# Agentic Ambient Intelligence

## Perception, Reasoning & Action

CV4Smalls Workshop at CVPR 2026
June 3, 2026

**Juan Carlos Niebles**
Director, AI Research
www.niebles.net
@jcniebles

# Solving Headaches

# Agentic Ambient Intelligence


# Part I

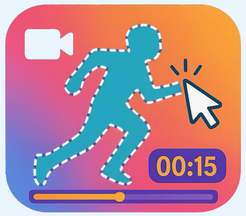

Should *this* board be installed horizontally, like the one I *just finished*?
Matched to previous installation type (Horizontal)
?


Wrap *this* cup and the ones I washed *5 minutes ago*.
KITCHEN

# Agentic Ambient Intelligence

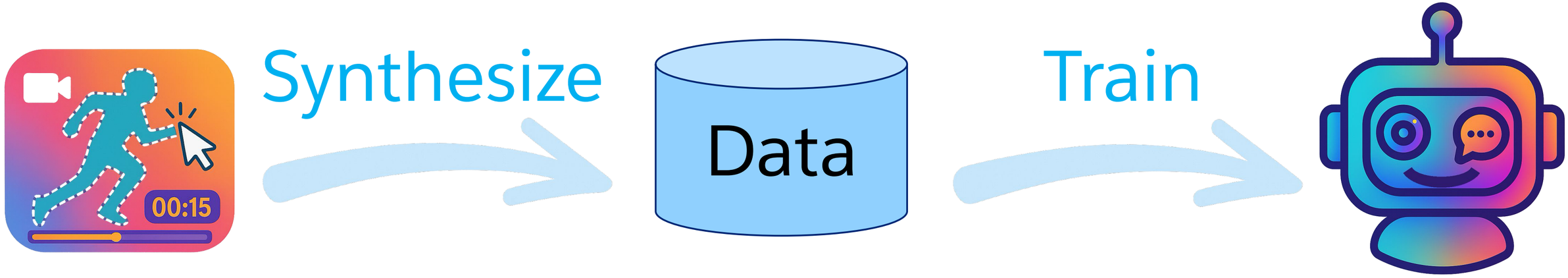
**Video QA with Space-time references**

**Strefer**
[ICCVW'25]



# Our work: Strefer

*sampled video frames*

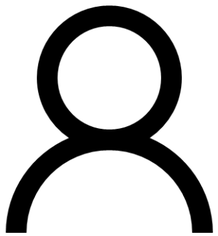

Strefer synthesizes instruction tuning data to empower Video LLMs to better interpret space-time information.

[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Strefer: new capabilities

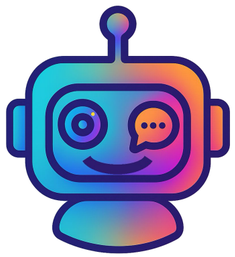
*sampled video frames*


interpret user queries with mask references



:

Could you describe his action in the video?



: He runs, jumps to catch the football with both hands, walks briefly holding it in one hand, then throws it.

[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Strefer: new capabilities

*sampled video frames*


interpret user queries with timestamp references

: What is happening to the boy wearing glasses from 00:03 to 00:09?

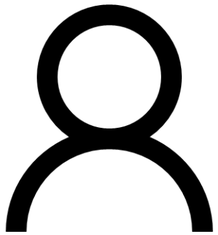
: The boy wearing glasses is not visible from 00:03 to 00:09.

[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Strefer: new capabilities

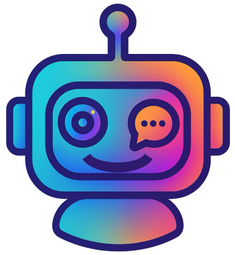
*sampled video frames*


interpret user queries with mask and timestamp references



:

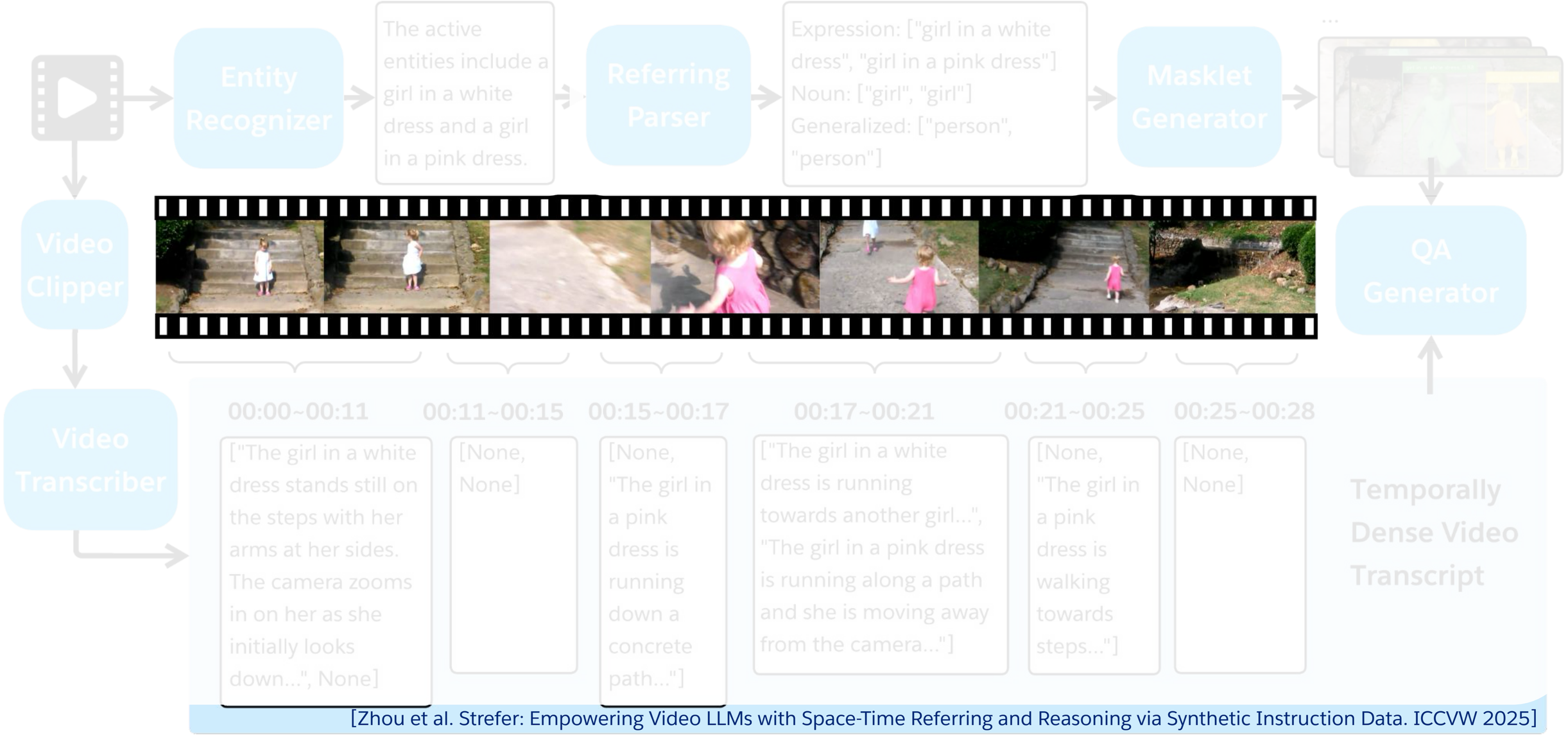

What did he do at 00:10?



: He runs, jumps to catch the football with both hands, walks briefly holding it in one hand, then throws it.

[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Data Synthesis Pipeline




[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Strefer-Synthesized Instruction Data: QA pairs

Strefer Input:


**Strefer output**

Instruction:

*(mask boundary visualized)*

What was the person doing between 00:51 and 01:07 in the video?

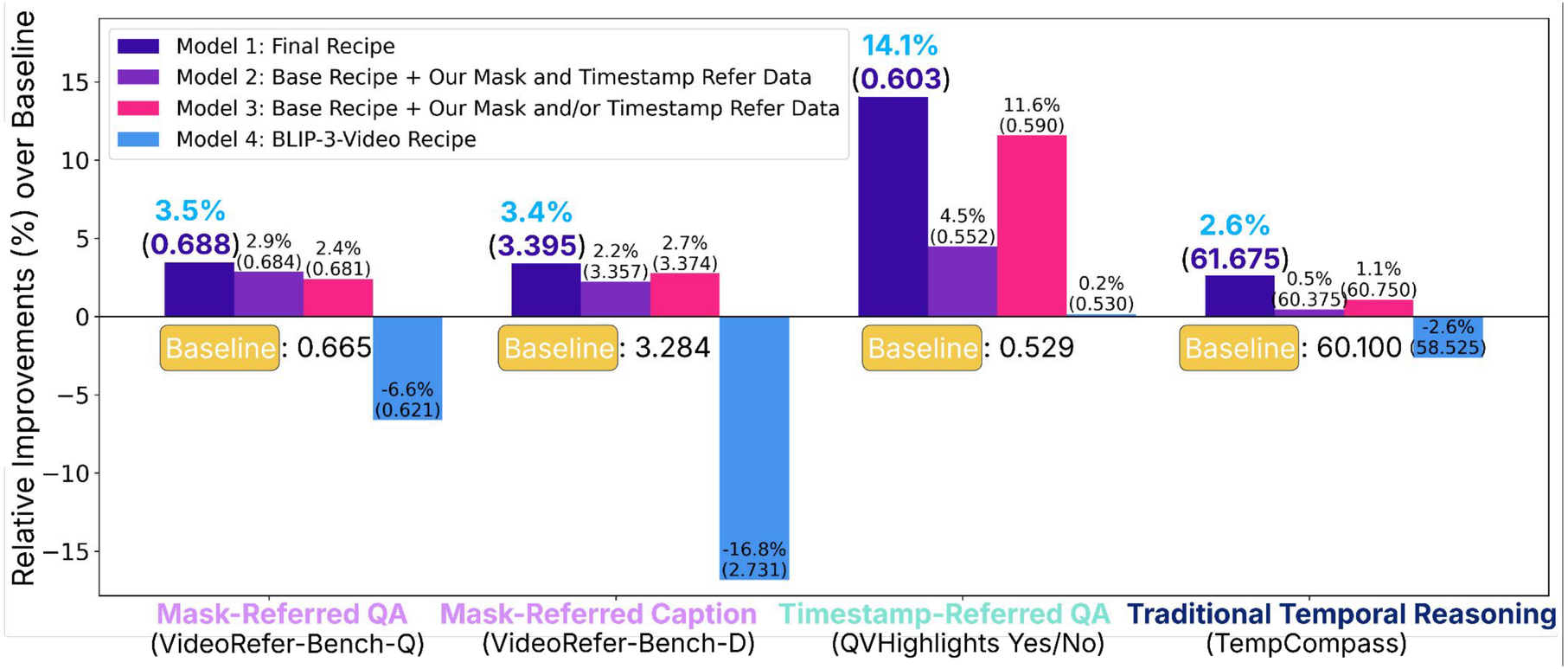
Response: Interacting with a toy guitar, moving it from a basket to the floor.

[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Model improvements

**947K+** samples generated from 4.2K NExT-QA videos.

**Just 545 new videos beyond baseline** boosted performance across benchmarks.




[Zhou et al. Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data. ICCVW 2025]

# Strefer: Summary

🔥 Key features:

▶️ 1. **Scalable**: Fully automatic, no reliance on legacy annotations.

▶️ 2. **Fine-grained** & **Space-time grounded**: Grounded metadata + Instruction data w/ multimodal prompts

▶️ 3. A modular system with a novel Referring Masklet Generation pipeline

https://strefer.github.io/

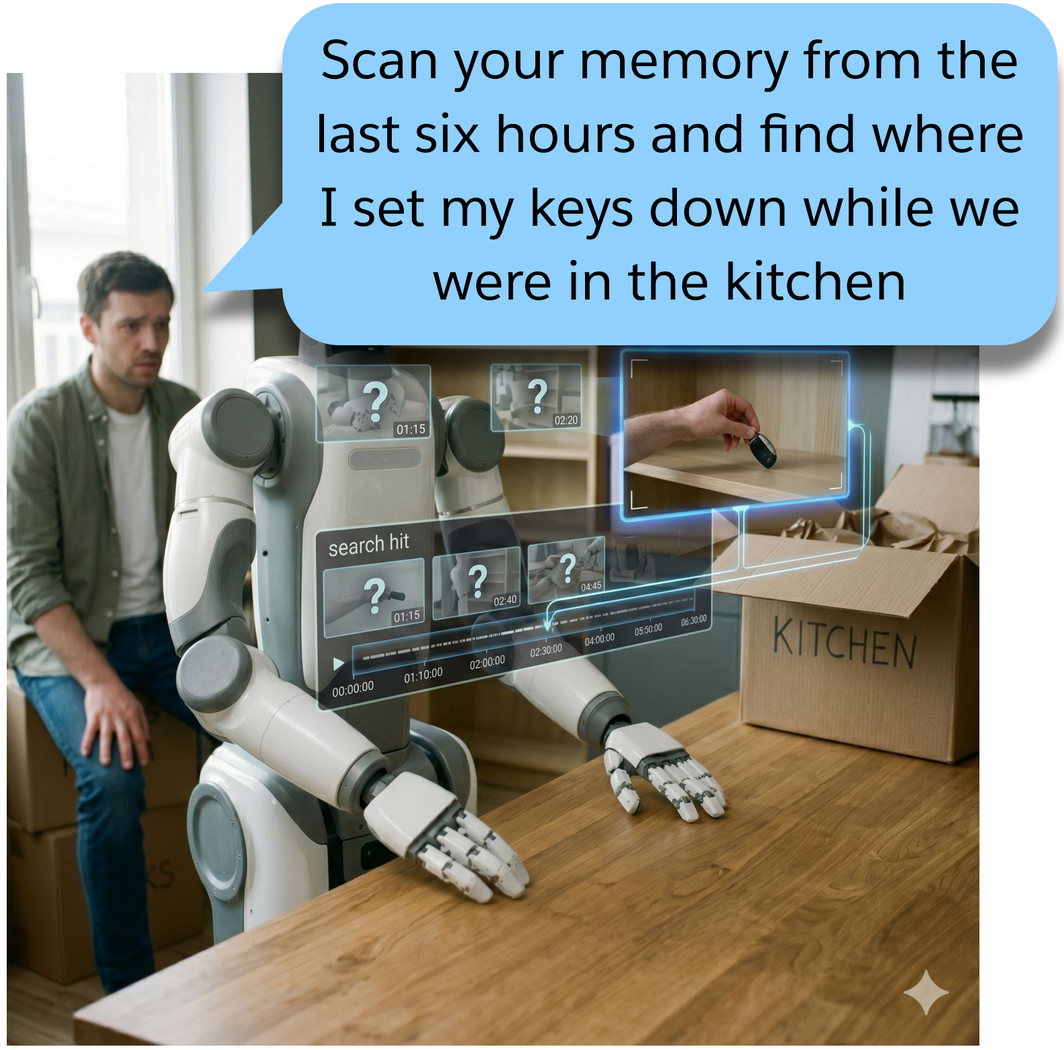
YouTube

Walkthrough

# Part I: Video QA with Space-time references


# Part II

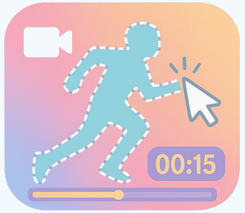

I have an extra dowel. Scan the last hour and tell me exactly which step I missed?
Step 4: Inserting Dowels 1-4
Step 4: Inserting Dowels 1-4
Step 9: Attaching Dowels 1-4
Step 9: Attaching drawer slides
Found missing dowel insertion point: Step 8, Part B. Visual Signature Match.


Scan your memory from the last six hours and find where I set my keys down while we were in the kitchen
search hit
01:15
02:20
02:40
04:45
00:00:00
01:10:00
02:00:00
02:30:00
04:00:00
05:50:00
06:30:00
KITCHEN

# Agentic Ambient Intelligence

Video QA with Space-time references

**Strefer**
[ICCVW'25]




**Reasoning over long videos**

**AVP**
[CVPR Findings '26]

# Existing Caption-based Agentic Frameworks

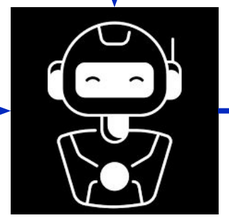
Step 1 : Passive perception via query-agnostic captioner




Caption | Caption | Caption | Caption | Caption → Caption Database

*Low efficiency, massive captioner computation overhead*

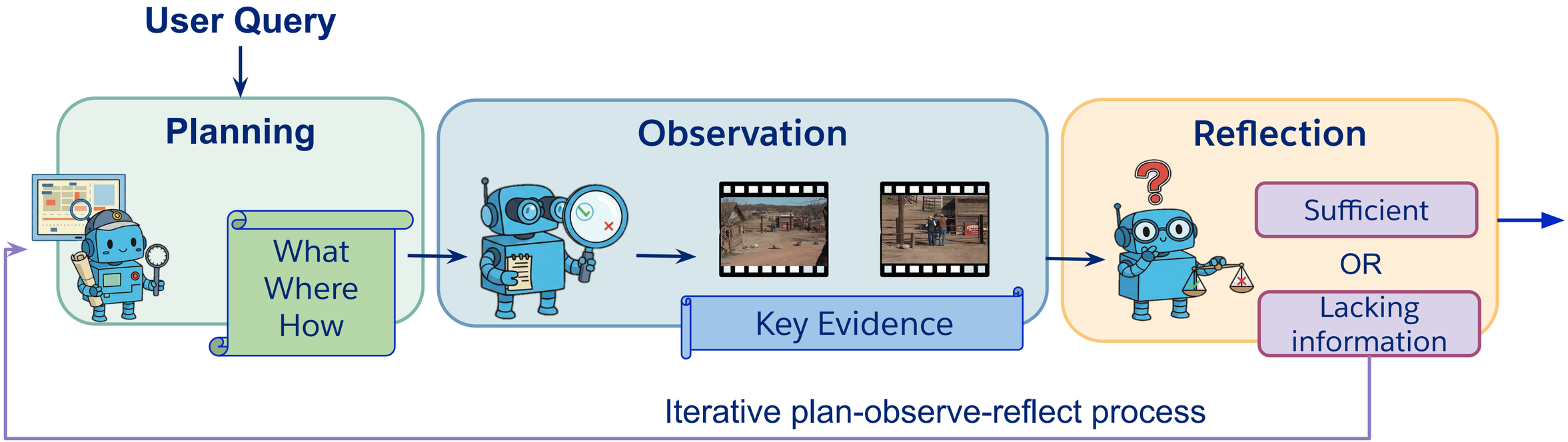
Step 2 : Evidence searching via caption database

Query →  → Localize Key Segment → Answer

*Imprecise Grounding based on text caption, losing visual details*

[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

# Our Work: Active Video Perception (AVP)

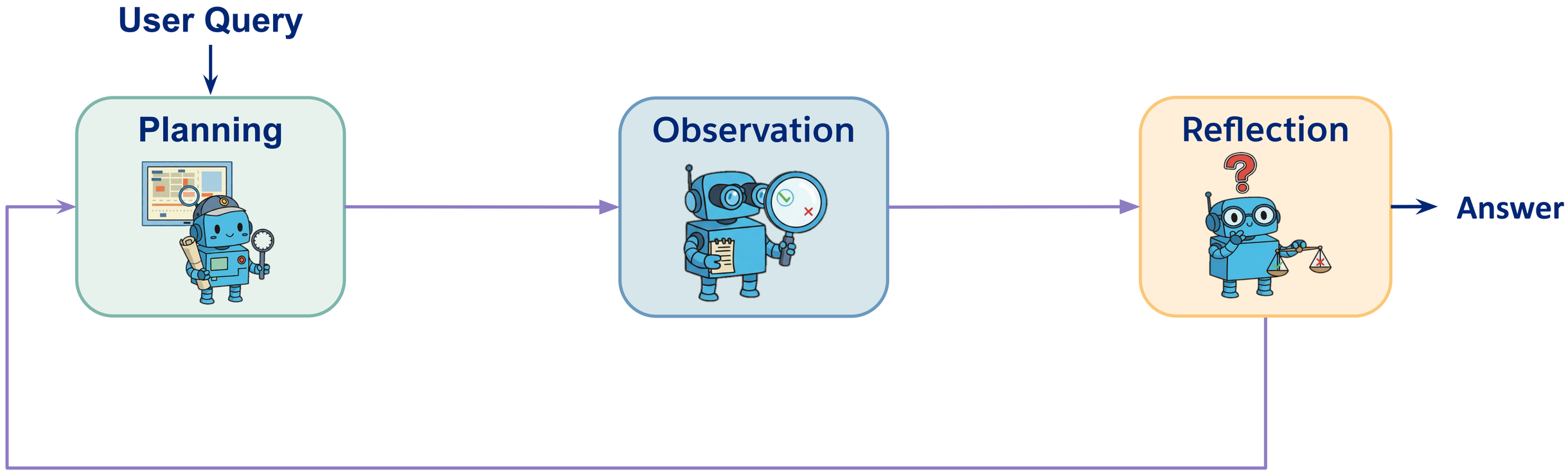



***Active* Evidence Seeking**

**Iterative Perception via Reflection**

[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

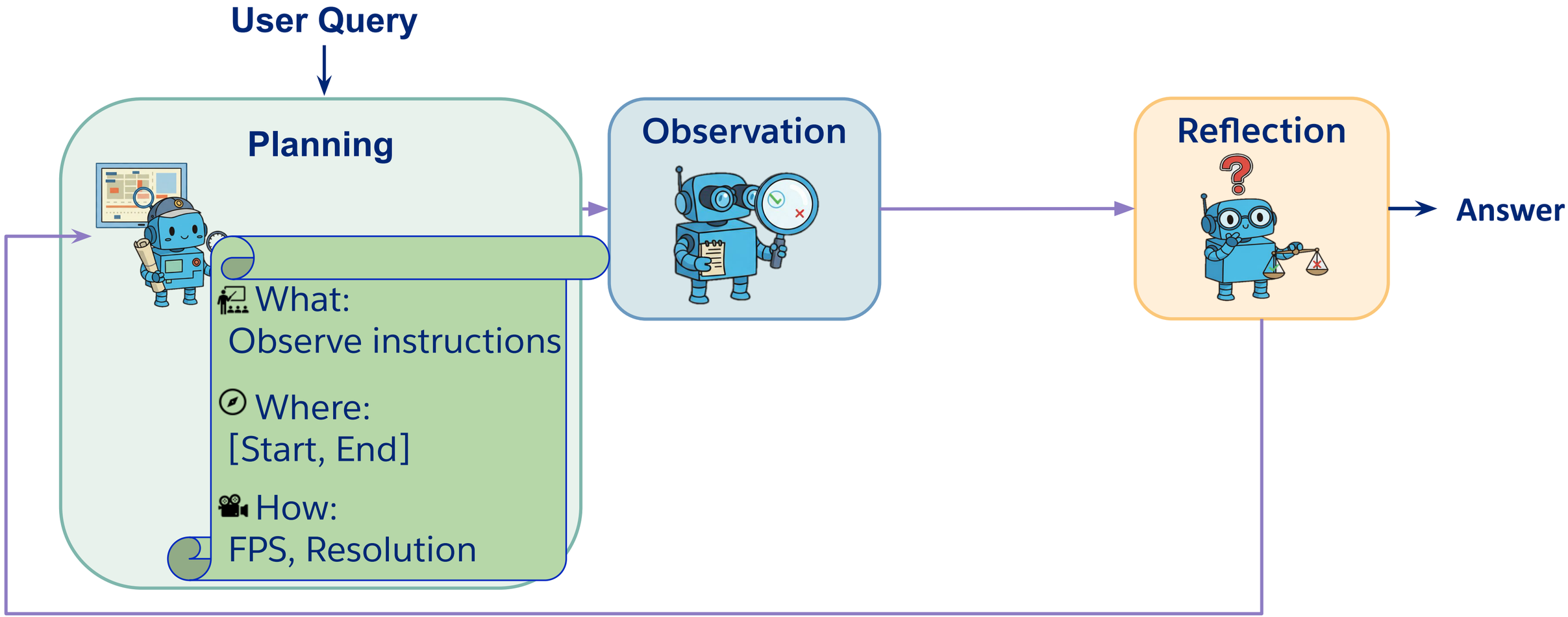
# Our Work: Active Video Perception (AVP)




[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

# Our Work: Active Video Perception (AVP)

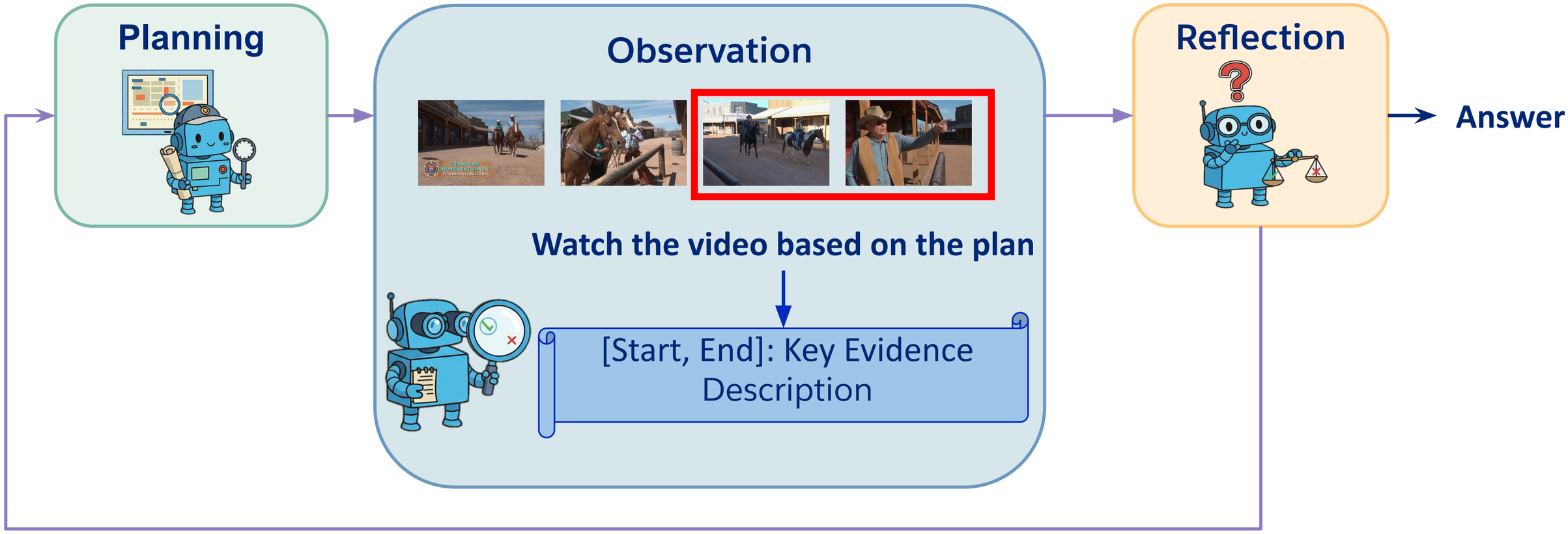
## Step 1 : Generate Plan




[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

# Our Work: Active Video Perception (AVP)

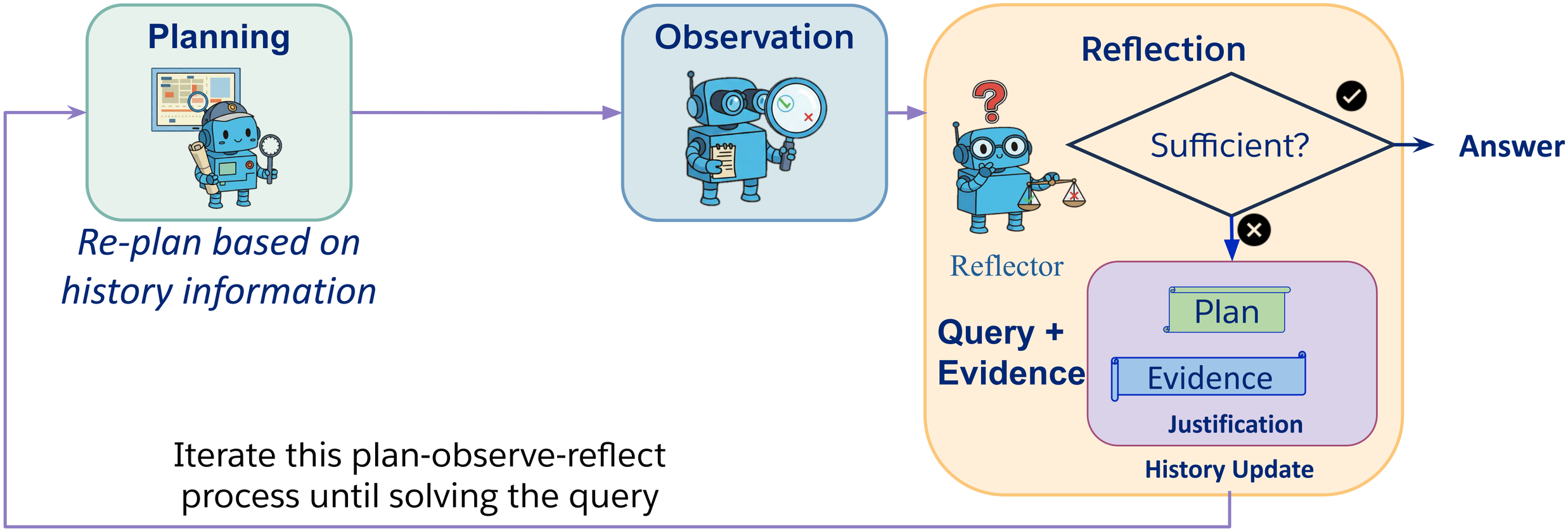
## Step 2 : Targeted Observation




[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

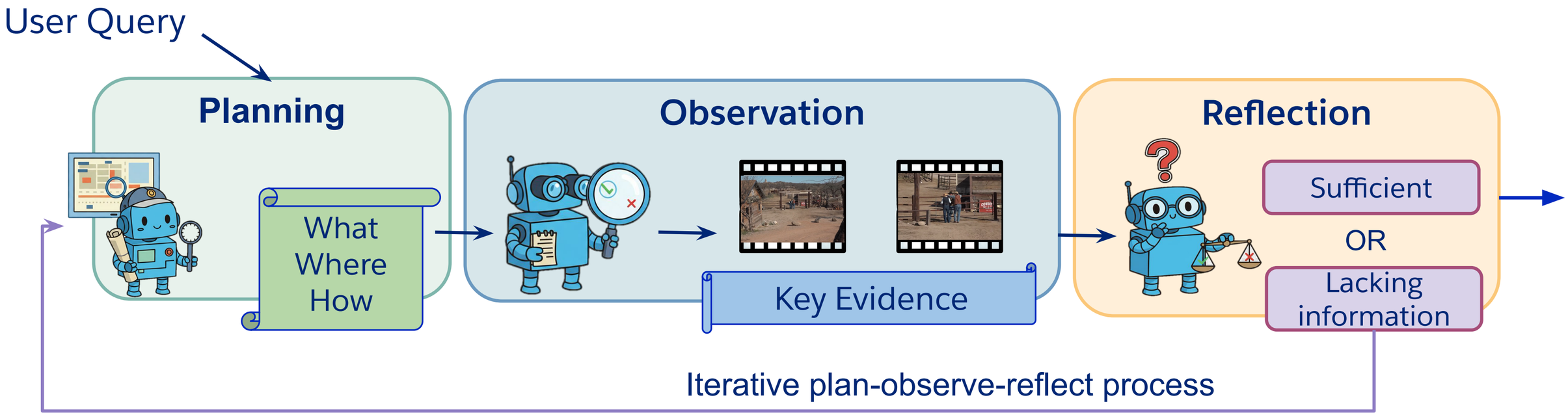
# Our Work: Active Video Perception (AVP)

## Step 3: Reflect on evidence




[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

# Our Work: Active Video Perception (AVP)



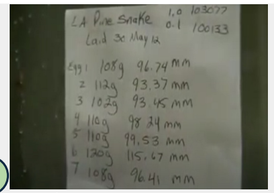

***Active* Evidence Seeking**

**Iterative Perception via Reflection**

[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

Query: After adding up all the millimeter totals on the sheet of paper illustrated at the timestamp 09:58, and then adding the average length of Louisiana Pine Snake hatchlings according to the video, how many total millimeters are there?
A. 2,217.41mm-4,130.04mm. B. 1,263.41mm-2,217.41mm. C. 693.41mm-1,912.63mm.
D. 1,912.63mm-2,217.41mm. E. 4,130.04mm-4,530.04mm.

## Round 1



**Plan**
**What**: extract the millimeter totals from the sheet of paper at 09:58 and find the average length of Louisiana Pine Snake hatchlings.
**Where**: [09:58, 09:59]
**How**: 2FPS, Medium spatial resolution

**Observe**

[09:58, 09:59]: A close-up of a handwritten sheet of paper listing measurements for seven Louisiana Pine Snake eggs. The millimeter totals listed are: 96.74, 93.37, 93.45, 98.24, 99.53, 115.67, and 96.41.

**Evidence**

**Reflect**

**Confidence**: 0.3
**Justification**: Need additional information about average length of Louisiana Pine Snake hatchlings

## Round 2

History (P, E, J)




**Re-Plan**
**What**: Perform a uniform scan of the entire video to find the 'average length of Louisiana Pine Snake hatchlings'.
**Where:** whole video
**How**: 0.5FPS, low spatial resolution

**Observe**

[Previous Evidence]

[17:47, 17:50]: The narrator states that the Louisiana Pine Snakes 'usually range from 4 to 5 feet in length.

**Reflect**

**Confidence**: 0.7
**Justification**: Lower bound: 693.41 mm + 1219.2 mm = 1912.61 mm - Upper bound: 693.41 mm + 1524 mm = 2217.41 mm
The calculated total range is 1912.61 mm to 2217.41 mm. This corresponds to option **D**.

[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

# AVP: Quantitative Results

| Methods | MINERVA | LVBench | MLVU | Video-MME | | LongVideoBench | |
|---|---|---|---|---|---|---|---|
| | *Overall* | *Overall* | *Test* | *Overall* | *Long* | *Val* | *Long* |
| Qwen-3-VL | - | 67.7 | 84.3 | 79.2 | - | - | - |
| GPT-4o | 45.5 | 48.9 | 54.9 | 71.9 | 65.3 | 66.7 | 60.9 |
| Gemini-2.5-Flash | 54.6 | 56.7 | 72.4 | 74.2 | 69.1 | 66.2 | 61.8 |
| Gemini-2.5-Pro | <u>61.8</u> | 67.4 | 79.6 | <u>82.4</u> | 77.6 | 69.8 | 66.6 |
| VideoAgent | - | 29.3 | 64.4 | - | 46.4 | - | - |
| VideoTree | 40.2 | 28.8 | 60.4 | 60.6 | 54.2 | - | - |
| SiLVR | 44.4 | - | 45.2 | 74.1 | <u>77.7</u> | - | - |
| VideoLucy | - | 58.8 | 76.1 | 72.5 | 66.8 | - | - |
| DeepVideoDiscovery | - | <u>74.2</u> | - | - | 67.3 | <u>71.6</u> | 68.6 |
| *Active Video Perception (Ours)* | | | | | | | |
| AVP w Gemini-2.5-Flash | 56.9 (+2.3) | 63.8 (+7.1) | 74.1 (+1.7) | 81.2 (+7.0) | 76.7 (+7.6) | 70.2 (+4.0) | 65.5 (+3.7) |
| AVP w Gemini-2.5-Pro | 65.6 (+3.8) | 74.8 (+7.4) | 84.3 (+4.7) | 85.3 (+2.9) | 81.9 (+4.3) | 73.4 (+3.6) | 70.0 (+3.4) |

[Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.]

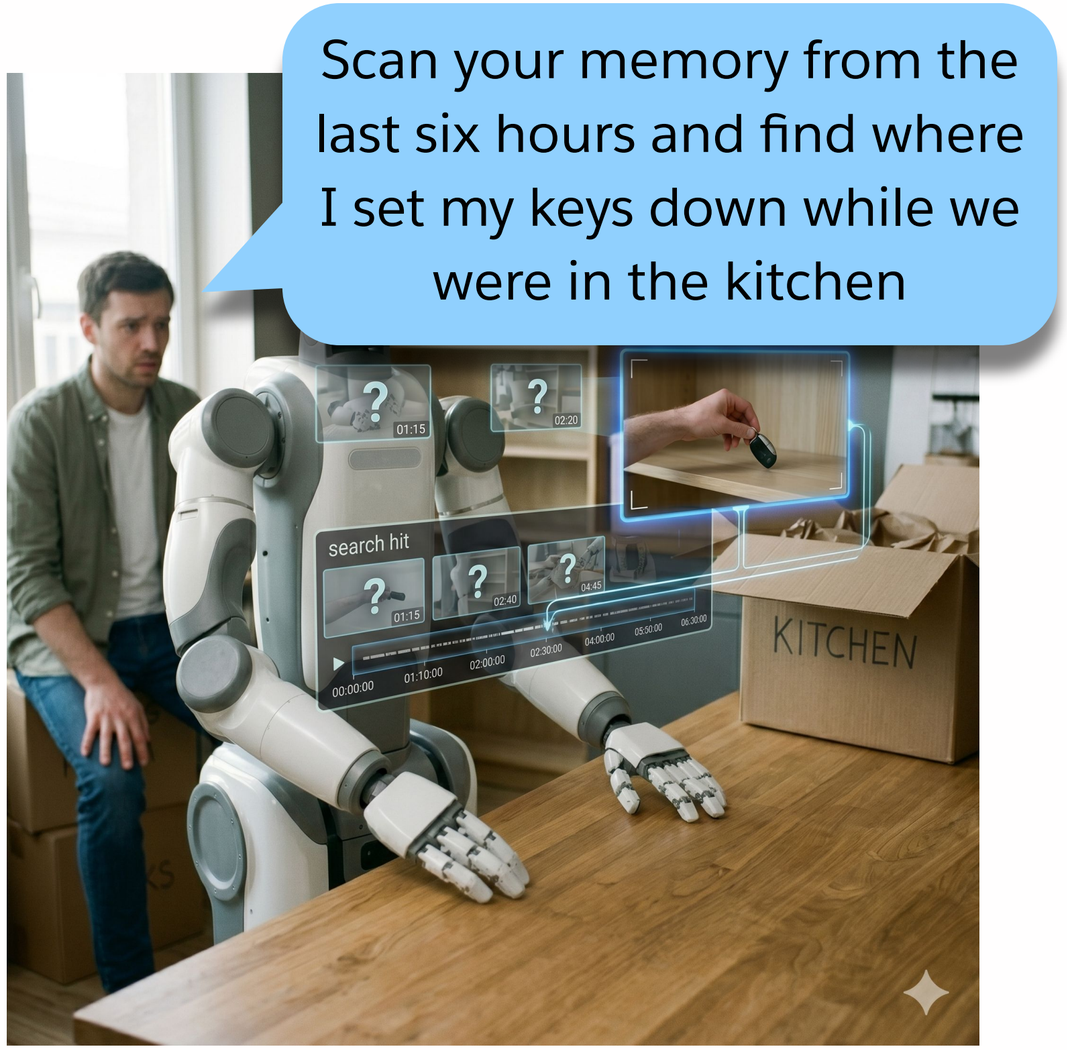
# AVP - Summary

- Active “Plan -> Observe -> Reflect” loop
- AVP achieves highest overall accuracy with significant improvements.
- AVP outperforms the best agentic method while requiring much less inference time and fewer input tokens.

**Poster - CVPR Findings**
**Sunday June 7th AM**

## Project page

# Part II: Active Perception

# Part III

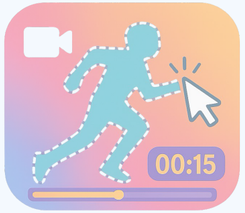

Remind me to wash my hands when I finish assembling this drawer.
Verify Part B
Verify Part B


Robot, come give me a hand as soon as I lift the next heavy object.
KITCHEN
BOOKS

# Agentic Ambient Intelligence

## Video QA with Space-time references

**Strefer**
[ICCVW'25]

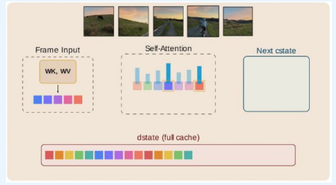



## Reasoning over long videos

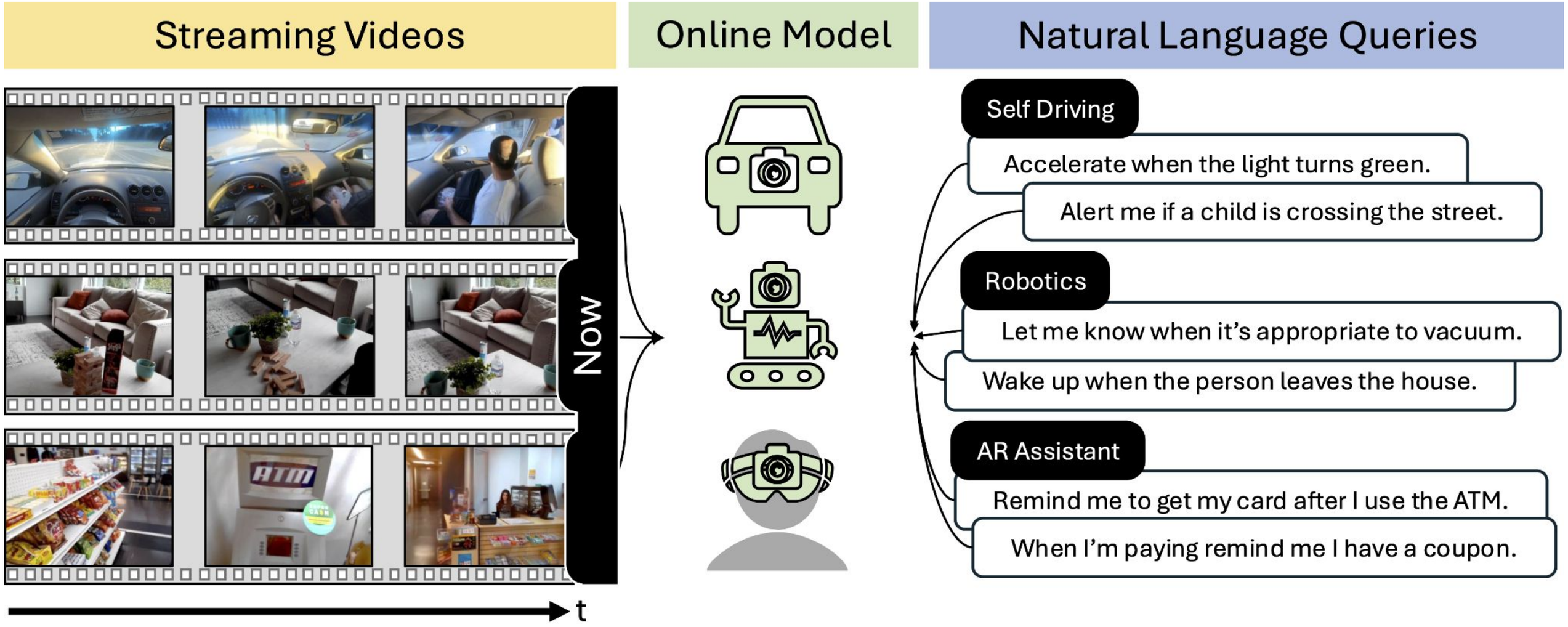
**AVP**
[CVPR Findings '26]

## Efficient Inference for Streaming Videos

**StateKV**
[arxiv'26]

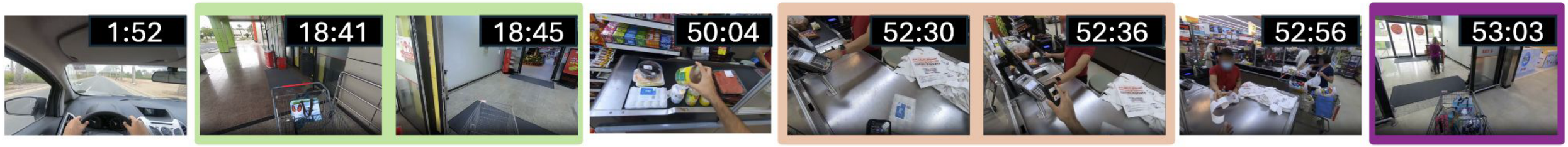


# Streaming Detection of Queried Event Start




[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.

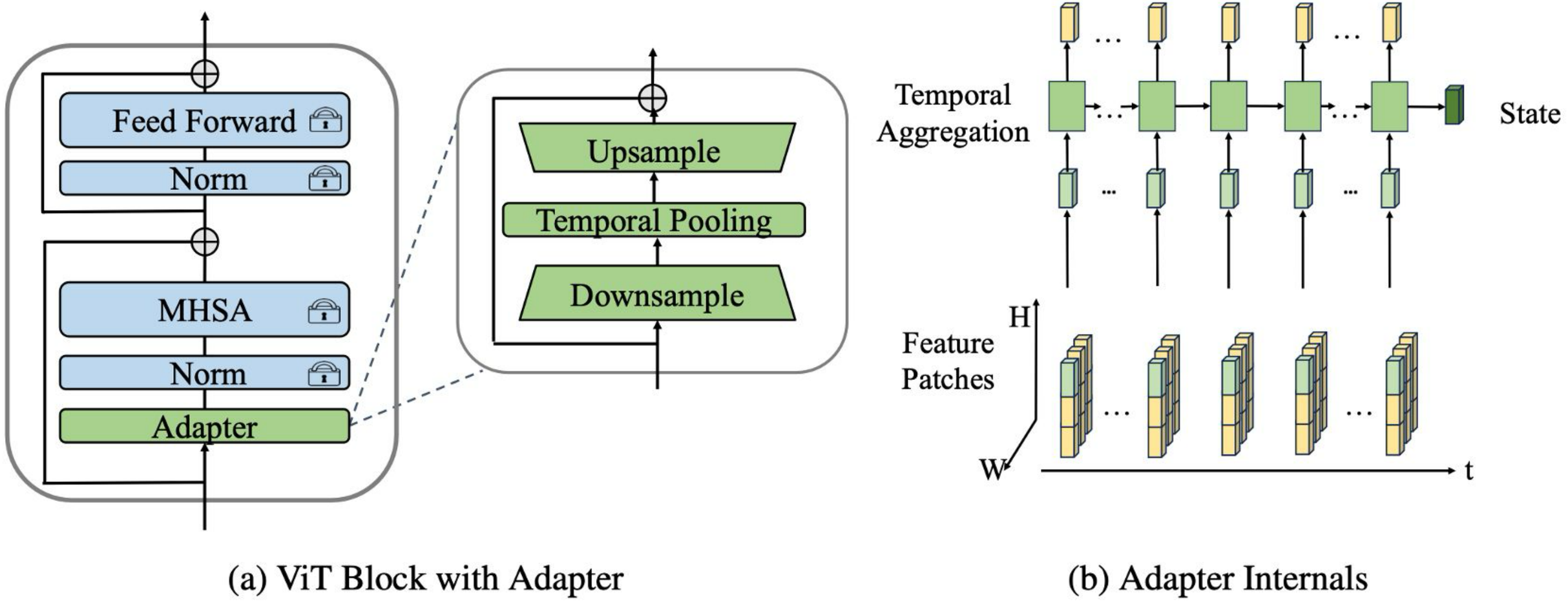

Next time I start doing my exercises, please remind me to drink some water.

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

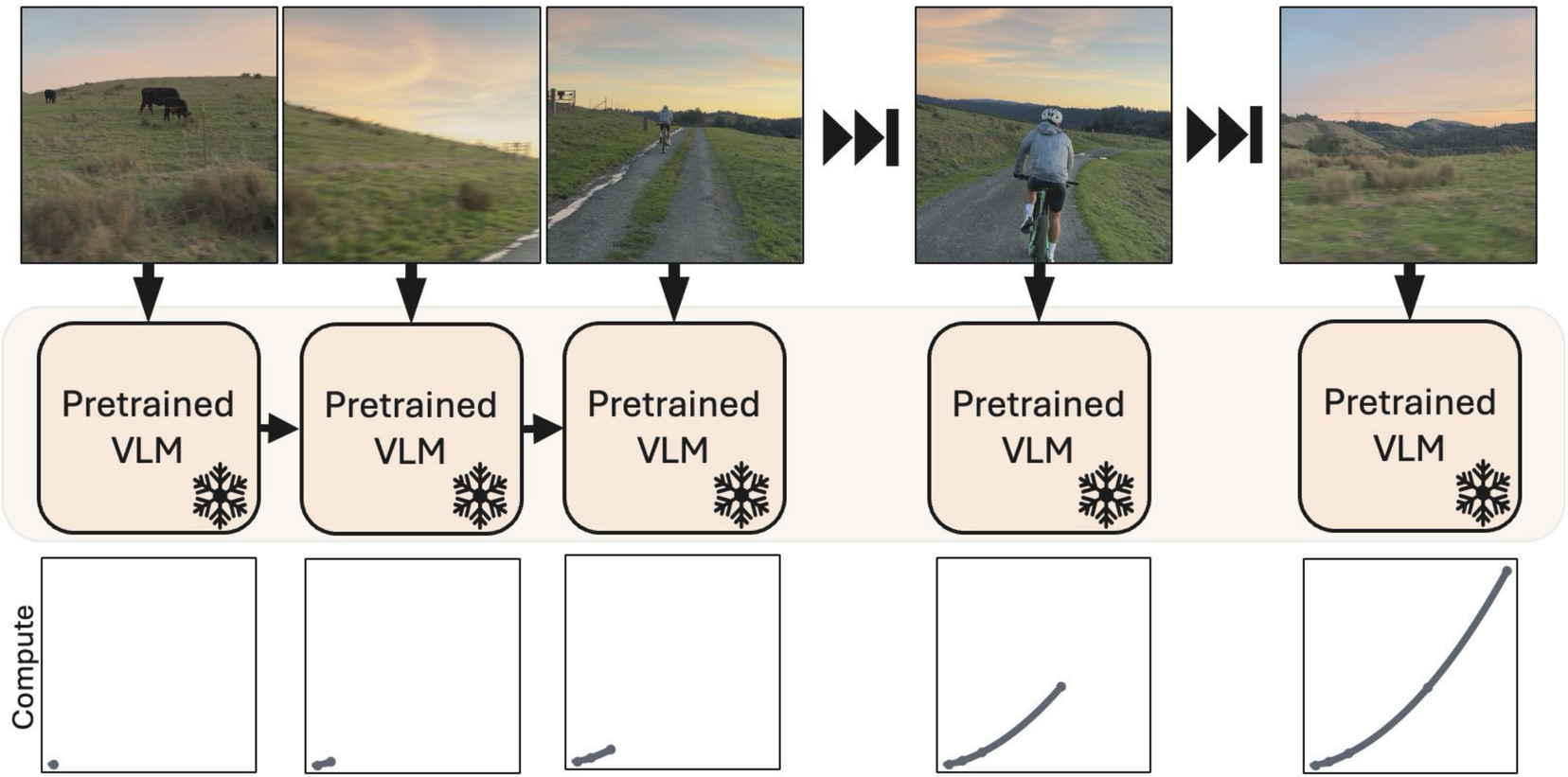



(a) ViT Block with Adapter

(b) Adapter Internals

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

# VLMs for Video Understanding



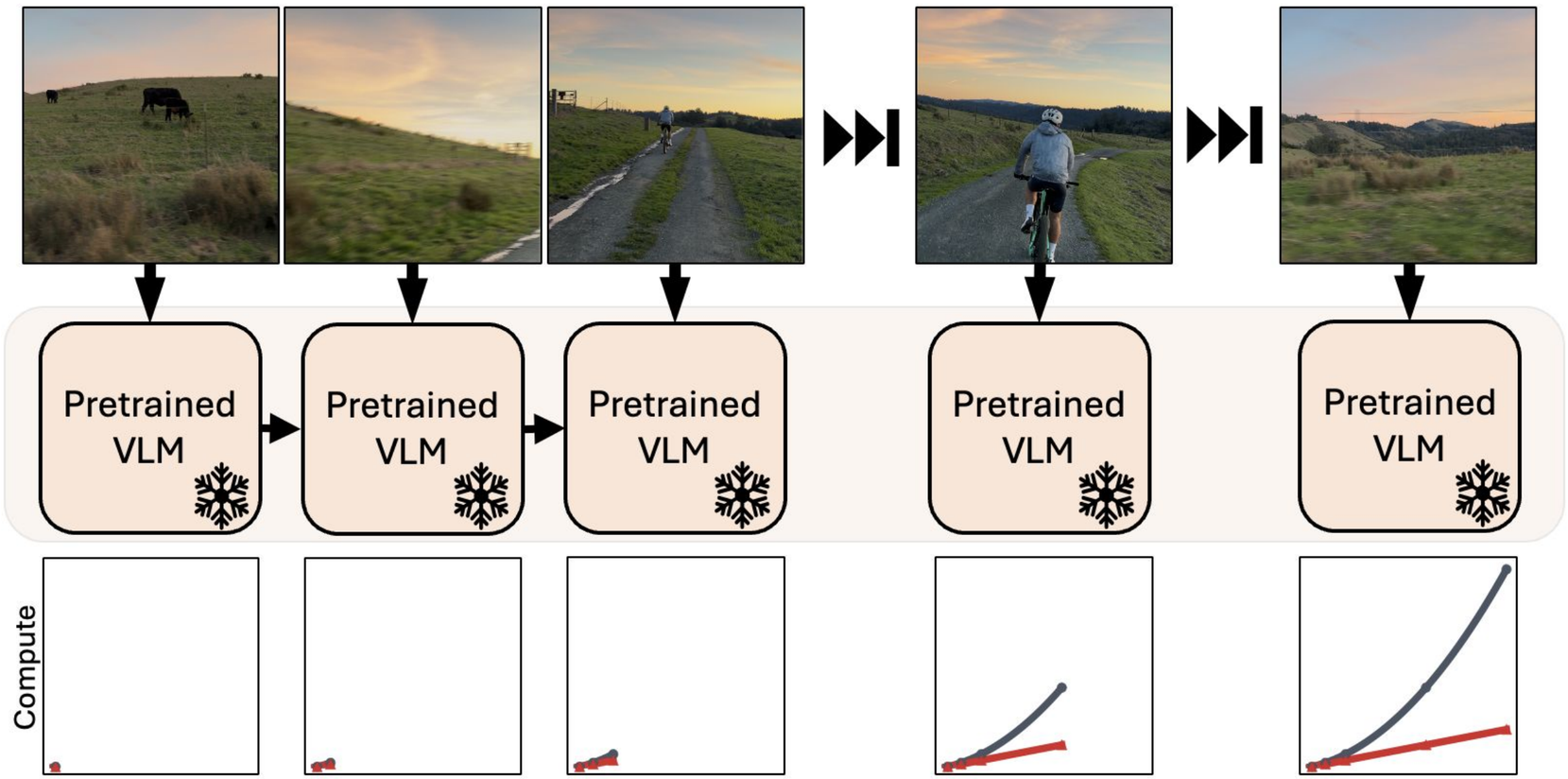

What we have

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

*dense self-attention over video tokens*

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

# StateKV: linear scaling of pretrained VLMs

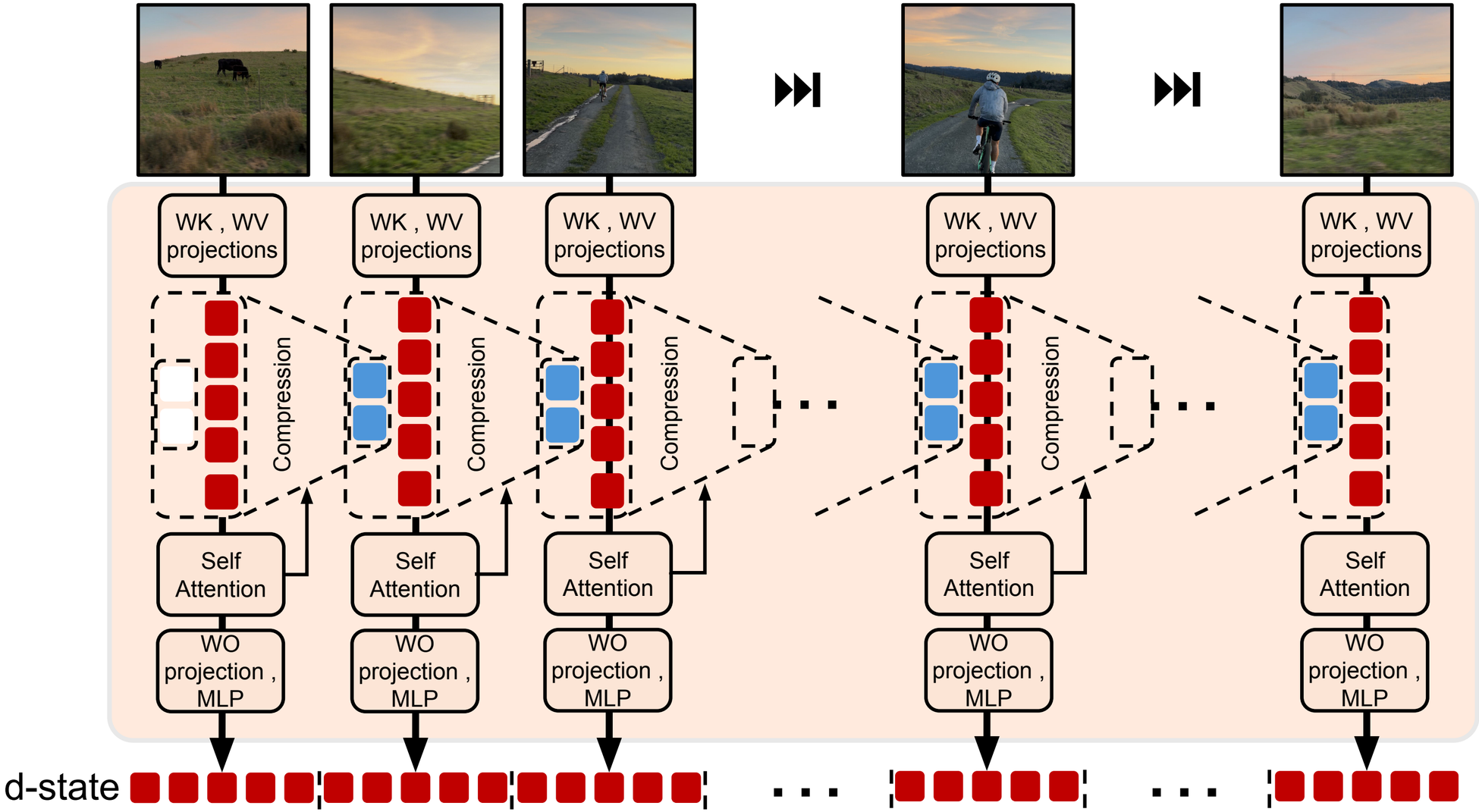
Stanford University

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. arxiv. 2026.]

# StateKV algorithm: compressed memory - `cstate`




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

# StateKV algorithm: update state recurrently

Stanford University

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

# StateKV algorithm: update state recurrently

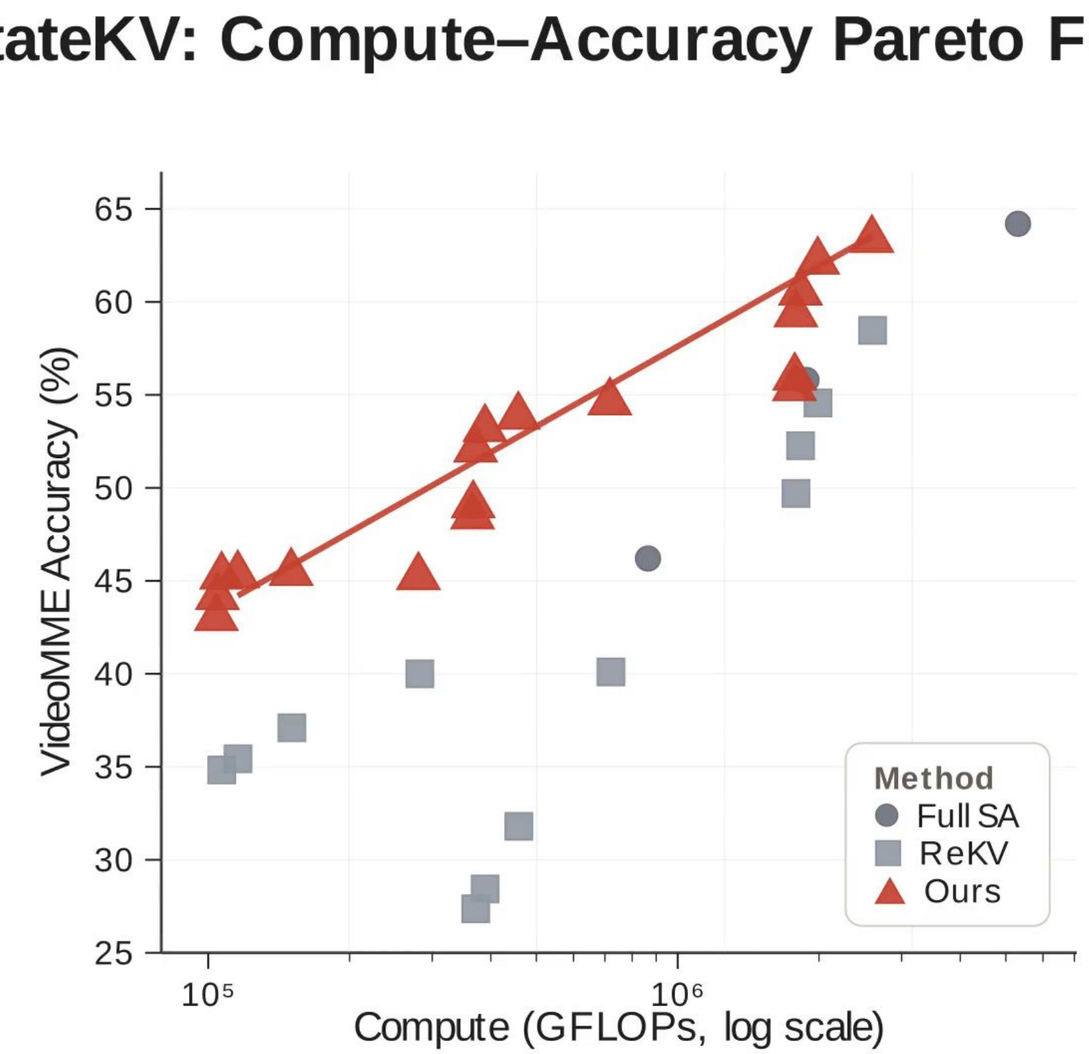

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. arxiv. 2026.]

# Accuracy–compute tradeoff

Stanford University

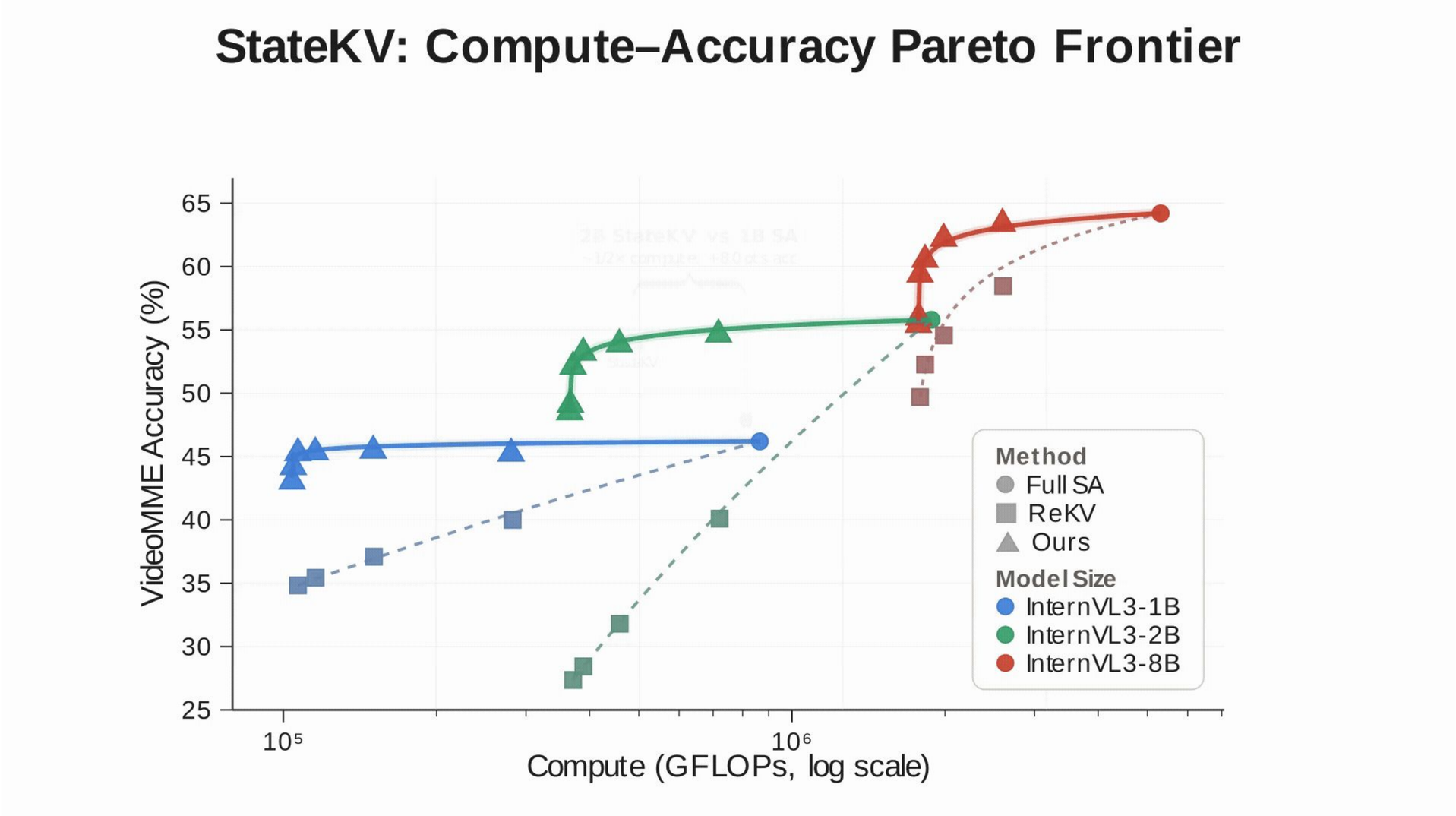
[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. arxiv. 2026.]

# Accuracy–compute tradeoff



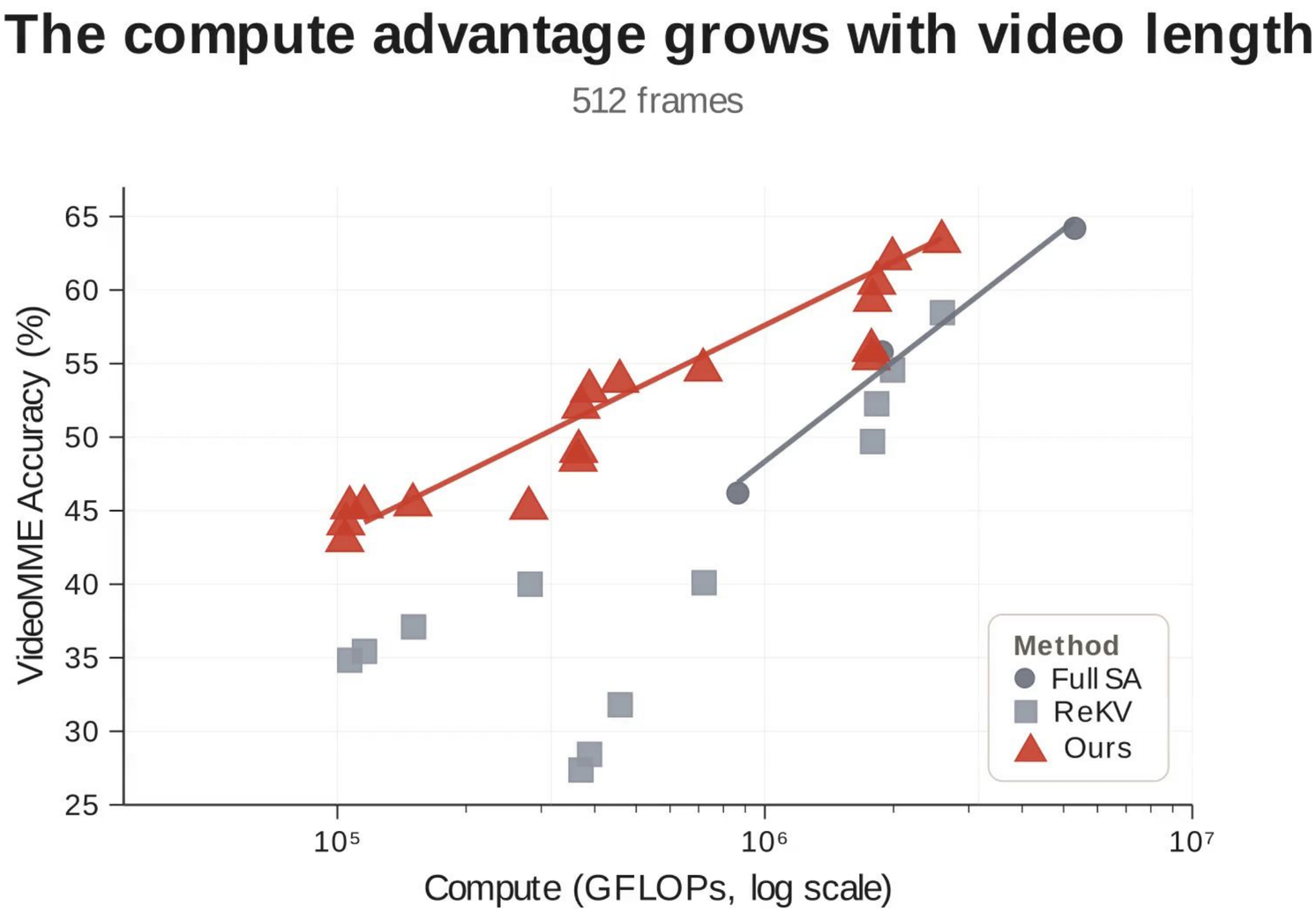

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. arxiv. 2026.]

# Accuracy–larger models with same compute




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

# The long-video regime



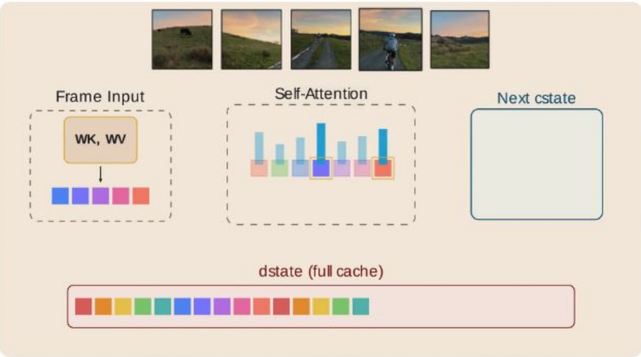

[C. Eyzaguirre et al. Linear Scaling Video VLMs for Long Video Understanding. arxiv. 2026.]

# 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

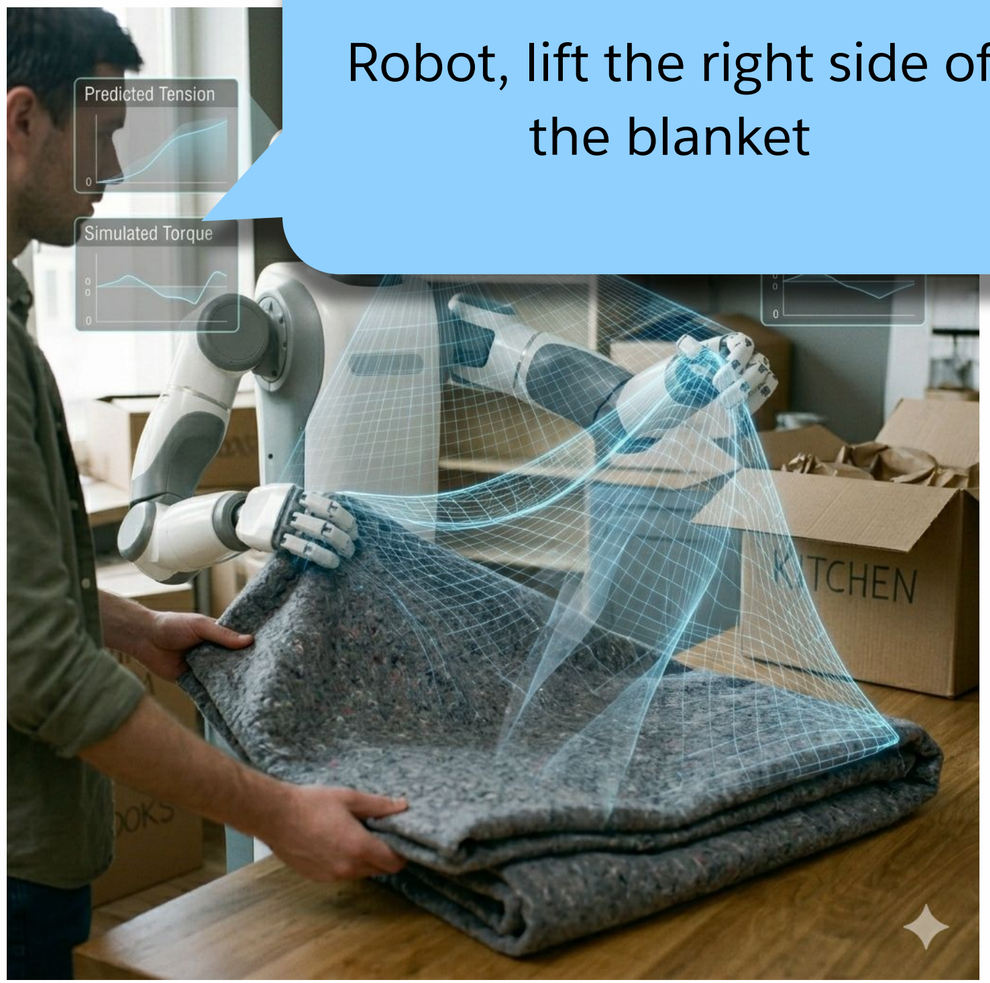
ceyzaguirre4.github.io/StateKV

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

# Part III: Linear Inference for Streaming Videos


# Part IV:

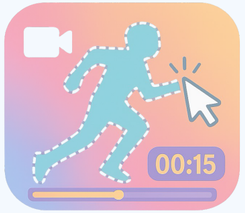

What should my next action be? Generate a video to show me
AR Next Step:
Fasten Handle
AR Imagined
Result
Assembly Flow: Step 21 of 32.


Robot, lift the right side of the blanket
Predicted Tension
Simulated Torque
KITCHEN

# Agentic Ambient Intelligence

## Video QA with Space-time references

**Strefer**
[ICCVW'25]

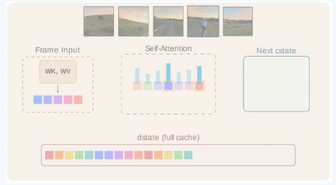



## Reasoning over long videos

**AVP**
[CVPR Findings '26]

## Efficient Inference for Streaming Videos

**StateKV**
[arxiv'26]




## Motion Guidance Generation

**FOFPred**
[CVPR Findings '26]

# Prior Work

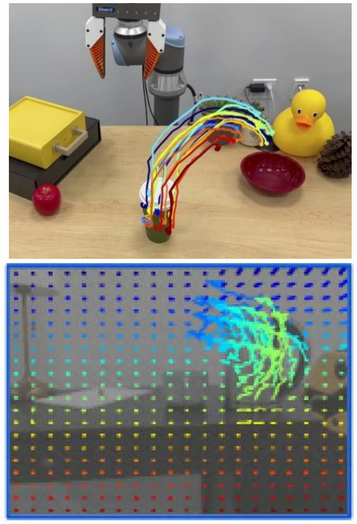
Grok

Veo

SOTA struggles at *text* to fine-grained motions;

Unless given manual motion guidance [1]

Specialized demos






VLA methods can learn such fine-grained motions;

But struggle at data scalability [2]

[1] “Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise”, CVPR 2025
[2] “Pixel Motion as Universal Representation for Robot Control”, Arxiv 2025

# Our Work

We tackle two key limitations:

1. Remove manual motion guidance dependency

2. Leverage internet scale training data

# FOFPred: Motion Guidance Generation

## Learn to Predict Explicit “Motion”

+

Generate motion for “assembling drawer into empty space”

→ **FOFPred** →

Future motion prediction

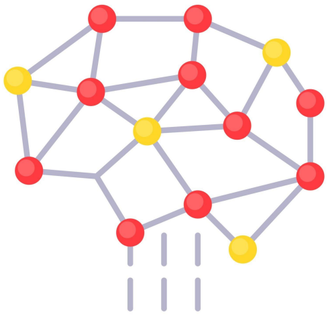
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# FOFPred: Motion Guidance Generation

We chose
**Future Optical Flow**
as our explicit way to reason about future motion

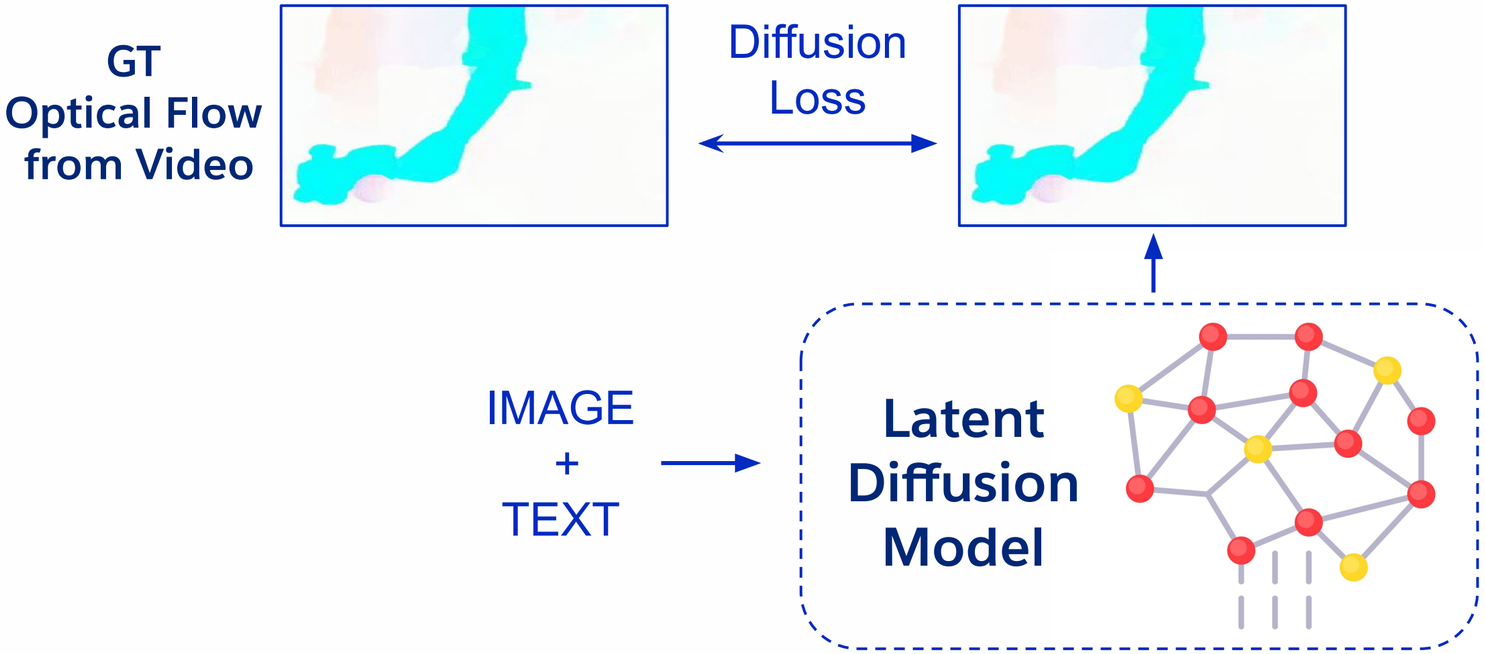
IMAGE
+
TEXT

Latent Diffusion Model



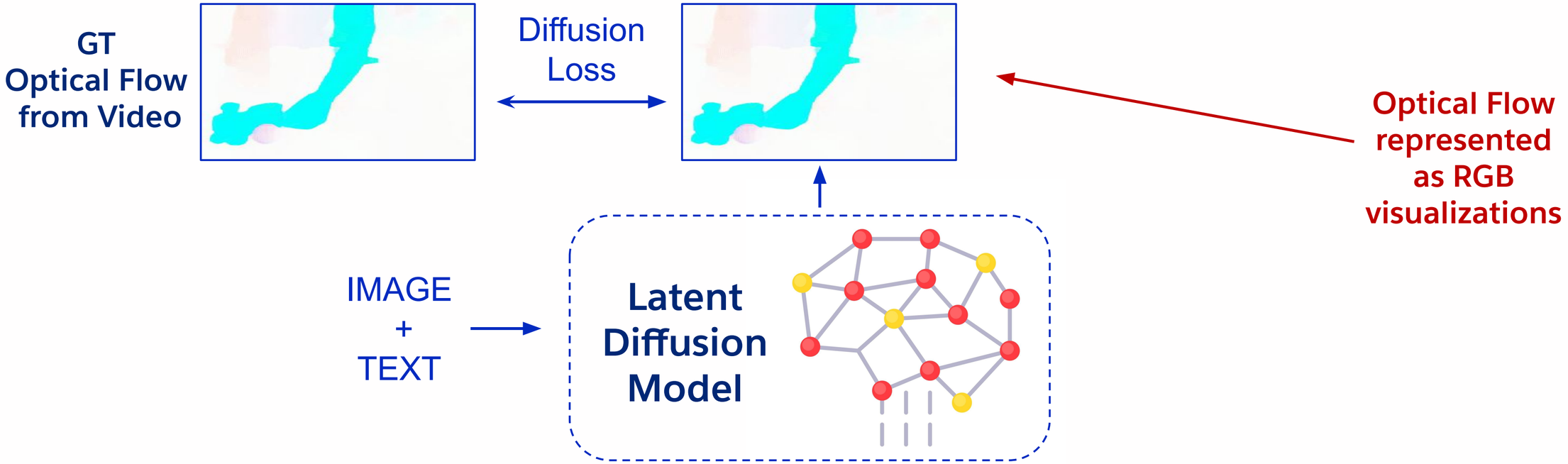
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# FOFPred: Motion Guidance Generation



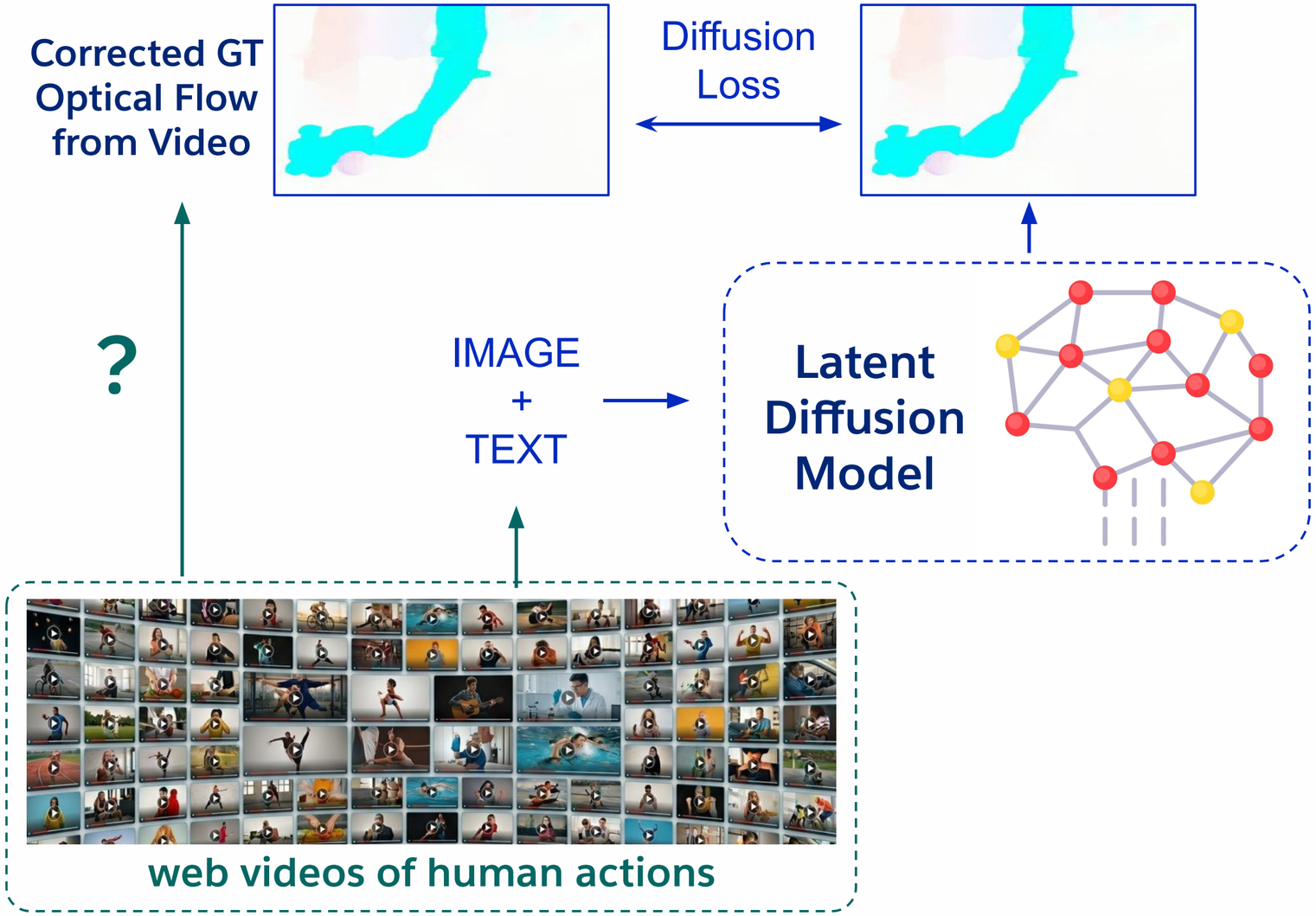

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# FOFPred: Motion Guidance Generation



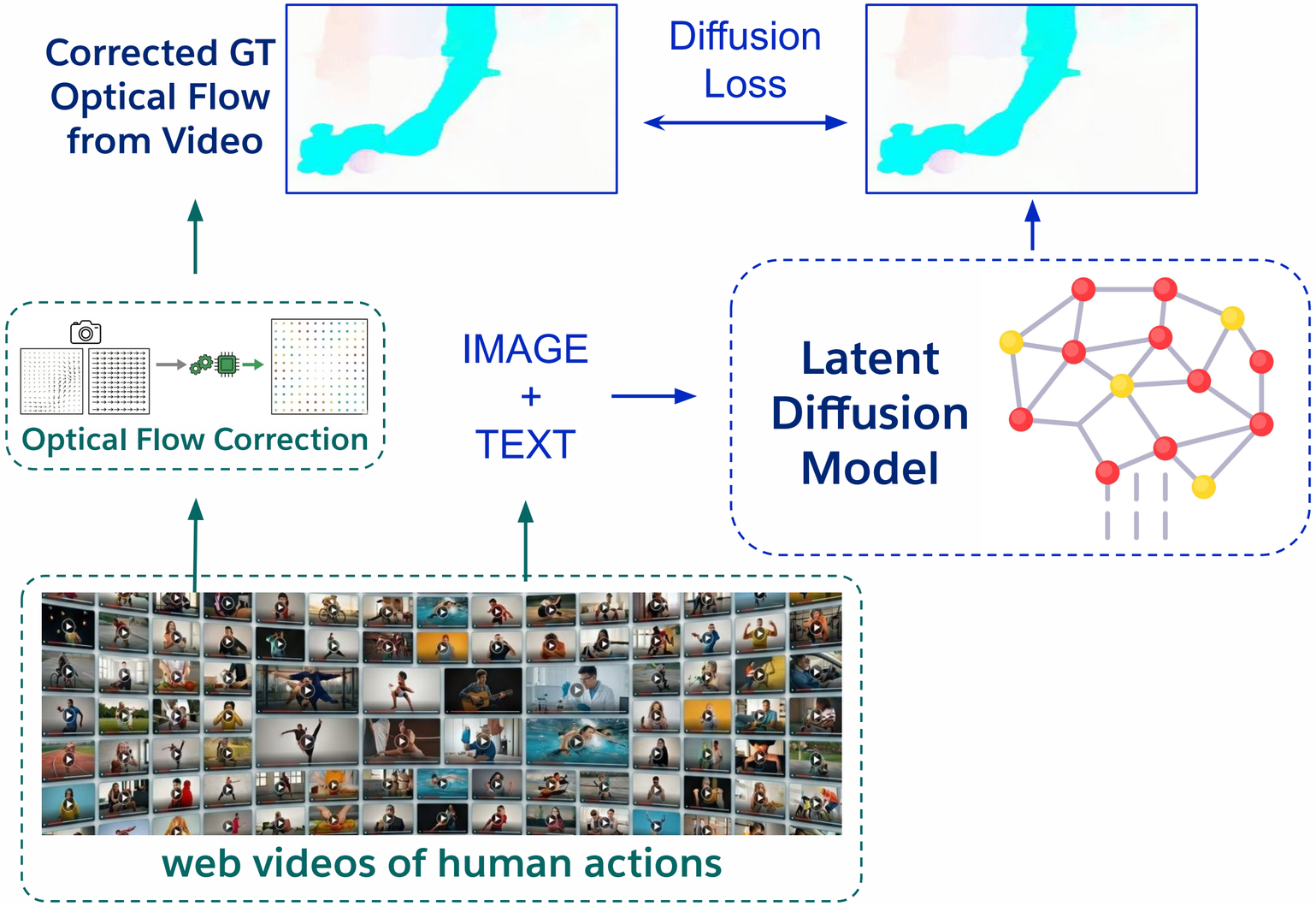

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Scalable Learning



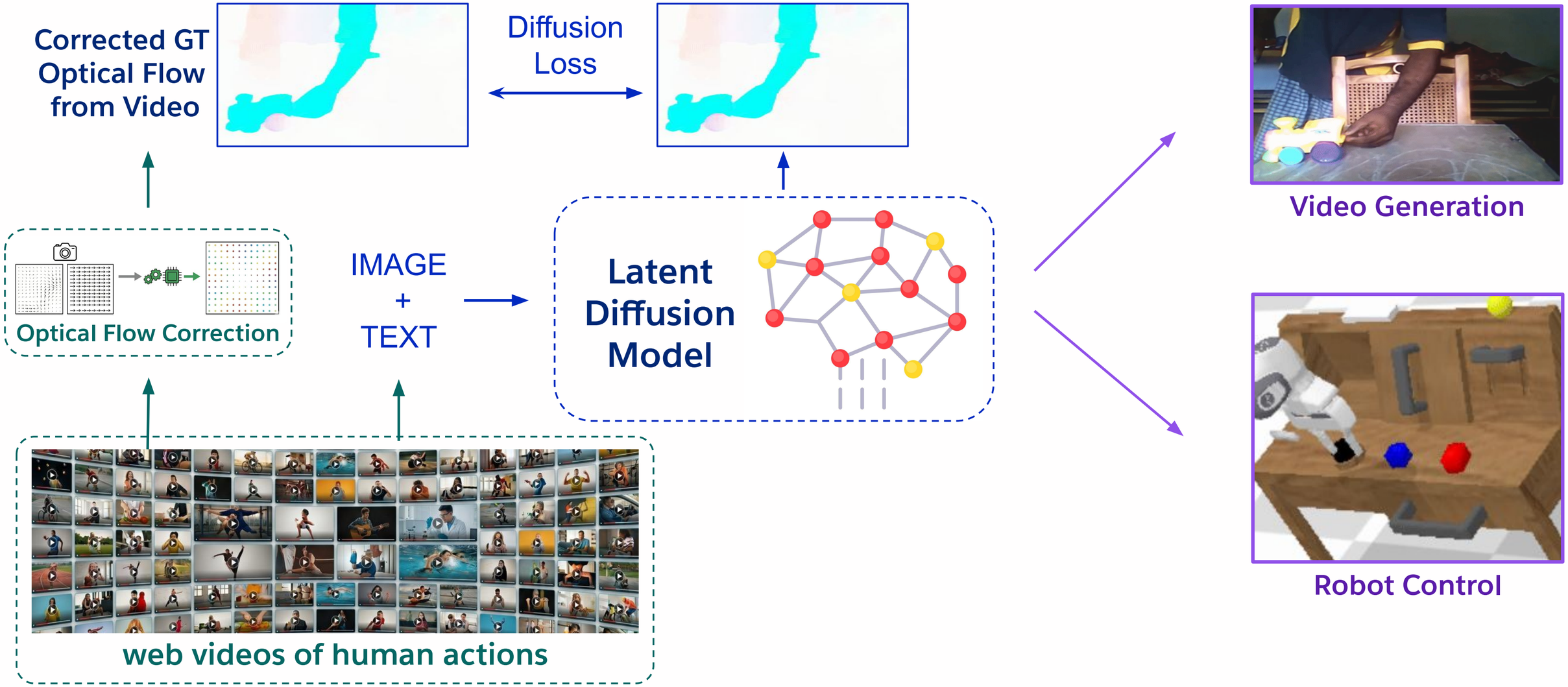

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Scalable Learning



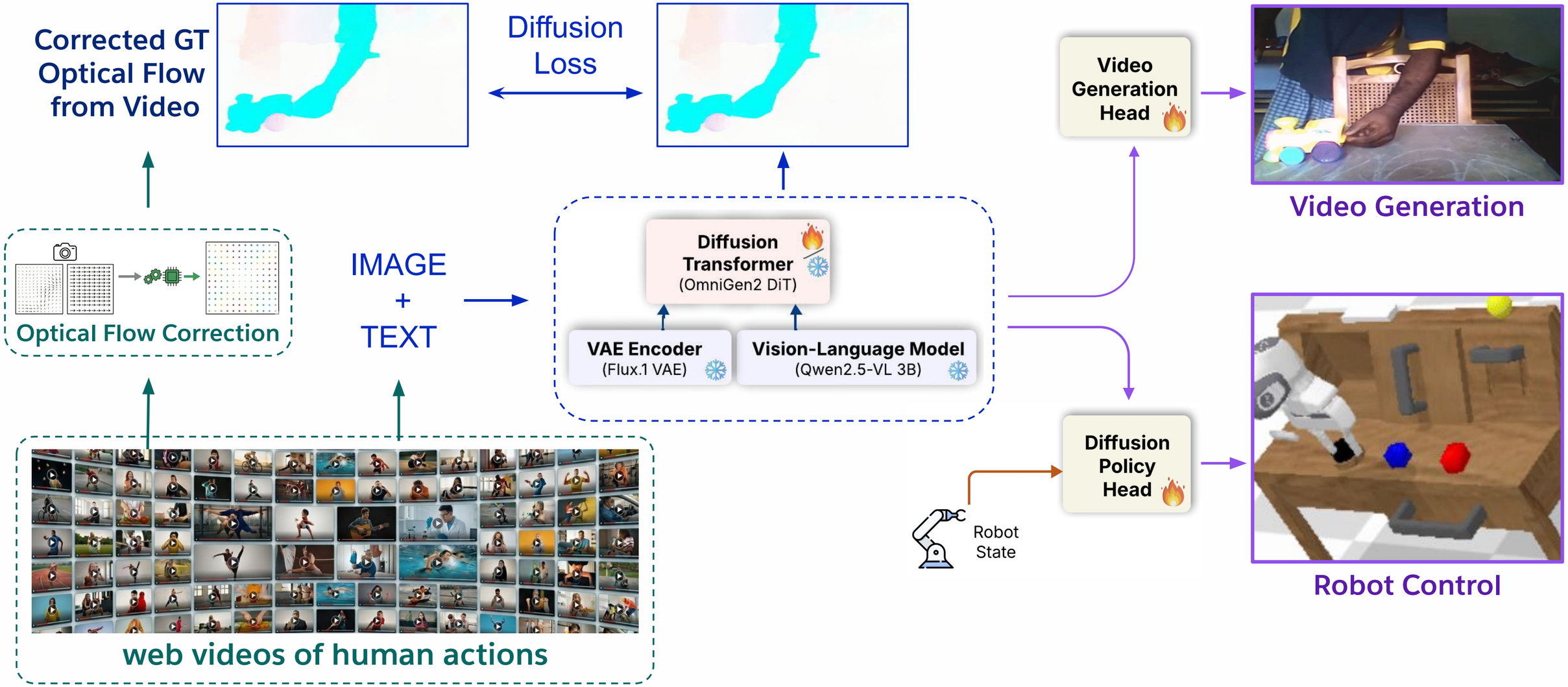

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Downstream Tasks



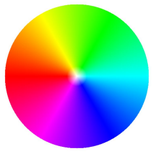

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# FOFPred: Detailed Architecture




[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

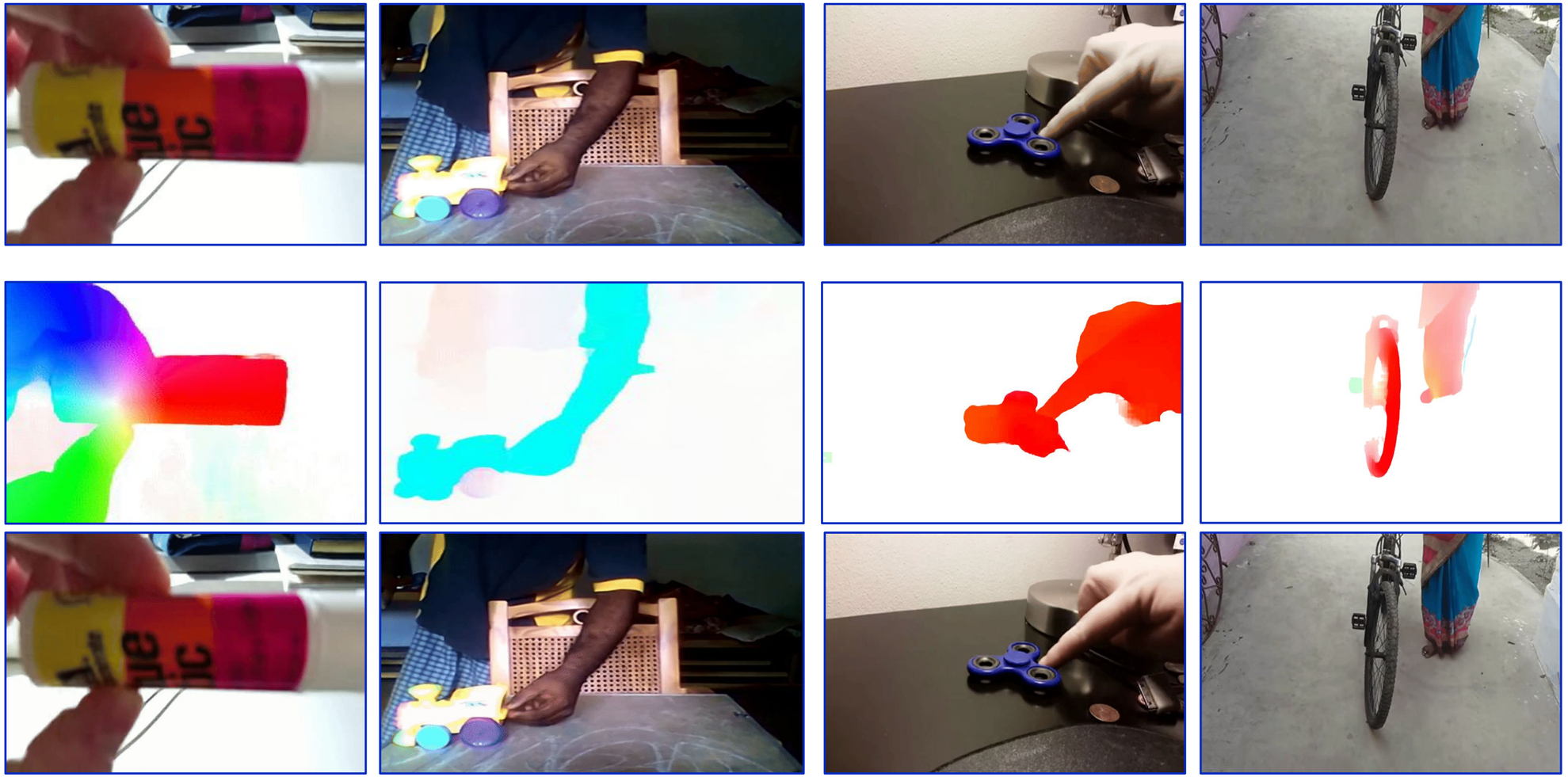
# Video Generation



OF direction color wheel

CogVideoX Baseline

Ours



NOTE: OF interpolated to match full video resolution

Moving glue stick away from camera

Pulling toy car from left to right

Push fidget spinner from left to right

Moving cycle towards camera

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Video Generation

| Method | SSIM ↑ | PSNR ↑ | LPIPS ↓ | FVD↓ | KVD↓ | MF ↑ [99] |
|---|---|---|---|---|---|---|
| Seer [27] | 41.8 | 10.71 | 58.8 | 287.46 | 81.31 | — |
| Dynamicrafter [95] | — | — | — | 204.11 | 31.81 | — |
| CosHand-I [77] | 61.5 | 16.87 | 31.3 | 91.18 | 19.24 | 0.432 |
| CosHand-A [77] | 53.1 | 14.92 | 40.8 | 90.30 | 13.68 | 0.570 |
| InterDyn [1] | 66.4 | 18.60 | 26.0 | 19.27 | 1.99 | 0.633 |
| InterDyn-R [1] | 68.0 | 19.04 | 25.2 | 22.22 | 2.09 | 0.641 |
| CogVideoX [98] | 67.2 | 21.51 | 30.3 | 78.47 | 12.46 | 0.594 |
| FOFPred (ours) | 68.4 (+1.2) | 22.26 (+0.75) | 28.5 (+1.8) | 75.39 (+3.08) | 11.38 (+1.08) | 0.662 (+0.068) |

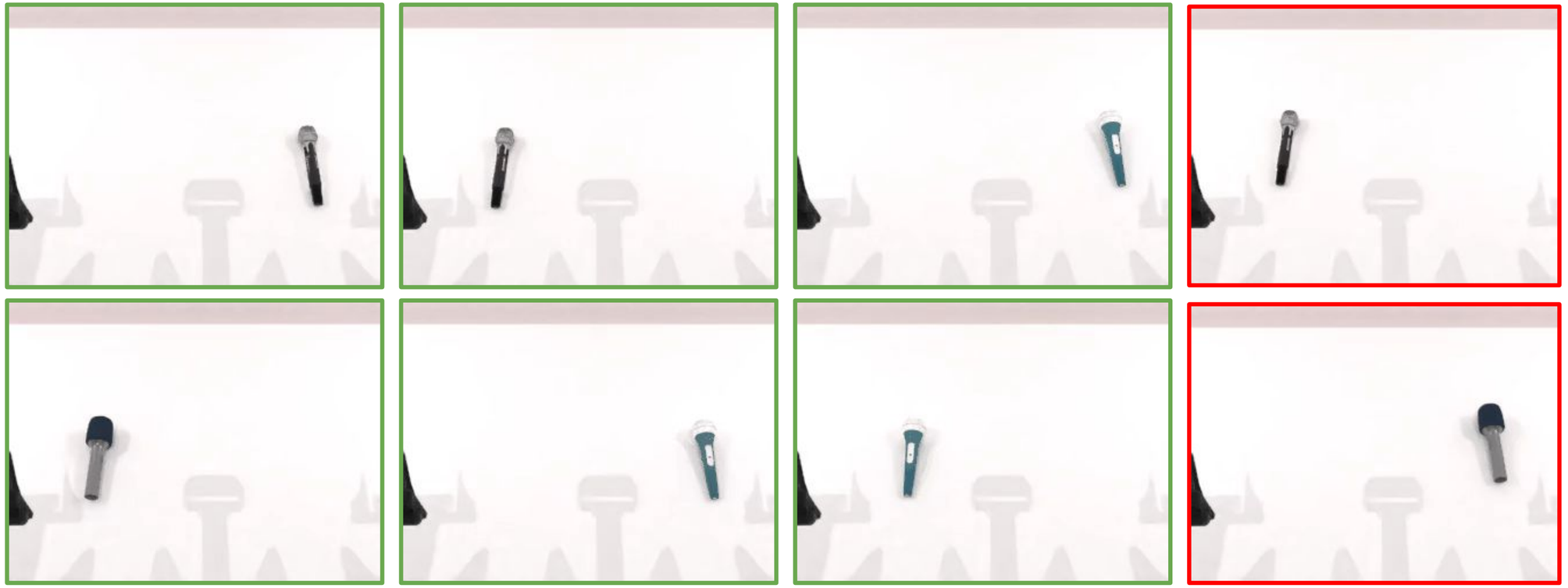
**SSv2 Video Generation Evaluation:** Over the CogVideoX baseline, our FOFPred framework shows consistent improvements in generation quality.

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Evaluations: RoboTwin

**RoboTwin 2.0 Benchmark**

- Complex Bimanual Manipulation
- Limited Training Data: only 50 demonstrations per task



*"Handover Mic"* Task. We visualize success and failure cases of our method.

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Evaluations: RoboTwin

| Method | Handover Block | Handover Mic | Pick Diverse Bottles | Pick Dual Bottles | Place Dual Shoes | Average |
|---|---|---|---|---|---|---|
| RDT [54] | 45 | 90 | 2 | 42 | 4 | 36.6 |
| ACT [108] | 42 | 85 | 7 | 31 | 9 | 34.8 |
| DP [18] | 10 | 53 | 6 | 24 | 8 | 20.2 |
| DP3 [103] | 70 | 100 | 52 | 60 | 13 | 59.0 |
| $\pi_0$ [9] | 45 | 98 | 27 | 57 | 15 | 48.4 |
| VPP [33] | 54 | 80 | 60 | 63 | 52 | 61.8 |
| FOFPred (ours) | 61 (+7) | 87 (+7) | 67 (+7) | 68 (+5) | 60 (+8) | 68.6 (+6.8) |

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# FOFPred: Summary

Predicting explicit motion representations improves Robot Control & Video Generation

We learn to predict such motion from internet-scale human action videos

This enables using natural language to guide motion generation

**Poster - CVPR Findings**
**June 6th AM**

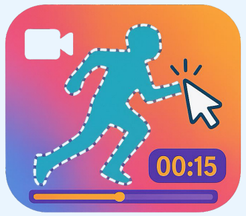
Code, Ckpt, & Demo Released
https://fofpred.github.io/

[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

# Agentic Ambient Intelligence

## Video QA with Space-time references

**Strefer**
[ICCVW'25]

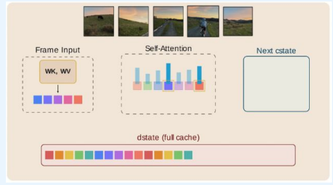



## Reasoning over long videos

**AVP**
[CVPR Findings '26]

## Efficient Inference for Streaming Videos

**StateKV**
[arxiv'26]




## Motion Guidance Generation

**FOFPred**
[CVPR Findings '26]

# Agentic Ambient Intelligence



salesforce
Thank
you

# Resources


1. Zhou et al. “Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data”. ICCVW, 2025.
2. Wang et al. “Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding”. CVPR Findings, 2026.
3. Eyzaguirre et al. “Streaming Detection of Queried Event Start” NeurIPS 2025.
4. Eyzaguirre et al. “Linear Scaling Video VLMs for Long Video Understanding”. arxiv 2026.
5. Ranasinghe et al. “Future Optical Flow Prediction Improves Robot Control and Video Generation”. CVPR Findings, 2026.


slid pers