounding (boxes) | ✓ | △ | △ | — | △ |
| **E-VQA (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ |

△ = *partial: boxes on sparse keyframes, limited reasoning-grounding coupling*

**E-VQA is the first task requiring reasoning and dense spatio-temporal grounding in a single, verifiable output.**

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

BENCHMARK

# ST-Evidence: human-annotated, human-verified

### 1 Sample

VLM-filtered QA pairs from six video benchmarks

### 2 Annotate

experts mark temporal segments + dense masks at 6 FPS

### 3 Verify

independent PhD-level review, flagged items re-annotated

**4,004**
human-verified questions

**~2,900**
videos across 6 sources

**6 FPS**
dense masklet annotation

**100%**
reviewed by researchers

*Sources: NExT-QA · Perception Test · STAR · CLEVRER · Ego4D · ViCaS*

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

BENCHMARK

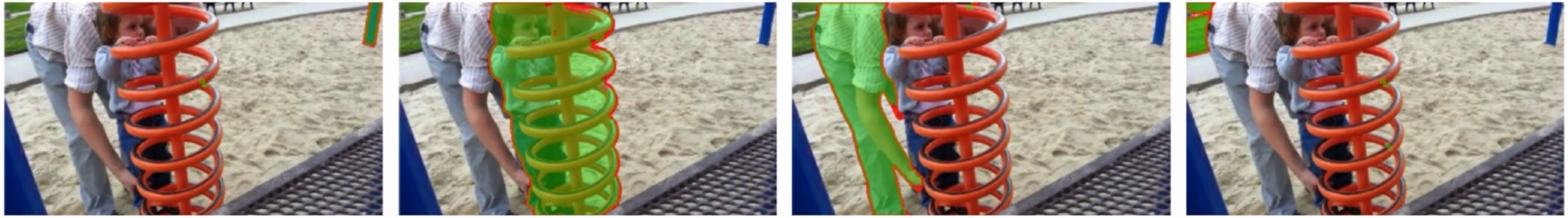
# One sample, two evaluation modes

**[Question]** How does the child make sure she does not fall?

**[Options]** **(A) hold onto the orange structure tightly** (B) wear professional equipment (C) hold the walker (D) soft surface (E) hold onto lady

**[Temporal Evidence]** (A) [[0.0, 4.6]] **(B) [[45.7, 49.6]]** (C) [[36.4, 39.9]] (D) [[40.1, 43.8]]

**[Spatial Evidence]**



A B C D

**Gen** produce the triplet
metrics: accuracy · tIoU / IoP · J&F

**MCQ** choose answer / segment / mask among candidates — accuracy

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

BENCHMARKING SOTA

# High QA accuracy does not certify perception

| Model | QA acc | T-Evid. acc | S-Evid. acc |
|---|---|---|---|
| *Random guess* | *20.0* | *25.0* | *25.0* |
| OpenAI-o3 † | 81.2 | 50.7 | 84.3 |
| Gemini-2.5-Flash † | 80.4 | **78.8** | 36.3 |
| Gemini-2.5-Pro † | **82.8** | 70.9 | 81.3 |
| LLaVA-OV-1.5-8B | 69.0 | 66.0 | 23.4 |
| Qwen2.5-VL-7B | 74.7 | 27.1 | 46.2 |
| Qwen3-VL-235B-A22B | 81.6 | 71.0 | **86.9** |
| UniPixel-7B (grounded) | 74.0 | 35.3 | 22.0 |

**Spatial evidence ≈ random**

Most open-source models pick the correct evidence mask at ~25% — chance level.

**Grounding is fragmented**

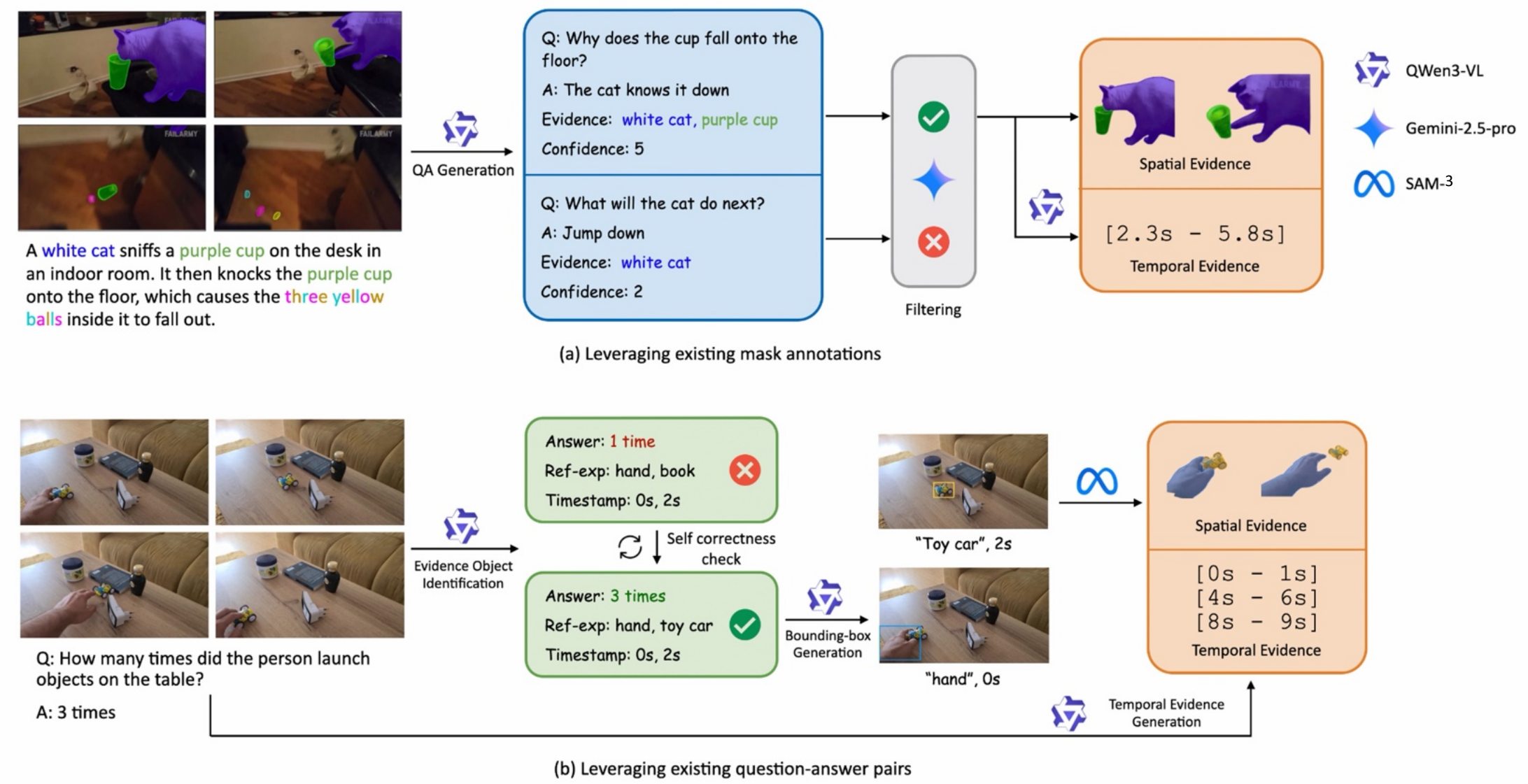
Gemini-2.5-Flash: best temporal (78.8) yet weak spatial (36.3); o3 shows the opposite.

*† closed-source. Full 16-model table in the paper.*

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

TRAINING DATA

# ST-Evidence-Instruct: 160k automatic evidence triplets

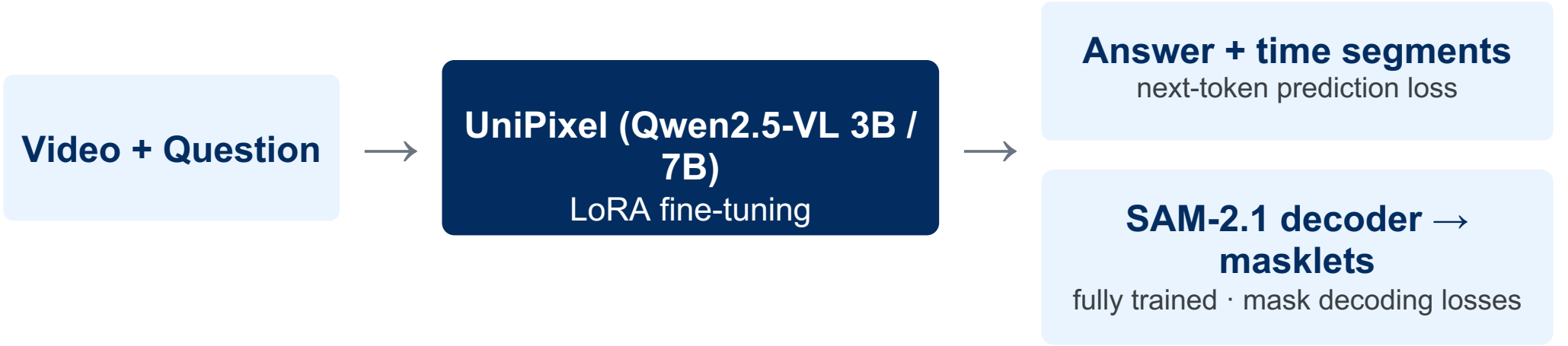



(a) mask-annotated videos → generate + cross-verify QA (b) QA-only videos → identify evidence objects → boxes → dense masks (SAM)

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

METHOD

# A simple, strong baseline: instruction-tune a grounded Video LLM




**Data mix**
ST-Evidence-Instruct + segmentation + general video data

**Joint objective**
answer, segments, and masks predicted in one pass

**No architecture changes**
gains isolate the contribution of our data

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

RESULTS · ST-EVIDENCE-GEN

# Large grounding gains over size-matched baselines

| Model | QA acc | T-Evid. (t-mean) | S-Evid. (J&F) |
|---|---|---|---|
| OpenAI-o3 † | 82.6 | 36.1 | 41.9 |
| Gemini-2.5-Pro † | **83.7** | **49.9** | 44.0 |
| Qwen3-VL-235B-A22B | 83.6 | 44.5 | 41.9 |
| Sa2VA-Qwen2.5-VL-7B | 76.8 | 0.0 | 26.2 |
| UniPixel-3B | 72.2 | 6.5 | 42.7 |
| UniPixel-7B | 75.0 | 1.9 | 40.3 |
| **Ours-3B** | 72.4 | **26.9 (+20.4)** | **52.7 (+10.0)** |
| **Ours-7B** | 76.0 | **29.1 (+27.2)** | **54.1 (+13.8)** |

**+27.2**

t-mean vs UniPixel-7B — temporal evidence, once absent, now competitive

**54.1**

J&F — best spatial evidence of all models, incl. Gemini-2.5-Pro (44.0)

*† closed-source; general Video LLMs use a proxy segmenter (UniPixel-3B) for masks. Gains vs. size-matched UniPixel.*

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

ANALYSIS · RETENTION

# Evidence tuning does not sacrifice general skills

| | | Referring Video Segmentation | | | General Video Understanding |
|---|---|---|---|---|---|
| **Method** | **Size** | **MeViS (J&F)** | **Ref-DAVIS17 (J&F)** | **ReVOS (J&F)** | **MVBench (acc)** |
| VideoLISA | 3.8B | 51.7 | 68.8 | — | — |
| VISA | 7B | — | 70.4 | 47.1 | — |
| Sa2VA | 4B | 52.1 | 73.8 | 53.2 | — |
| UniPixel | 3B | 59.7 | 74.2 | 62.1 | 62.5 |
| UniPixel | 7B | 61.7 | 76.4 | 63.9 | 64.3 |
| **Ours** | 3B | **61.8** (+2.1) | **74.0** (−0.2) | **62.7** (+0.6) | **62.3** (−0.2) |
| **Ours** | 7B | **62.5** (+0.8) | **77.8** (+1.4) | **64.2** (+0.3) | **65.6** (+1.3) |

*Gains shown vs. the size-matched UniPixel baseline. Ours-7B sets the best score in every column.*

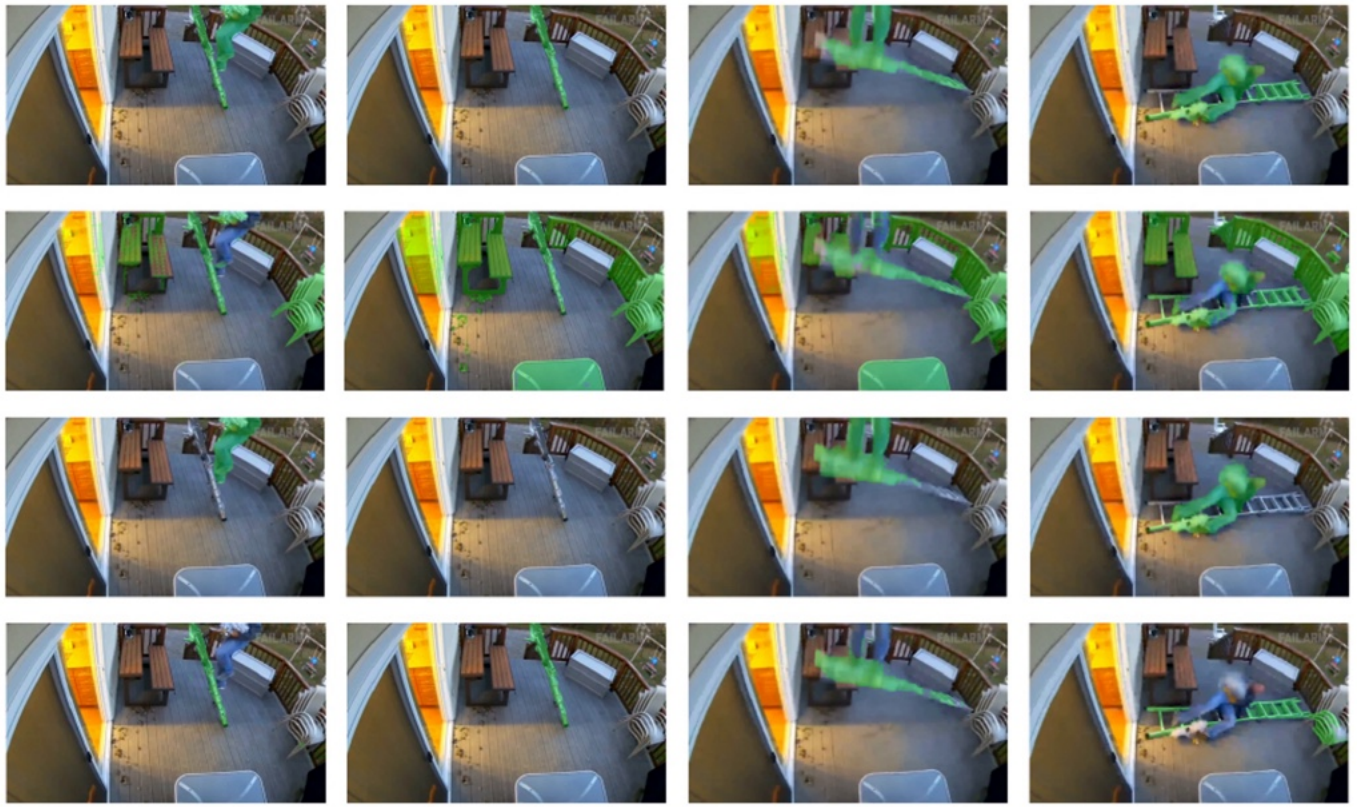
[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

VISUALIZATION

# Qualitative Example

Why did the person fall down? (GT time evidence: [8.6s, 11s])

A) ladder slips backward  B) The ladder collapses under him



**Ours-7B**
[8.5s, 10.3s]
A

**UniPixel-7B**
[0s - 5s]
B

**QWen3-VL-8B**
[9s, 11s]
A

**QWen2.5-VL-7B**
[8s, 9s]
A

*Ours tracks both the person and the slipping ladder over the correct causal interval — baselines miss the interaction, ground background objects, or mislocate the time window.*

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

TAKEAWAYS

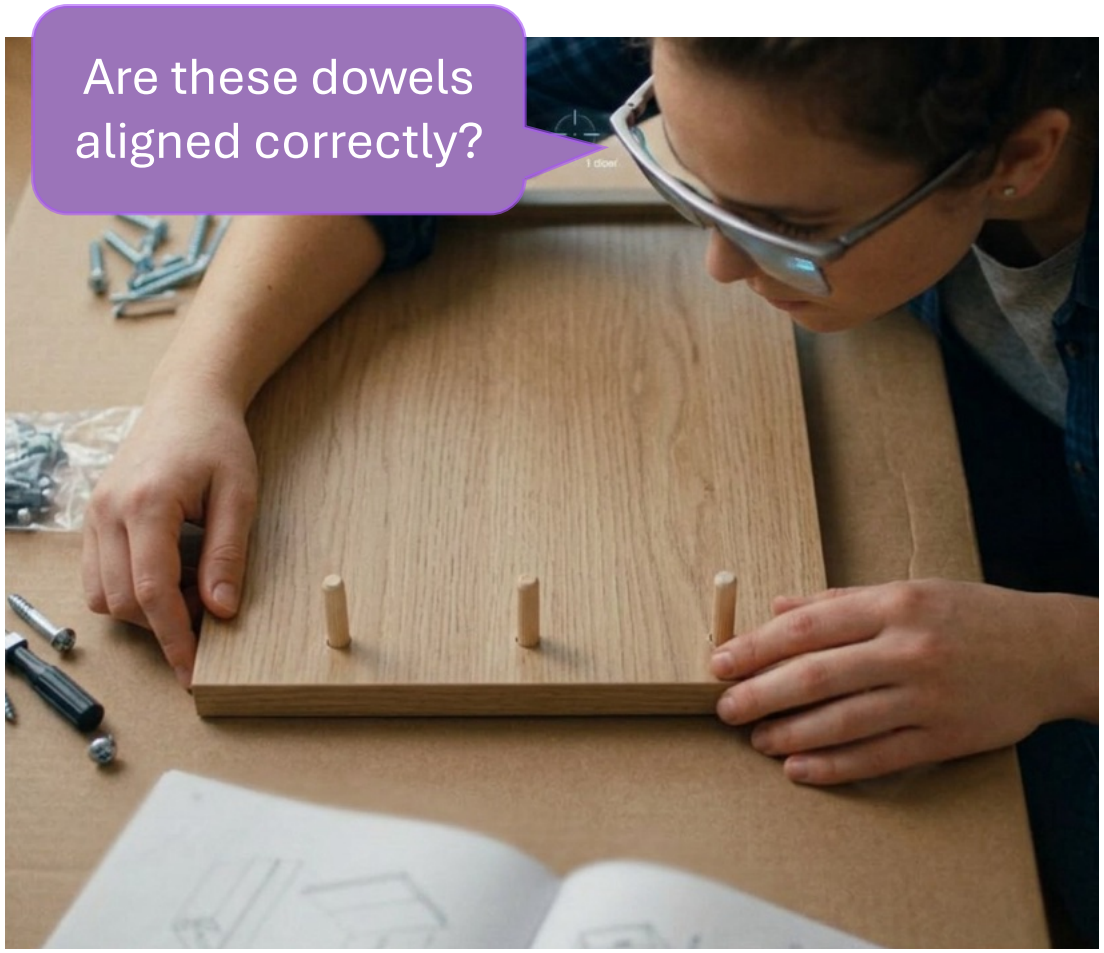
# Evidence-backed answers make Video LLMs more verifiable

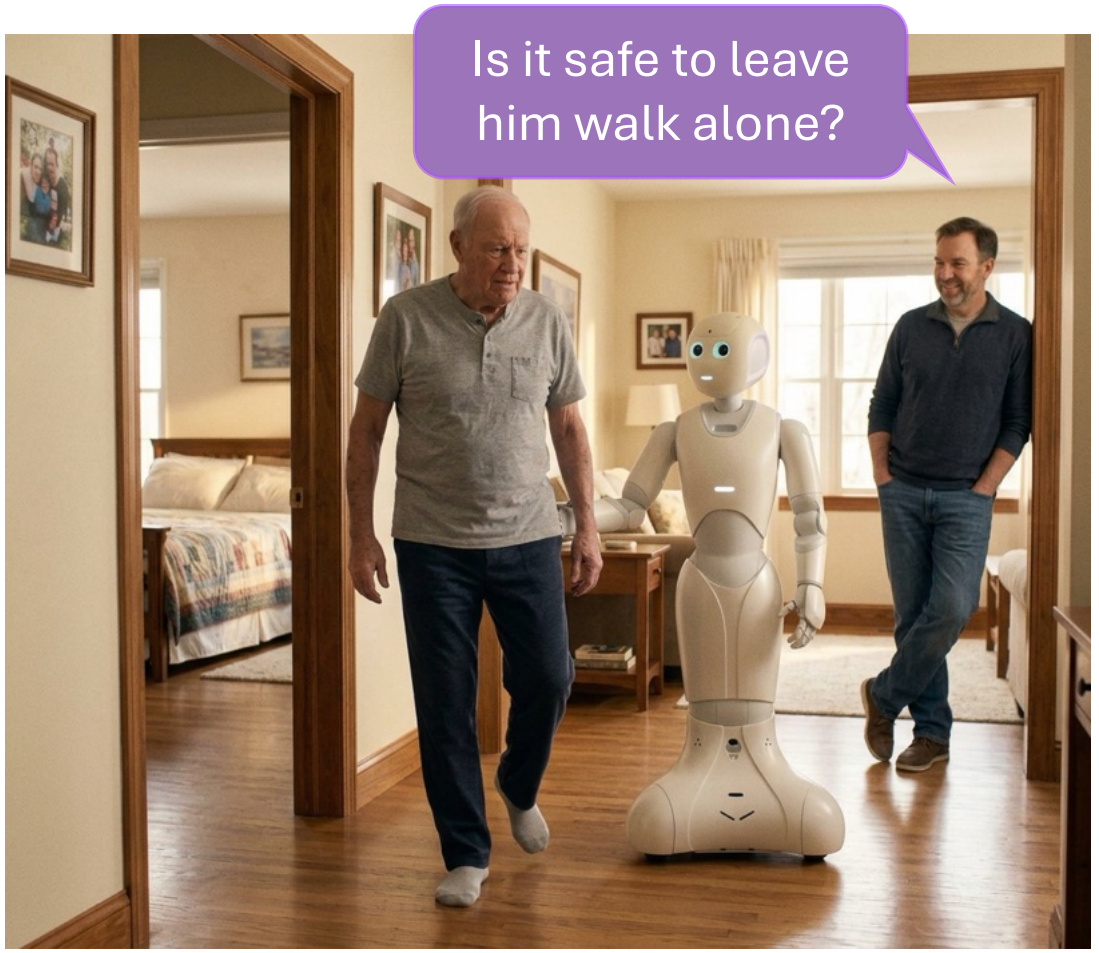
- **E-VQA** a task that couples reasoning with dense, verifiable spatio-temporal evidence
- **ST-Evidence** the first human-verified benchmark — QA accuracy is decoupled from the ability to provide evidence
- **ST-Evidence-Instruct** 160k automatic triplets — +27.2 t-mean, +13.8 J&F over size-matched baselines

🔗 **benchmark/code/data/models**

[S. Wang et al. Evidence-Backed Video Question Answering. ECCV 2026]

# Part II: Evidence-backed reasoning



# Perception, Evidence & Efficiency for AI in the Physical World

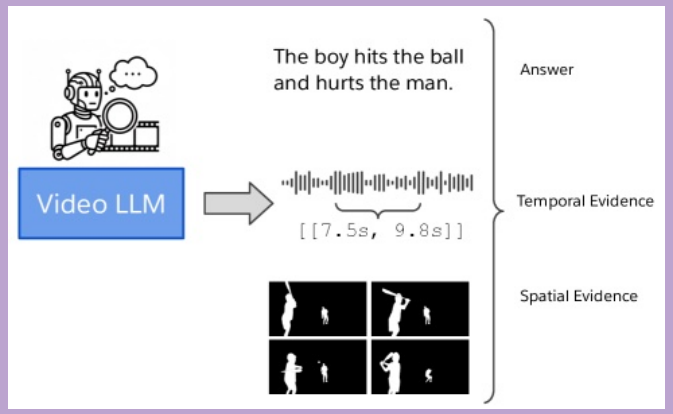
## Predictive Motion Understanding

[UniEgoMotion ICCV'25]




## Evidence-backed Reasoning

[E-VQA ECCV'26]




## Streaming Efficiency

# Perception, Evidence & Efficiency for AI in the Physical World

## Predictive Motion Understanding

[UniEgoMotion ICCV'25]



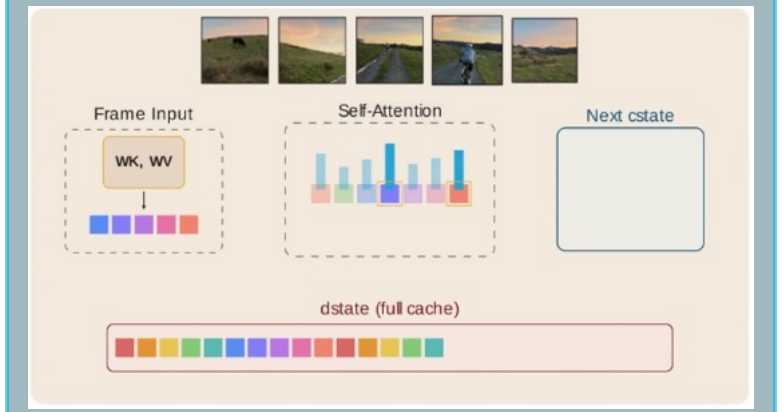

## Evidence-backed Reasoning

[E-VQA ECCV'26]

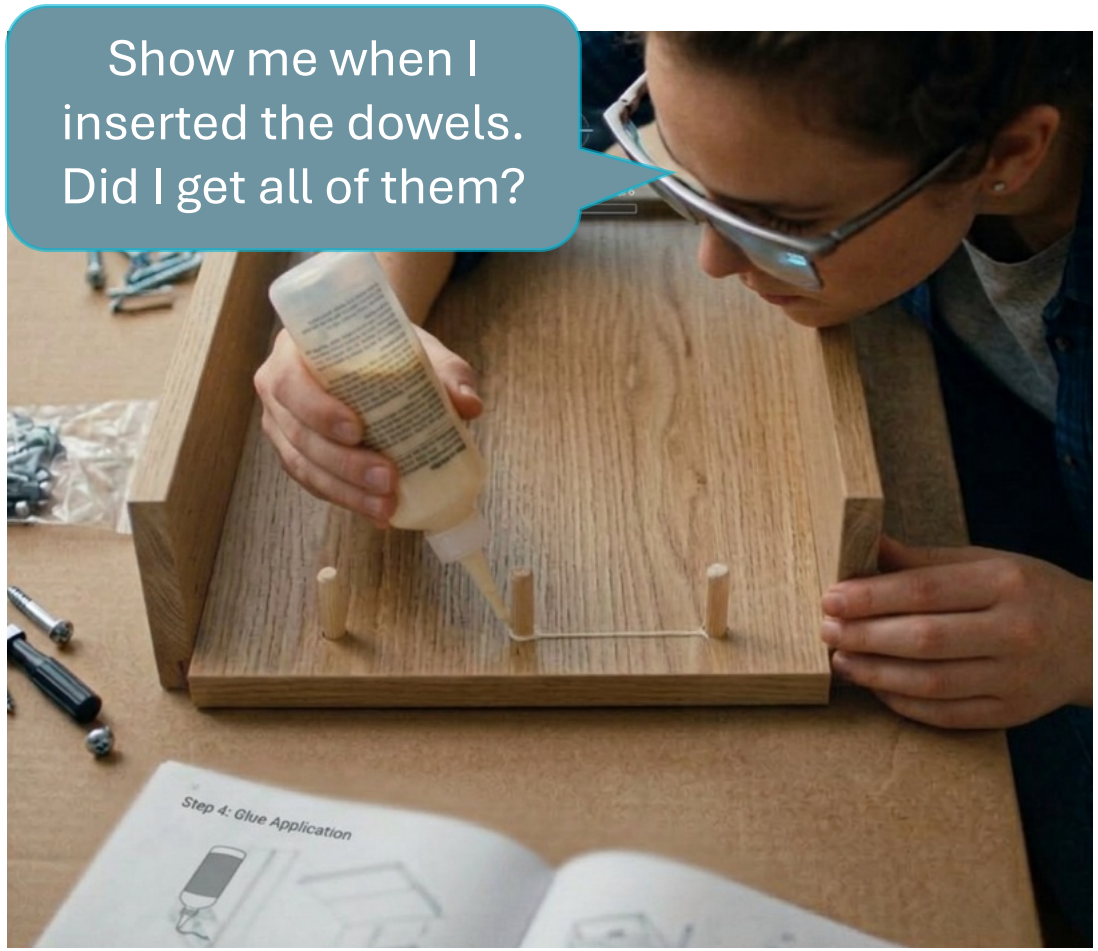


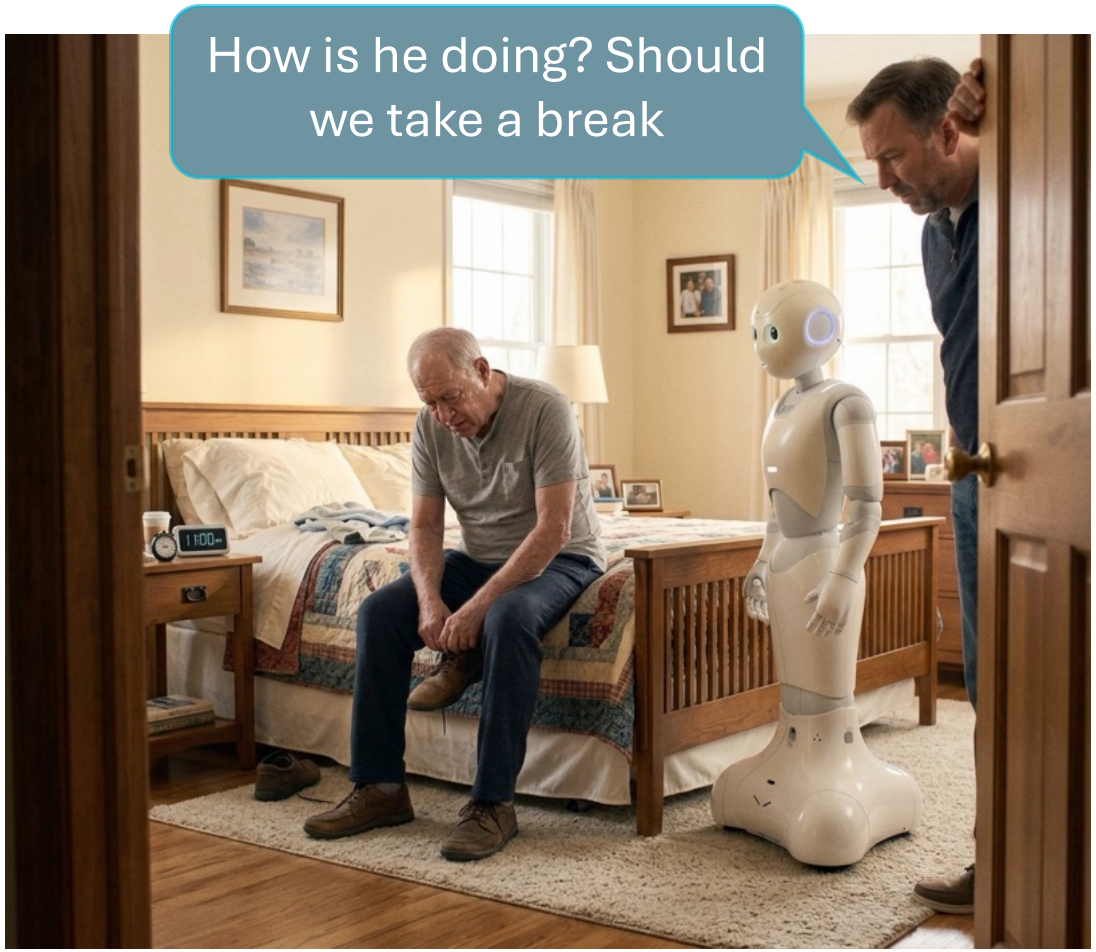

## Streaming Efficiency

[StateKV ECCV'26]

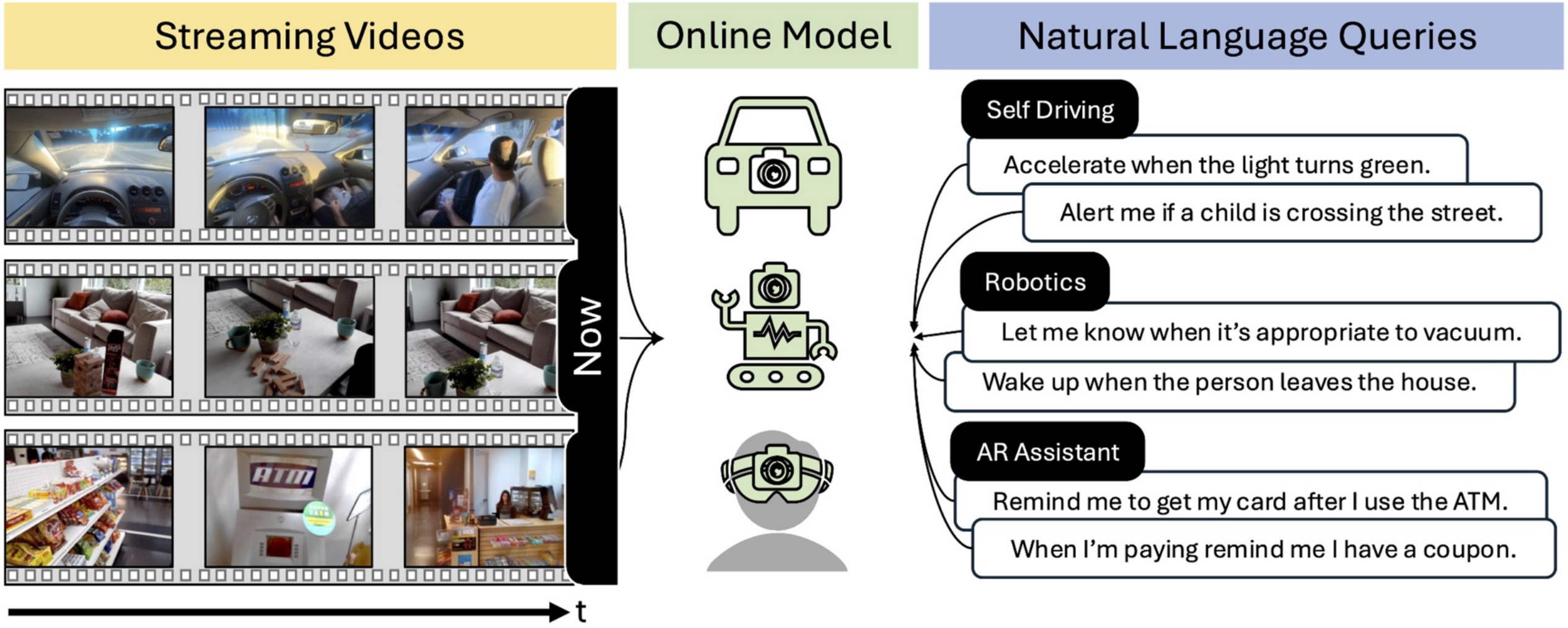


# Part III: Efficiency



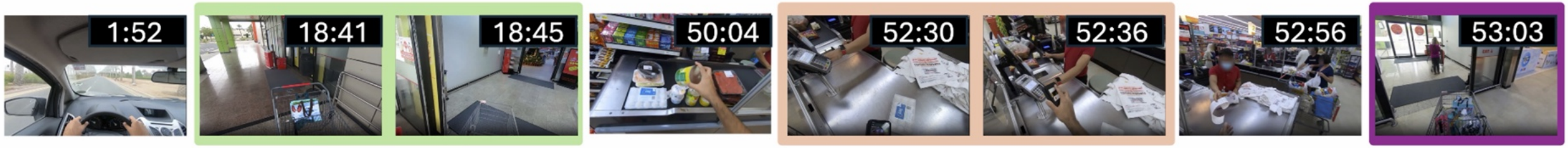



# Streaming Detection of Queried Event Start




[C. Eyzaguirre et al. Streaming Detection of Queried Event Start. NeurIPS 2024]

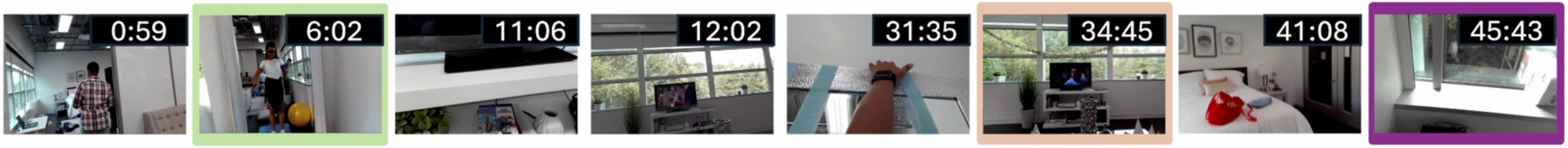
# “Proactive” Dataset: Streaming Detection of Queries (EgoSDQES)




Next time I enter a supermarket, please remind me to sanitize my hands.

Remind me to use my loyalty card when I start to pay at the billing counter.

When I start to exit the supermarket, remind me to return the shopping trolley.



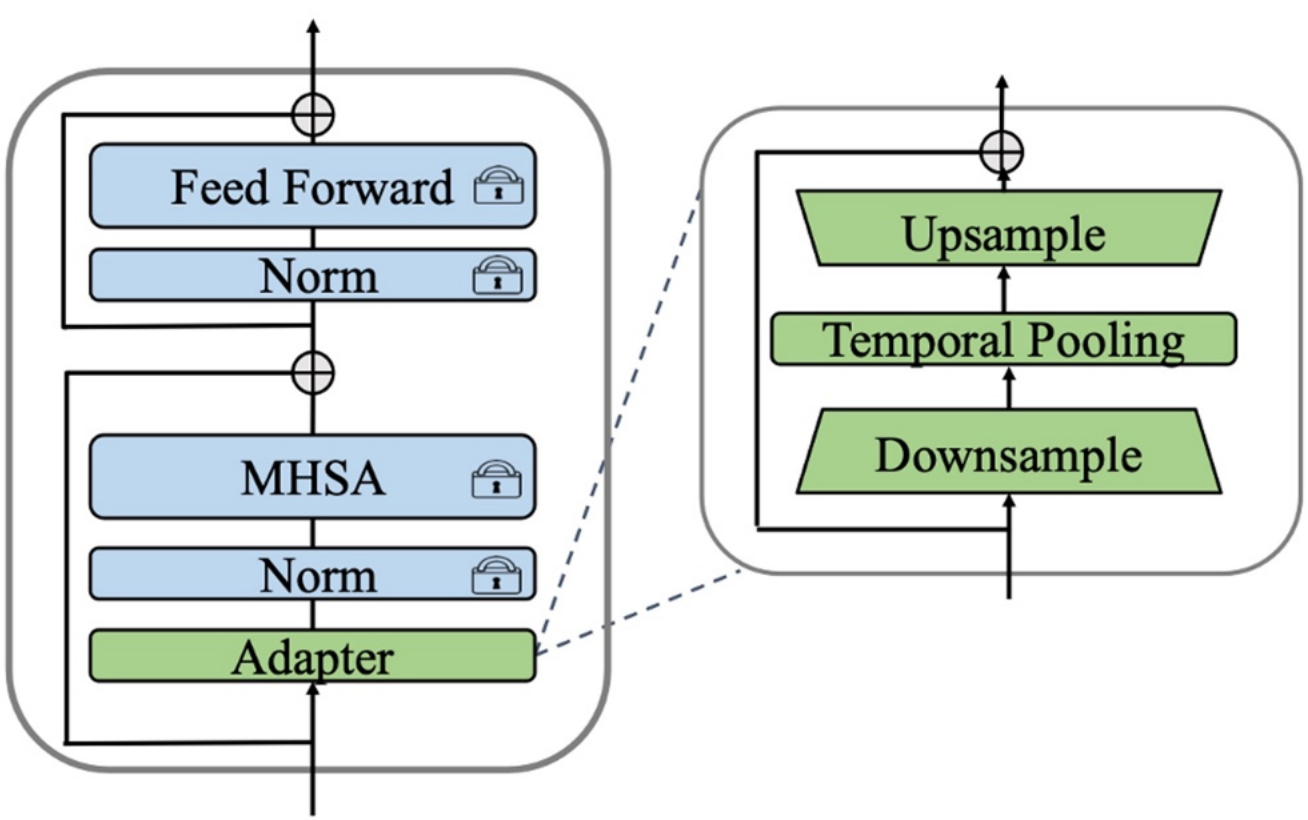

Next time I start doing my exercises, please remind me to drink some water.

When I start watching television **after** decorating the house, remind me to adjust the volume.

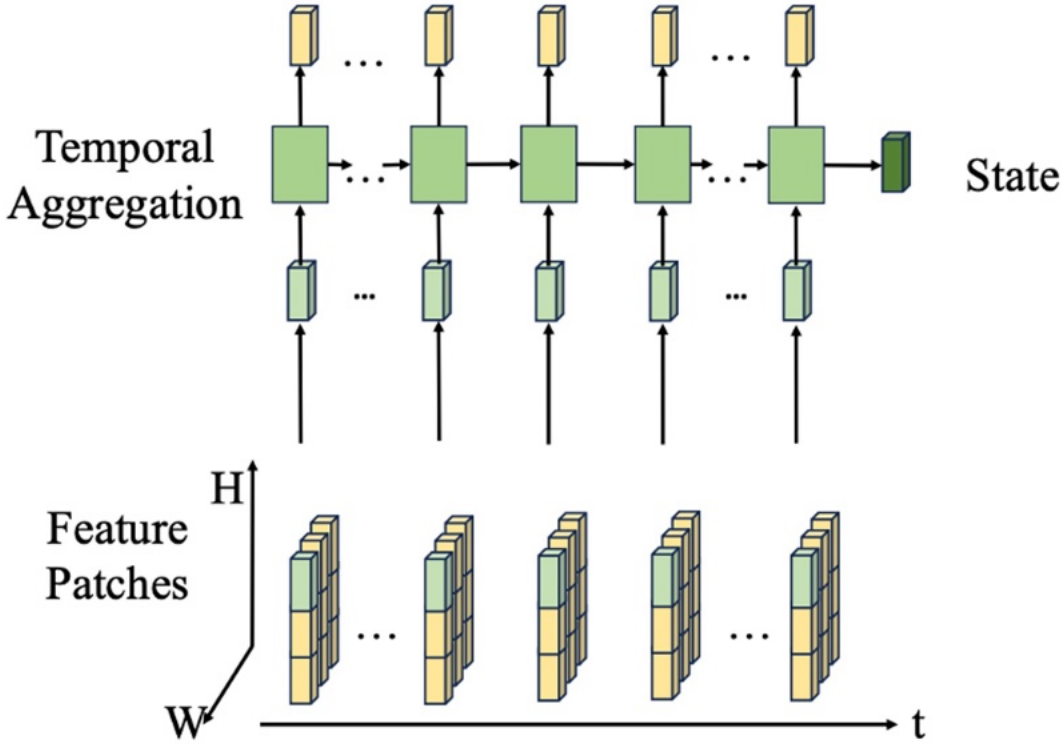
As soon as I throw away the trash, remind me to wash my hands.

[C. Eyzaguirre et al. Streaming Detection of Queried Event Start. NeurIPS 2024]

# Simple Architecture: CLIP plus temporal state



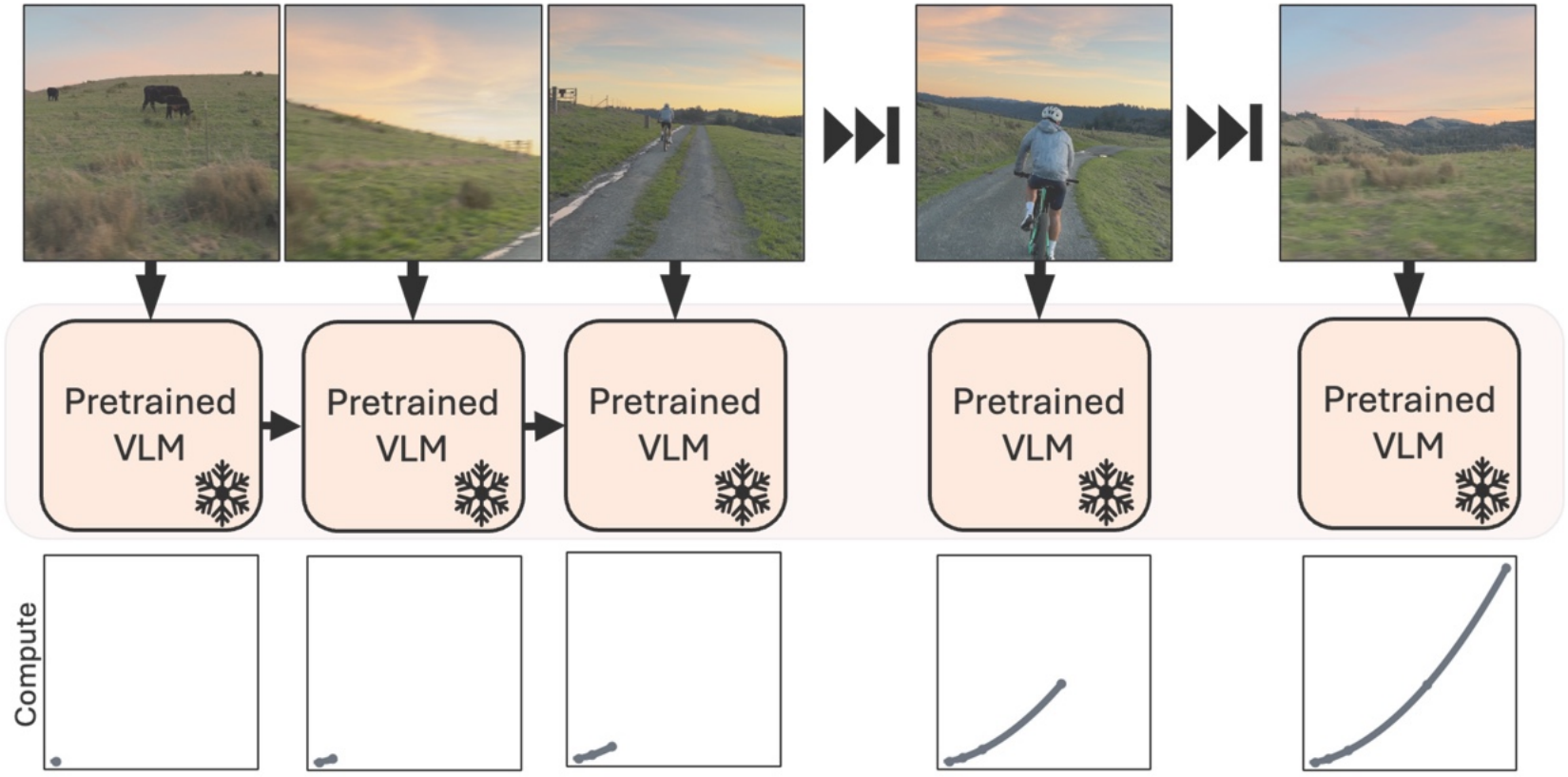

(a) ViT Block with Adapter




(b) Adapter Internals

[C. Eyzaguirre et al. Streaming Detection of Queried Event Start. NeurIPS 2024]

# VLMs for Video Understanding



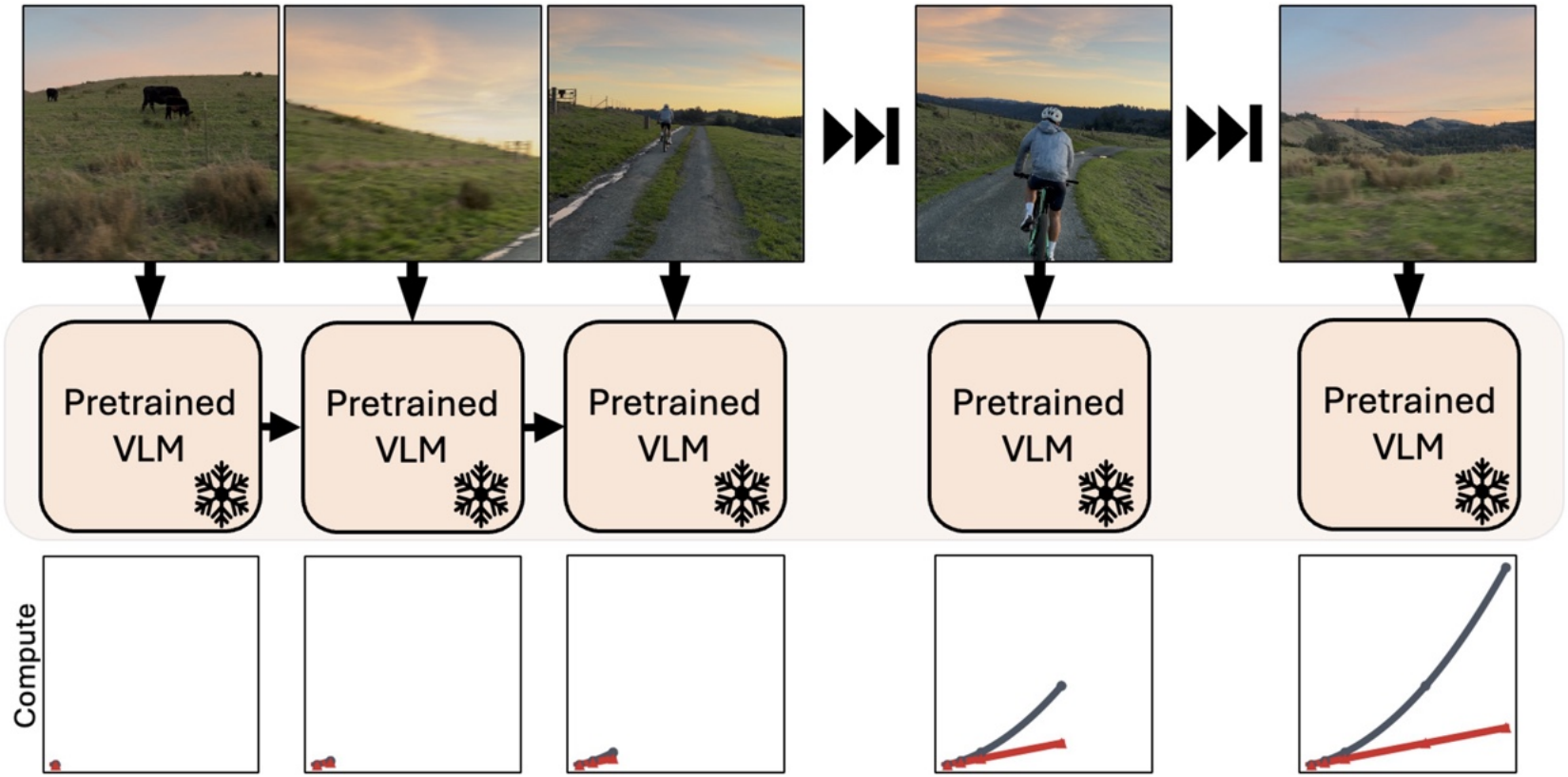

What we have

**$O(N^2)$**

*dense self-attention over video tokens*

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# StateKV: linear scaling of pretrained VLMs




What we have

**O(N²)**

*dense self-attention over video tokens*

What we need

**O(N)**

*constant cost per frame · linear total*

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# StateKV: linear scaling of pretrained VLMs

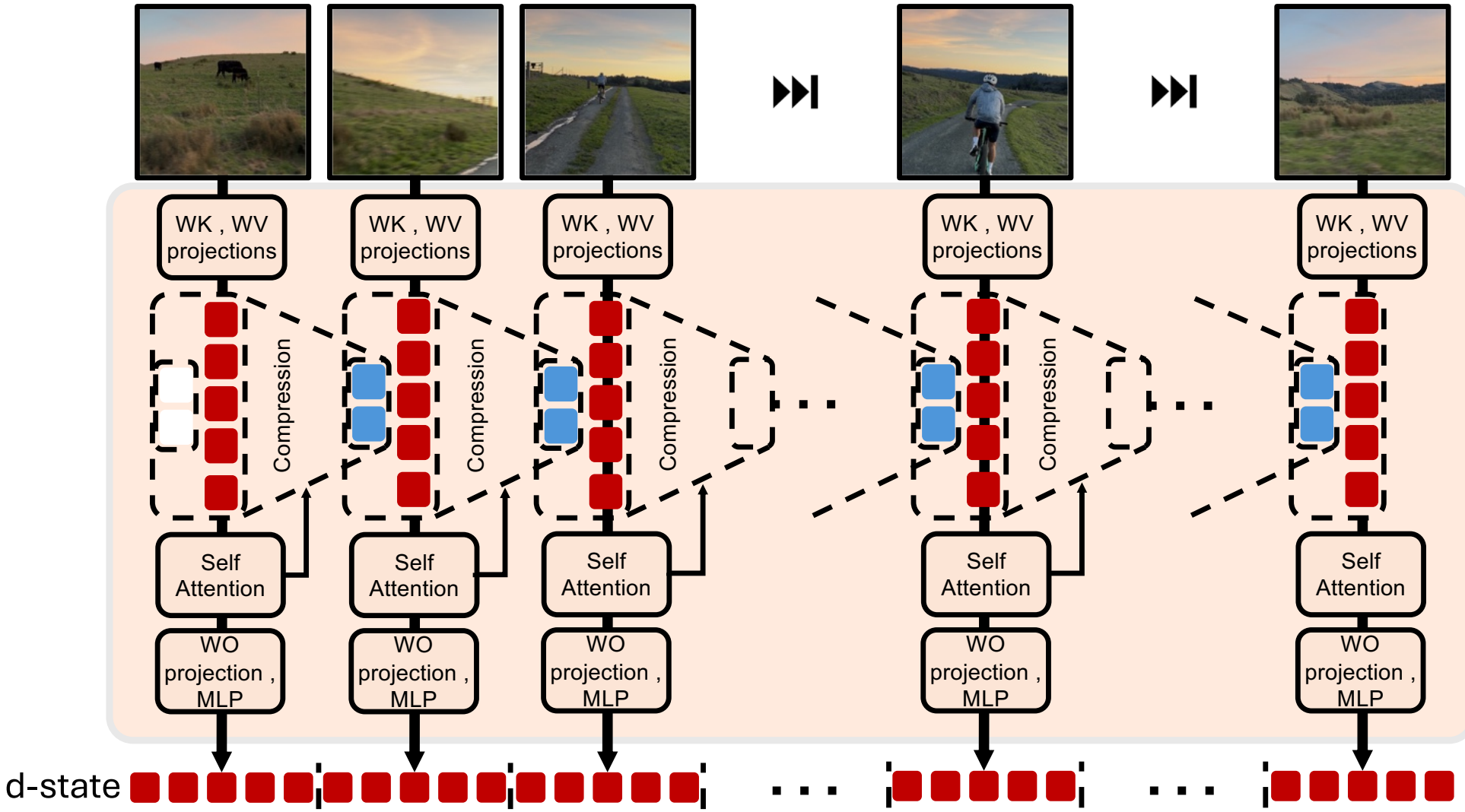
Two key assumptions that we verify empirically (see paper):

1. Most of the past does not matter
    a. most cross-frame attention is concentrated on a surprisingly small subset of tokens

2. The important tokens change slowly
    a. instead of recomputing an optimal memory from scratch, we can maintain and update a compact state as the video progresses

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# StateKV algorithm: compressed memory - cstate




[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# StateKV algorithm: update state recurrently



[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

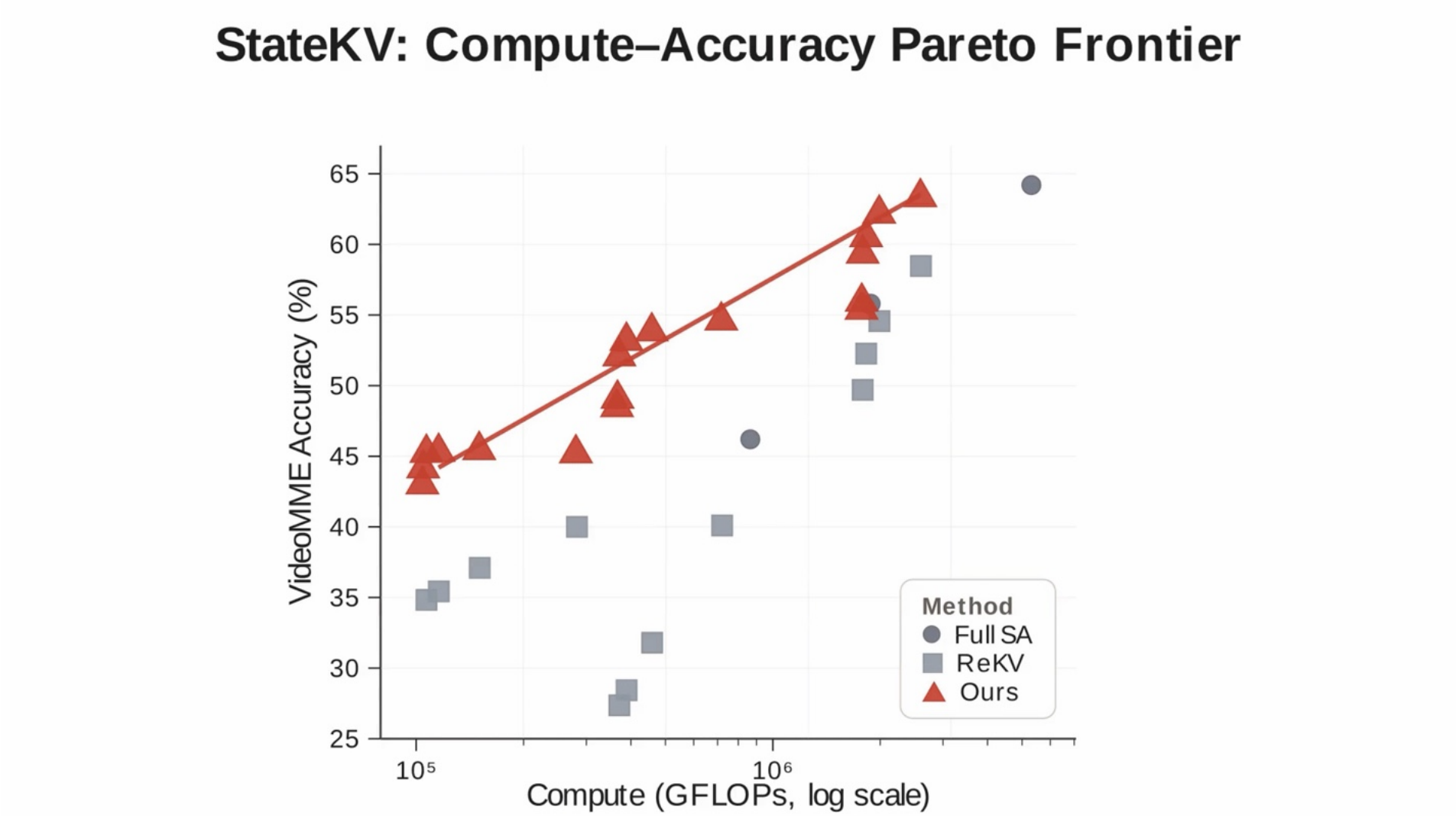
# StateKV algorithm: update state recurrently


[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# Accuracy–compute tradeoff

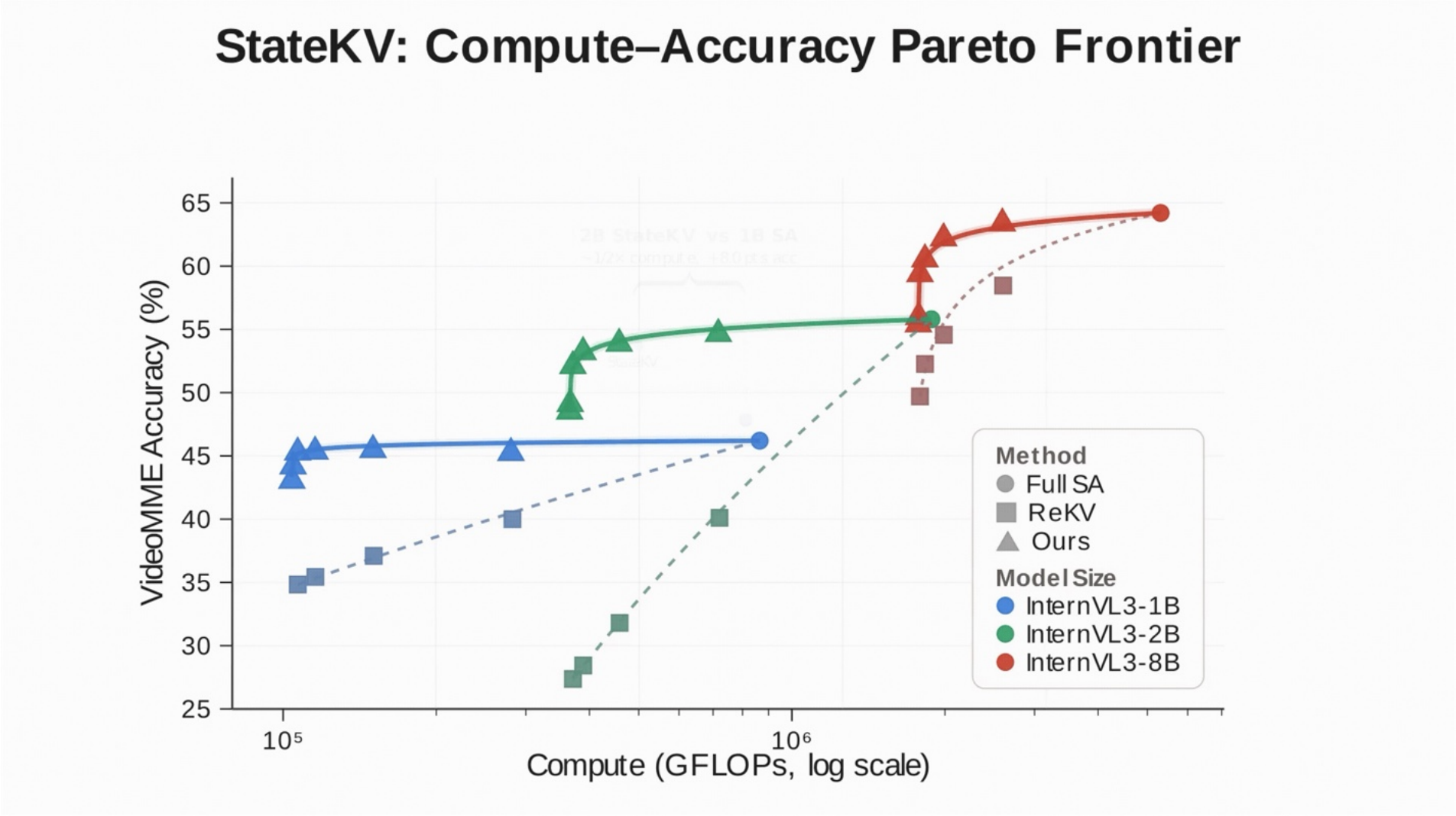
[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# Accuracy–compute tradeoff



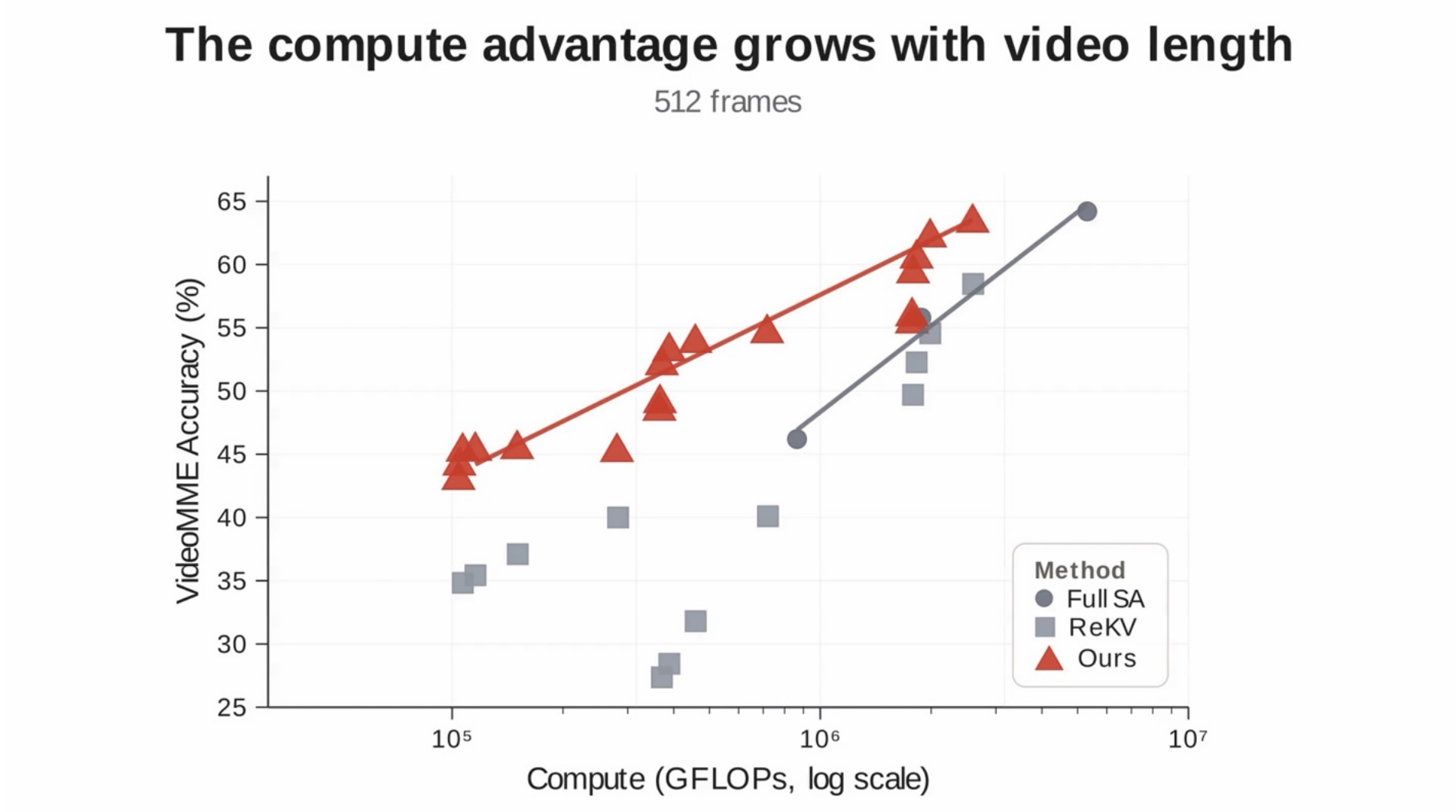

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# Accuracy–larger models with same compute




[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

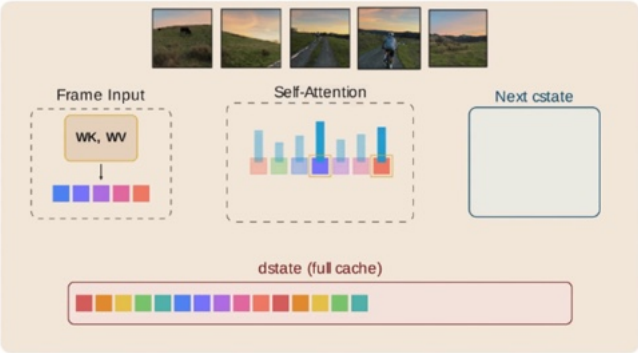
# The long-video regime




[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# StateKV - Summary

1. StateKV enables long-video VLM **inference to scale linearly**
2. It preserves most of full self-attention **performance**
3. Enables running larger models at a the cost of smaller models

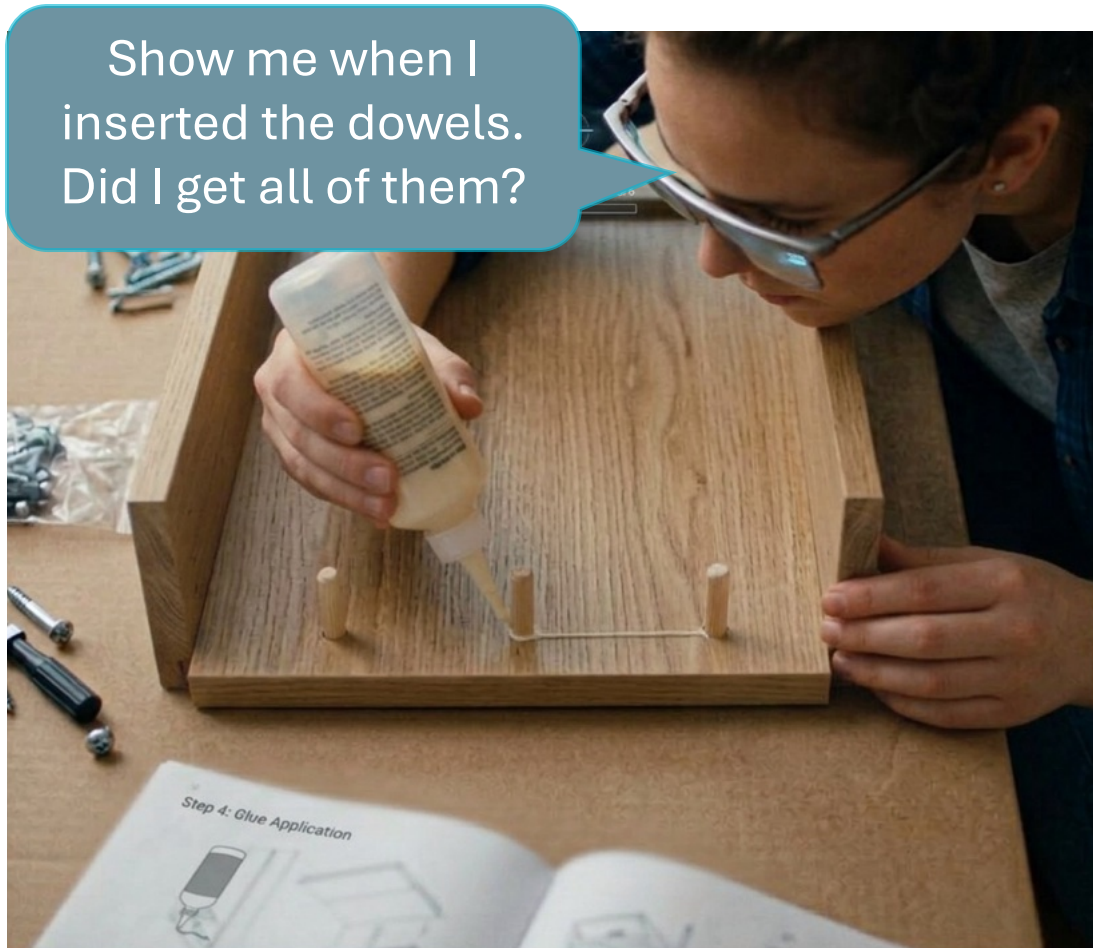


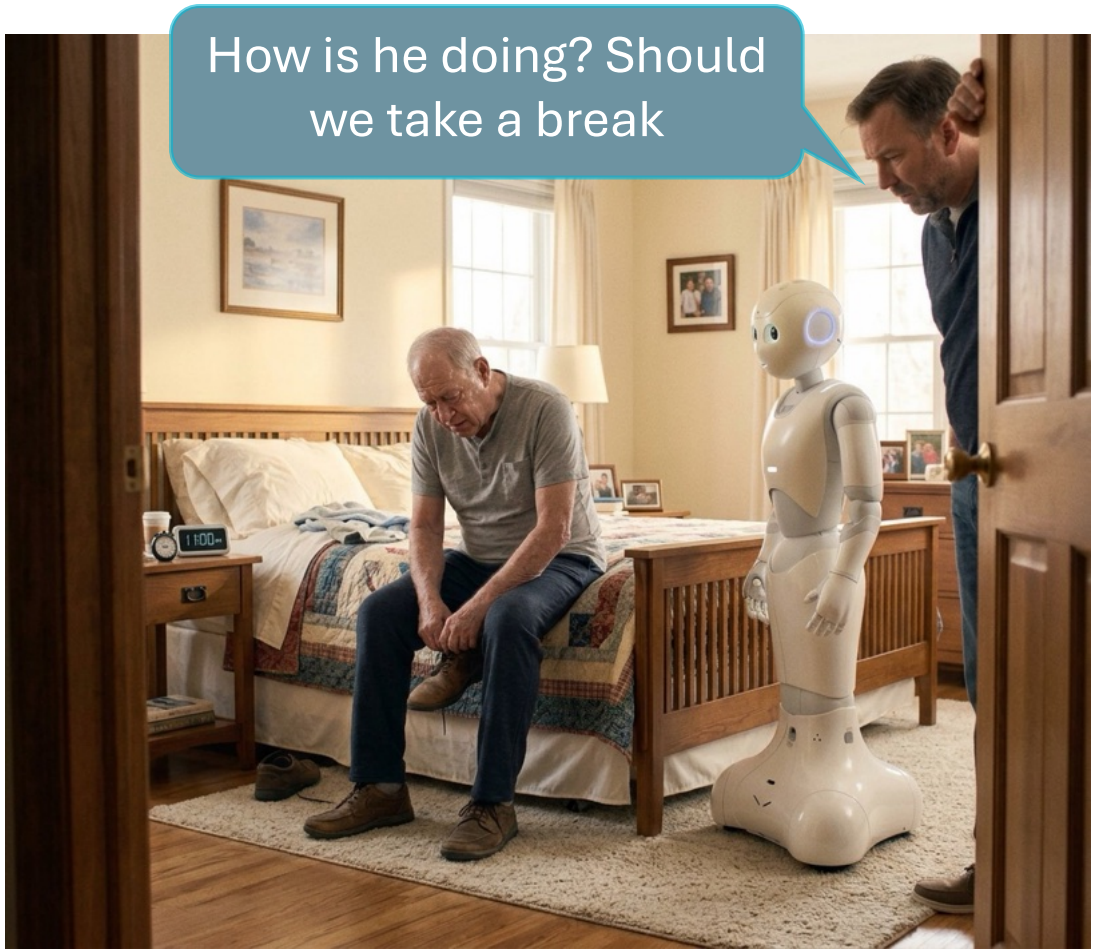

Blog / paper/ code

ceyzaguirre4.github.io/StateKV

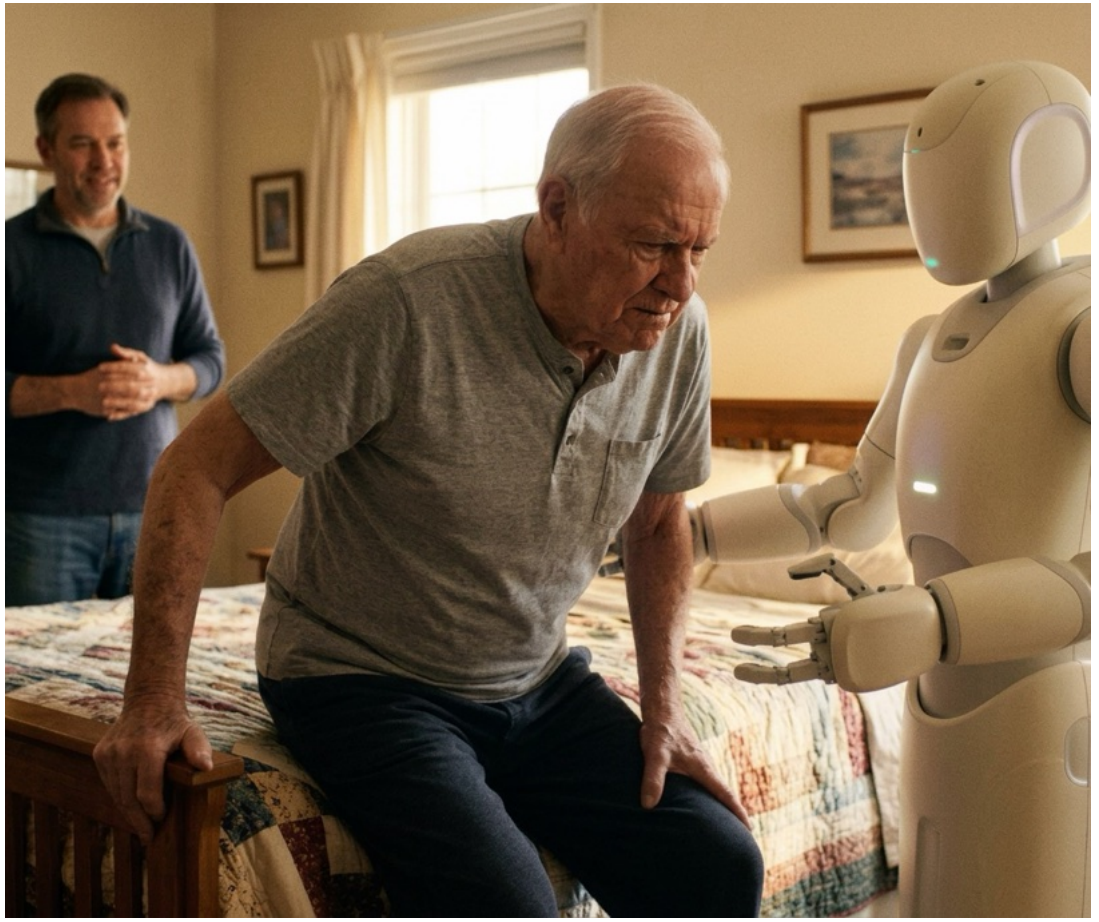
[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. ECCV 2026.]

# Part III: Efficiency



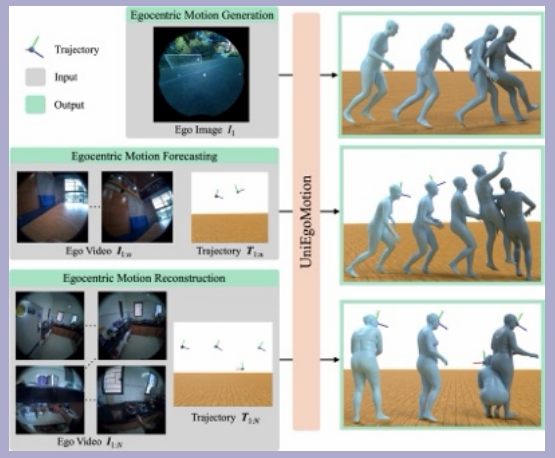
# Perception, Trust & Efficiency for AI in the Physical World



# Perception, Evidence & Efficiency for AI in the Physical World

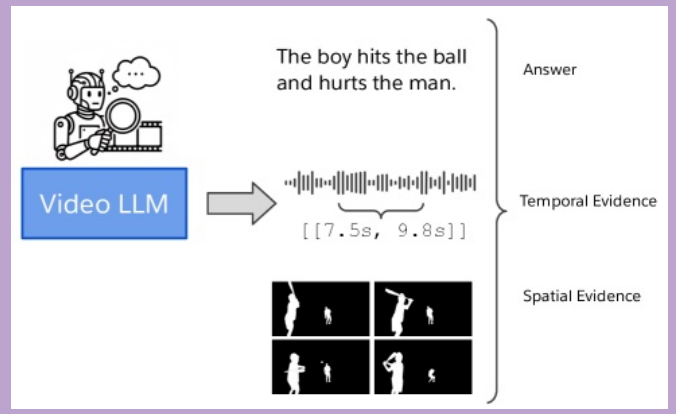
## Predictive Motion Understanding

[UniEgoMotion ICCV'25]



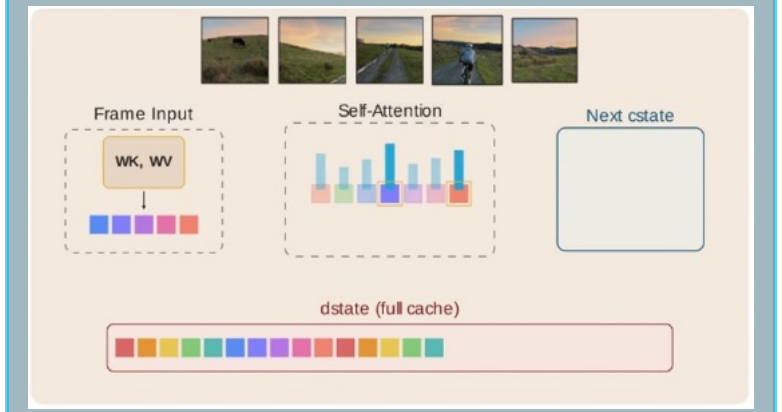

## Evidence-backed Reasoning

[E-VQA ECCV'26]




## Streaming Efficiency

[StateKV ECCV'26]



# Samsung Research America is Hiring

Real-world AI—from ambient intelligence to wearables to robotics—requires rethinking how AI works with edge devices and people's lives.

🏢 **Mountain View:** Head of Vision Intelligence

🏙️ **Toronto:** Head of Toronto AI Center

Scan the QR code to join us.

# Thank you!

🔗 **slides/code/data/models/papers**


1. Patel et al. **“UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation”**. ICCV 2025
2. Wang et al. **“Evidence-Backed Video Question Answering”**. ECCV 2026
3. Eyzaguirre et al. **“Linear Scaling Video VLMs for Long Video Understanding”**. ECCV 2026