

Agentic Ambient Intelligence

Efficient Understanding & Action

VITA Workshop at CVPR
June 3, 2026

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



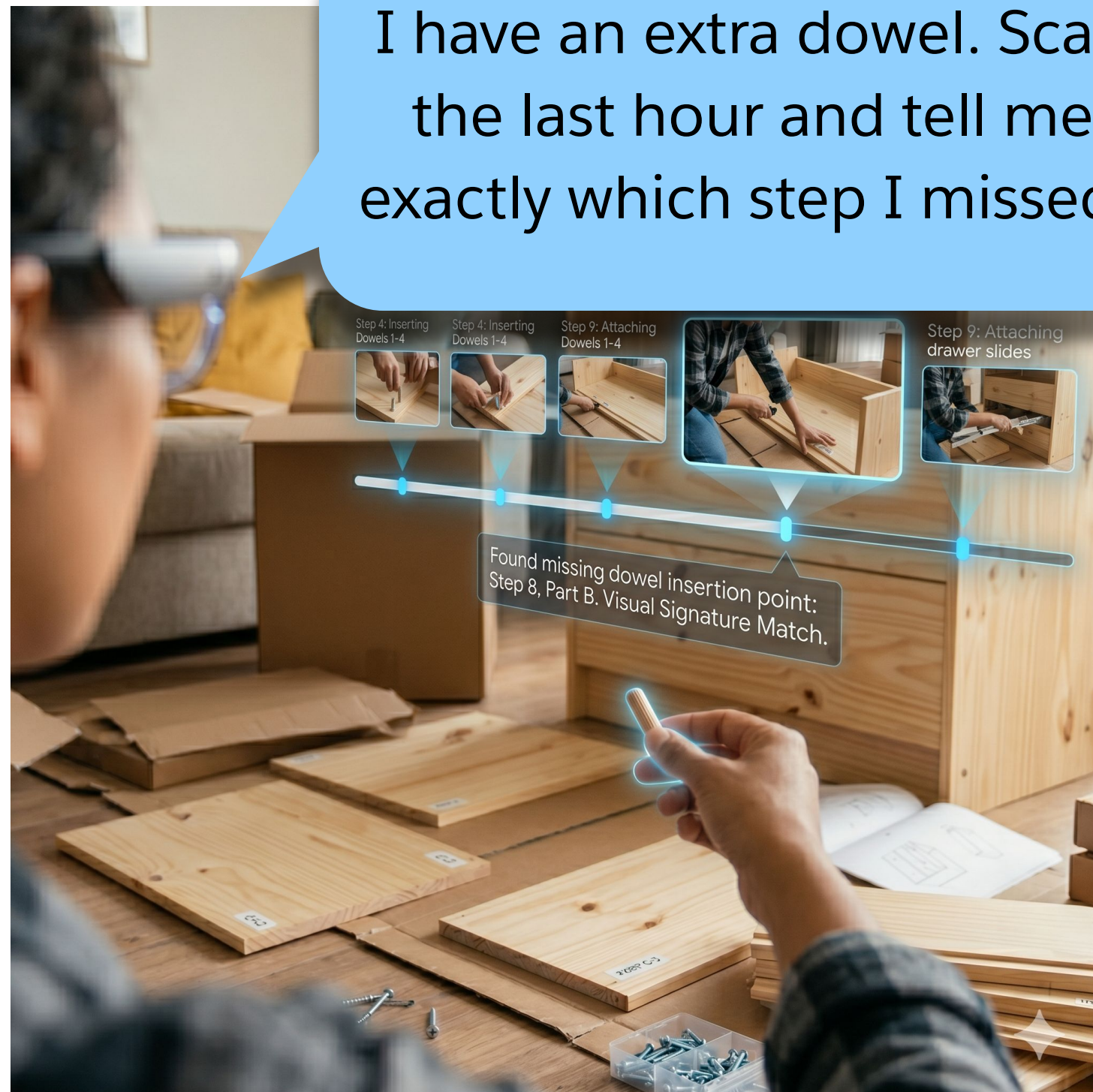
Agentic Ambient Intelligence



Part I



I have an extra dowel. Scan the last hour and tell me exactly which step I missed?



Scan your memory from the last six hours and find where I set my keys down while we were in the kitchen

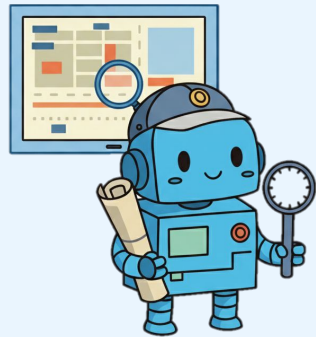


Agentic Ambient Intelligence



Reasoning over long videos

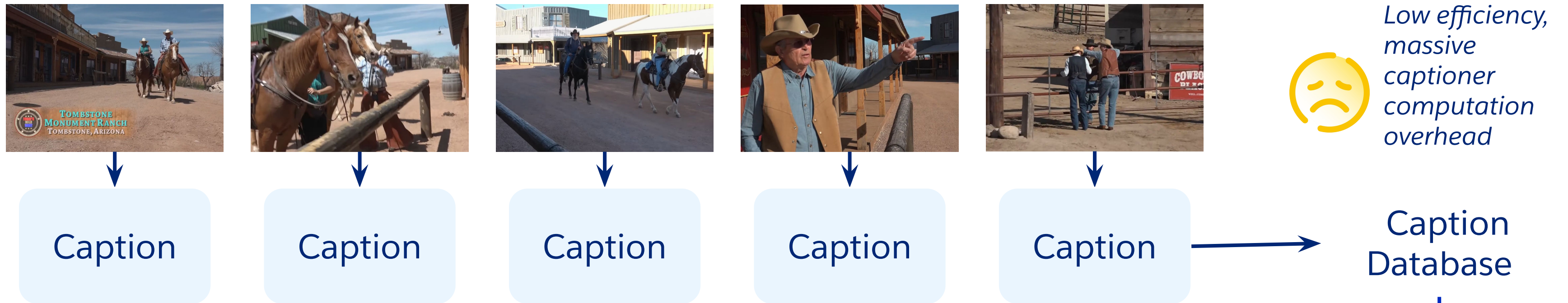
AVP
[CVPR Findings '26]



Existing Caption-based Agentic Frameworks



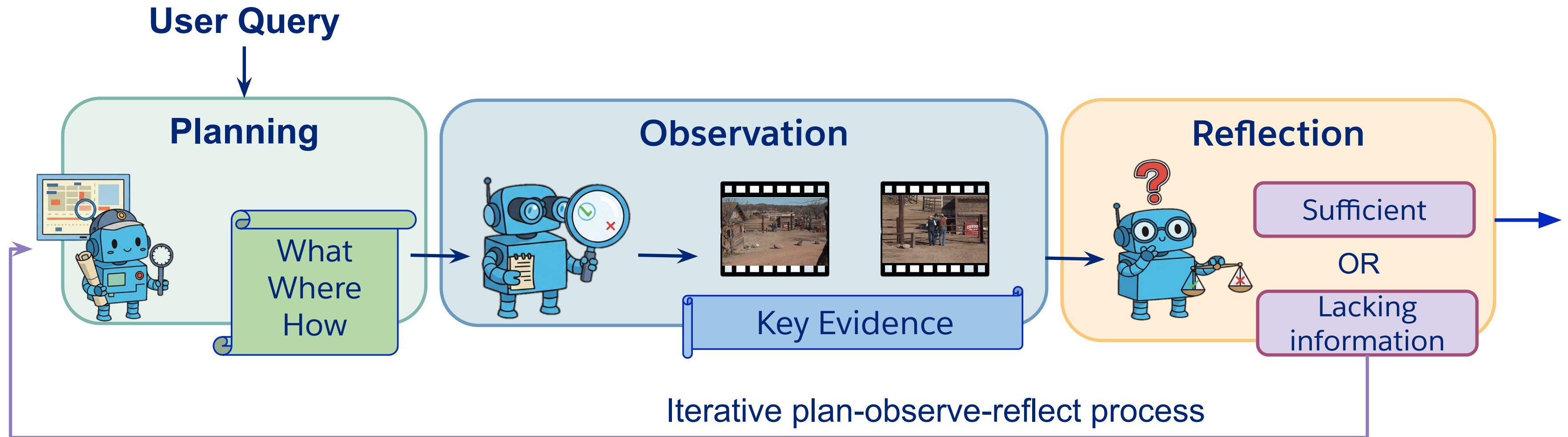
Step 1 : Passive perception via query-agnostic captioner



Step 2 : Evidence searching via caption database



Our Work: Active Video Perception (AVP)

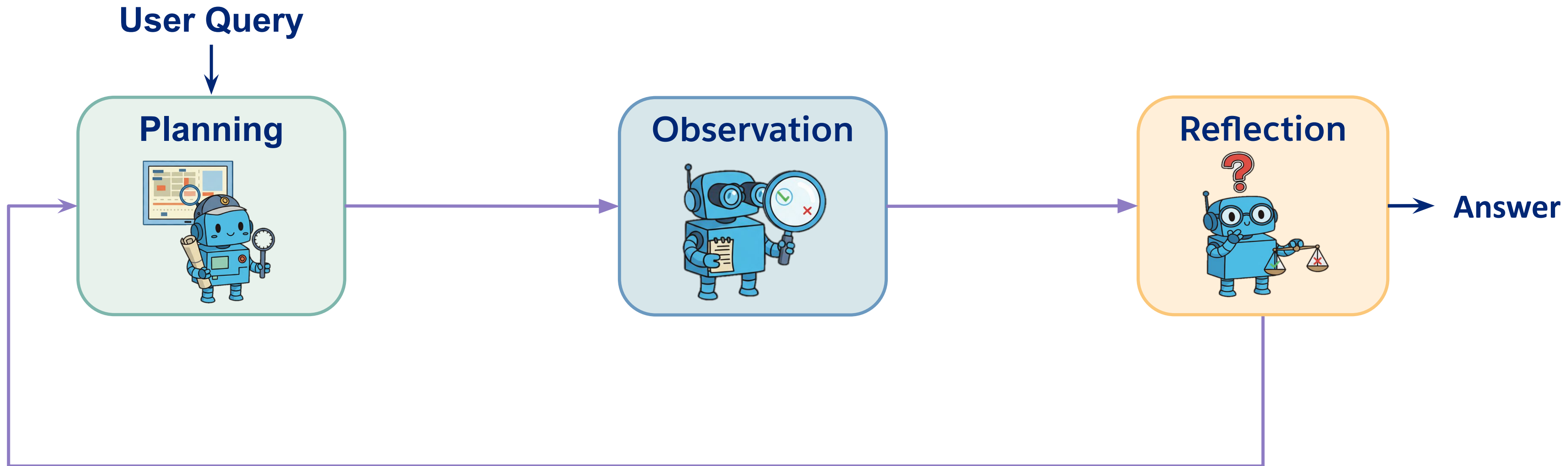


Active Evidence Seeking



Iterative Perception via Reflection

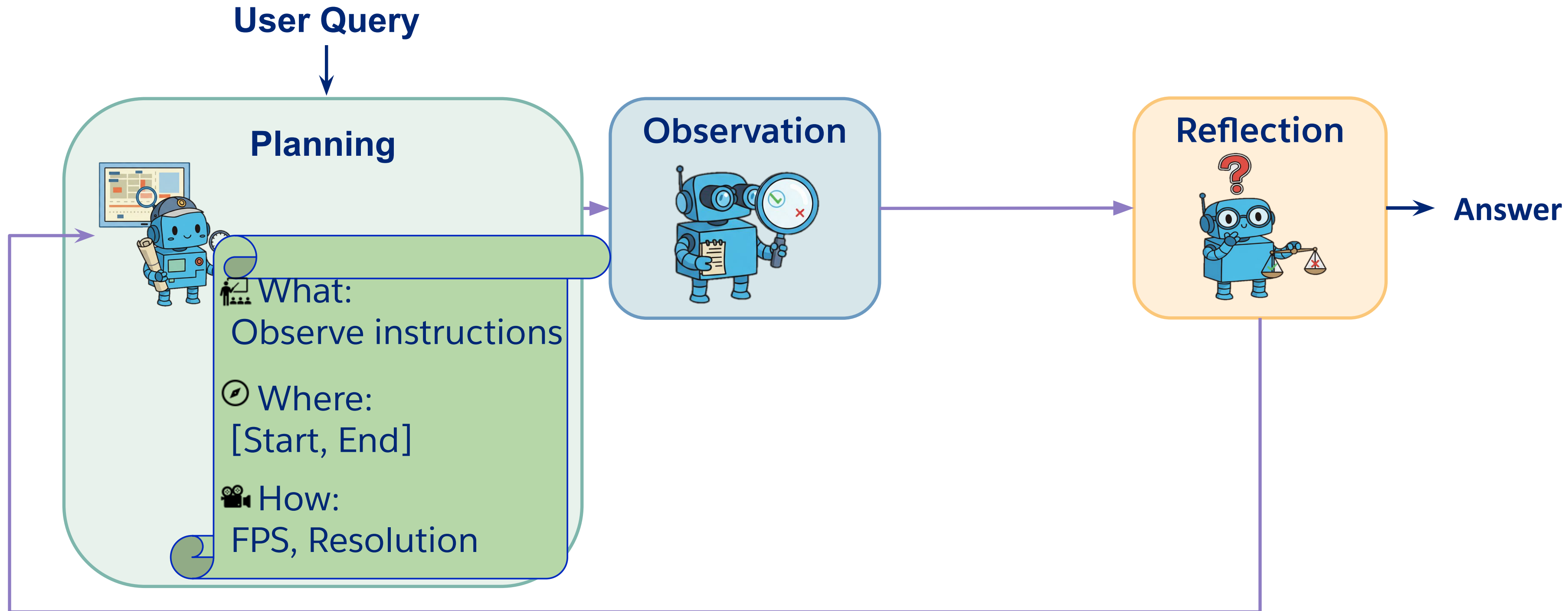
Our Work: Active Video Perception (AVP)



Our Work: Active Video Perception (AVP)



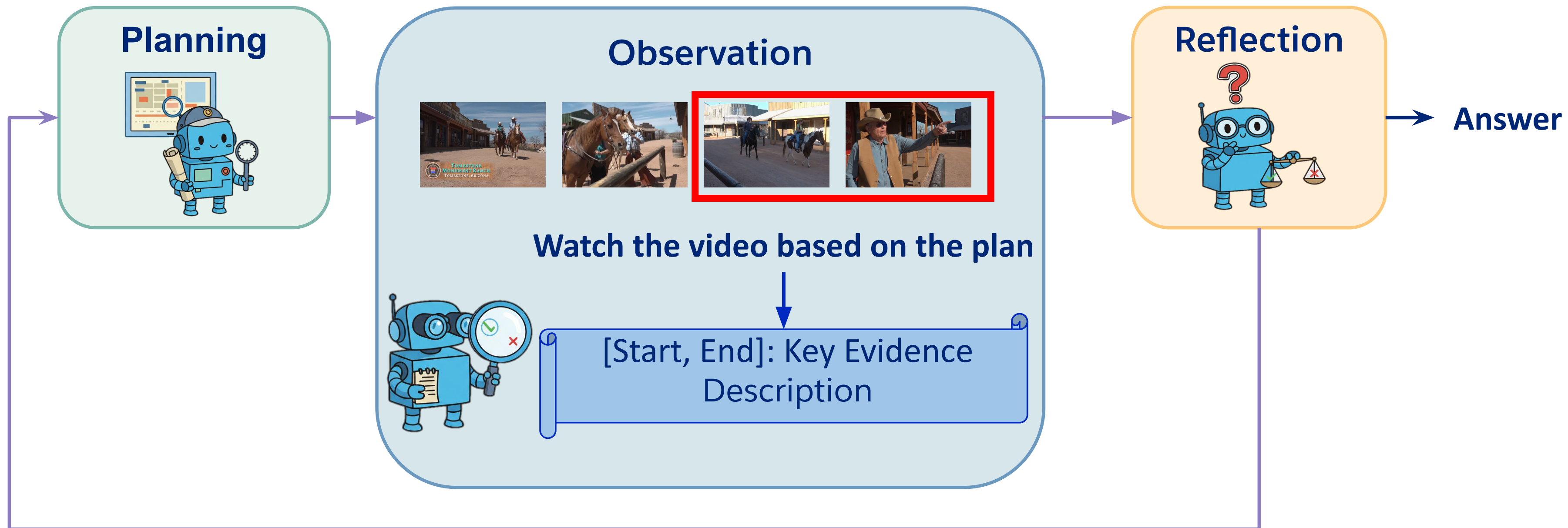
Step 1 : Generate Plan



Our Work: Active Video Perception (AVP)



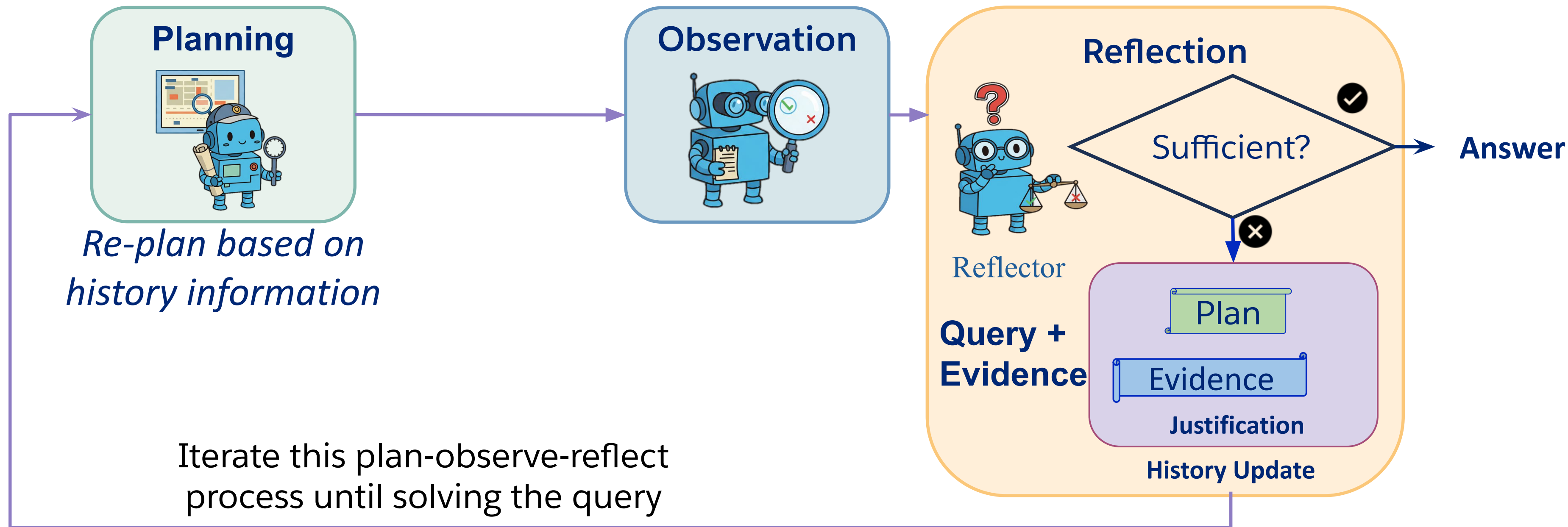
Step 2 : Targeted Observation



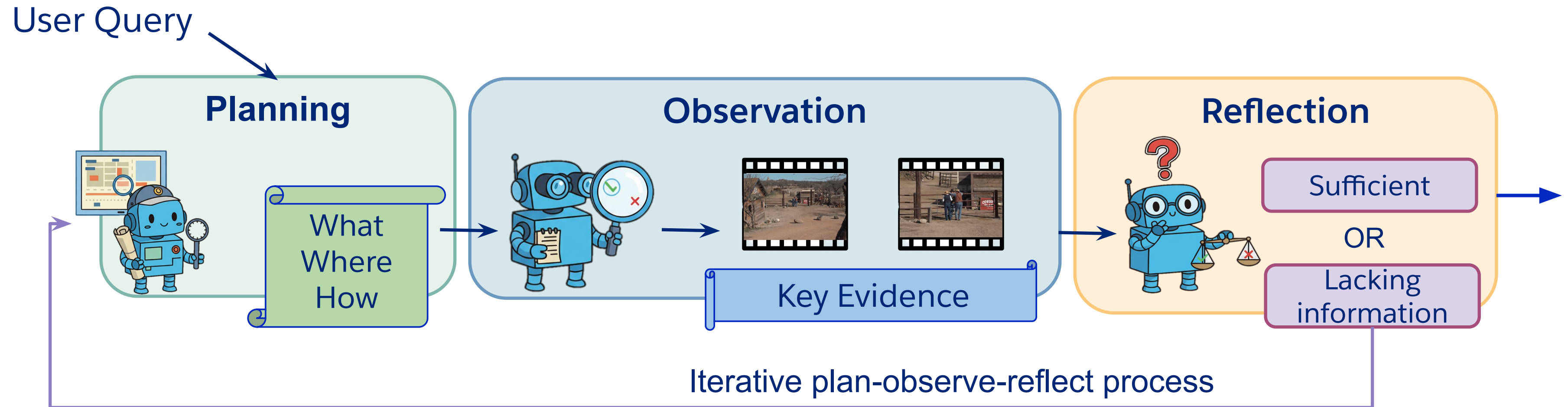
Our Work: Active Video Perception (AVP)



Step 3: Reflect on evidence



Our Work: Active Video Perception (AVP)

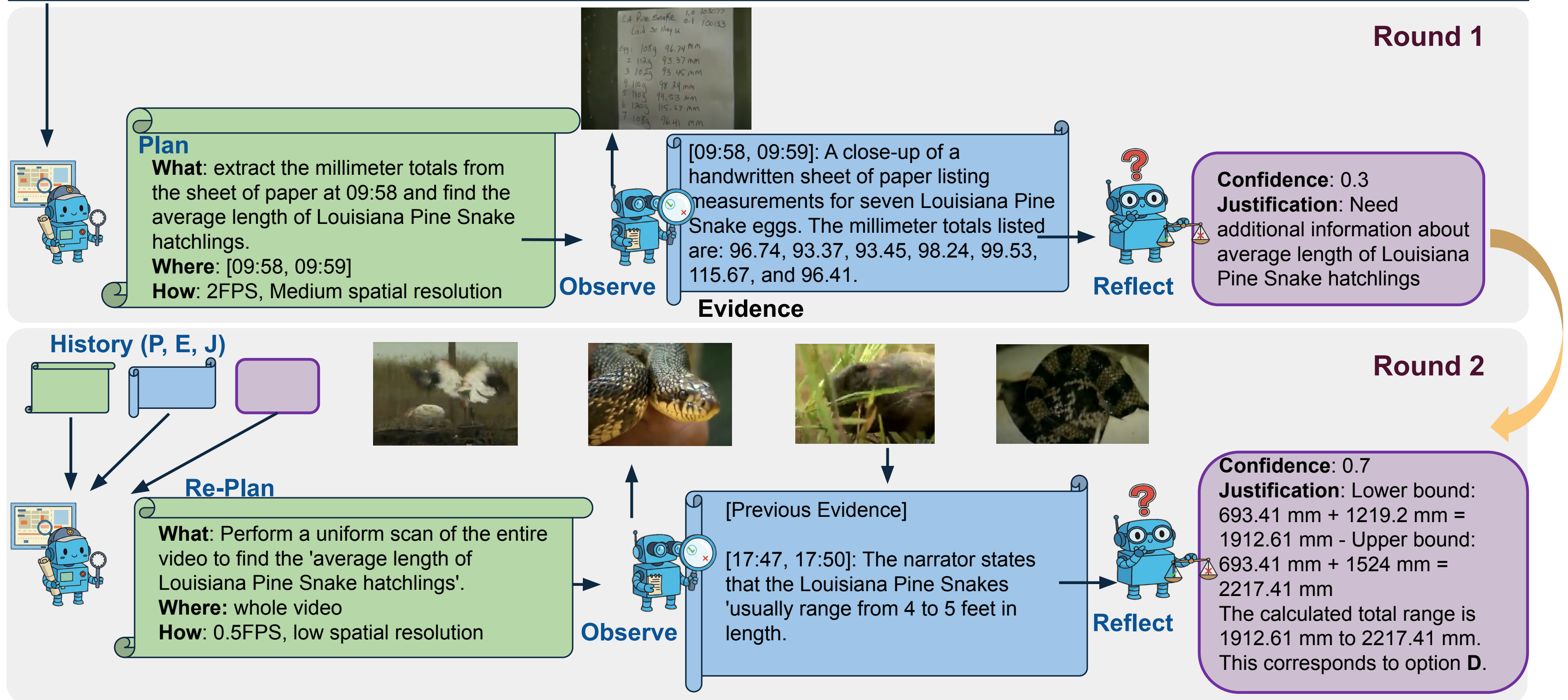


Active Evidence Seeking



Iterative Perception via Reflection

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.



AVP: Quantitative Results



Methods	MINERVA	LVBench	MLVU	Video-MME		LongVideoBench	
	<i>Overall</i>	<i>Overall</i>	<i>Test</i>	<i>Overall</i>	<i>Long</i>	<i>Val</i>	<i>Long</i>
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
<i>Active Video Perception (Ours)</i>							
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)

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.

Project page

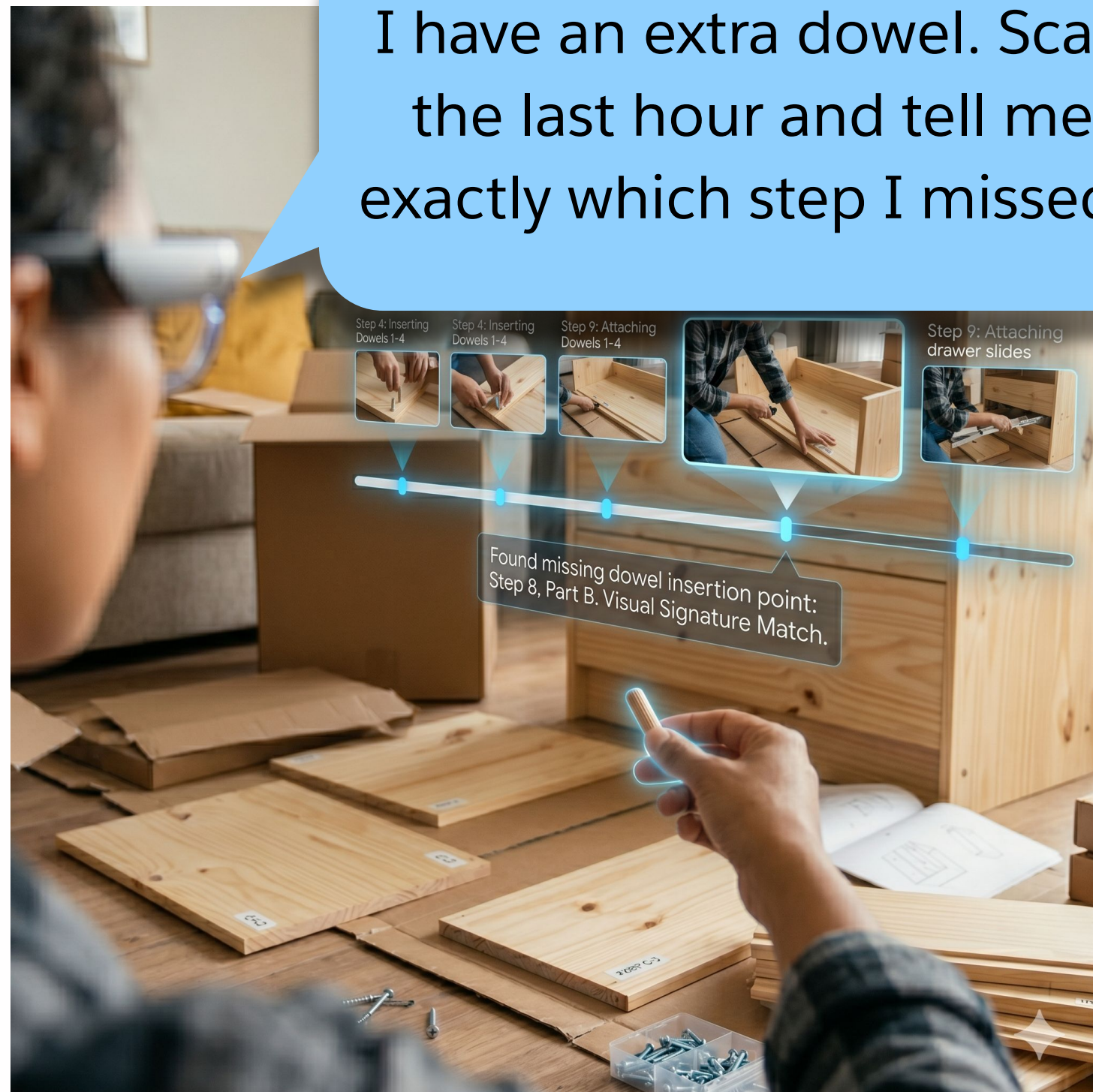


**Poster - CVPR Findings
Sunday June 7th AM**

Part I: Active Perception



I have an extra dowel. Scan the last hour and tell me exactly which step I missed?



Scan your memory from the last six hours and find where I set my keys down while we were in the kitchen



Part II:

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



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

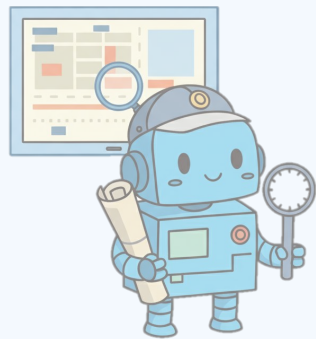


Agentic Ambient Intelligence



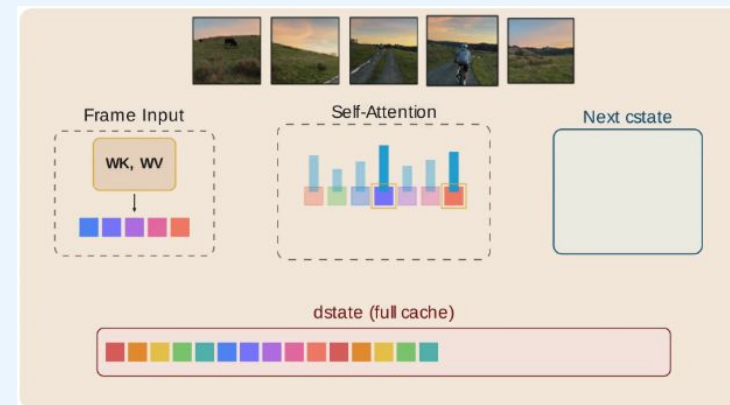
Reasoning over long videos

AVP
[CVPR Findings '26]



Fast inference for streaming videos

StateKV
[arxiv '26]



Streaming Detection of Queried Event Start

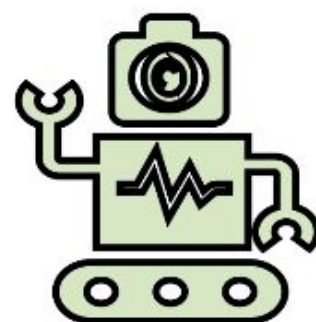
Streaming Videos

Online Model

Natural Language Queries

Now

t



Self Driving

Accelerate when the light turns green.

Alert me if a child is crossing the street.

Robotics

Let me know when it's appropriate to vacuum.

Wake up when the person leaves the house.

AR Assistant

Remind me to get my card after I use the ATM.

When I'm paying remind me I have a coupon.

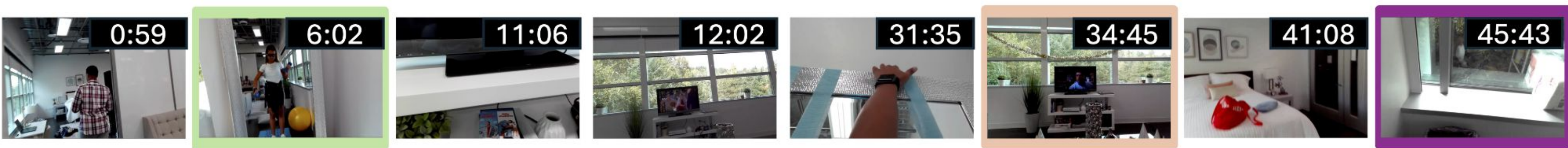
“Proactive” Dataset: Streaming Detection of Queries (EgoSDQES)



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.

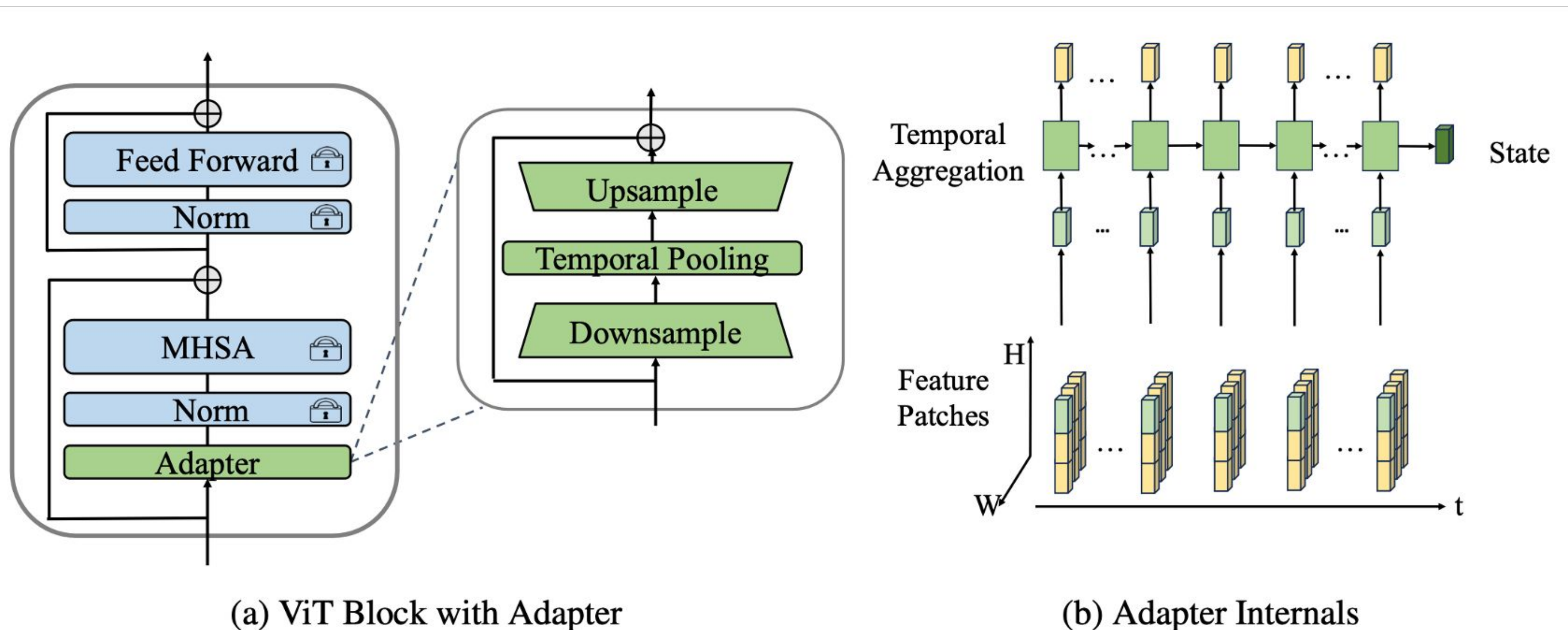


Next time I start doing my exercises, please remind me to drink some water.

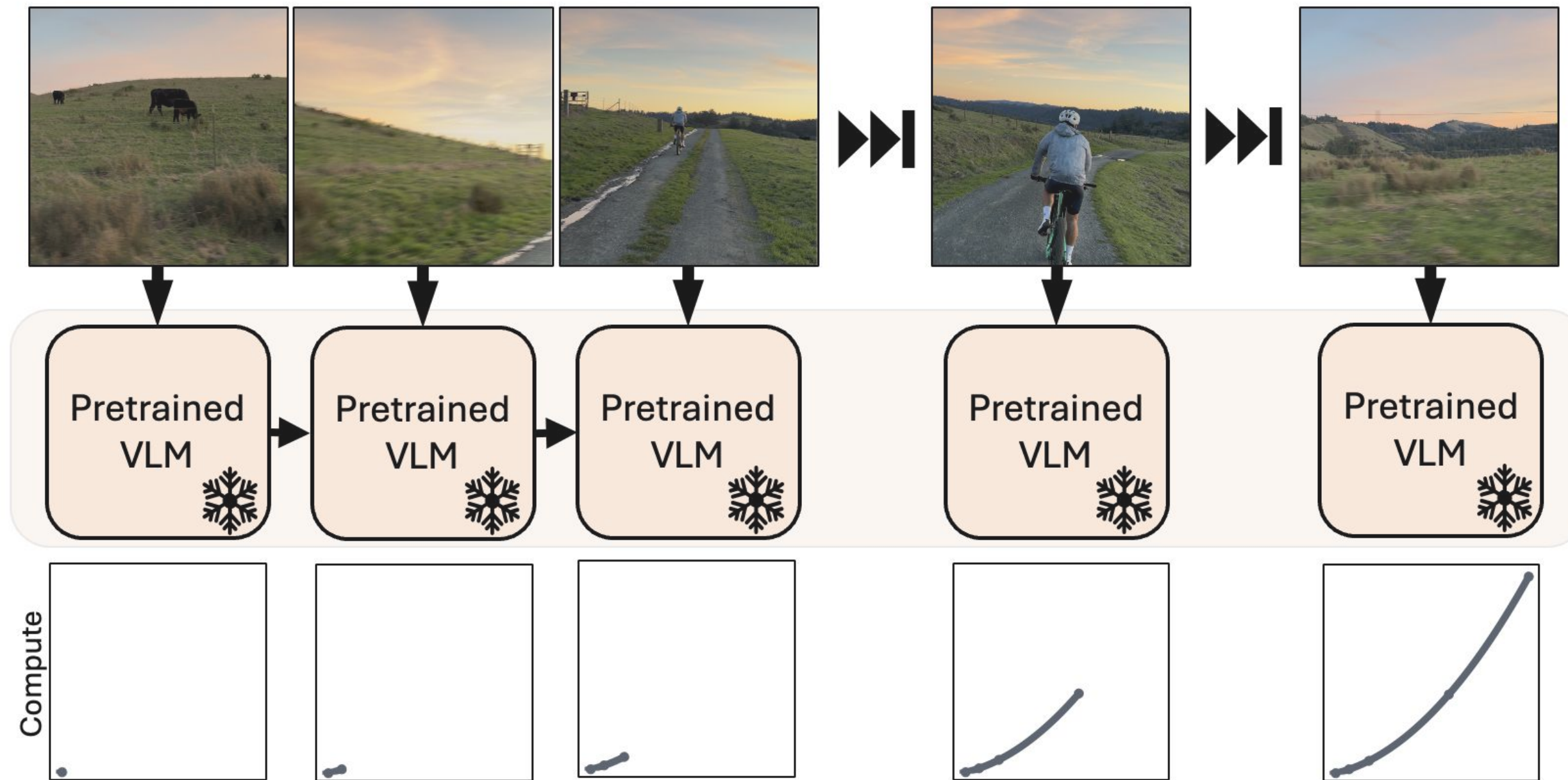
When I start watching television after decorating the house, remind me to adjust the volume.

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

Simple Architecture: CLIP plus temporal state



VLMs for Video Understanding

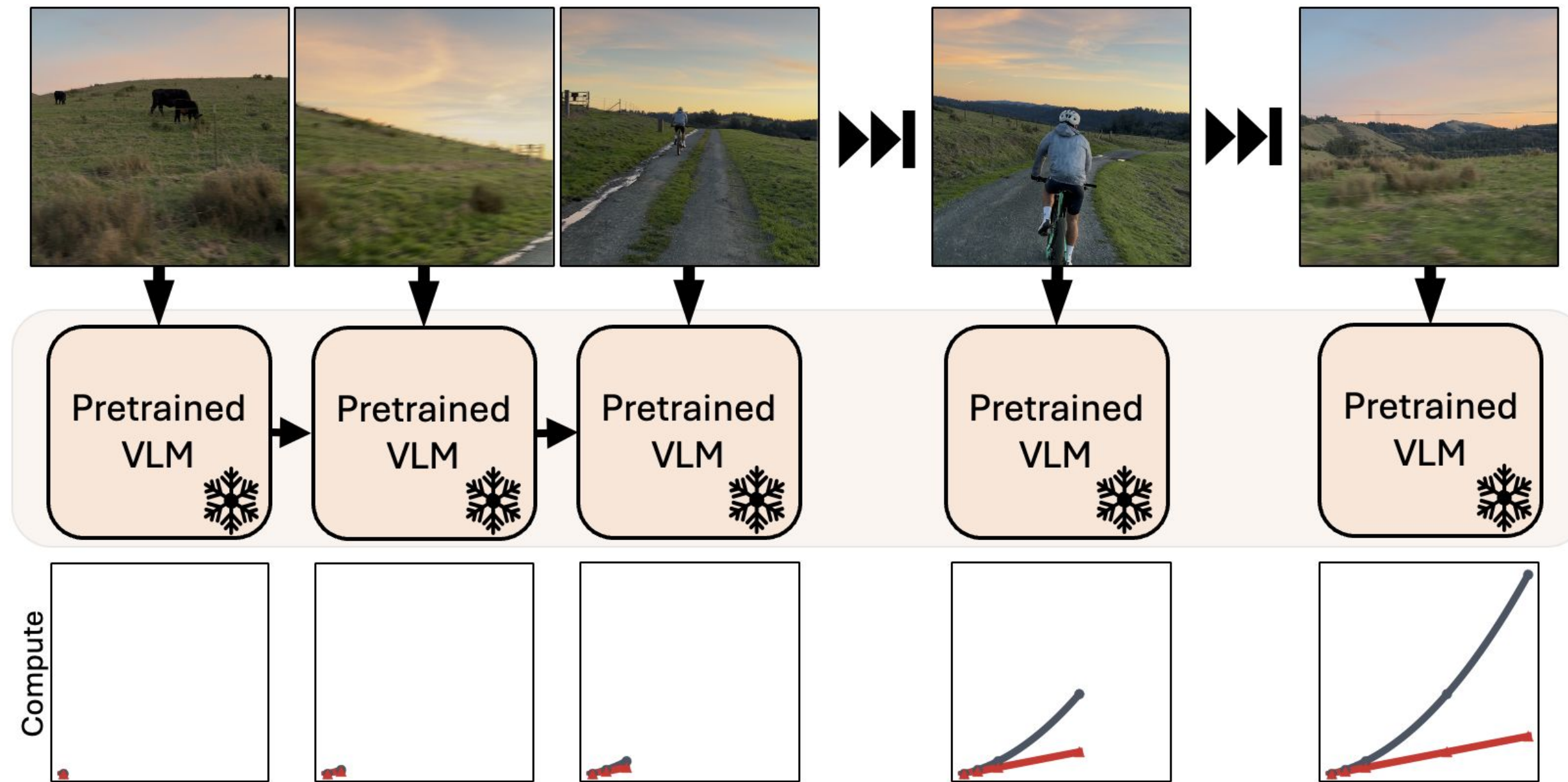


What we have

$$O(N^2)$$

dense self-attention over video tokens

StateKV: linear scaling of pretrained VLMs



What we have

$O(N^2)$

dense self-attention over video tokens

What we need

$O(N)$

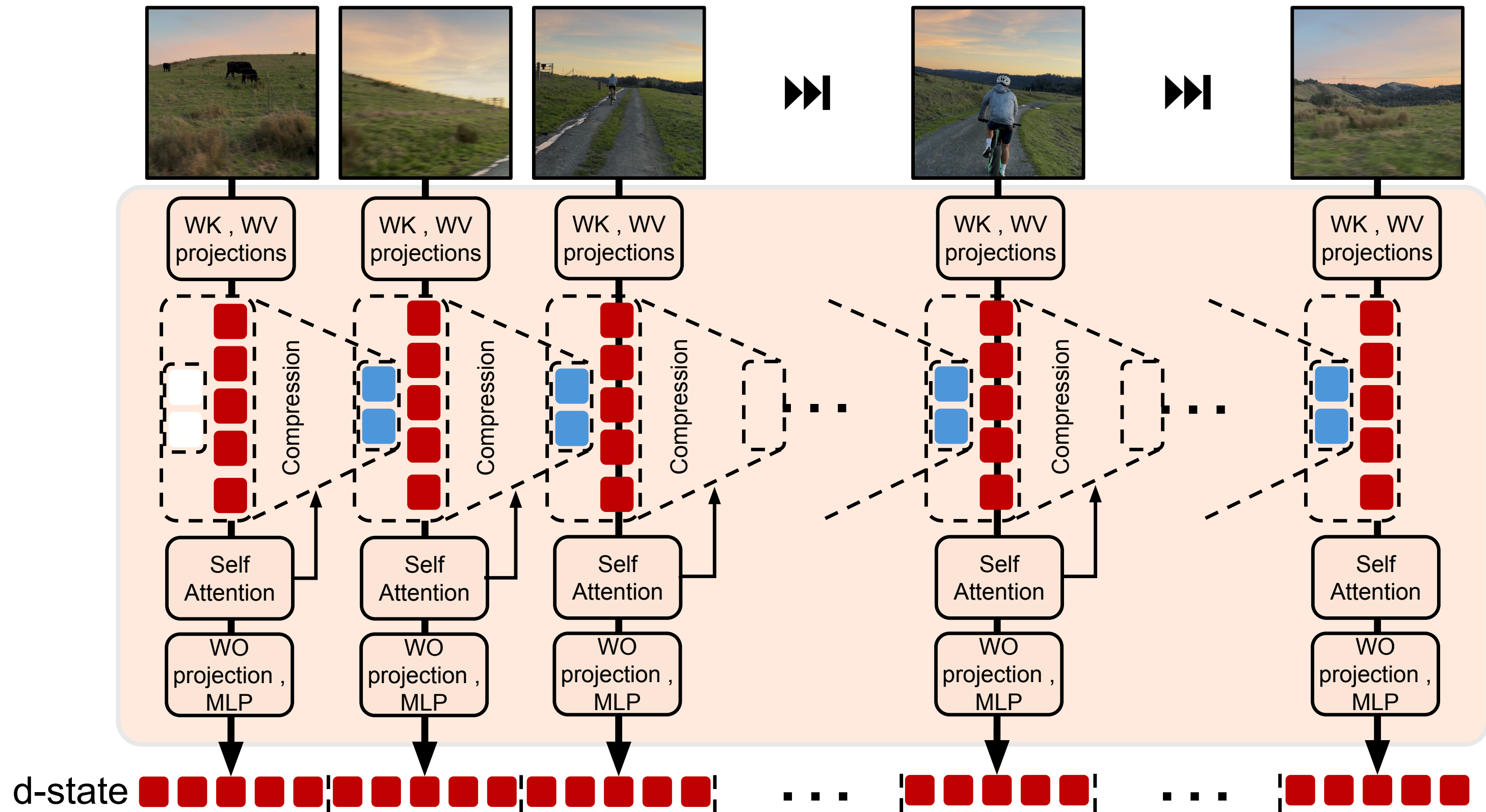
constant cost per frame · linear total

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

StateKV algorithm: compressed memory - cstate



StateKV algorithm: update state recurrently

StateKV algorithm: update state recurrently

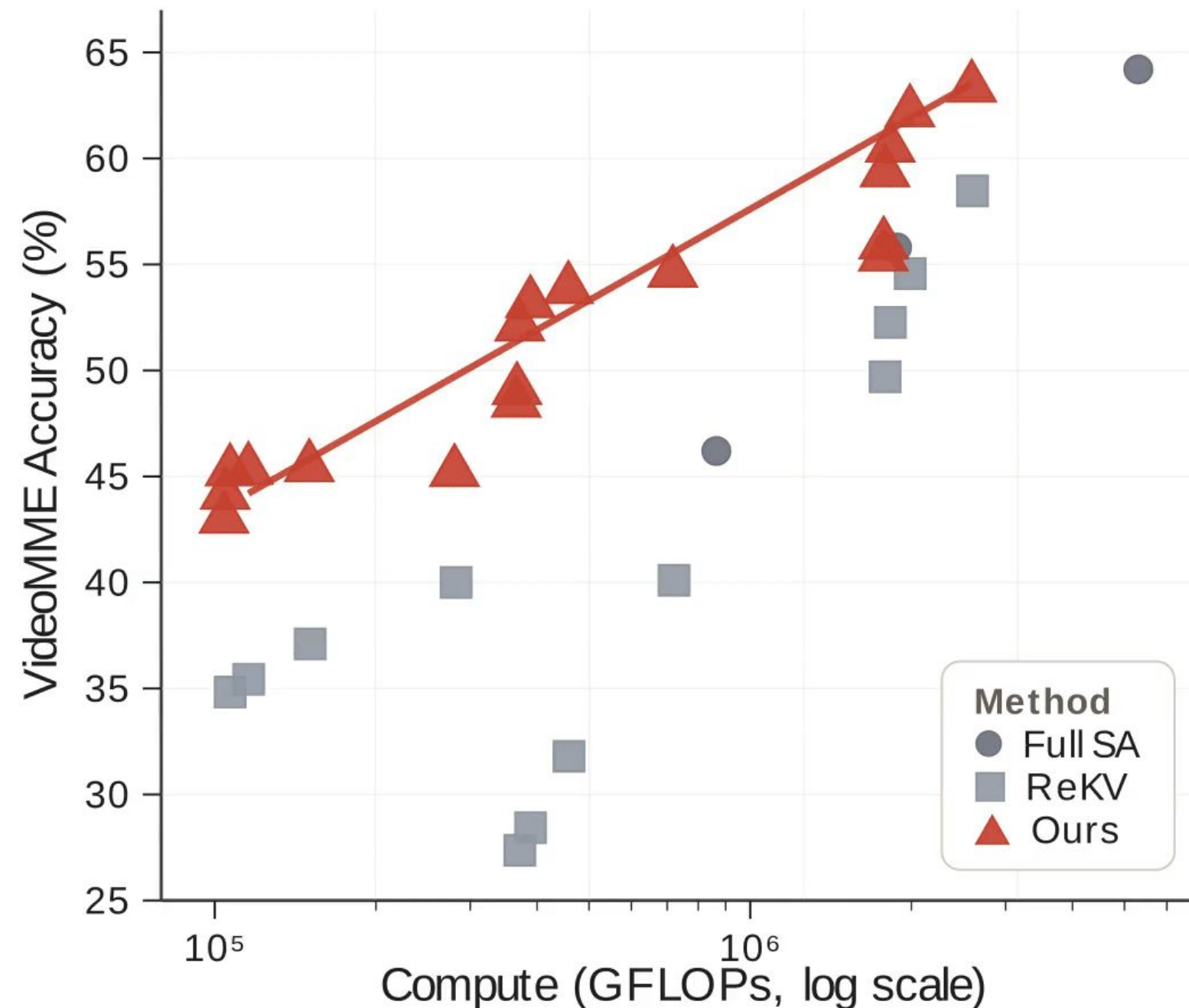


Accuracy–compute tradeoff



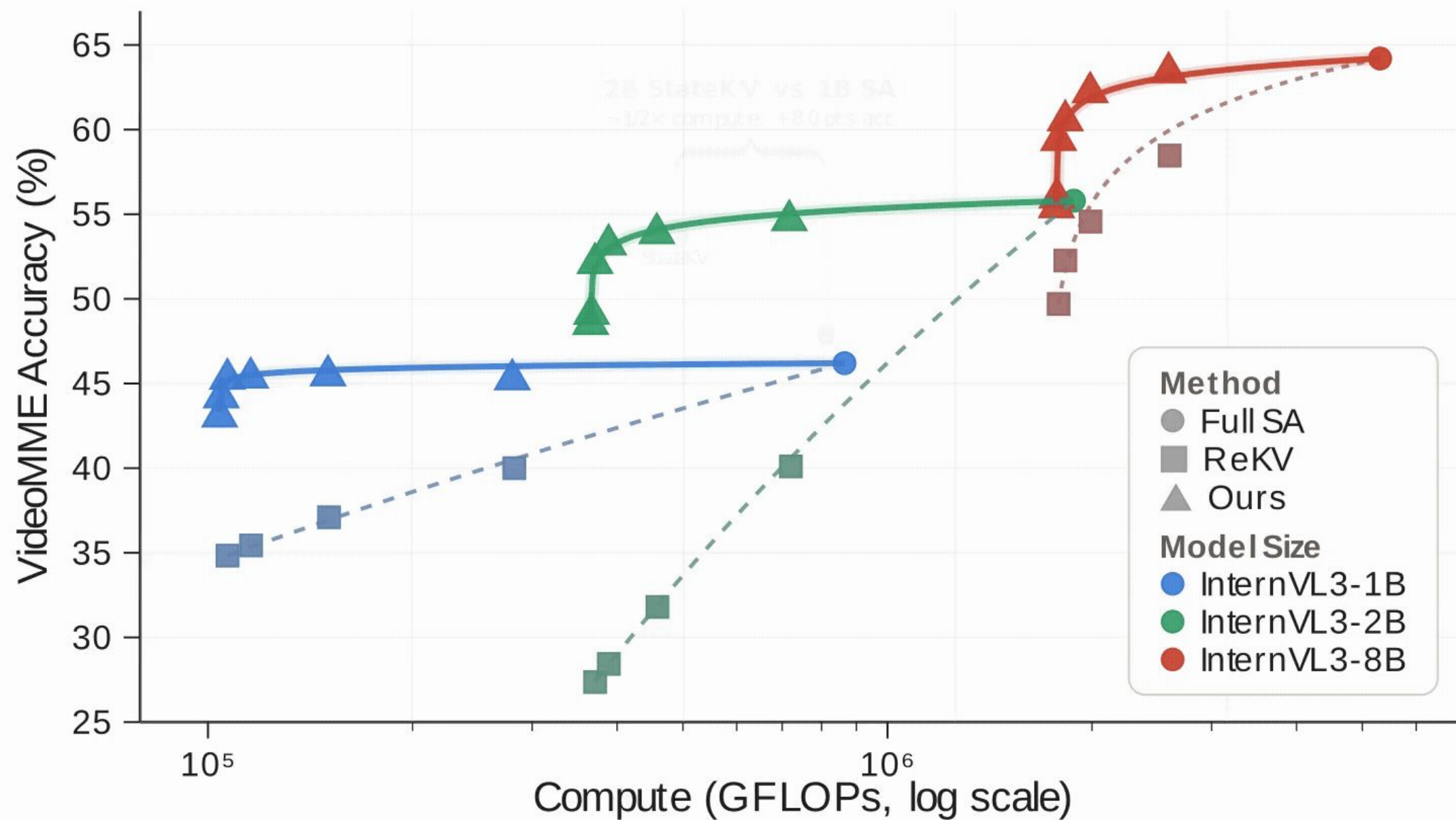
Accuracy–compute tradeoff

StateKV: Compute–Accuracy Pareto Frontier



Accuracy—larger models with same compute

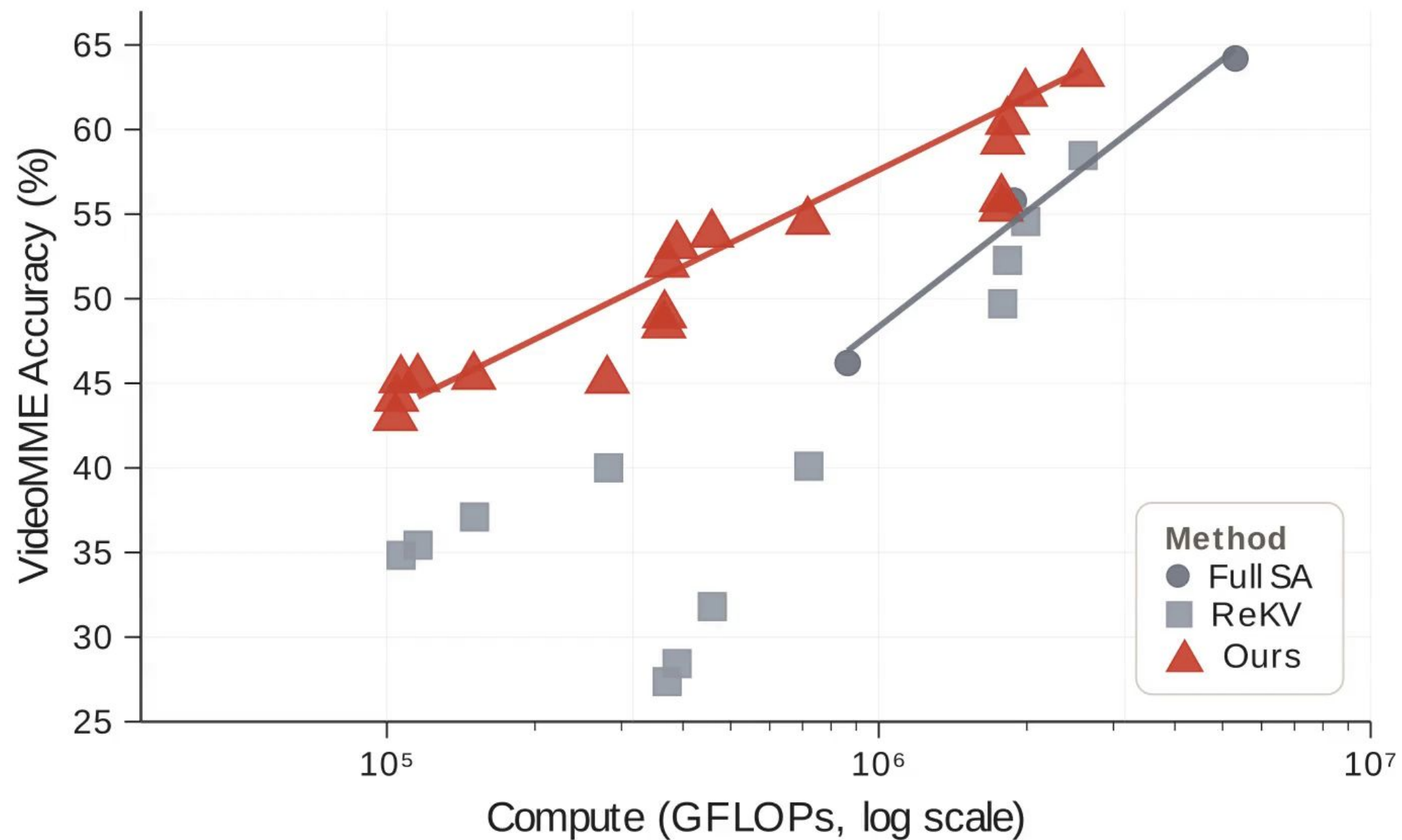
StateKV: Compute–Accuracy Pareto Frontier



The long-video regime

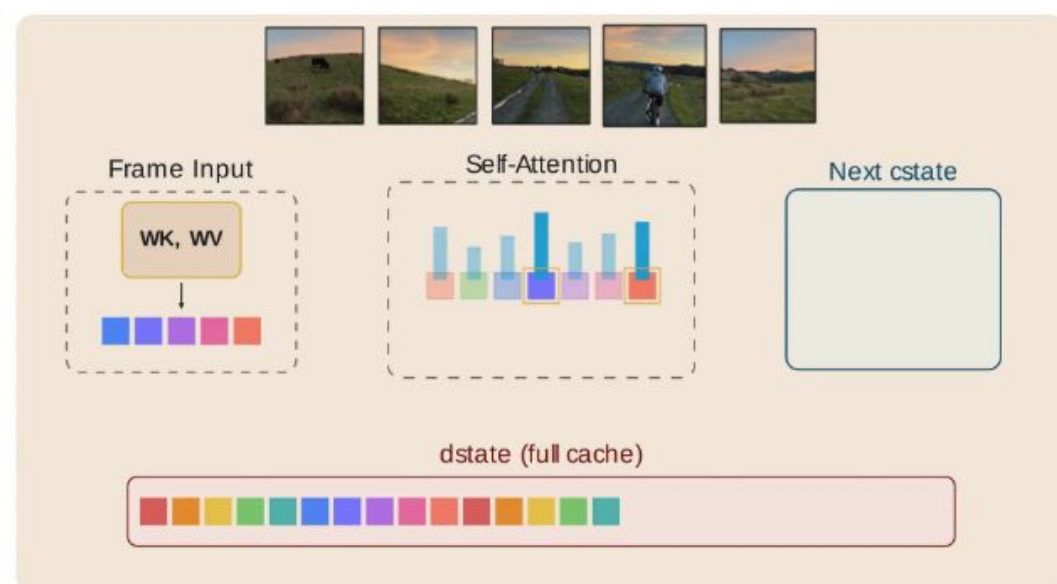
The compute advantage grows with video length

512 frames



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



Part II: Linear Inference for Streaming Videos



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



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



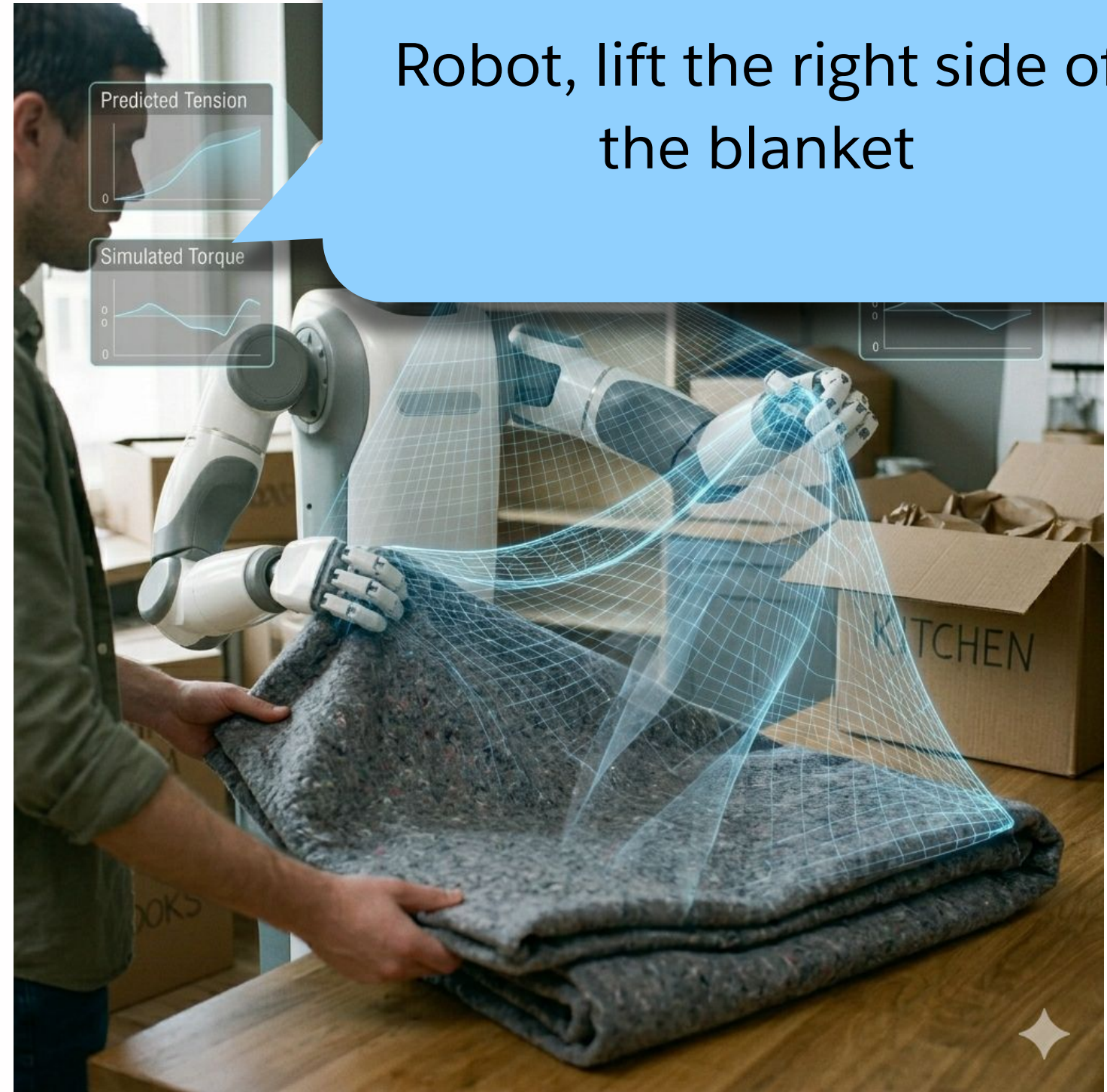
Part III:



What should my next action be? Generate a video to show me



Robot, lift the right side of the blanket

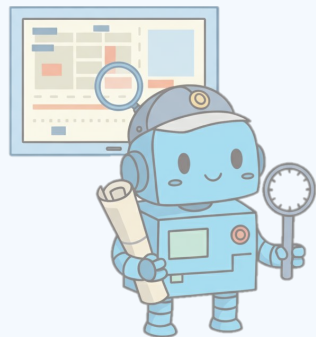


Agentic Ambient Intelligence



Reasoning over long videos

AVP
[CVPR Findings '26]



Fast inference for streaming videos

StateKV
[arxiv '26]



Motion Guidance Generation

FOFPred
[CVPR Findings '26]



Prior Work



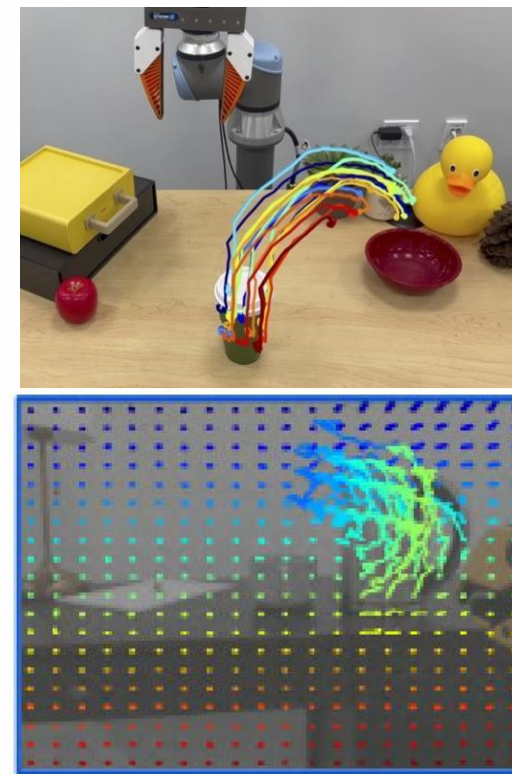
Grok



Veo



Specialized demos



SOTA struggles at *text* to fine-grained motions;
Unless given manual motion guidance [1]

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:

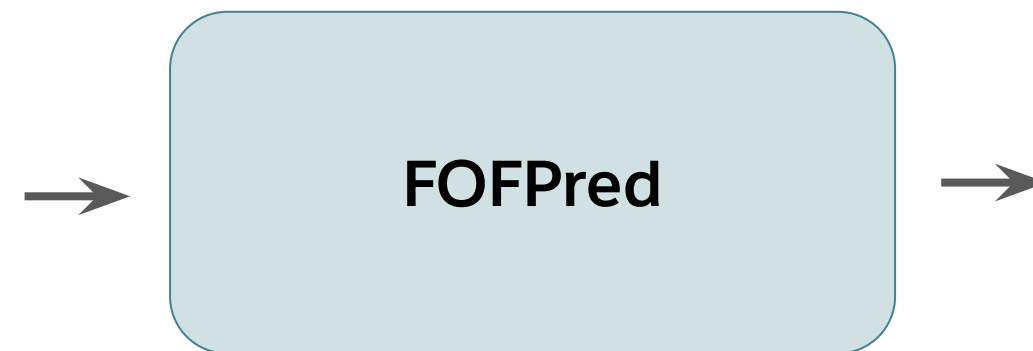
1. Remove manual motion guidance dependency
2. Leverage internet scale training data

FOFPred: Motion Guidance Generation



Learn to Predict Explicit “Motion”

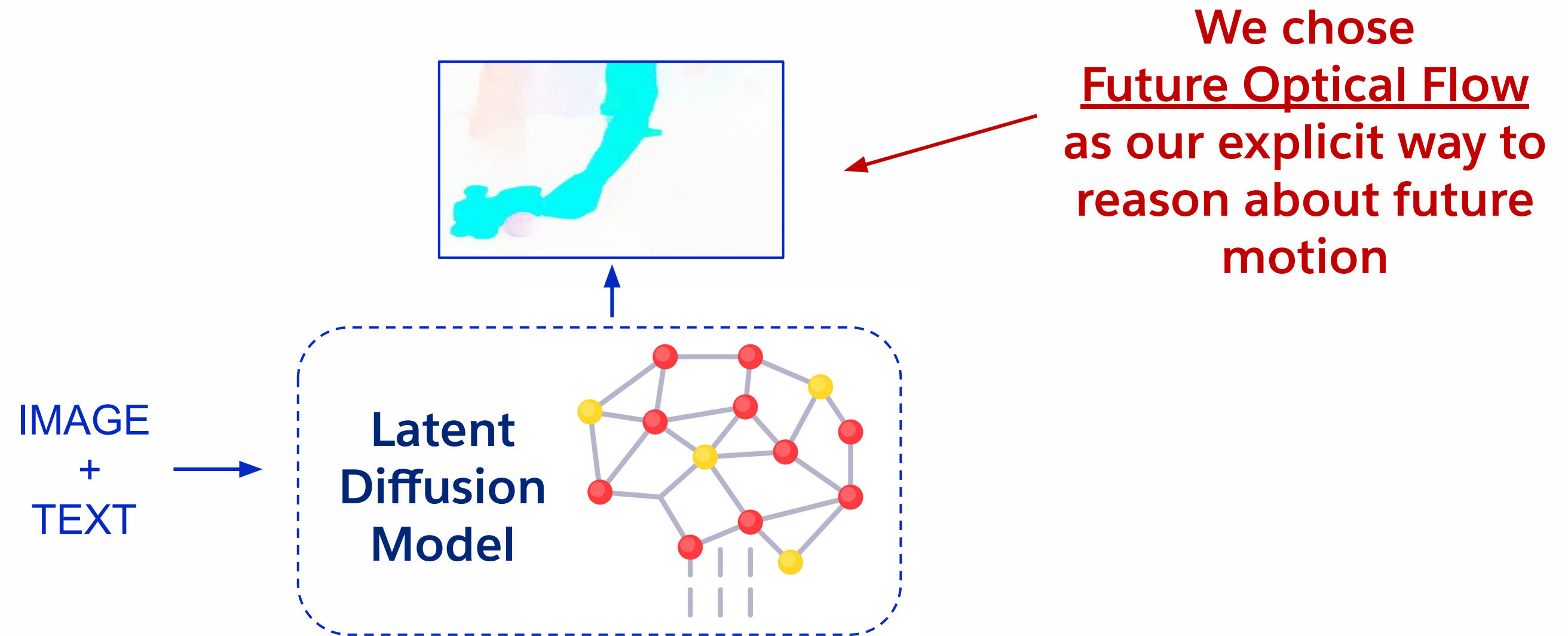
Future motion prediction



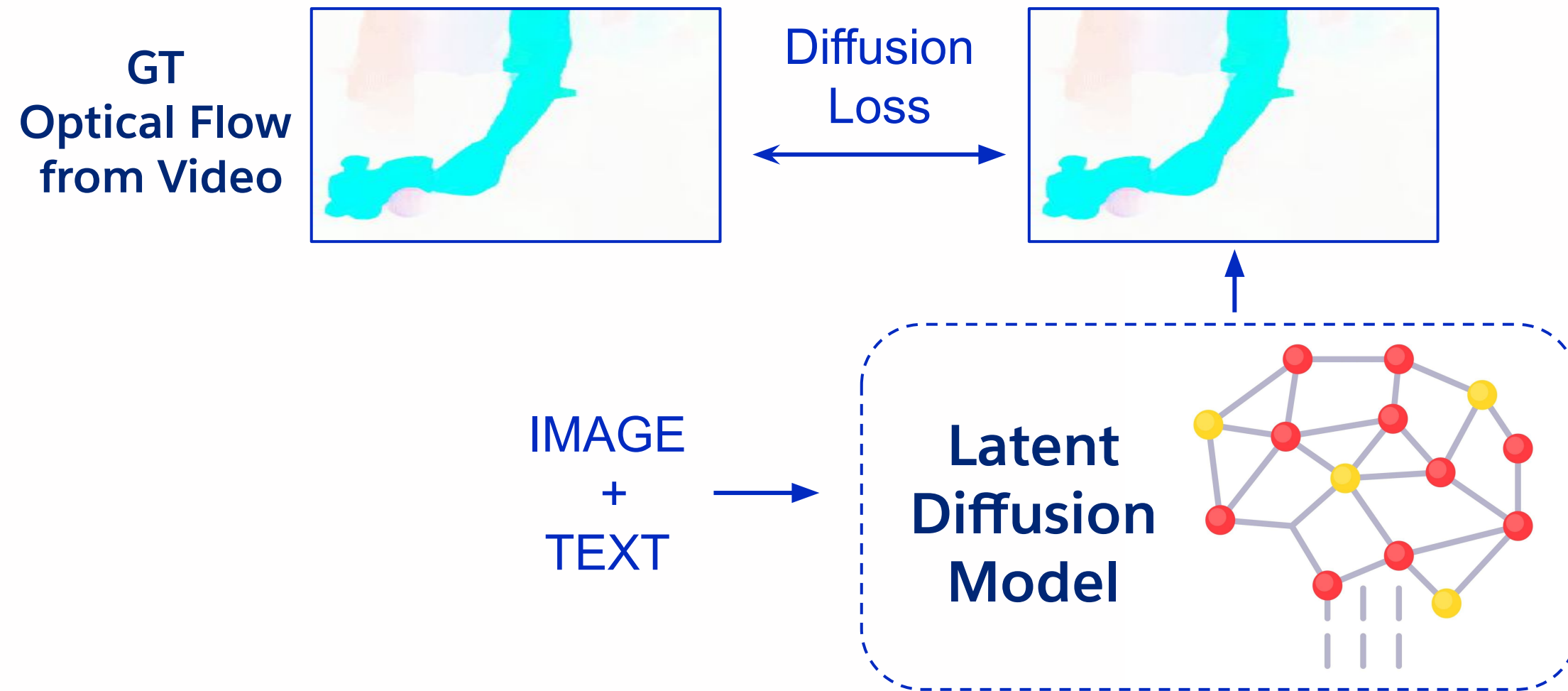
+

Generate motion for
“assembling drawer
into empty space”

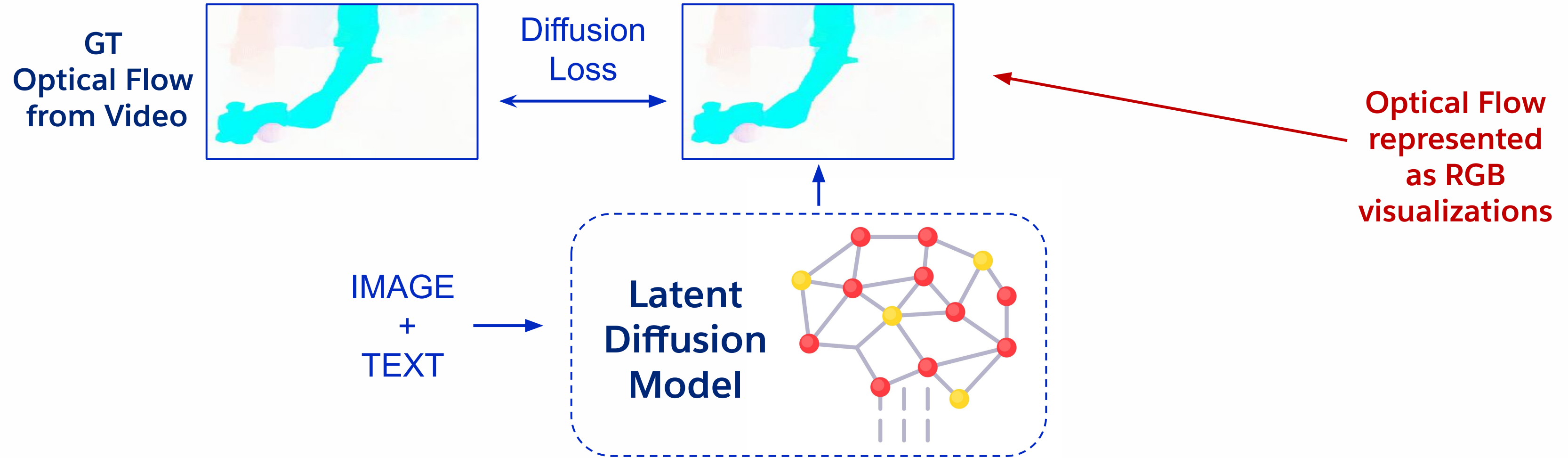
FOFPred: Motion Guidance Generation



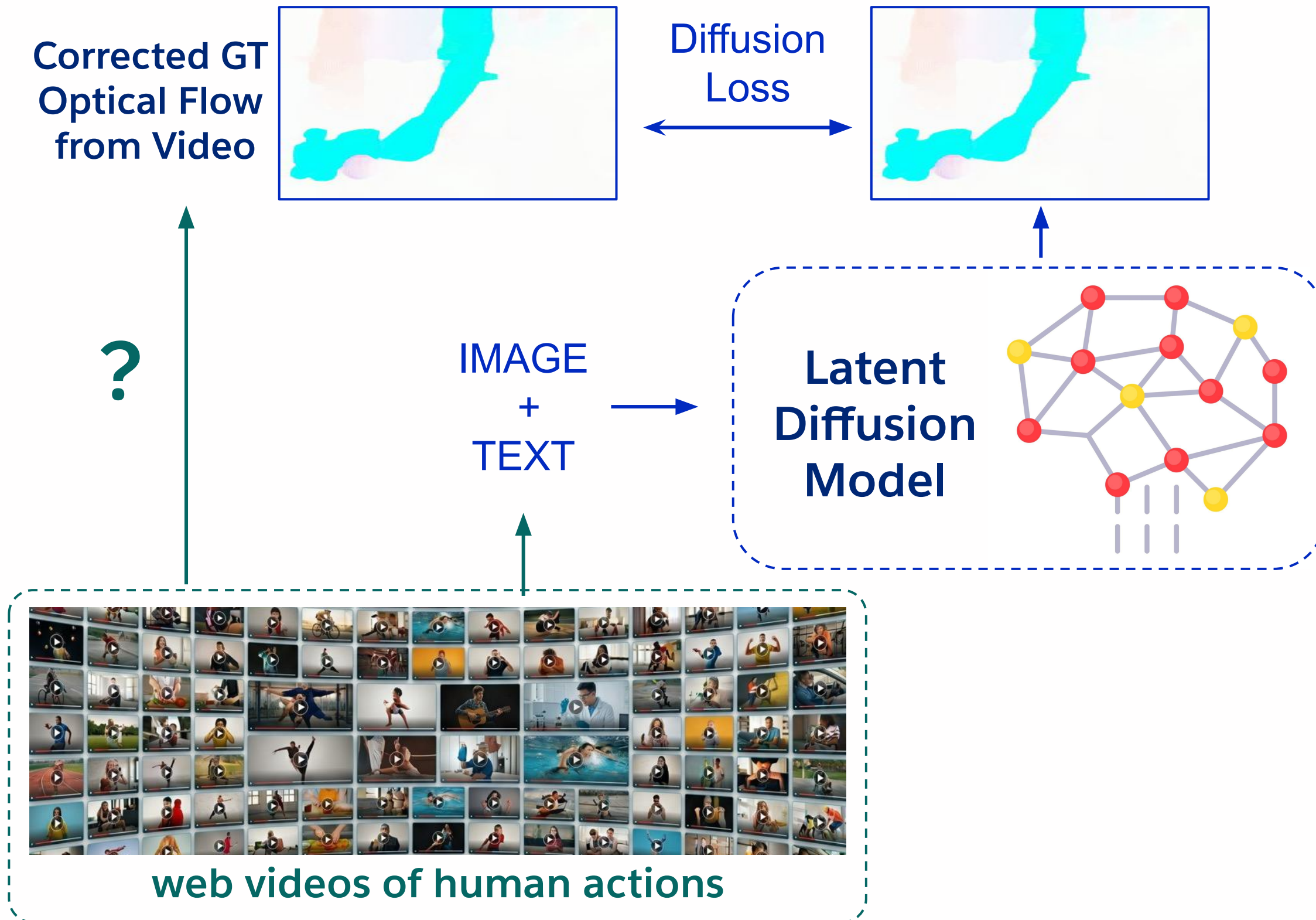
FOFPred: Motion Guidance Generation



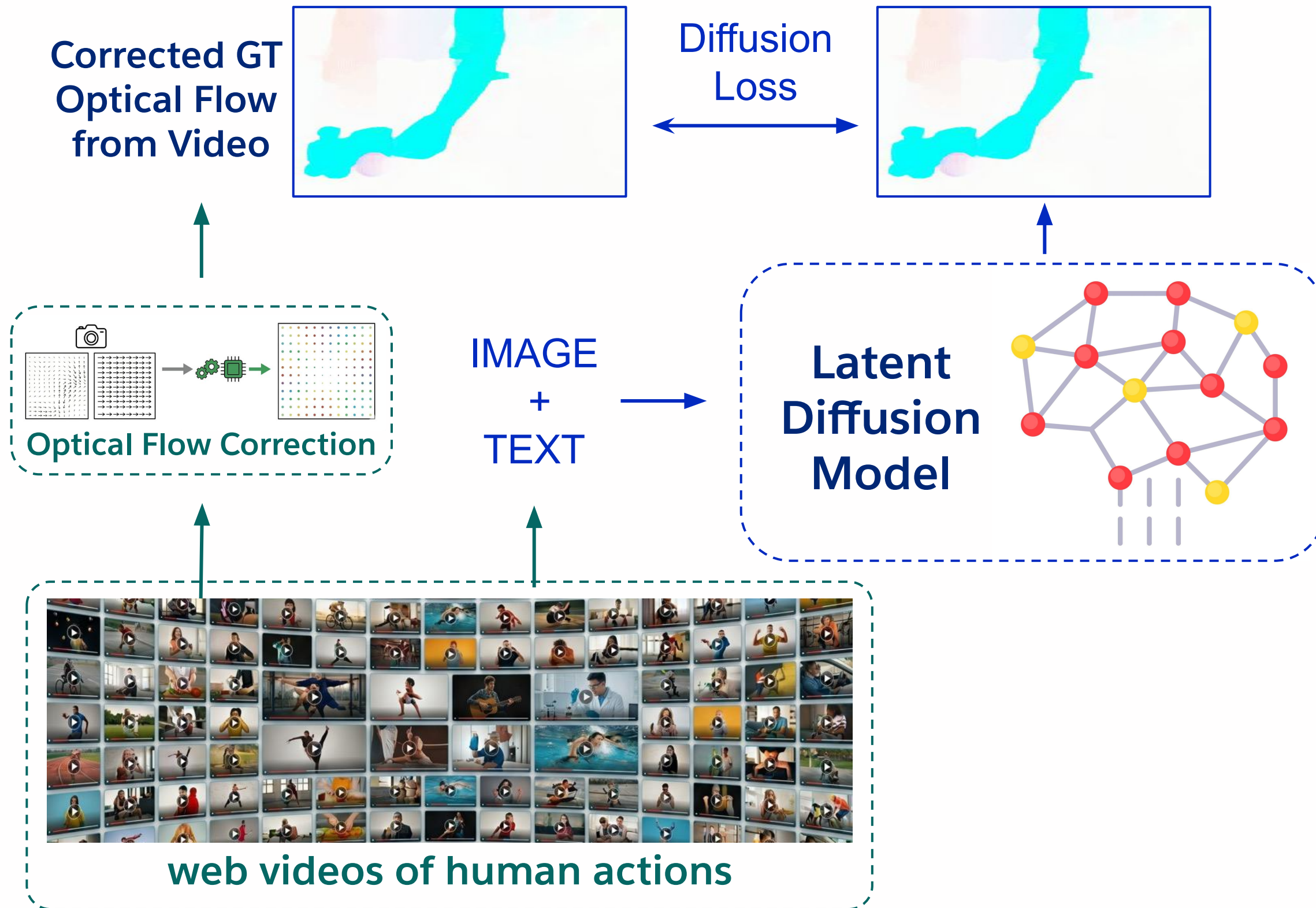
FOFPred: Motion Guidance Generation



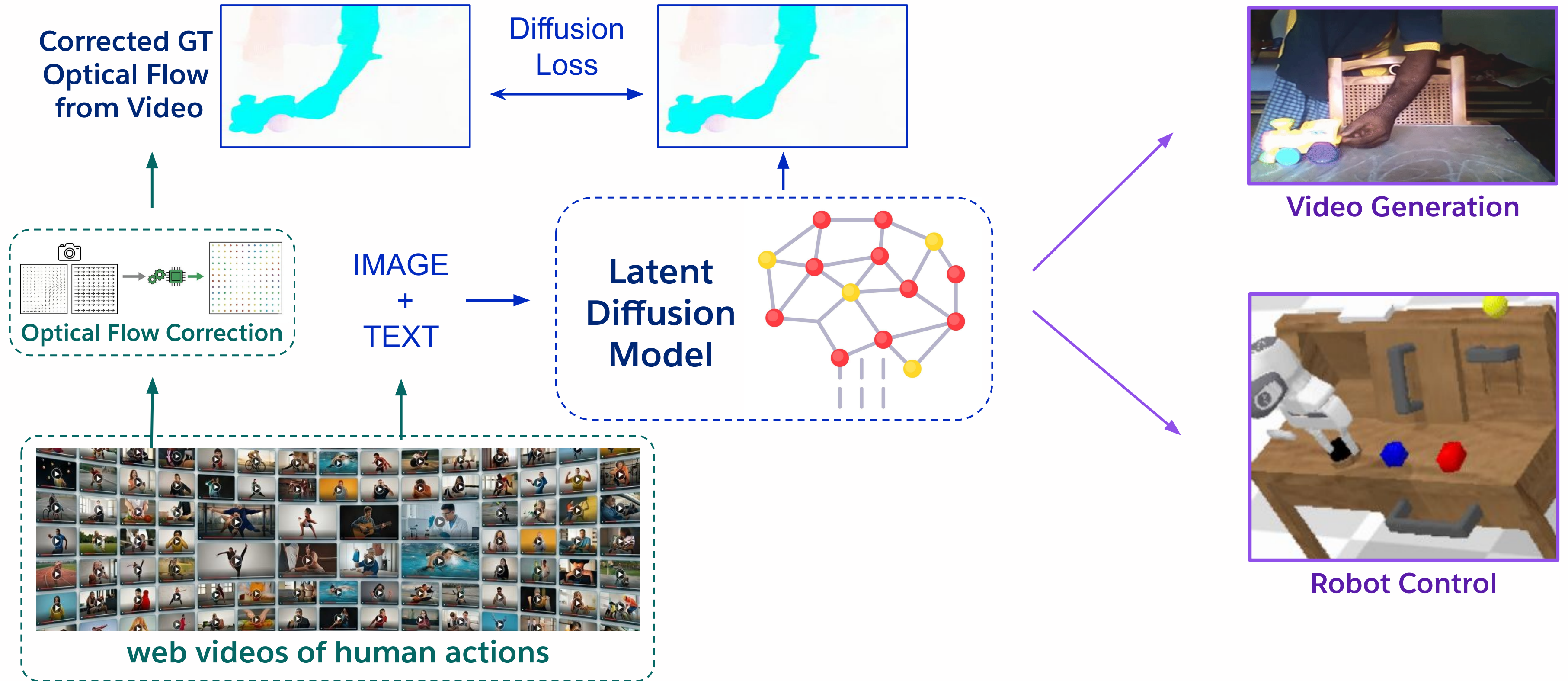
Scalable Learning



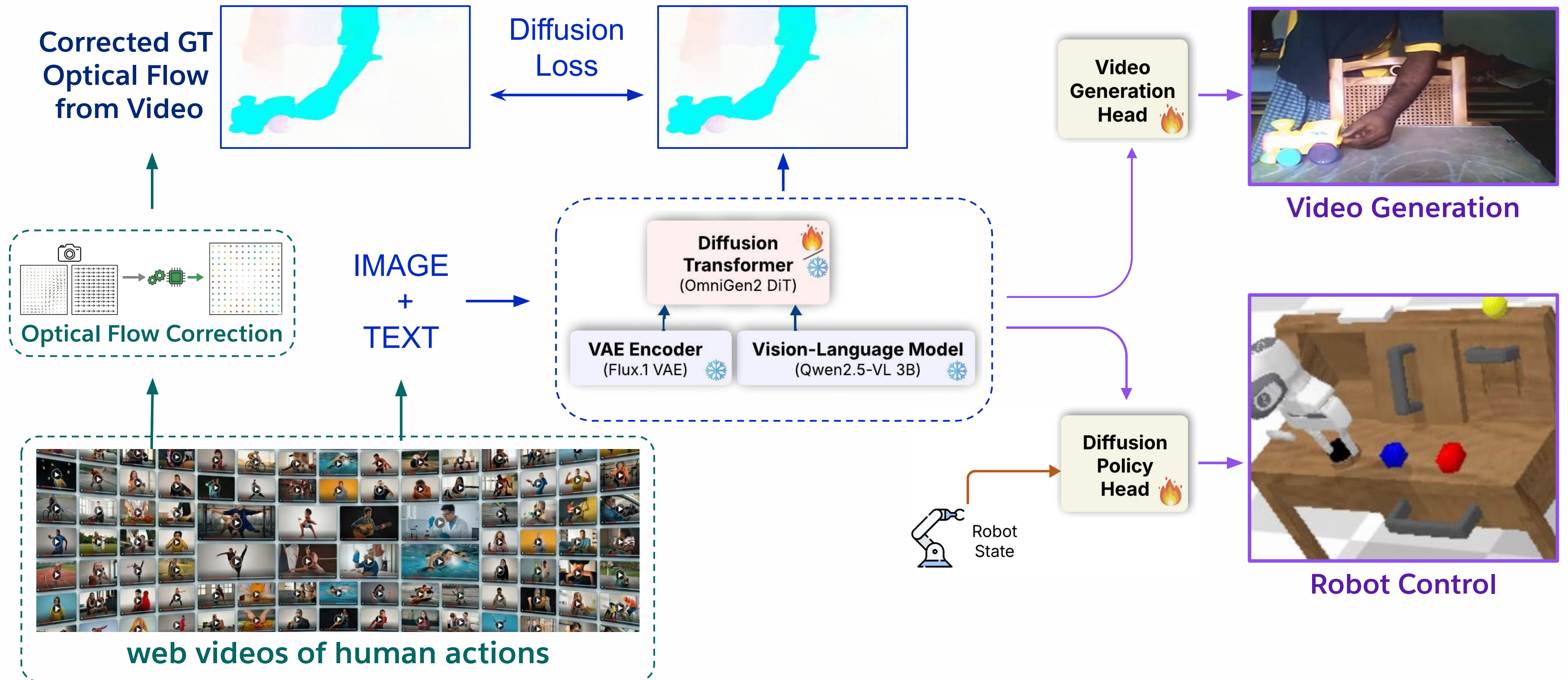
Scalable Learning



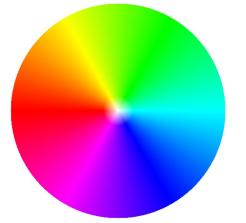
Downstream Tasks



FOFPred: Detailed Architecture



Video Generation

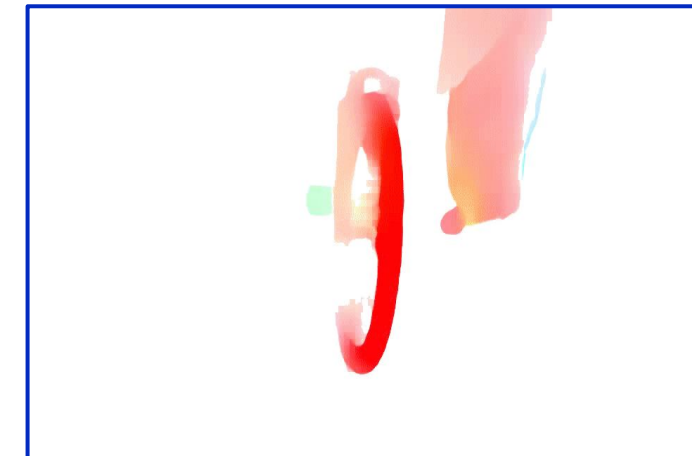
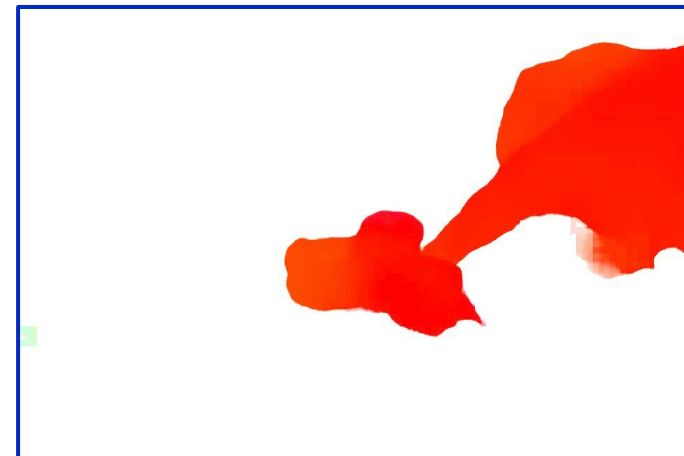


OF direction color wheel

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

Video Generation



Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow	FVD \downarrow	KVD \downarrow	MF \uparrow [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)

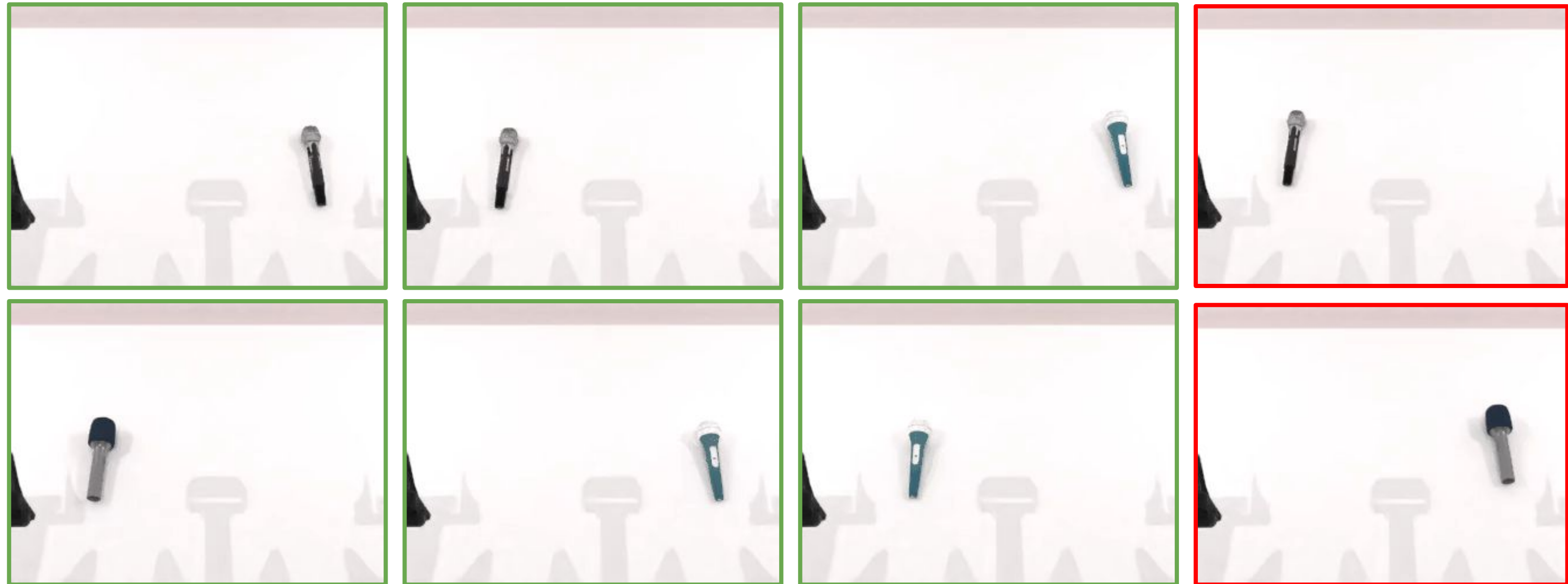
SSv2 Video Generation Evaluation: Over the CogVideoX baseline, our FOFPred framework shows consistent improvements in generation quality.

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
π_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)

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

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

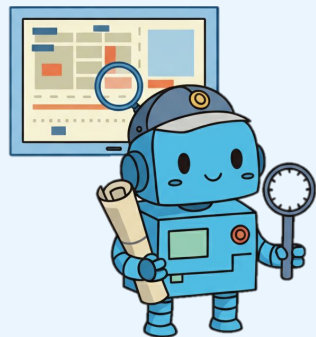


Agentic Ambient Intelligence



Reasoning over long videos

AVP
[CVPR Findings '26]



Fast inference for streaming videos

StateKV
[arxiv '26]



Motion Guidance Generation

FOFPred
[CVPR Findings '26]



Agentic Ambient Intelligence





Thank
you

Four yellow starburst graphics are scattered around the text: one to the left of "you", one below the "T" of "Thank", one to the right of "Thank", and one to the right of "you".

Resources



1. Zhou et al. “Strefer: Empowering Video LLMs with Space-Time Referring and Reasoning via Synthetic Instruction Data”. ICCVW, 2025.
2. Panagopoulou et al. “Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D”. EMNLP 2025.
3. Wang et al. “Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding”. CVPR Findings, 2026.
4. Ranasinghe et al. “Future Optical Flow Prediction Improves Robot Control and Video Generation”. CVPR Findings, 2026.



Motion Guidance Generation

