Agentic Ambient Intelligence

Perception, Reasoning & Action

ICLR Workshops
April 25, 2026

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

Agentic Ambient Intelligence

Part I

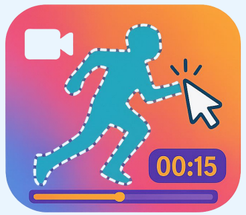

Should *this* board be installed horizontally, like the one I *just finished*?
Matched to previous installation type (Horizontal)
?


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

Agentic Ambient Intelligence

Video QA with Space-time references

Strefer
[ICCVW'25]



Existing Video LLMs

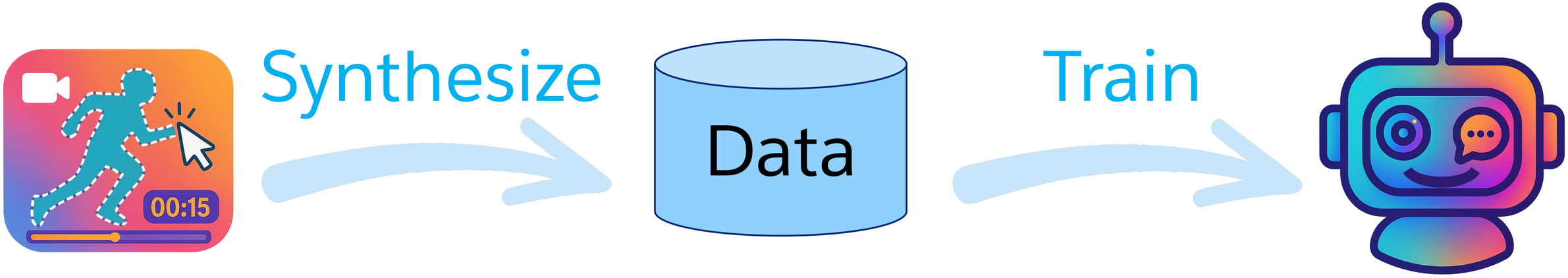
sampled video frames

What happens to him at 00:10?

I don’t know because I struggle to interpret spatial and timestamp references.

Our work: Strefer

sampled video frames



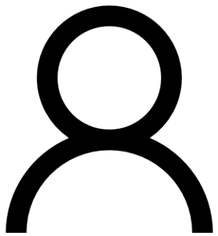
Strefer synthesizes instruction tuning data to empower Video LLMs to better interpret space-time information.

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

Strefer: new capabilities

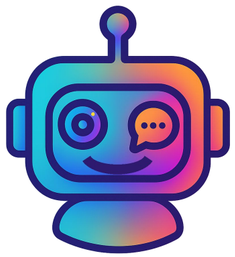
sampled video frames

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?

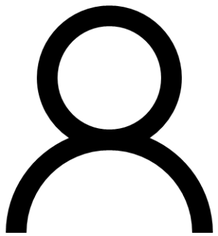
: The boy wearing glasses is not visible from 00:03 to 00:09.

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

Strefer: new capabilities

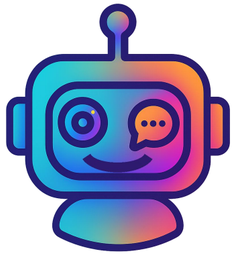
sampled video frames

interpret user queries with mask and timestamp references



:

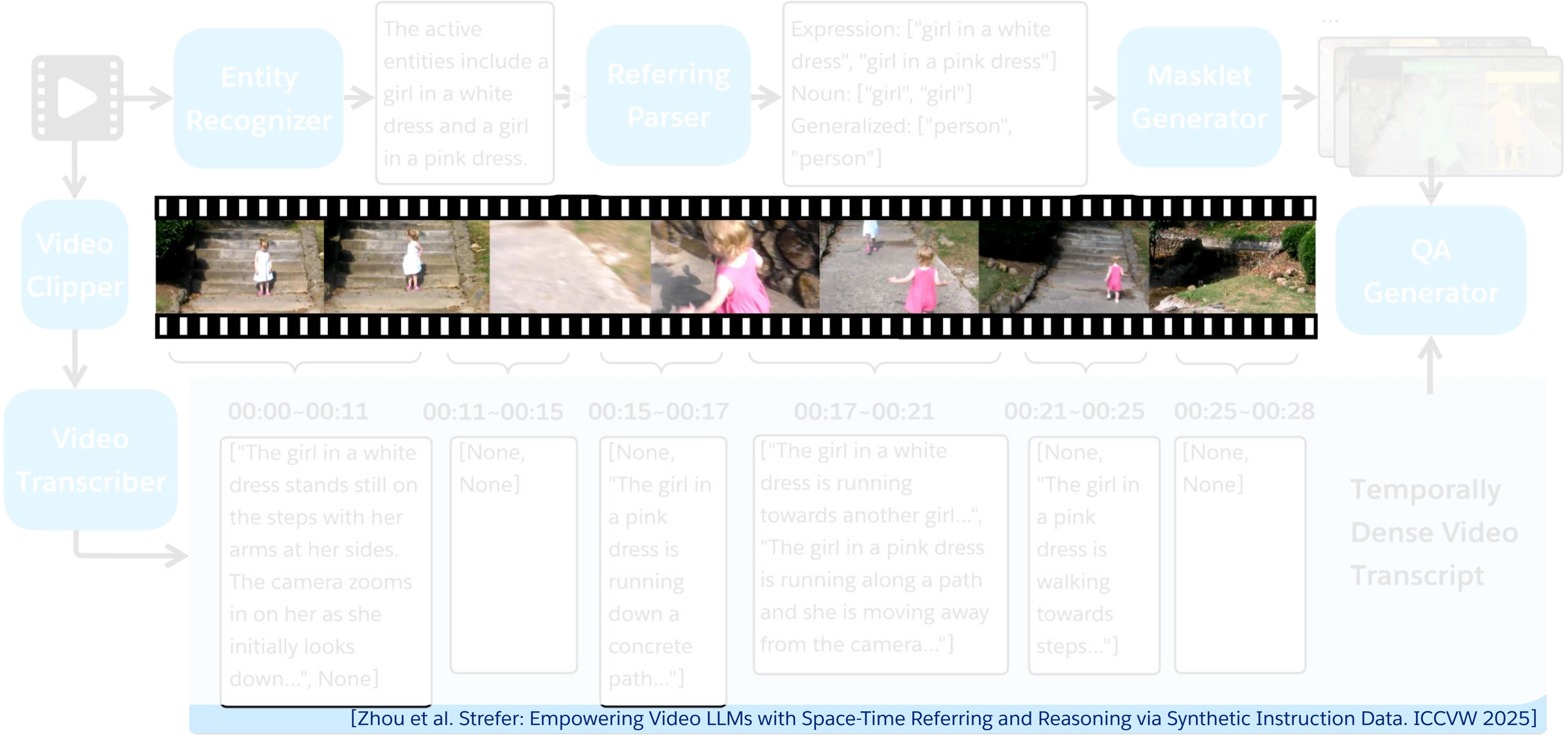
What did he do at 00:10?



: 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?

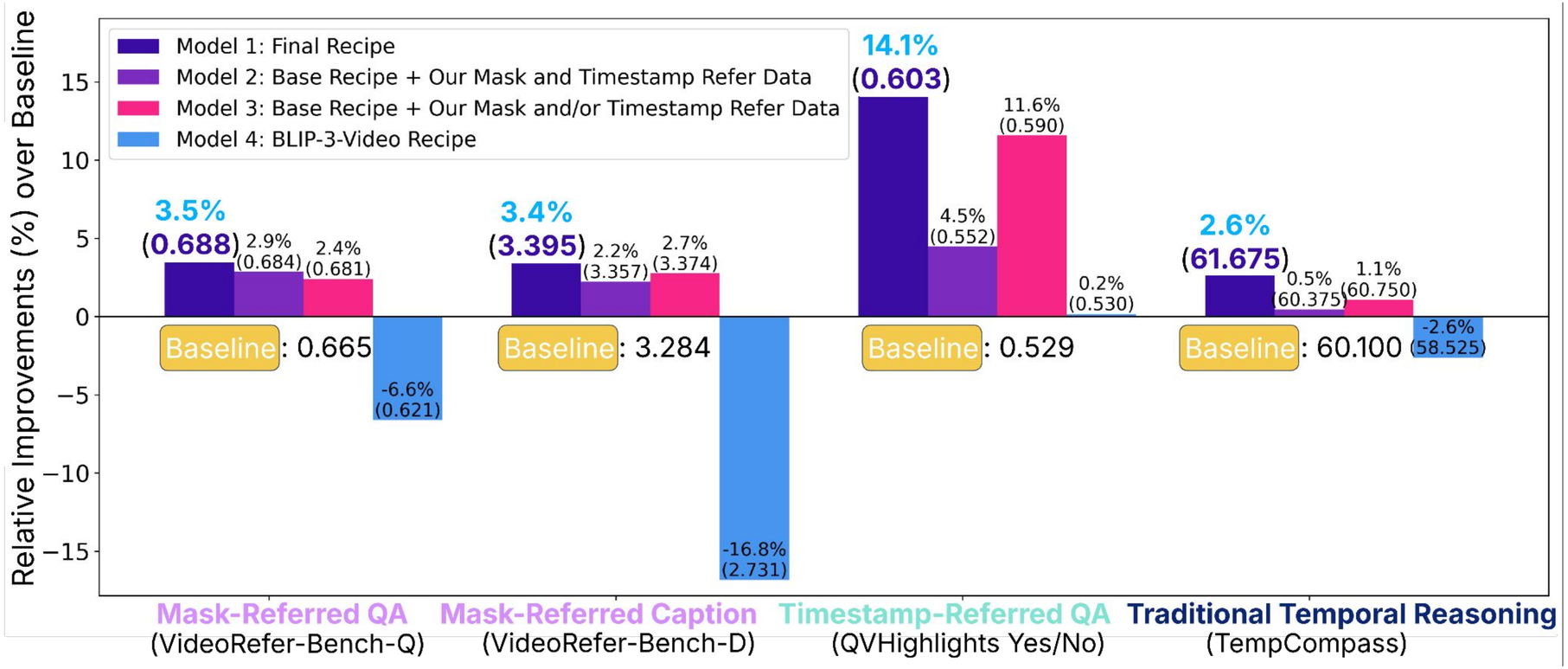
Response: Interacting with a toy guitar, moving it from a basket to the floor.

[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

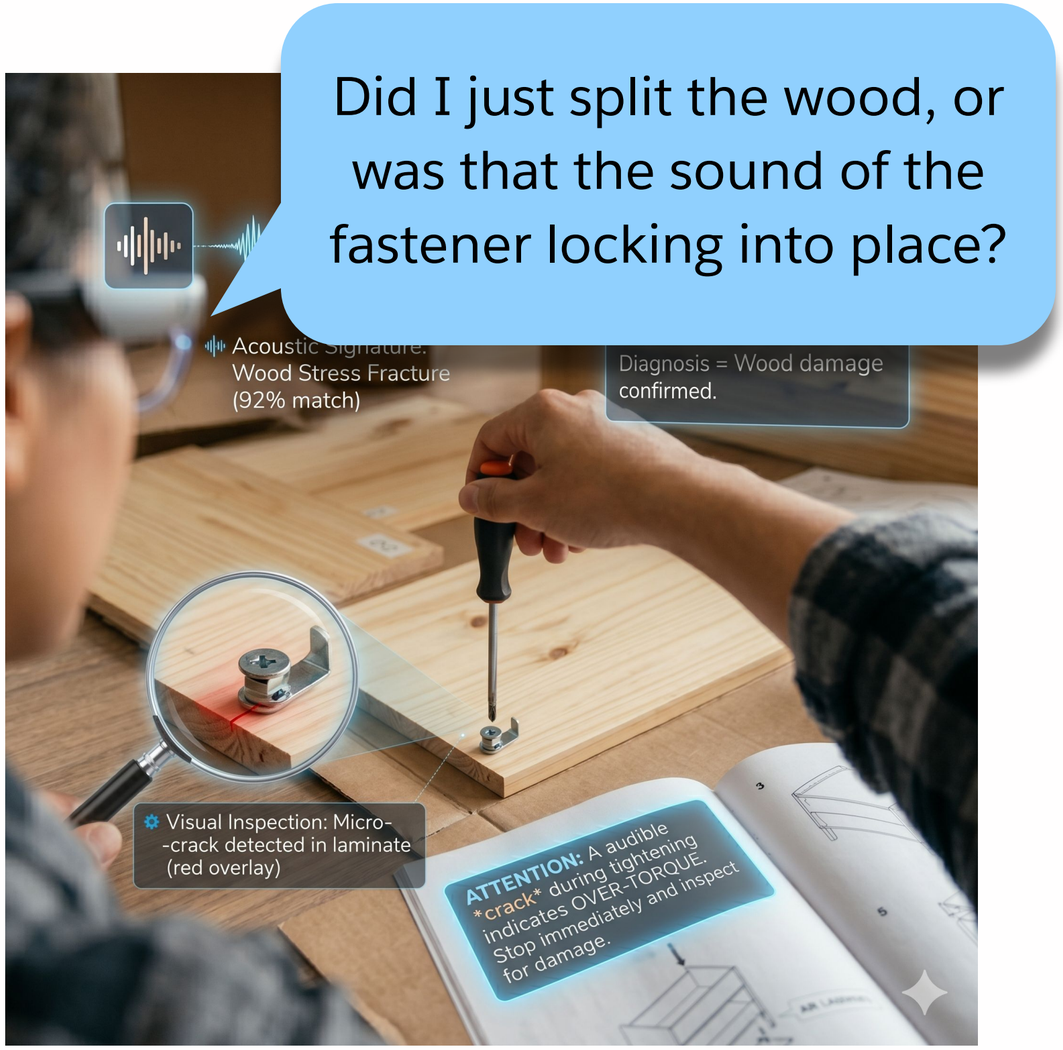
Walkthrough

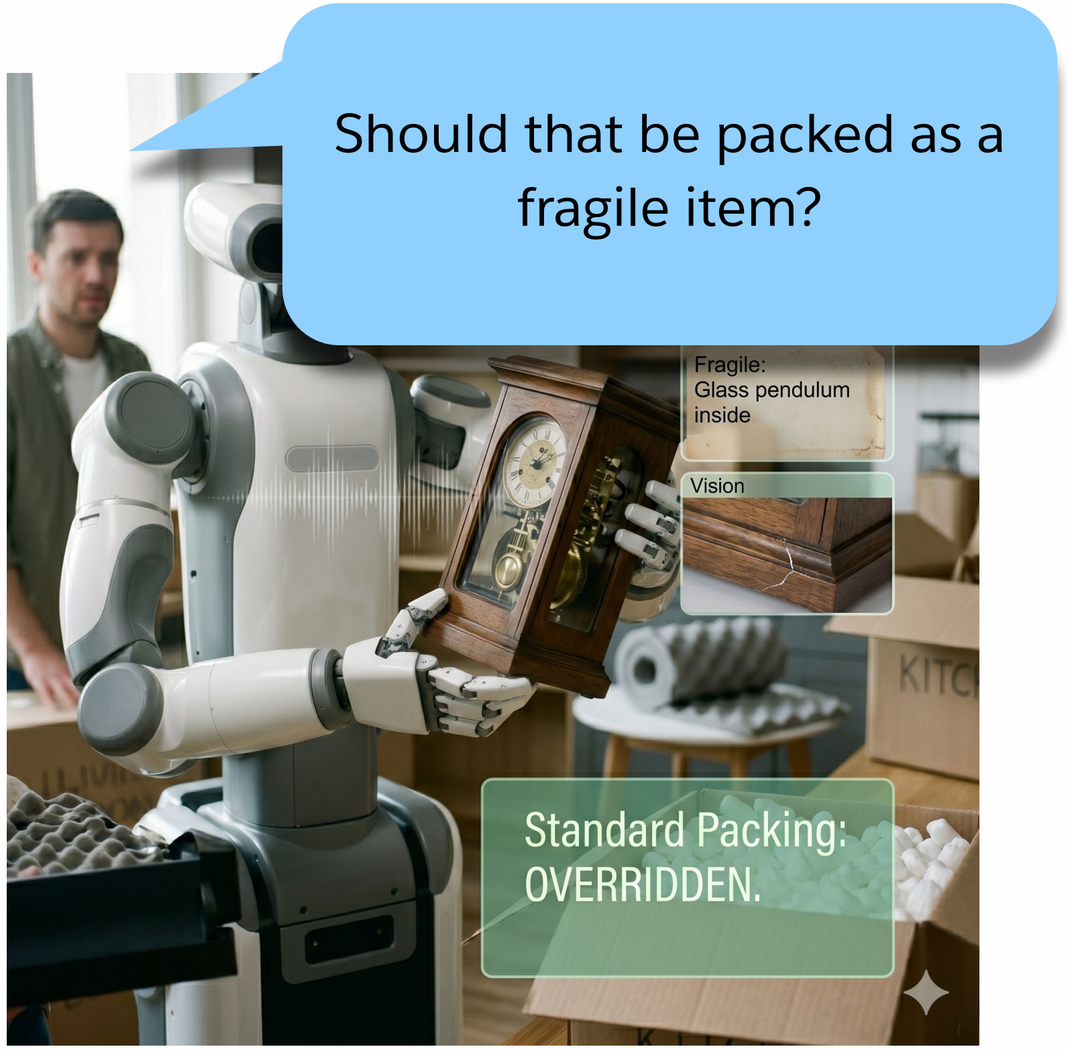
Part I


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

Part II


Did I just split the wood, or was that the sound of the fastener locking into place?
Acoustic Signature: Wood Stress Fracture (92% match)
Diagnosis = Wood damage confirmed.
Visual Inspection: Micro--crack detected in laminate (red overlay)
ATTENTION: A audible *crack* during tightening indicates OVER-TORQUE. Stop immediately and inspect for damage.


Should that be packed as a fragile item?
Fragile: Glass pendulum inside
Vision
Standard Packing: OVERRIDDEN.

Agentic Ambient Intelligence

Video QA with Space-time references

Strefer
[ICCVW'25]

Cross-modal Reasoning

Contra4
[EMNLP'25]

Contrastive Cross-Modal Reasoning

Real-world systems receive multiple sensory inputs: images, audio, video, 3D scans

Task: select the most relevant modality to a particular query

How do can we evaluate this?

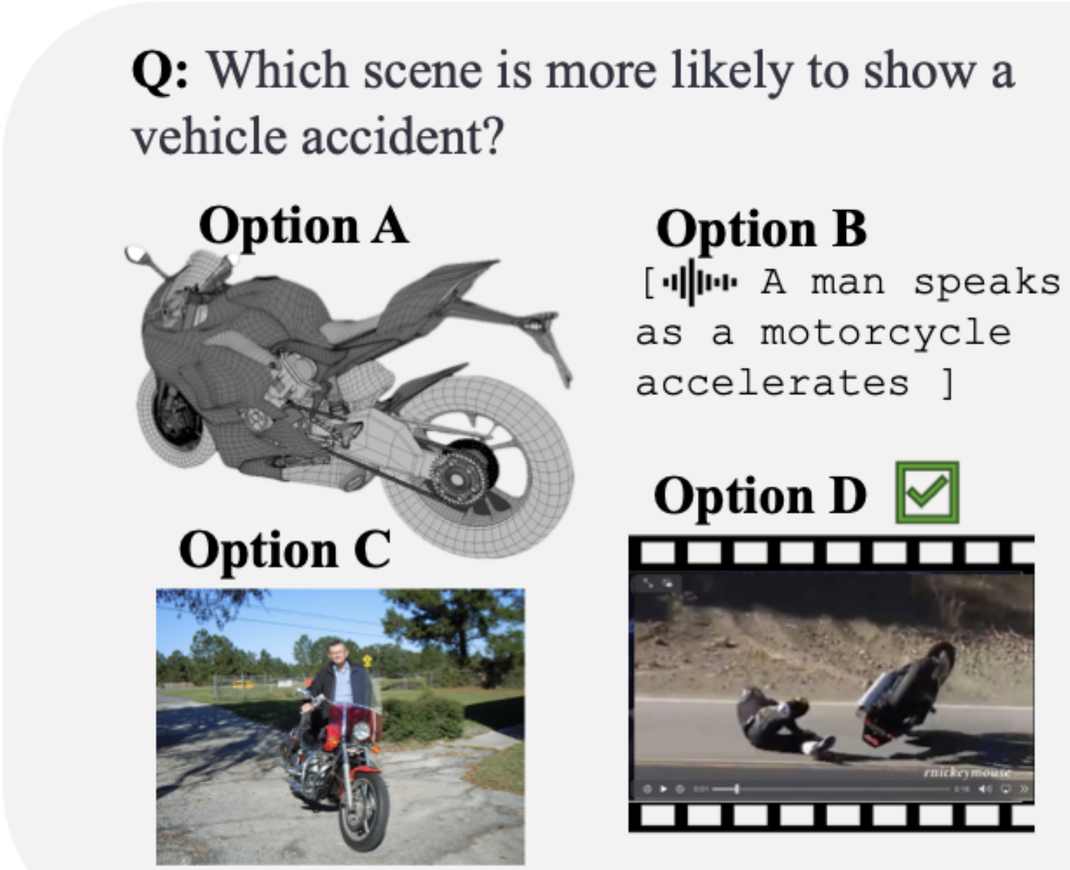
[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Benchmarking cross-modal reasoning

Task Definition: Given inputs from multiple modalities and a query, the model must decide which modality best corresponds to the query.

Challenge: Collecting data queries involving multiple modalities is scarce and is costly to manually annotate.

Key idea: Reuse single modality benchmarks, by finding related data points via language



[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Synthetic Data Generation

Build contrastive cross-modal reasoning data with captions as a universal connector.

Step 1. Sampling

3D Captions
A digital camera… A small purple…

Image Captions
A group of people… A white statue…

Audio Captions
A woman talks while… Loud intermittent…

Video Captions
A person eats a… A baby with pink…

Random Sampling
- A pyramid-shaped…
- A car motor revs…
- A catcher in a…
- A boy looking at…

Similarity Sampling
- A white fridge…
- A machine whirls.
- A refrigerator…
- He opens the fridge…

Step 2. Question Generation

In-Context-Examples LLM Prompt

Scene A. a shattered piece of paper, resembling a broken phone and a flying newspaper
Scene B. tourists walking near a catholic church in Mexico on a sunny summer day
Generated Question: Which input evokes a sense of chaos and abandonment?
⋮
Scene A. People putting their suitcases onto a train.
Scene B. A train approaching on the tracks and a car reversing and revving as it drives away
Generated Question: **Which scene shows a train?**

Step 3. Answer Generation

In-Context-Examples LLM Prompt

Scene A. People putting their suitcases onto a train.
Scene B. A train approaching on the tracks and a car reversing and revving as it drives away
Question: Which scene shows a train? Answer: **Scene A**

Step 4. Mixture-of-Models Round-Trip-Consistency

Scenes | Question | Answer

- LLM A: Scene B ≠ Scene A ✗
- LLM B: Scene A = Scene A ✓
- LLM C: Scene A = Scene A ✓

Majority Filter Pass | Unanimous Filter Fail

Permuted Scenes | Question | Answer

- LLM A: Scene A ≠ Scene B ✗
- LLM B: Scene B = Scene B ✓
- LLM C: Scene B = Scene B ✓

Permute Majority Filter Pass | Permute Unanimous Filter Fail

[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

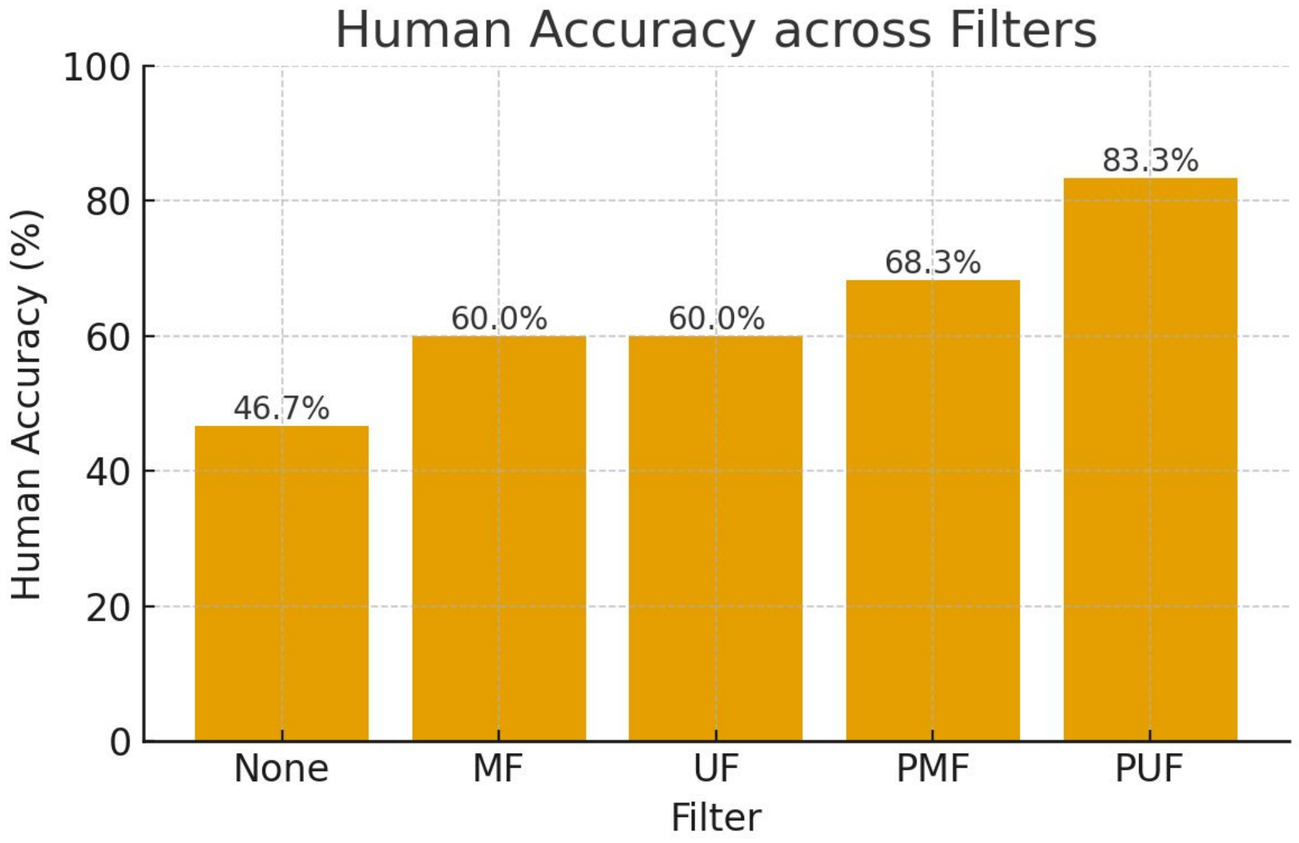
Contra4: MoM-RTC Accuracy

Dataset Scale

- 174k automatically annotated training samples
- 2.3k manually annotated test samples

Filtering methods

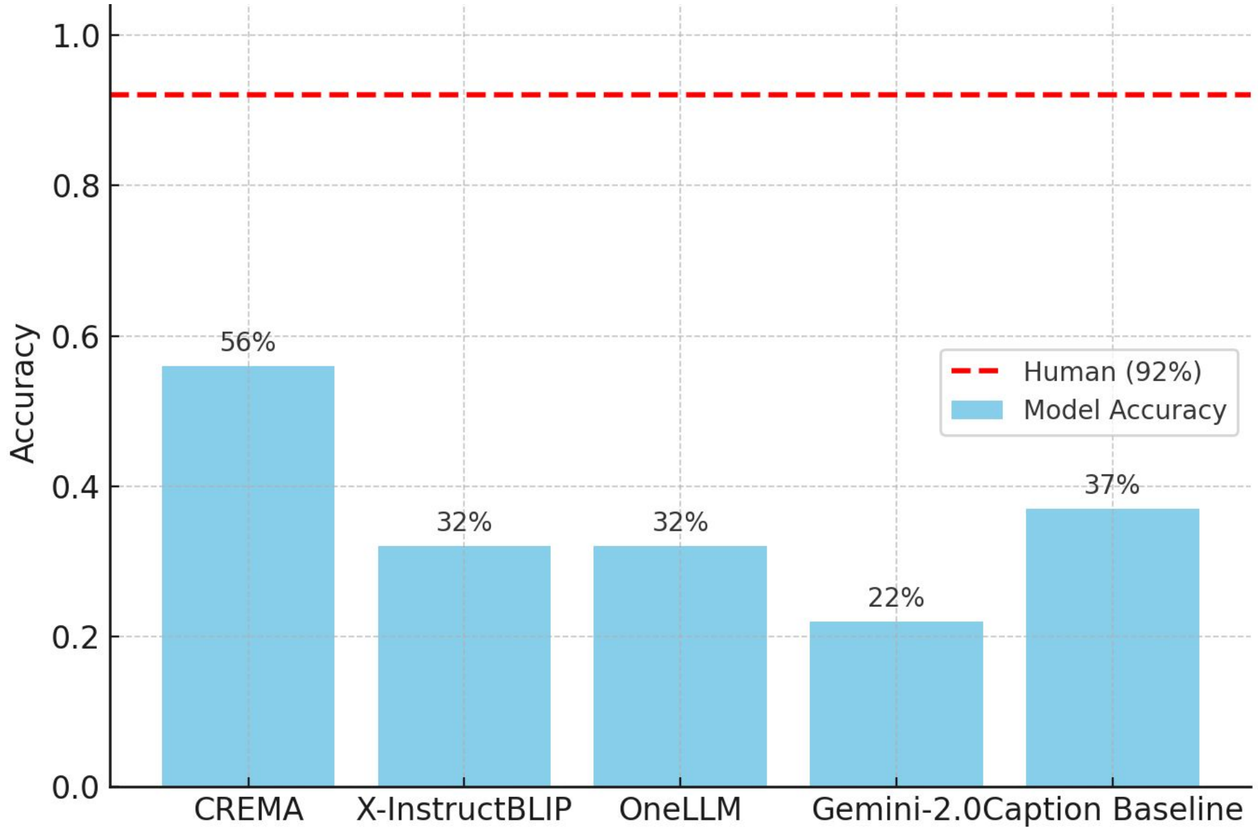
- Stricter filtering methods with choice permutations and a mixture of model agreement improves data quality.



[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Model Performance

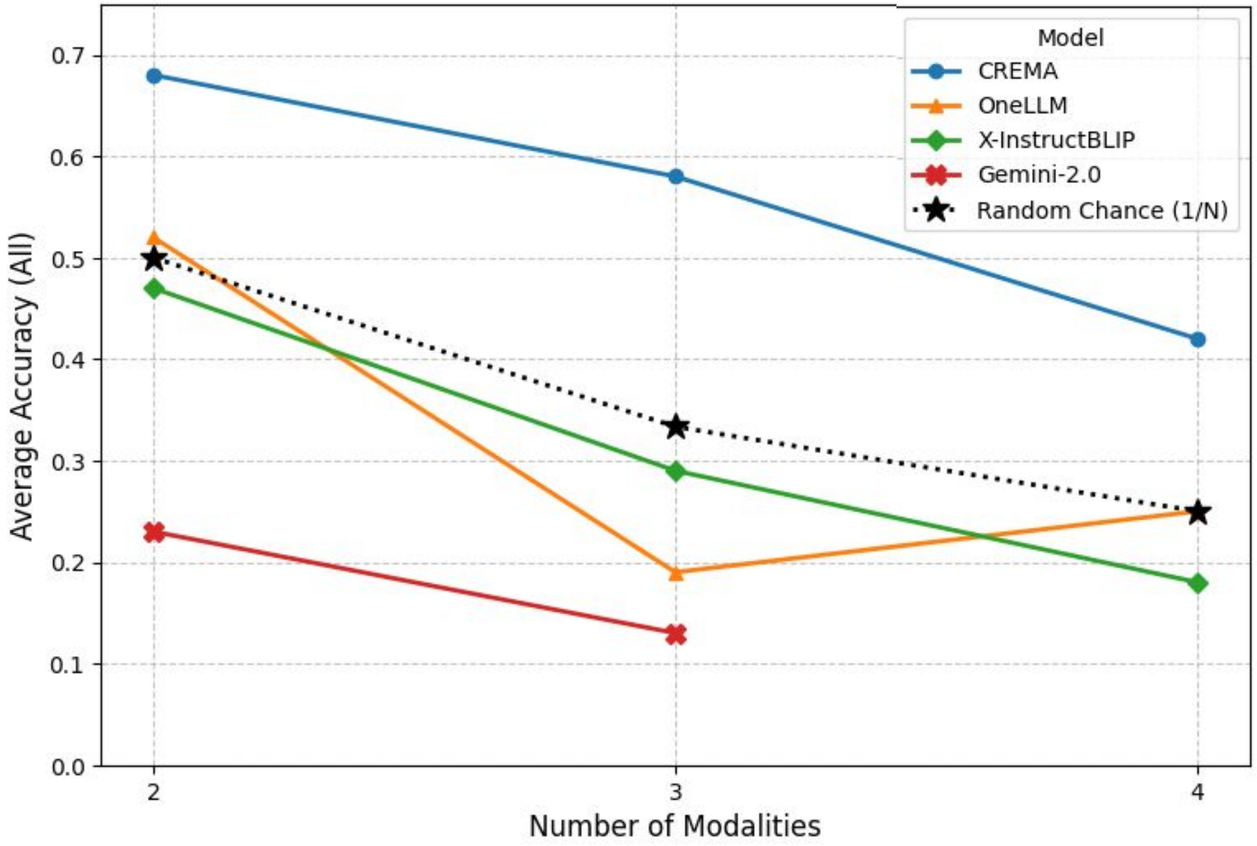
Models struggle significantly with cross-modal reasoning.



[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Model Performance

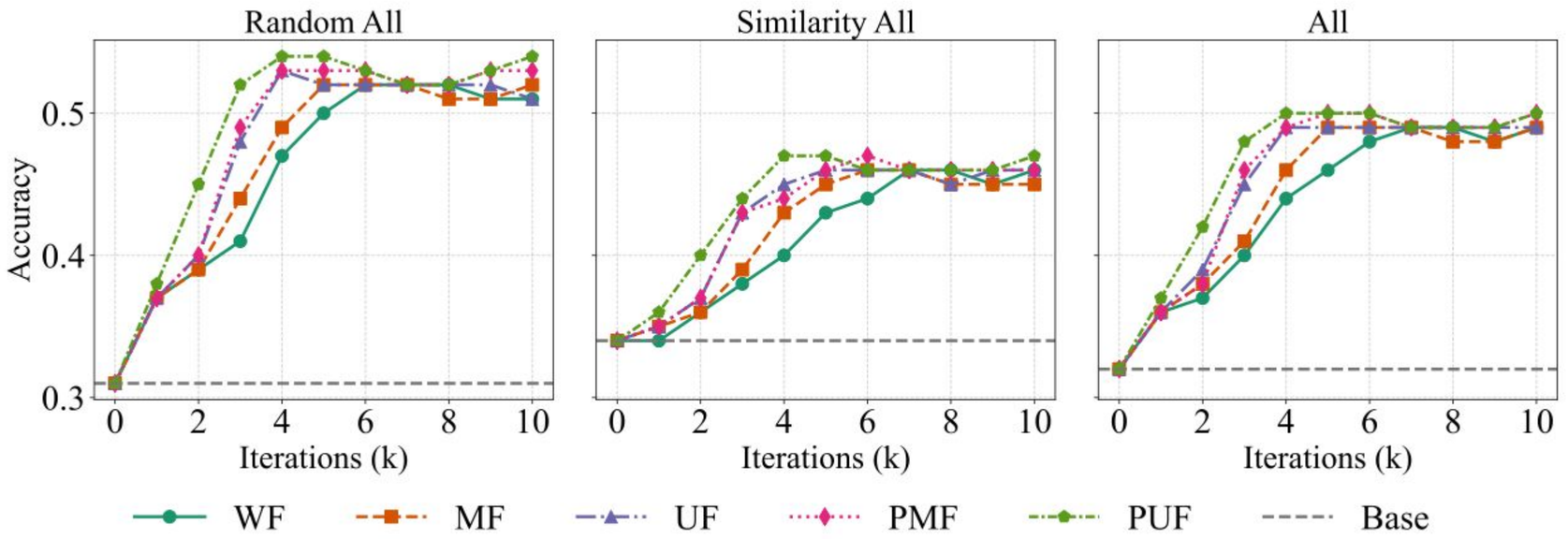
The number of modalities in the question is inversely correlated to performance.



[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Model Finetuning

Finetuning shows moderate improvements.



[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

Contra4: Summary

New Task:

- Contrastive Cross-Modal Reasoning.

Dataset:

- Construct large-scale synthetic training data and manually annotated test sets.

Evaluation:

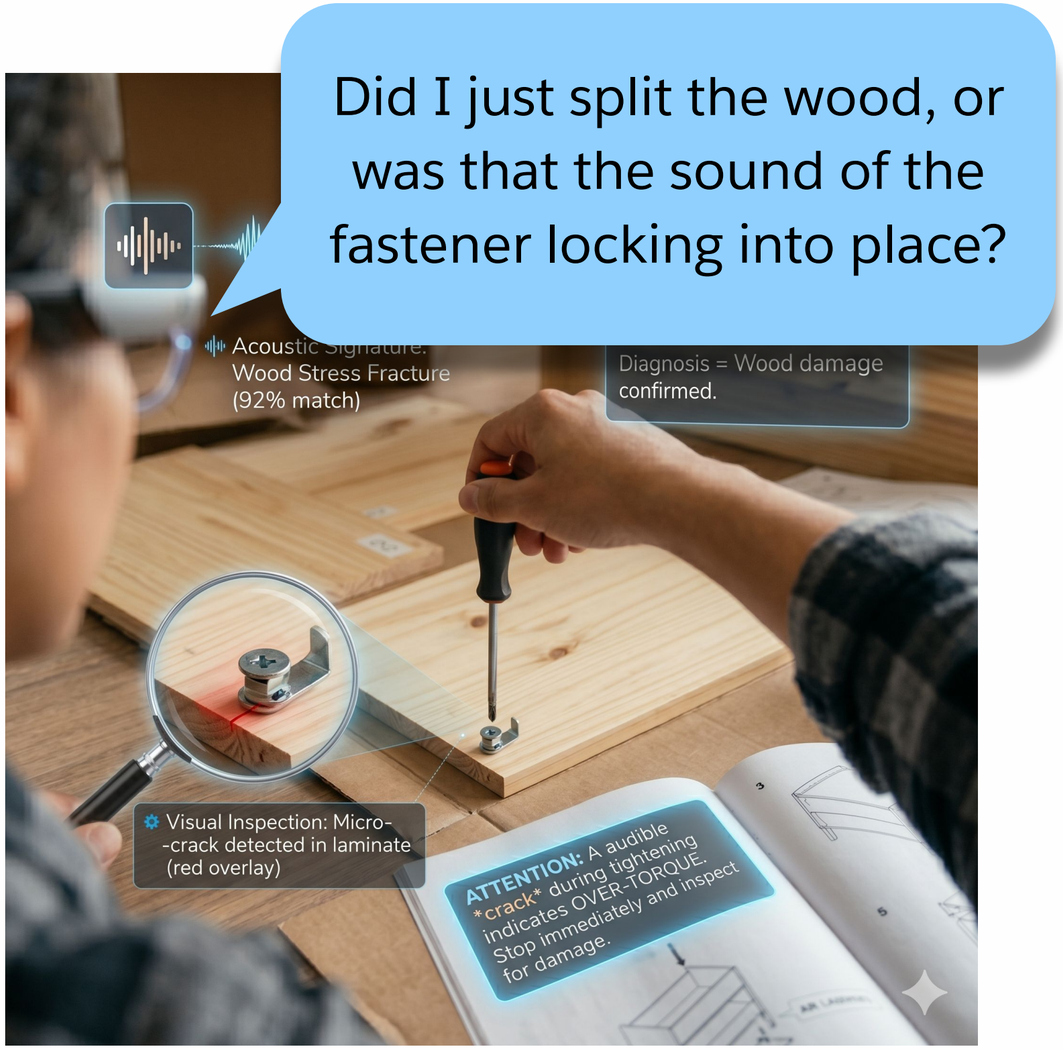
- Show that state-of-the-art models perform poorly on this task.
- Fine-tuning on Synthetic Data Provides moderate gains, but significant improvements are still needed.

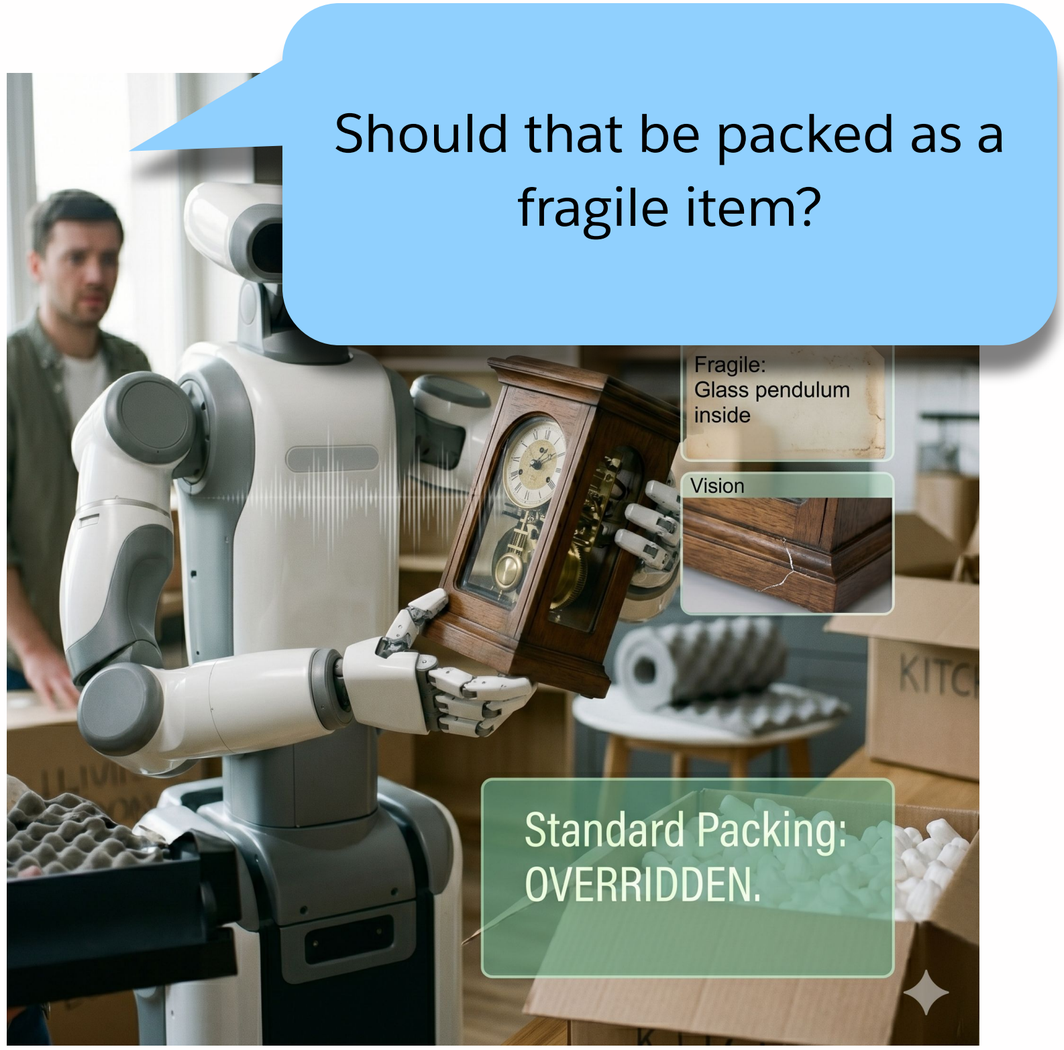
Paper, Data:

https://artemisp.github.io/contra4-web/

[A. Panagopoulou et. al. Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D. EMNLP 2025.]

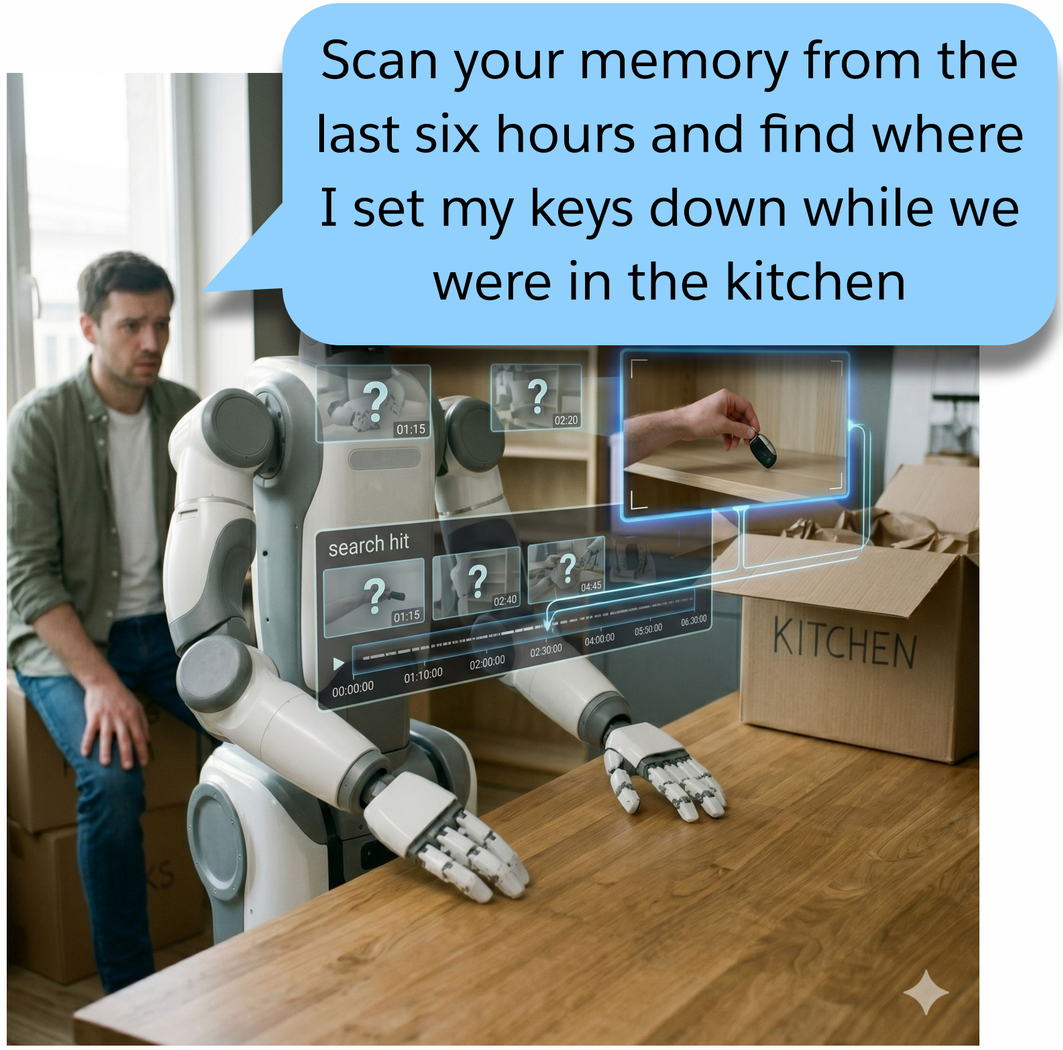
Part II: Cross-modal reasoning





Part III


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]

Cross-modal Reasoning

Contra4
[EMNLP'25]

Reasoning over long videos

AVP
[CVPR Findings '26]

Existing Caption-based Agentic Frameworks

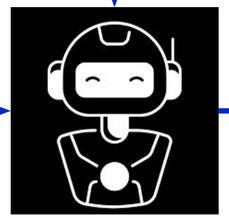
Step 1 : Passive perception via query-agnostic captioner




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

Low efficiency, massive captioner computation overhead

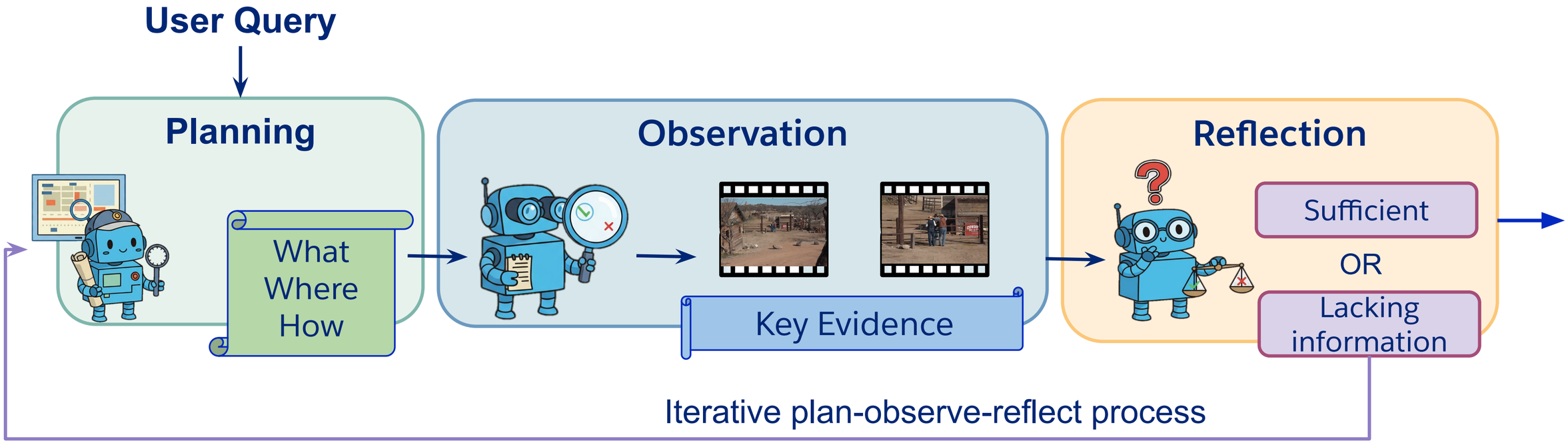
Step 2 : Evidence searching via caption database

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)

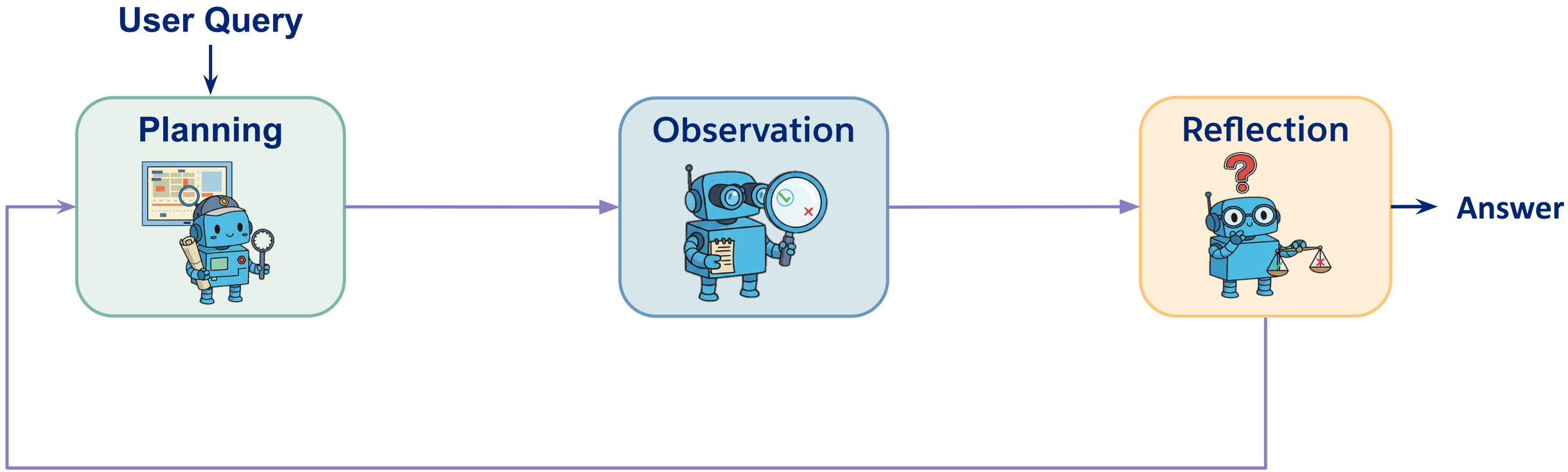


***Active* Evidence Seeking**

Iterative Perception via Reflection

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

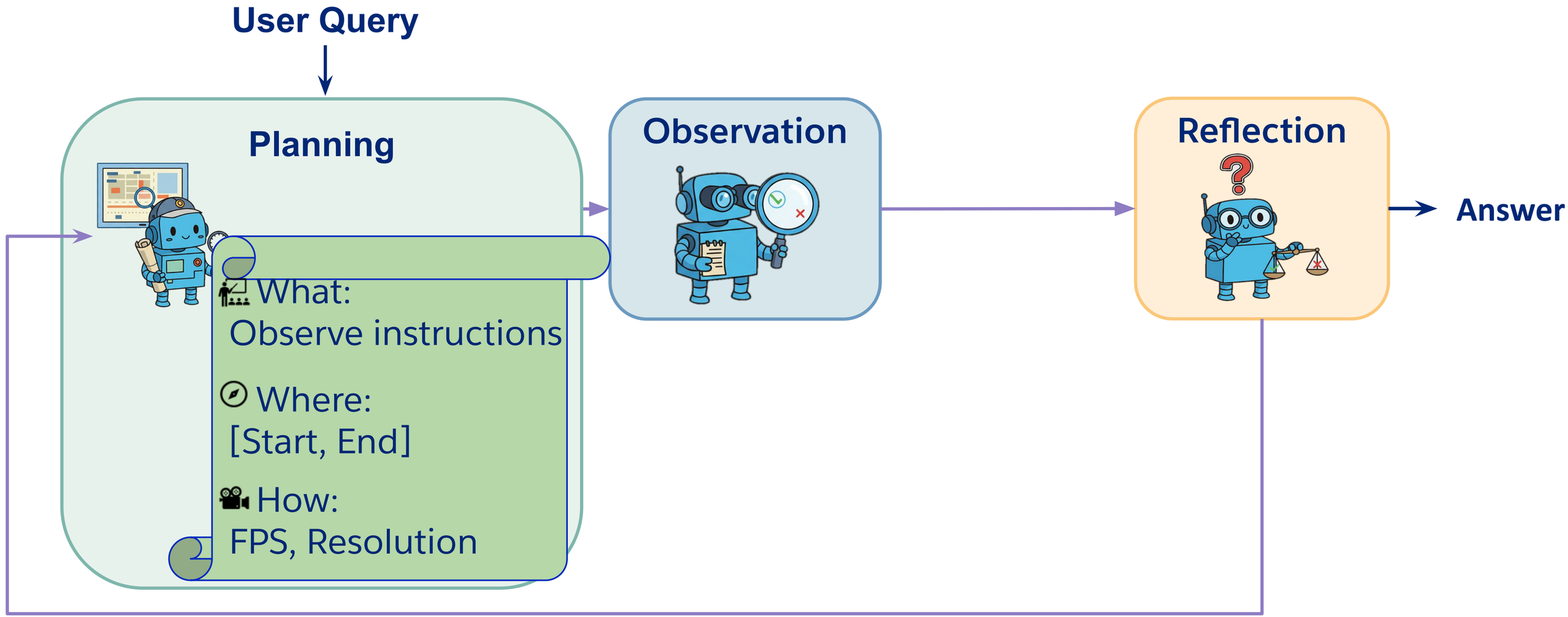
Our Work: Active Video Perception (AVP)



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

Our Work: Active Video Perception (AVP)

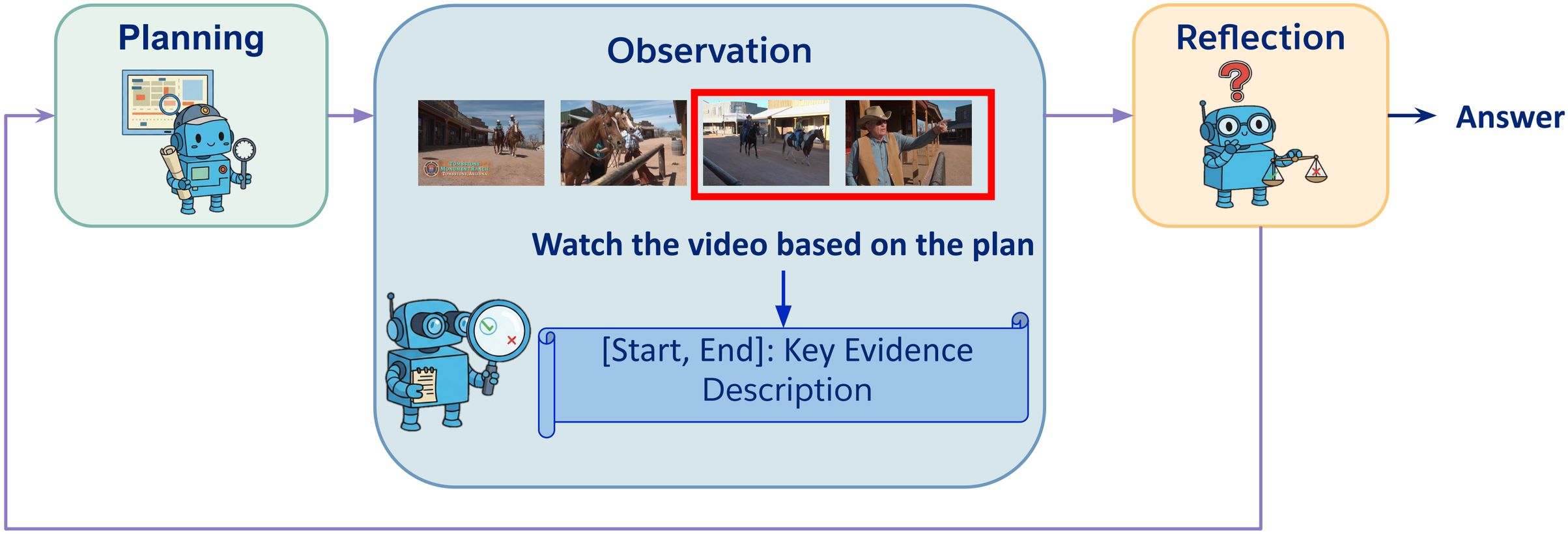
Step 1 : Generate Plan



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

Our Work: Active Video Perception (AVP)

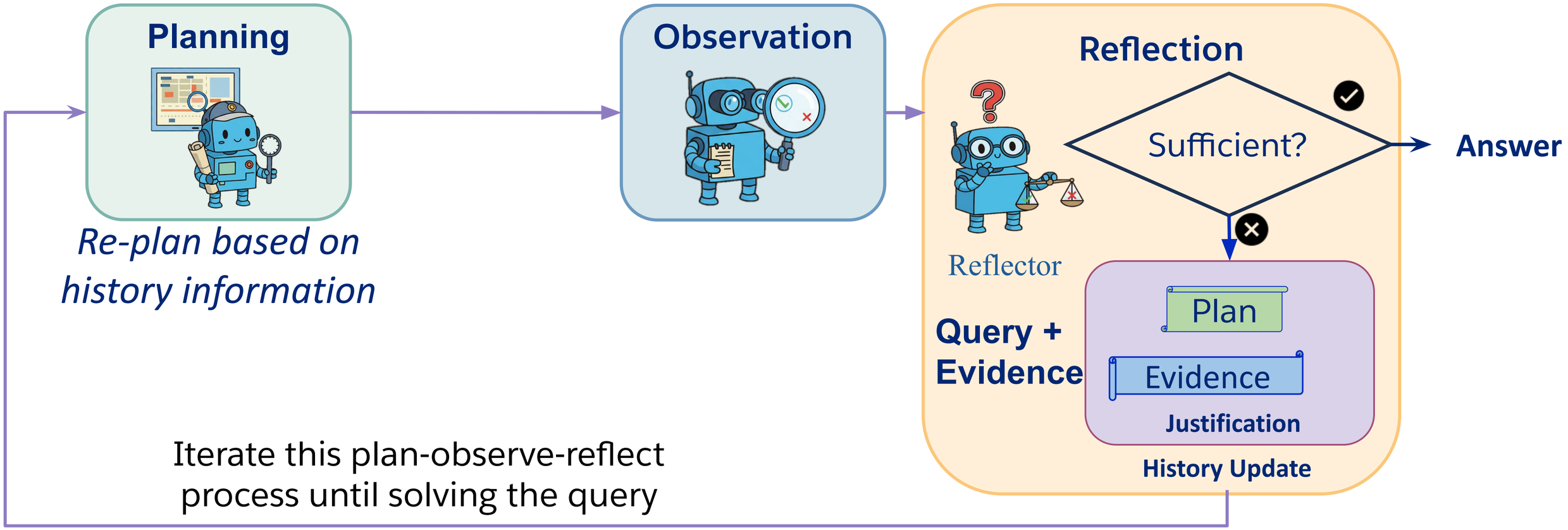
Step 2 : Targeted Observation



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

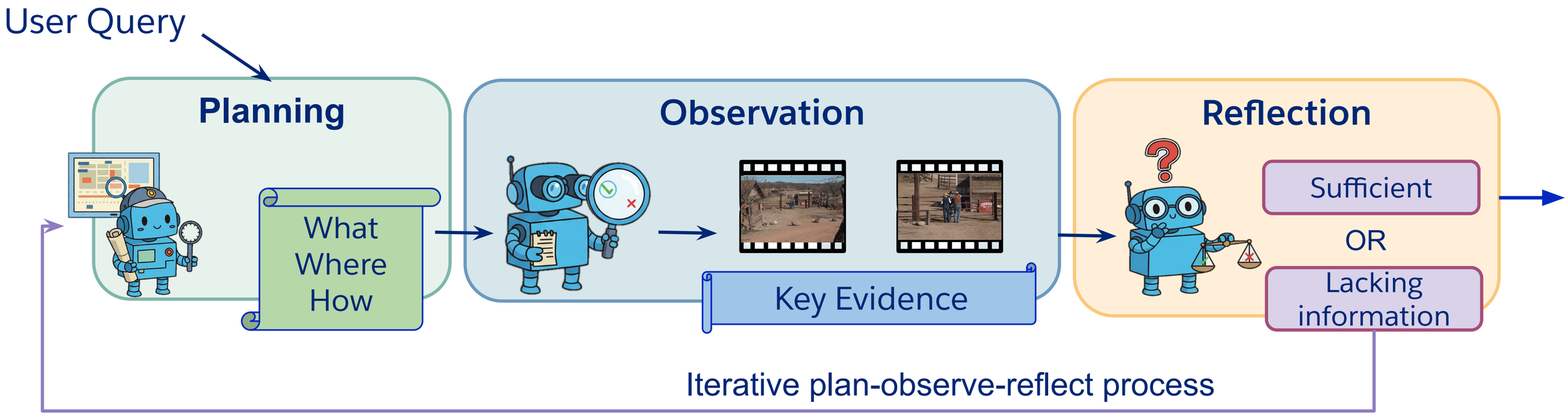
Our Work: Active Video Perception (AVP)

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)



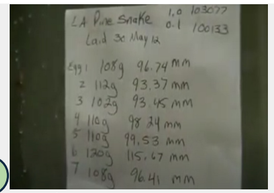
***Active* Evidence Seeking**

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

Evidence
[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.

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

Evidence
[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**.

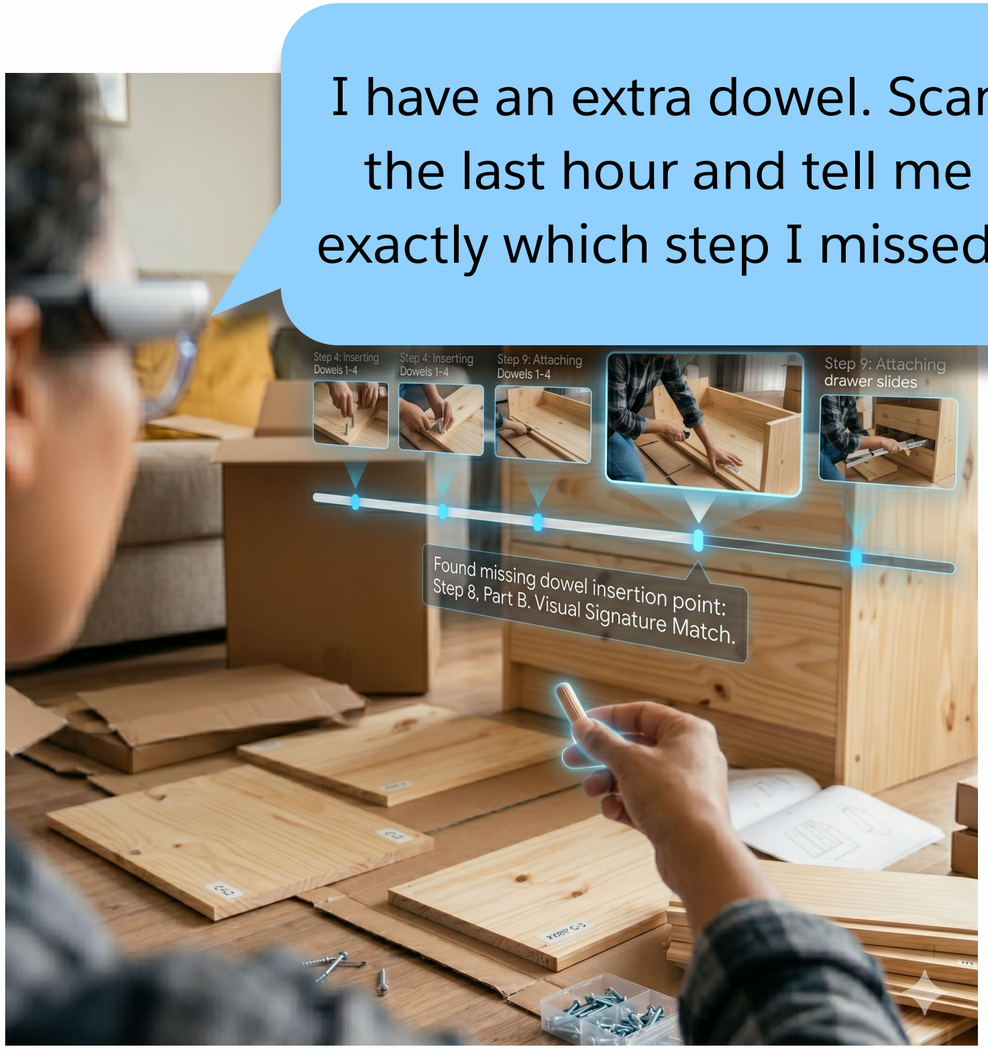
Wang et al. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. CVPR Findings 2026.

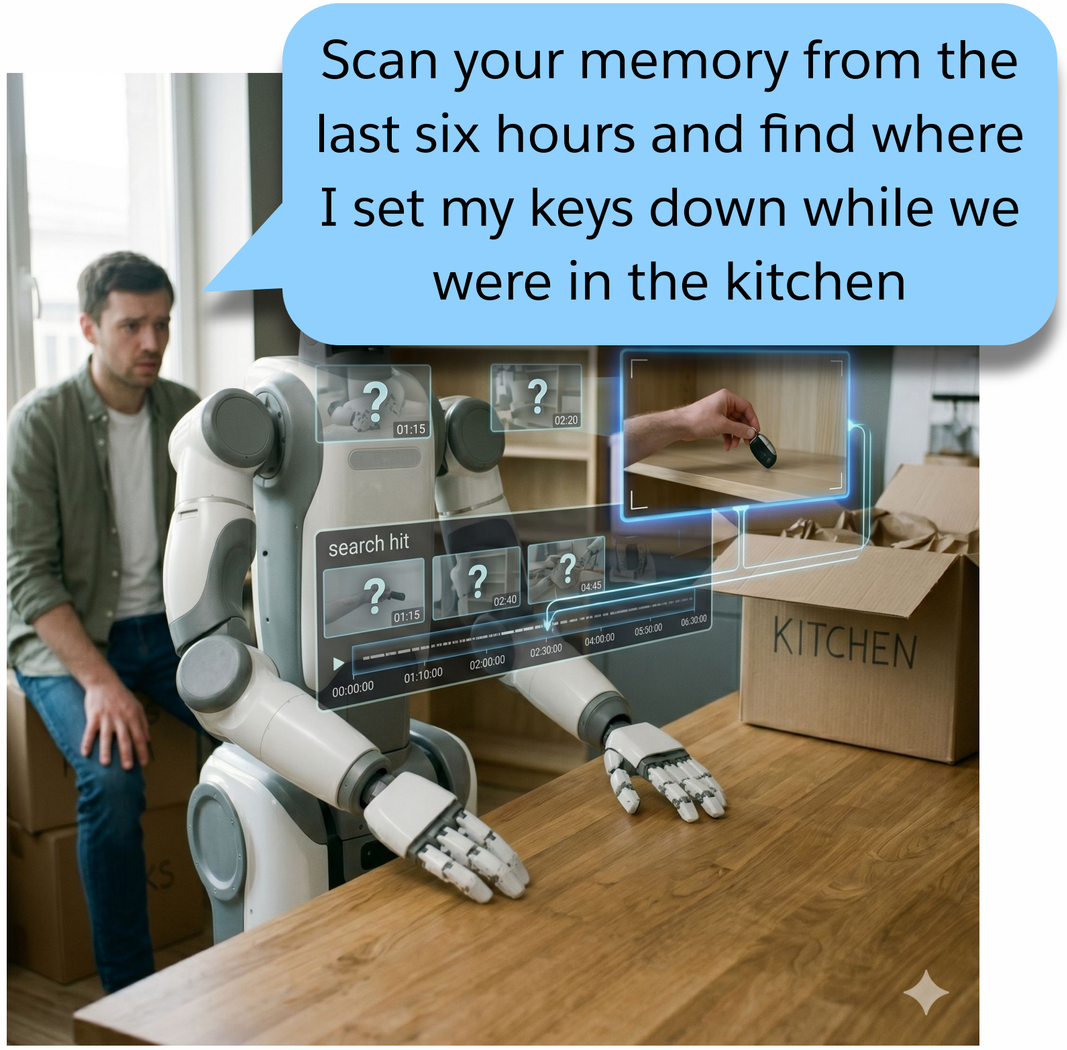
AVP: Quantitative Results

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	61.8	67.4	79.6	82.4	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	77.7	-	-
VideoLucy	-	58.8	76.1	72.5	66.8	-	-
DeepVideoDiscovery	-	74.2	-	-	67.3	71.6	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.

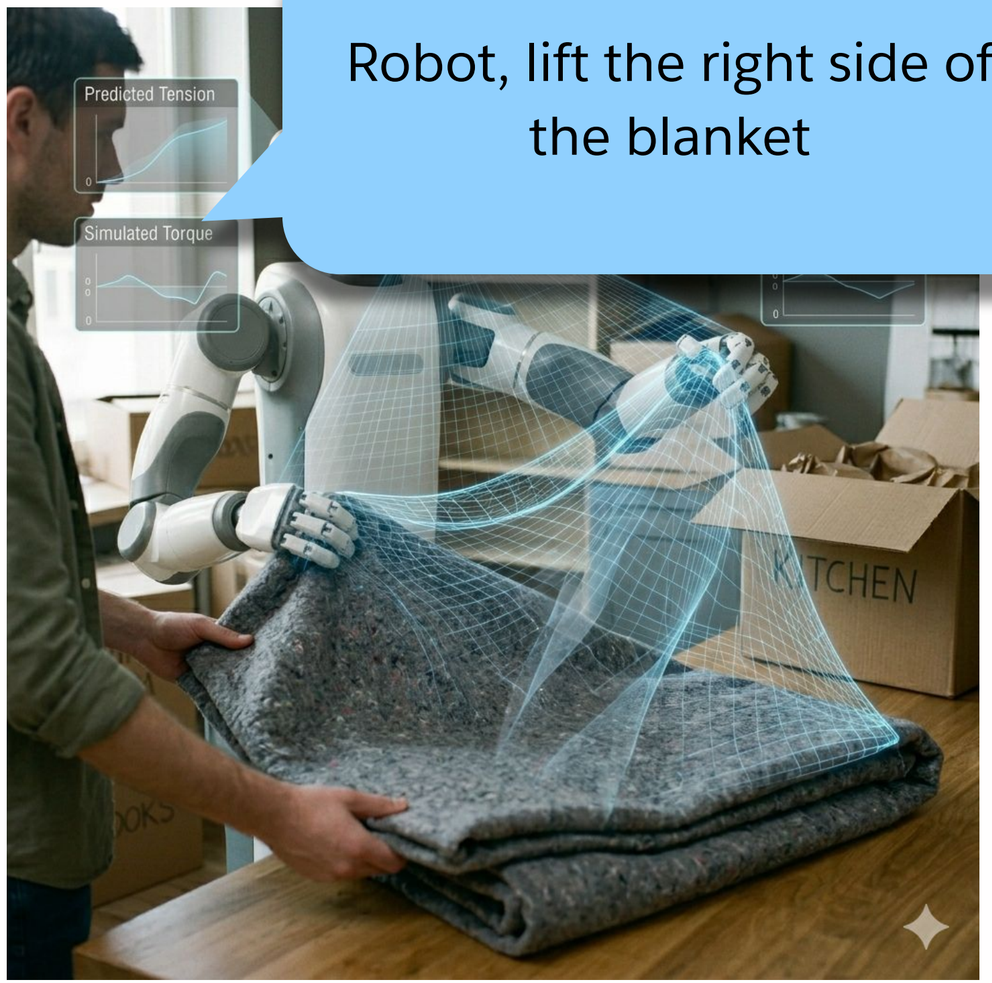
Part III: Active Perception





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]

Cross-modal Reasoning

Contra4
[EMNLP'25]

Reasoning over long videos

AVP
[CVPR Findings '26]

Motion Guidance Generation

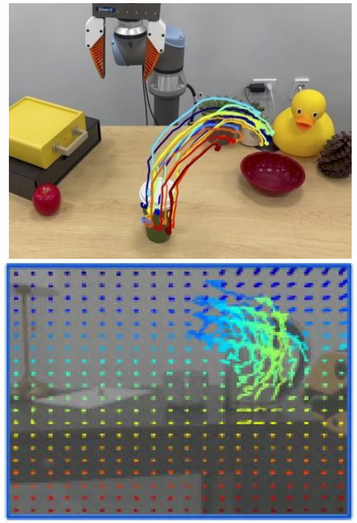
FOFPred
[CVPR Findings '26]

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”

+

Generate motion for “assembling drawer into empty space”

→ **FOFPred** →

Future motion prediction

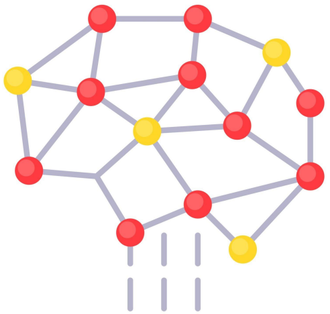
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

FOFPred: Motion Guidance Generation

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

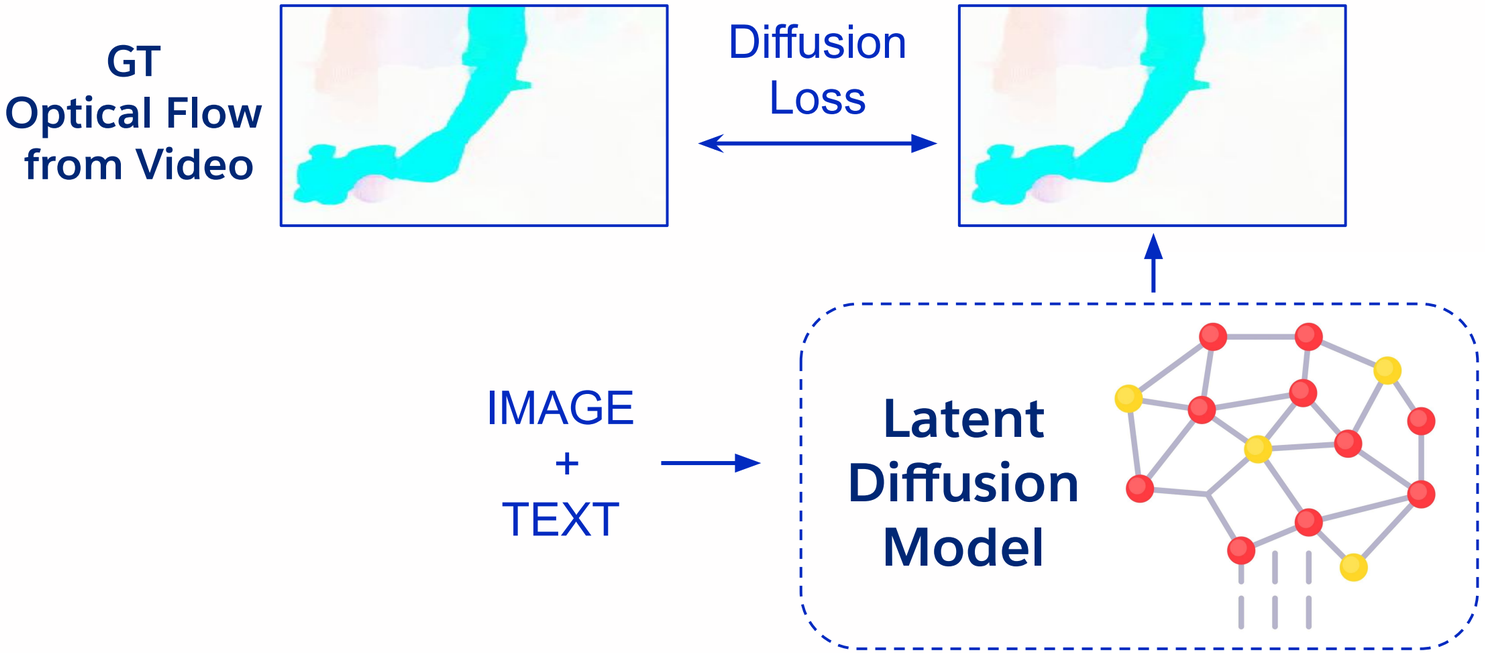
IMAGE
+
TEXT

Latent Diffusion Model



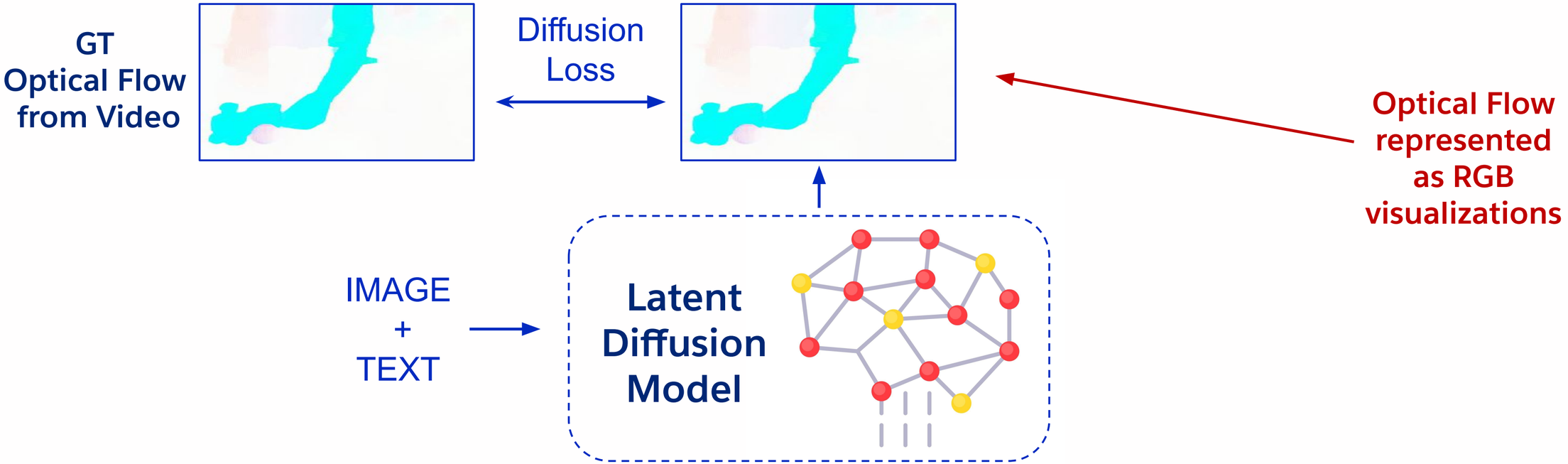
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

FOFPred: Motion Guidance Generation



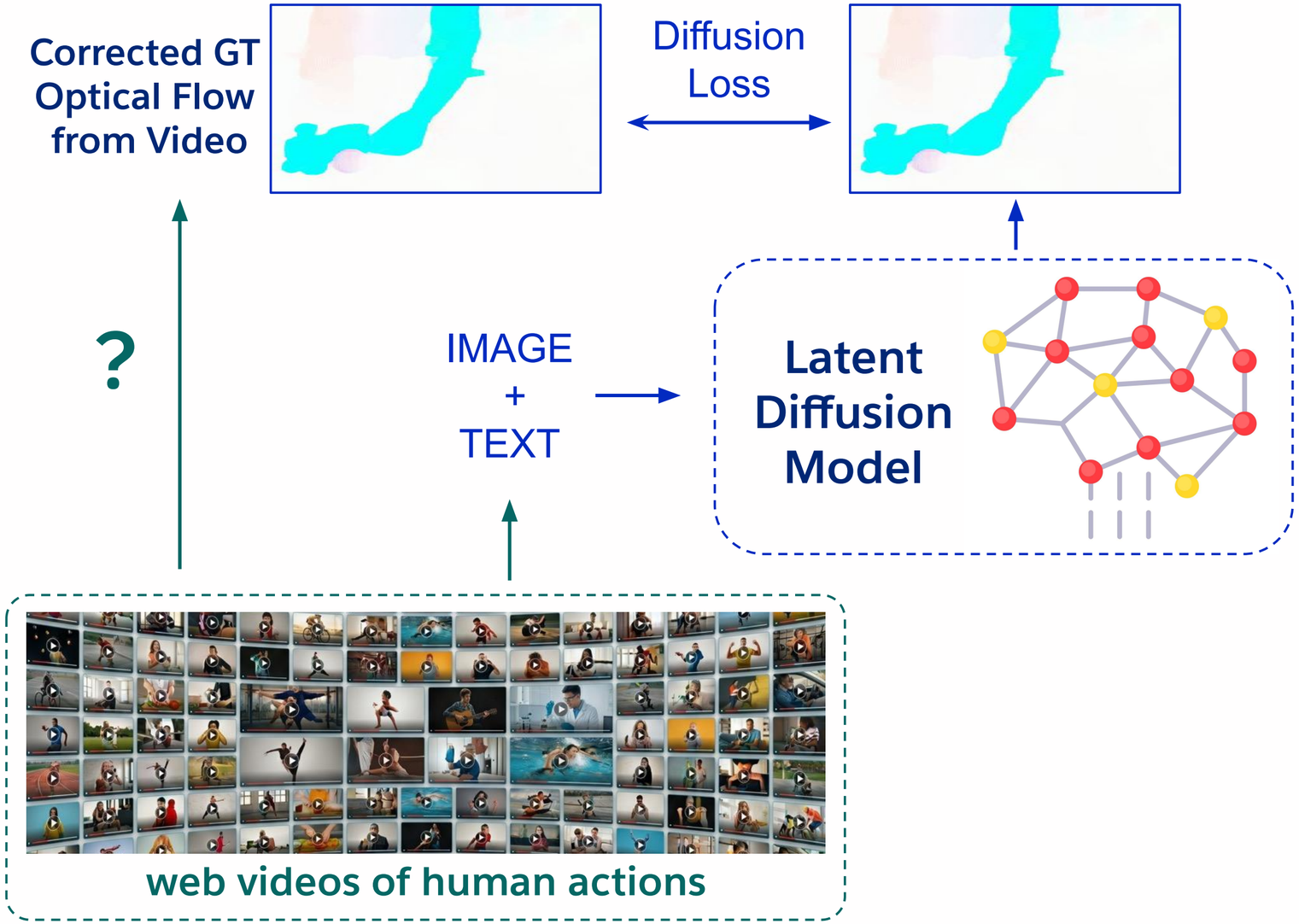
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

FOFPred: Motion Guidance Generation



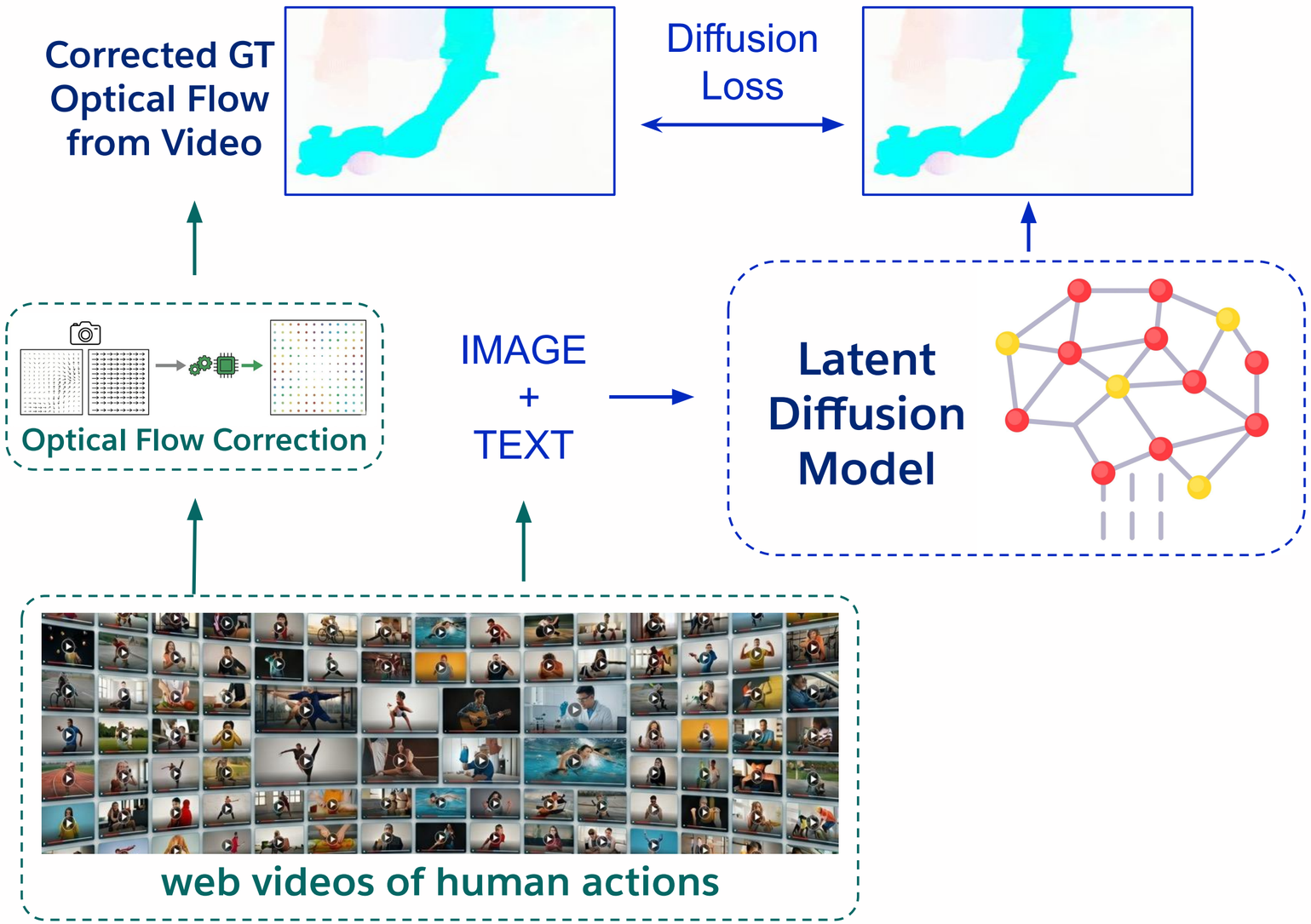
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

Scalable Learning



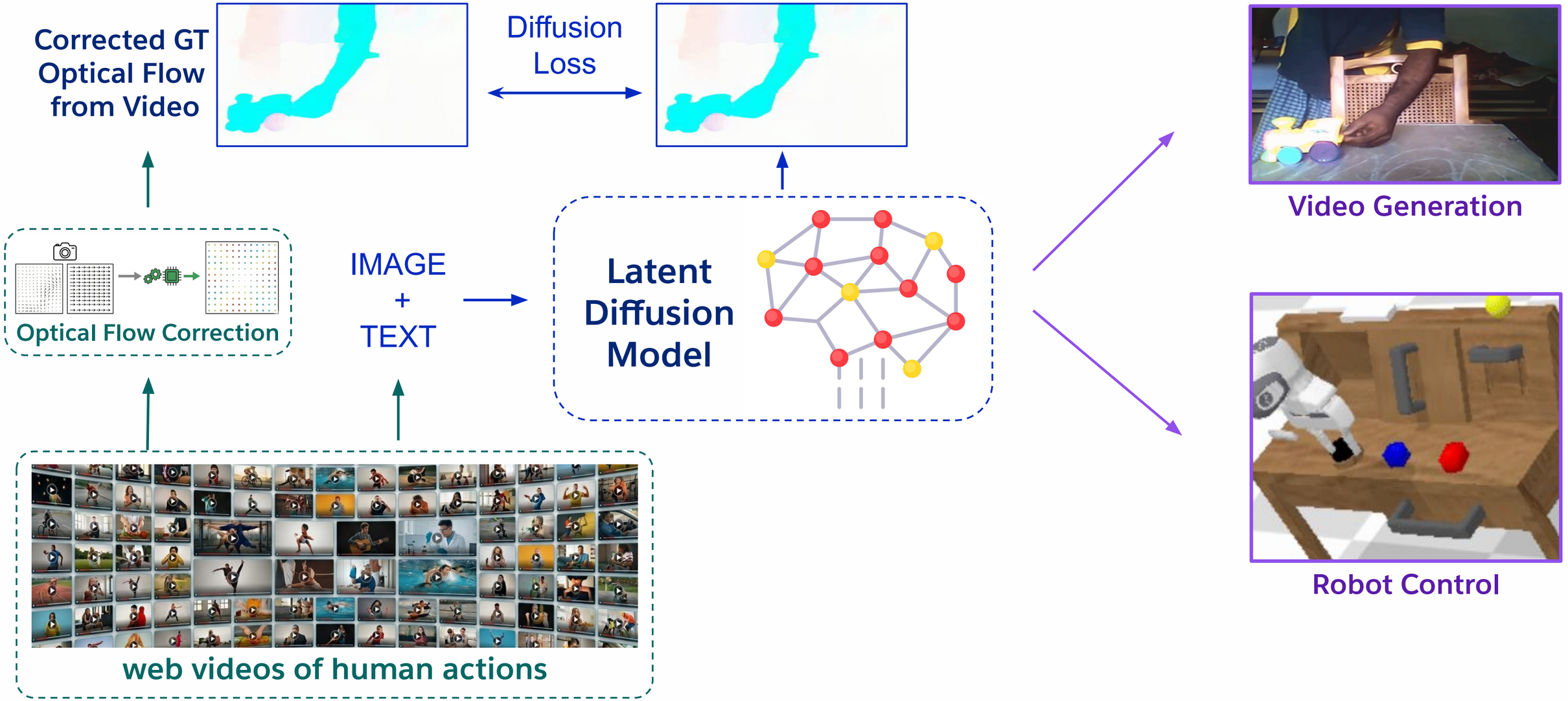
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

Scalable Learning



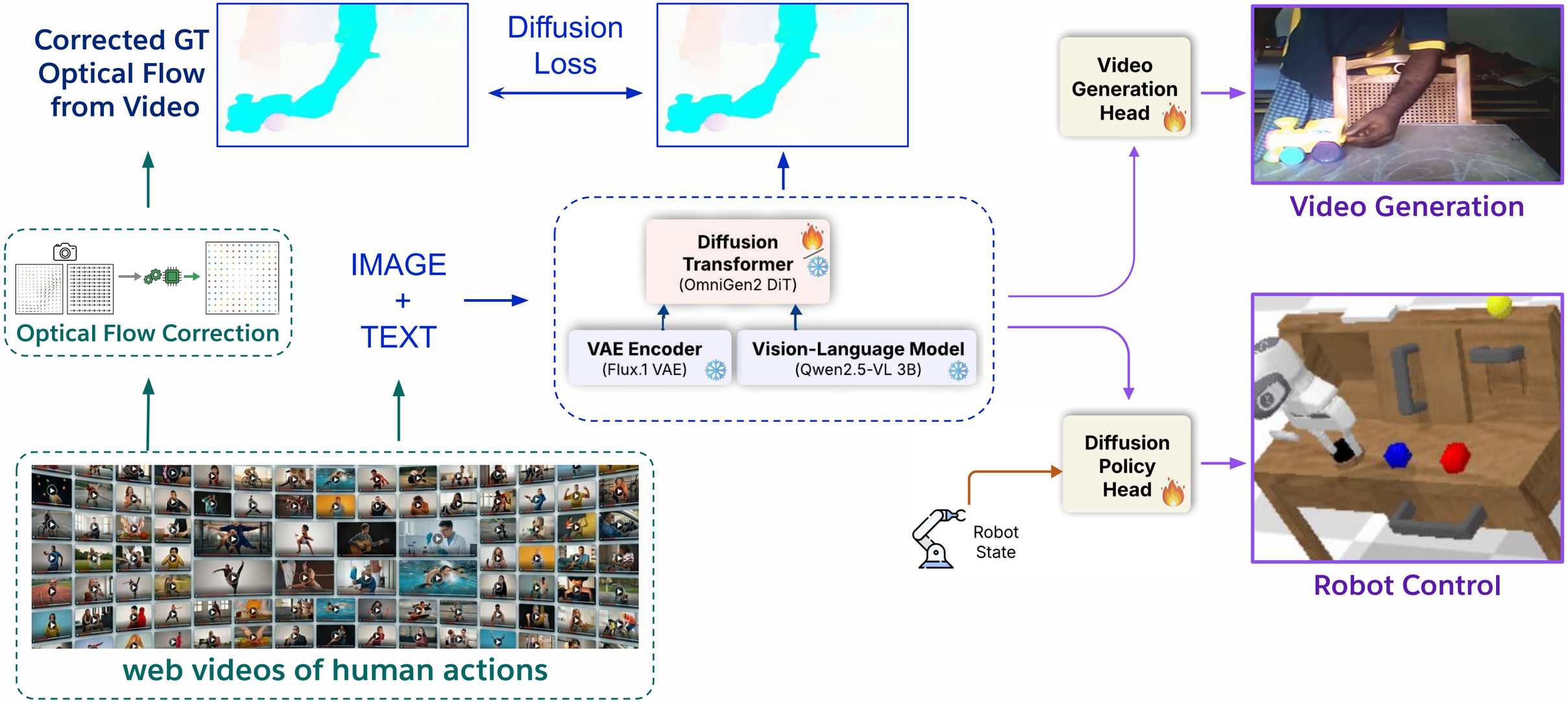
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

Downstream Tasks



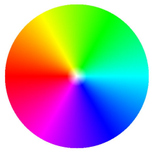
[K. Ranasinghe et al. Future Optical Flow Prediction Improves Robot Control and Video Generation. CVPR Findings 2026]

FOFPred: Detailed Architecture



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

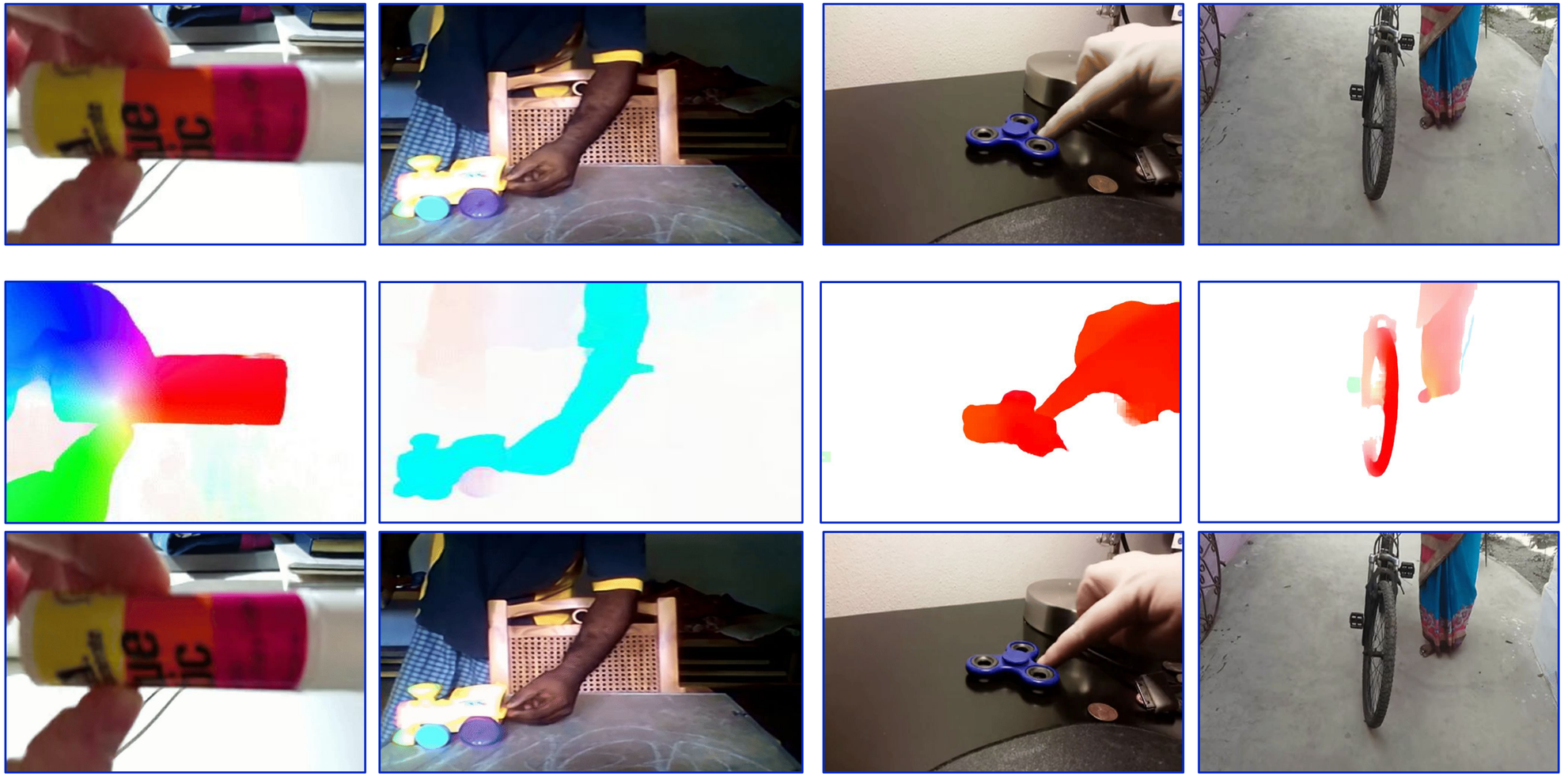
Video Generation



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)

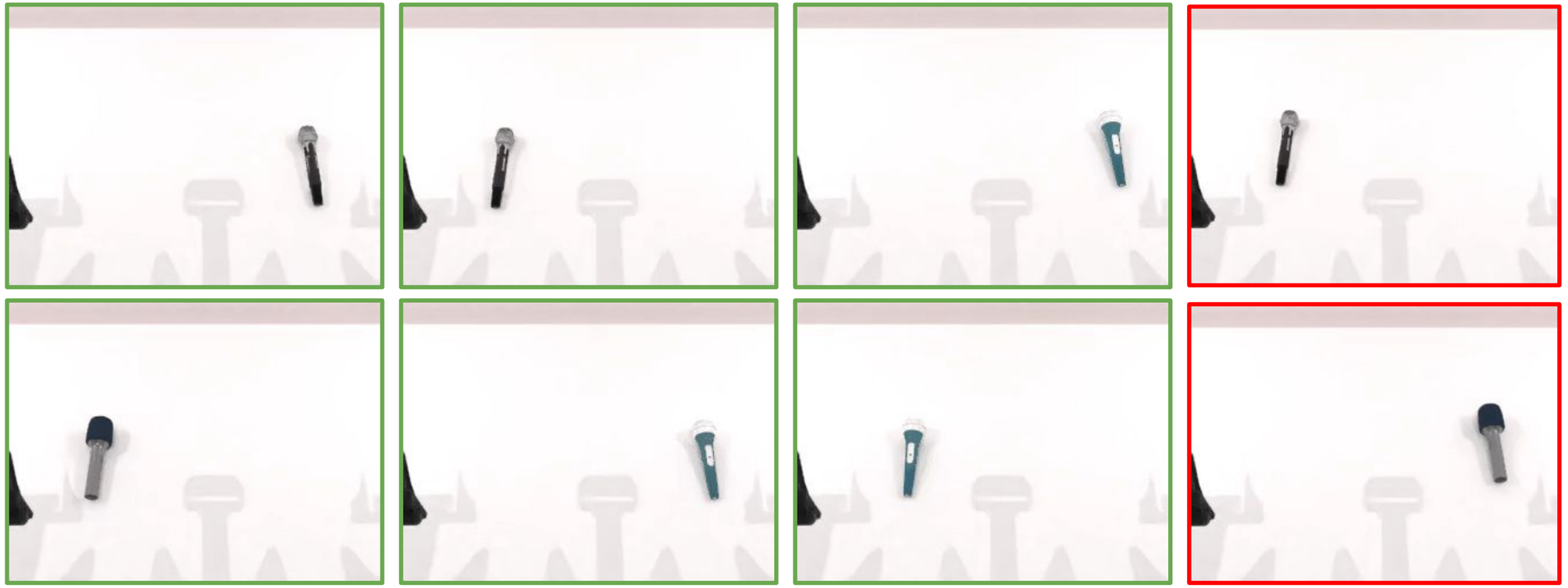
SSv2 Video Generation Evaluation: Over the CogVideoX baseline, our FOFPred framework shows consistent improvements in generation quality.

[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
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)

[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

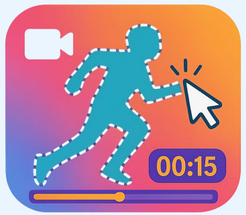
Code, Ckpt, & Demo Released
https://fofpred.github.io/

[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]



Cross-modal Reasoning

Contra4
[EMNLP'25]

Reasoning over long videos

AVP
[CVPR Findings '26]

Motion Guidance Generation

FOFPred
[CVPR Findings '26]

Agentic Ambient Intelligence

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.

slid bers


salesforce
Thank
you

Motion Guidance Generation