

# Toward Robust and Scalable Temporal Understanding in Video Large Language Models

Seoul National University  
**Presenter:** Minjoon Jung

*2026.6.11*



# Outline of the Presentation

- 1. Introduction**
- 2. On the Consistency of Temporal Comprehension in Video Large Language Models [1]**
- 3. EgoExo-Con: Exploring View-Invariant Video Temporal Understanding [2]**
- 4. EvoGround: Self-Evolving Video Agents for Video Temporal Grounding [3]**
- 5. Conclusion**

[1] Jung et al., On the Consistency of Video Large Language Models in Temporal Comprehension. CVPR 2025

[2] Jung et al., EgoExo-Con: Exploring View-Invariant Video Temporal Understanding. arXiv 2025

[3] Jung et al., EvoGround: Self-Evolving Video Agents for Video Temporal Grounding. arXiv 2025

# 1 Introduction

# Introduction

- Video uniquely introduces *continuous temporal dynamics* beyond spatial structure.

## 1. Image (Vision)

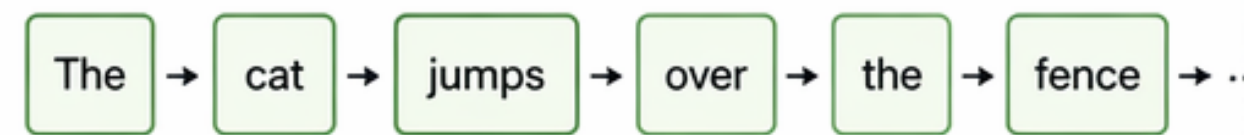
Spatial only



A single snapshot.  
No notion of time.

## 2. Text

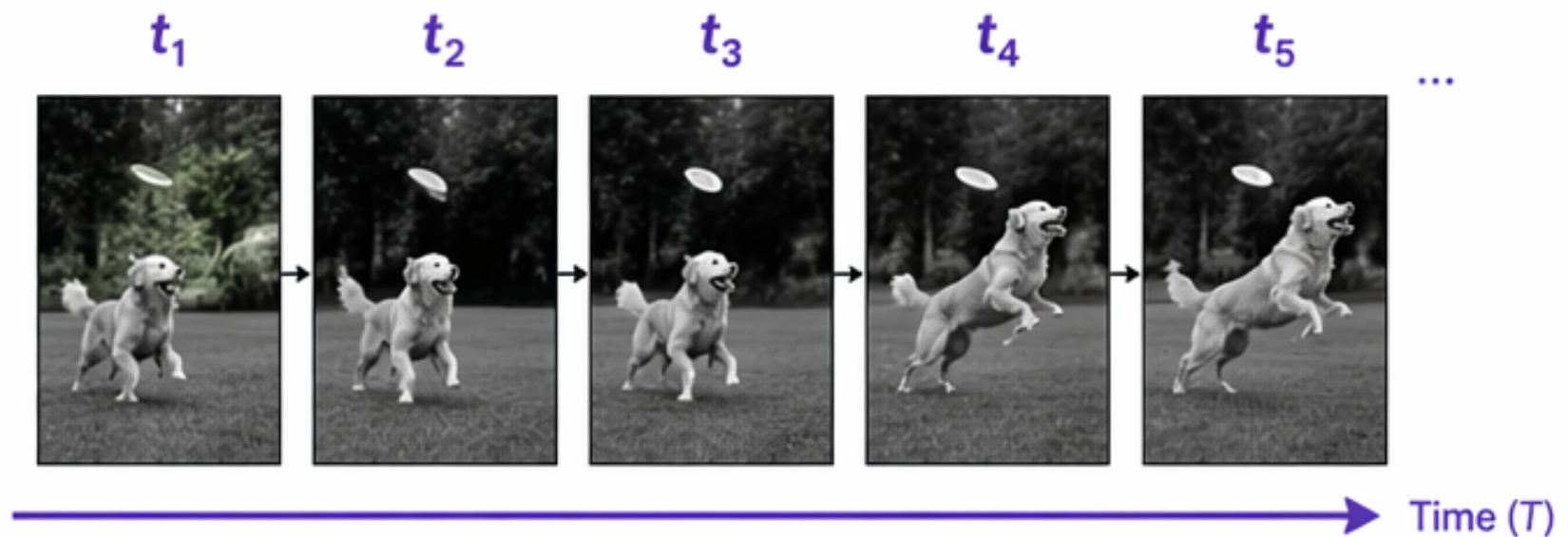
Discrete symbolic sequence



A sequence of tokens.  
Order exists, but no inherent visual time.

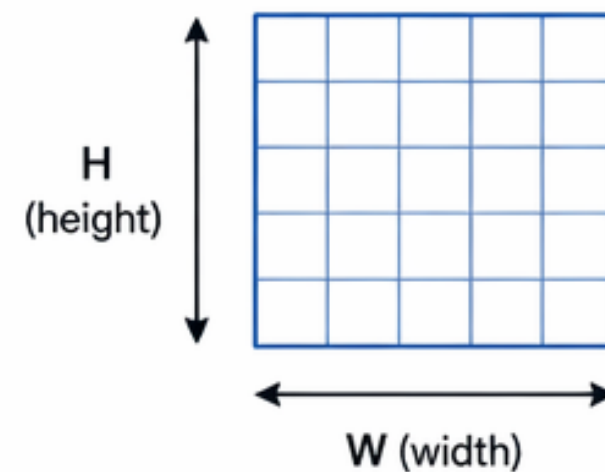
## 3. Video

Spatiotemporal signal ( $H \times W \times T$ )

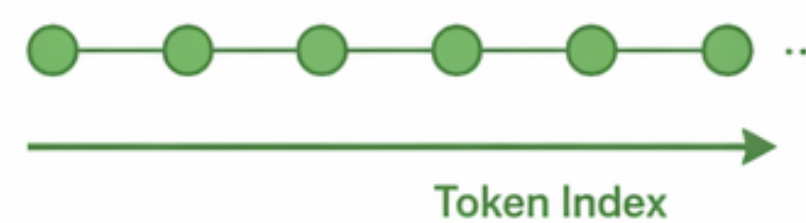


Continuous evolution of visual content over time.  
Captures motion, causality, and persistence.

## Image: 2D

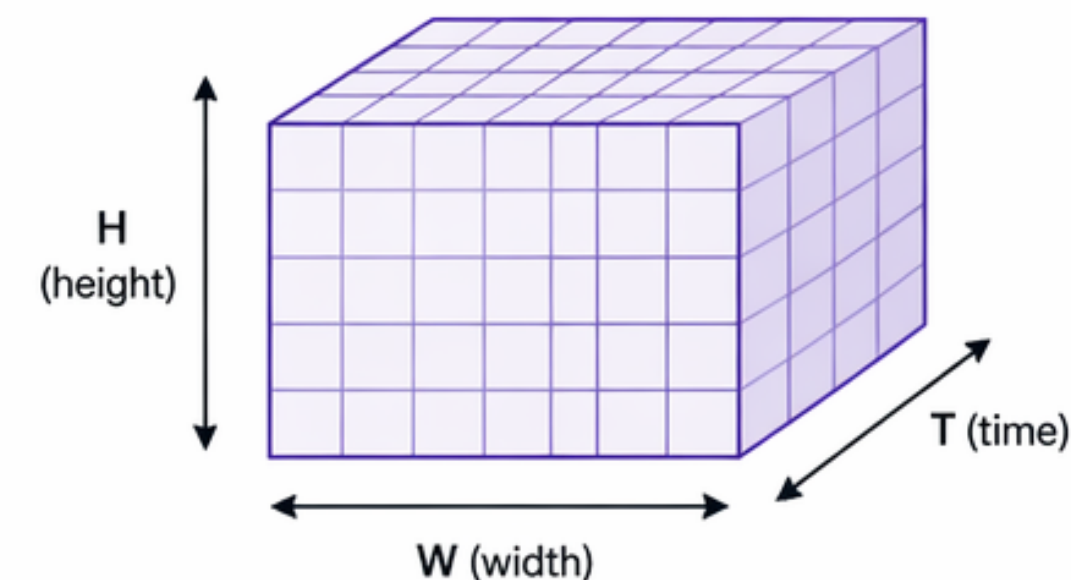


## Text: 1D (Discrete)



Discrete steps in an abstract  
symbolic space.

## Video: 3D (Spatiotemporal)



Adds a temporal axis  
to spatial structure,  
forming a 3D  
spatiotemporal volume.

# Introduction

- **Current Paradigm: Temporal Grounding (What is optimized?)**

**Query**

A man with a long beard and a hat is seen holding a compass.



**Target Moment: 10s - 25s**



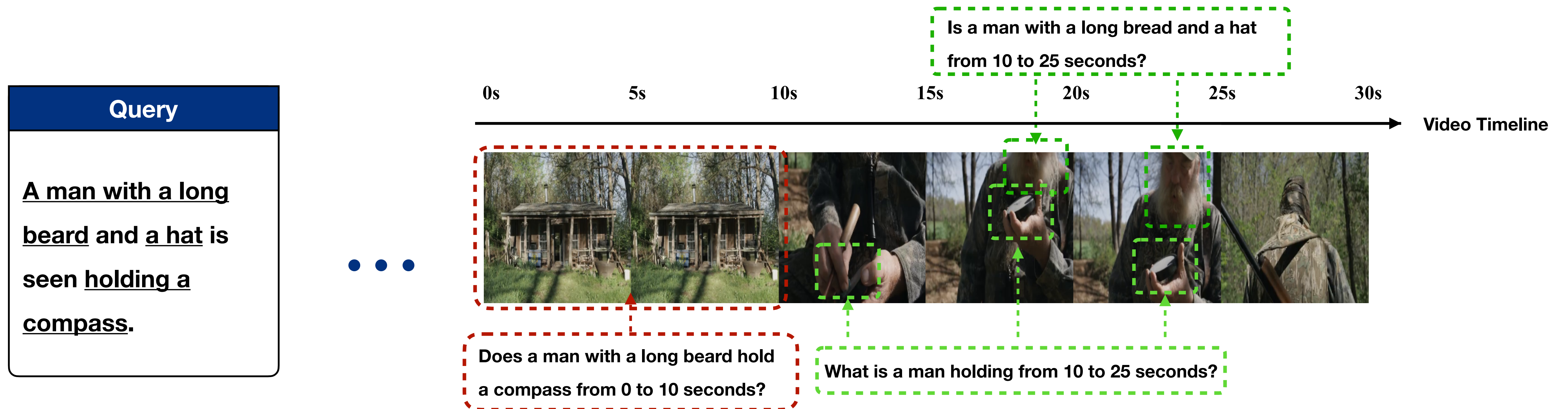
**Model Output (Grounding):**

**The query happens from 10 to 25 seconds.**

**Models achieve correct temporal grounding!** 🙌

# Introduction

## ● Missing Paradigm: From Grounding to Verification (What is not verified?)



**Model Output (Grounding):**  
The query happens from 10 to 25 seconds.

Does a model reliably capture all semantic details over time?



# Introduction

- **In this talk, we aim to answer the following questions:**
  - **How can we ensure that a model's reasoning is grounded in true visual evidence?**
  - **How can we reliably measure and improve such capabilities?**
  - **How can we develop video large language models in a robust and scalable manner?**

# On the Consistency of Video Large Language Models in Temporal Comprehension



Minjoon Jung  
Seoul National University



Junbin Xiao  
National University of Singapore



Byoung Tak Zhang  
Seoul National University



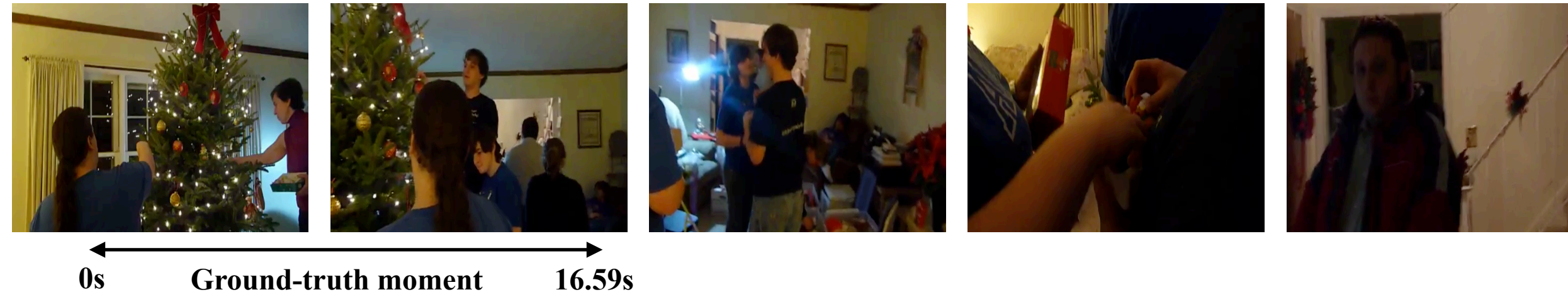
Angela Yao  
National University of Singapore



# Introduction



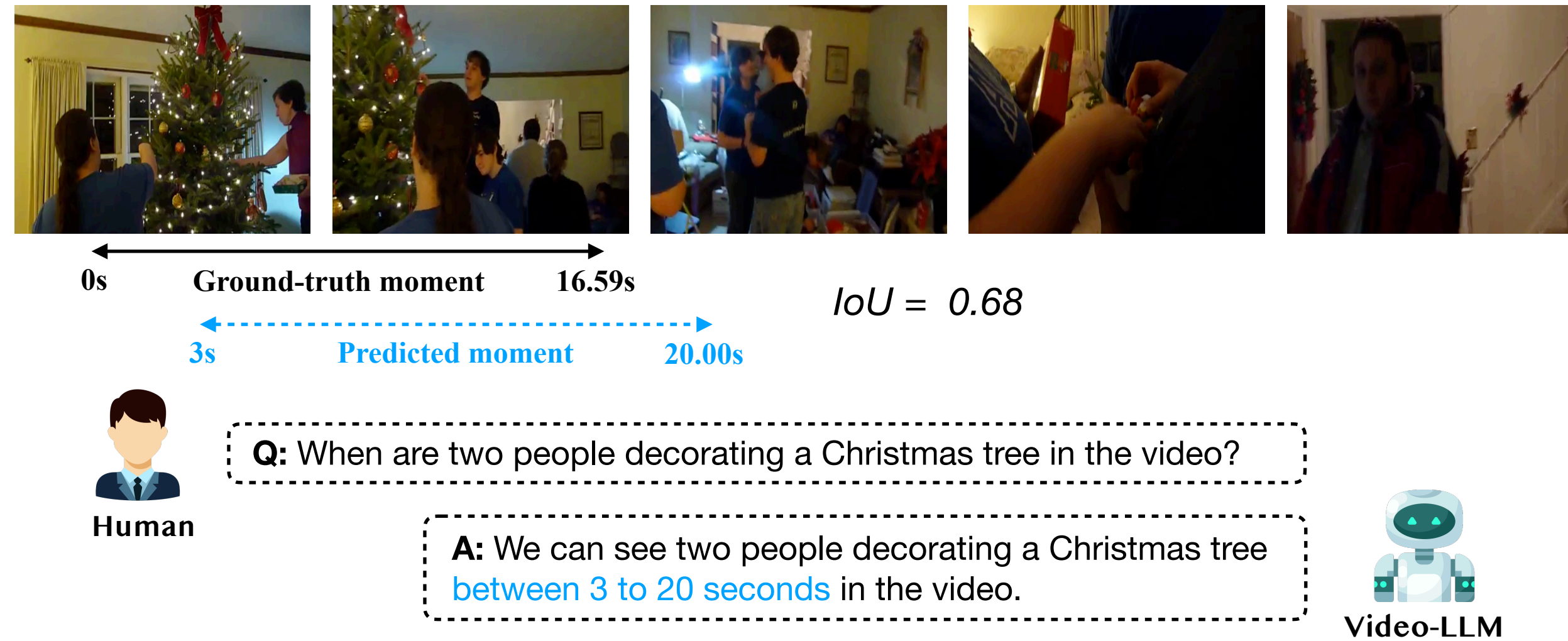
# Introduction



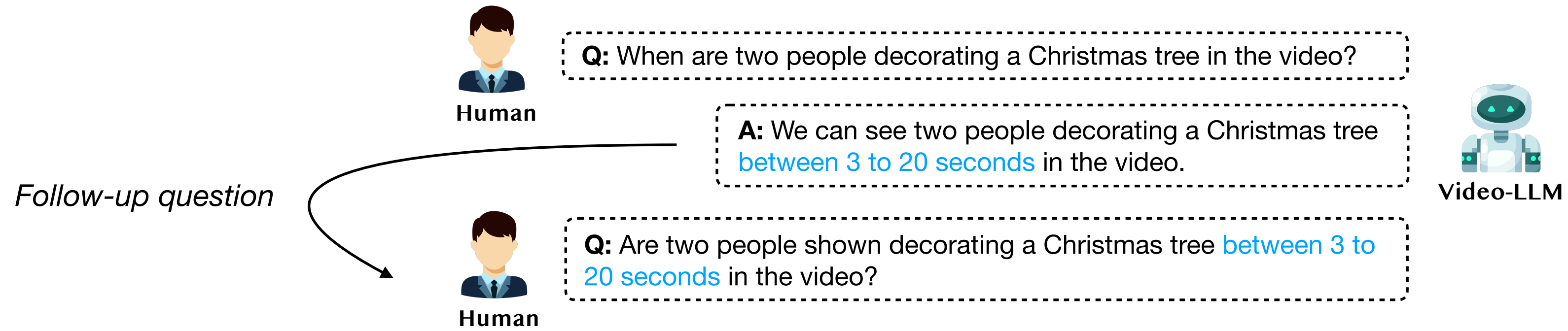
Human

Q: When are two people decorating a Christmas tree in the video?

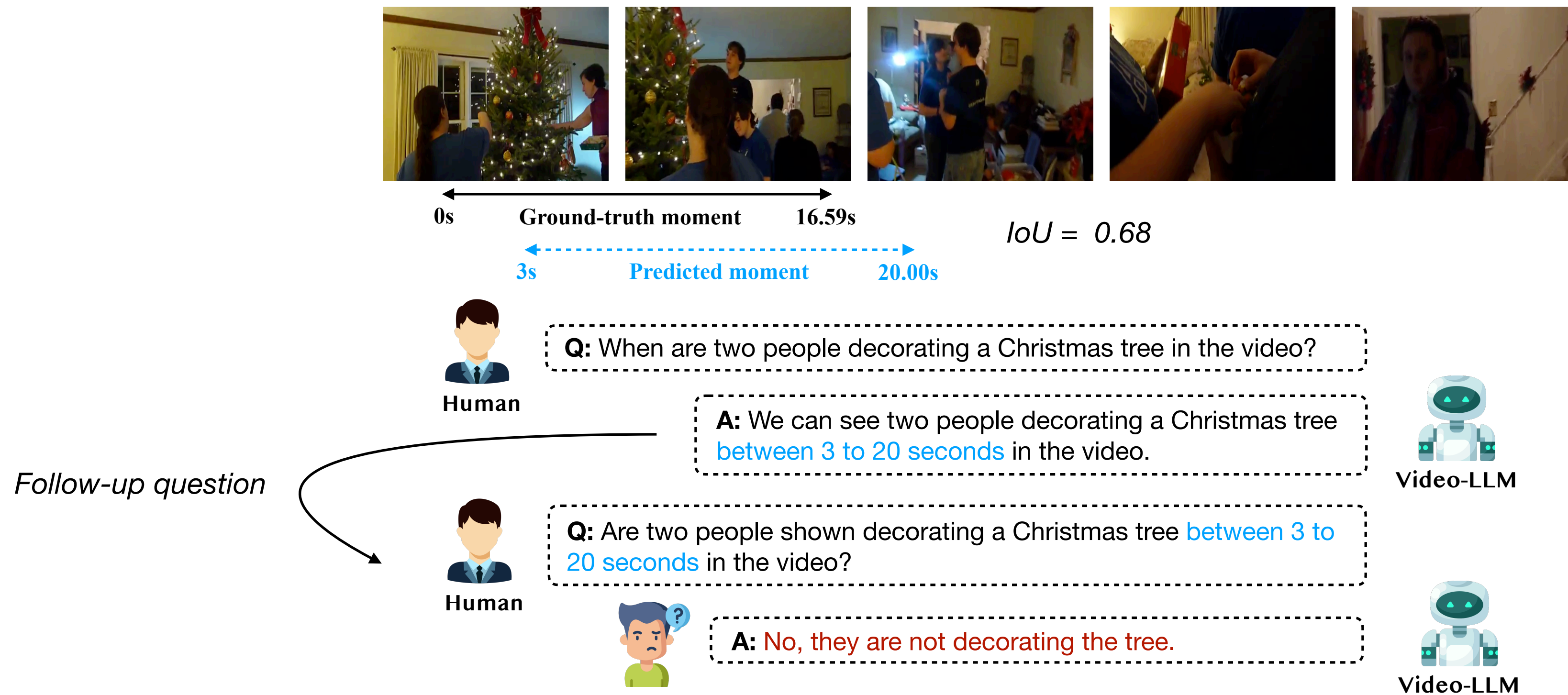
# Introduction



# Introduction

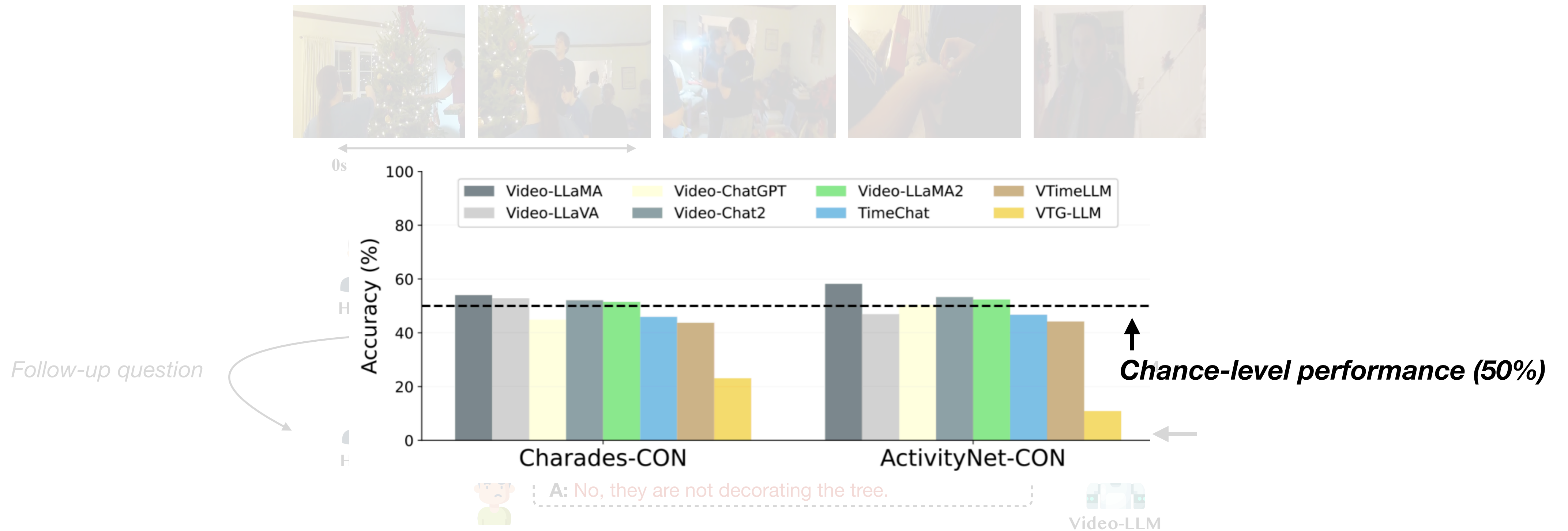


# Introduction



***Even if a model correctly captures video moments, it may fail to confirm its initial response!***

# Introduction



**Existing models are struggling to maintain consistent temporal understanding!**

# Dataset

## Grounding

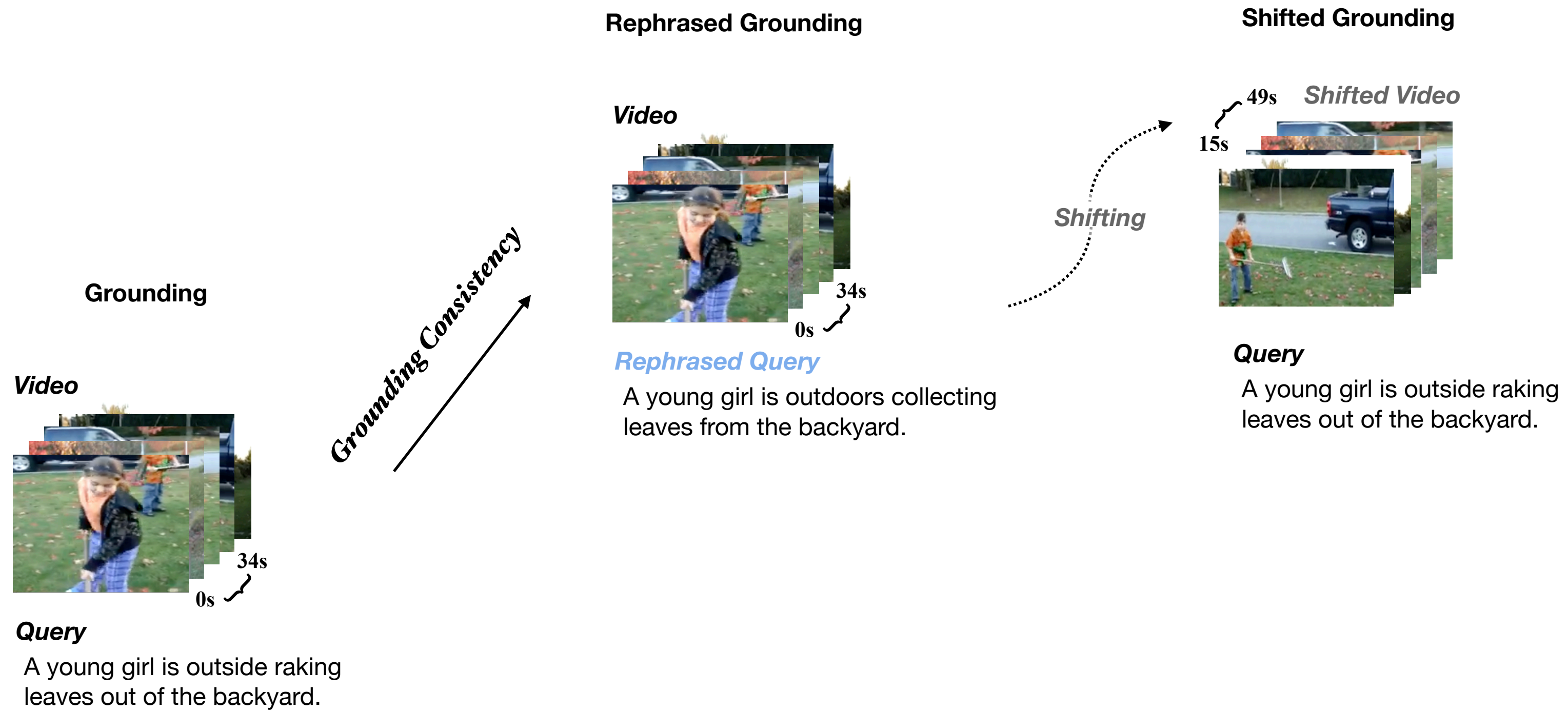
### Video



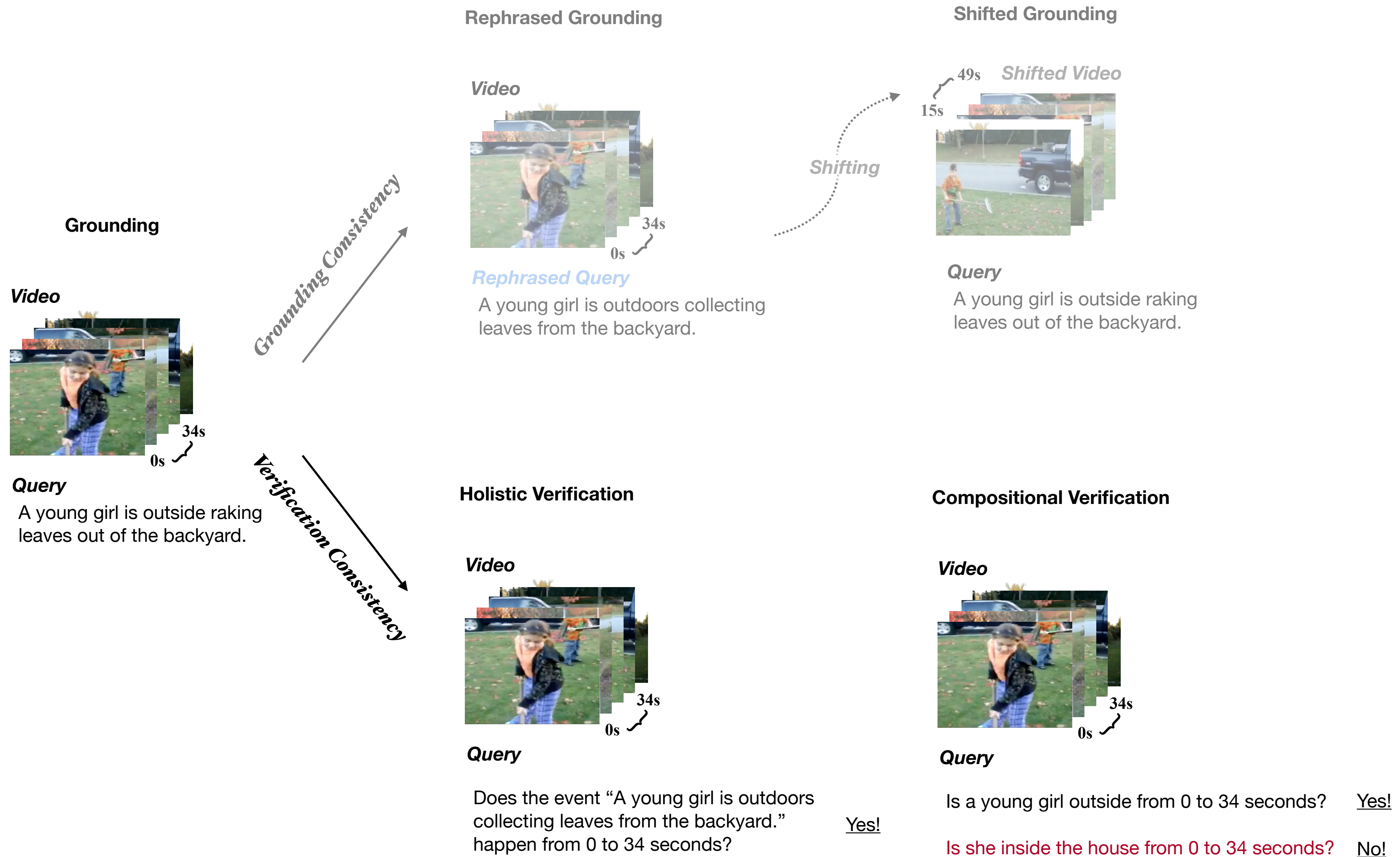
### Query

A young girl is outside raking leaves out of the backyard.

# Dataset



# Dataset



# Experiments

**Query:** A small group of people is seen standing together and speaking.



0s **Ground-truth moment** 8.47s

## Video Temporal Grounding

**Q:** When does the event 'A small group of people is seen standing together and speaking.' happen in the video?



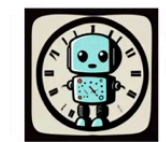
Video-LLaMA

**A:** The start time is 0.0 second and the end time is 4.4 second.



Video-ChatGPT

**A:** The event 'A small group of people is seen standing together and speaking.' occurs at 0:00 - 0:05.



TimeChat

**A:** The given query happens in 0.0 to 10.0 seconds.



VTimeLLM

**A:** The event 'A small group of people are seen standing together and speaking.' can be seen 0 to 6 seconds.



Gemini 1.5

**A:** The event 'A small group of people is seen standing together and speaking.' happens between 00:00 and 00:33.

## Holistic Verification

**Q:** Does the event 'A small group of people is seen standing together and speaking.' present from  $m$  in the video?



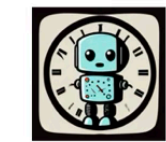
Video-LLaMA

**A:** Yes, the people are seen standing together and speaking.



Video-ChatGPT

**A:** The event 'A small group of people are seen standing together and speaking.' occurs at 0:00 - 0:05.



TimeChat

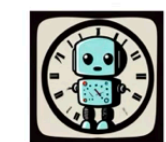
**A:** No, the event is not present from 0.0 to 10.0 seconds.

**Q:** Does the event '**A large group of people is seen sitting apart and speaking loudly.**' happen from  $m$  in the video?



Video-ChatGPT

**A:** Yes, the people are sitting apart from 0 to 5 seconds in the video.



TimeChat

**A:** No, the people are standing.



Gemini 1.5

**A:** No, the event 'A large group of people is seen sitting apart and speaking loudly' does not happen from 0 to 33 seconds. The video shows a man playing a drum.

## Compositional Verification

**Q:** Are they standing together from  $m$  in the video?



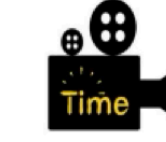
Video-LLaMA

**A:** Yes, they are standing together.



Video-ChatGPT

**A:** Yes, the people are standing together from 0 to 5 seconds in the video.



VTimeLLM

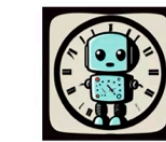
**A:** Yes, the people are standing together in the video.

**Q:** Are they heard singing from  $m$  in the video?



VTimeLLM

**A:** Yes, they are heard singing from 0 to 6 seconds in the video.



TimeChat

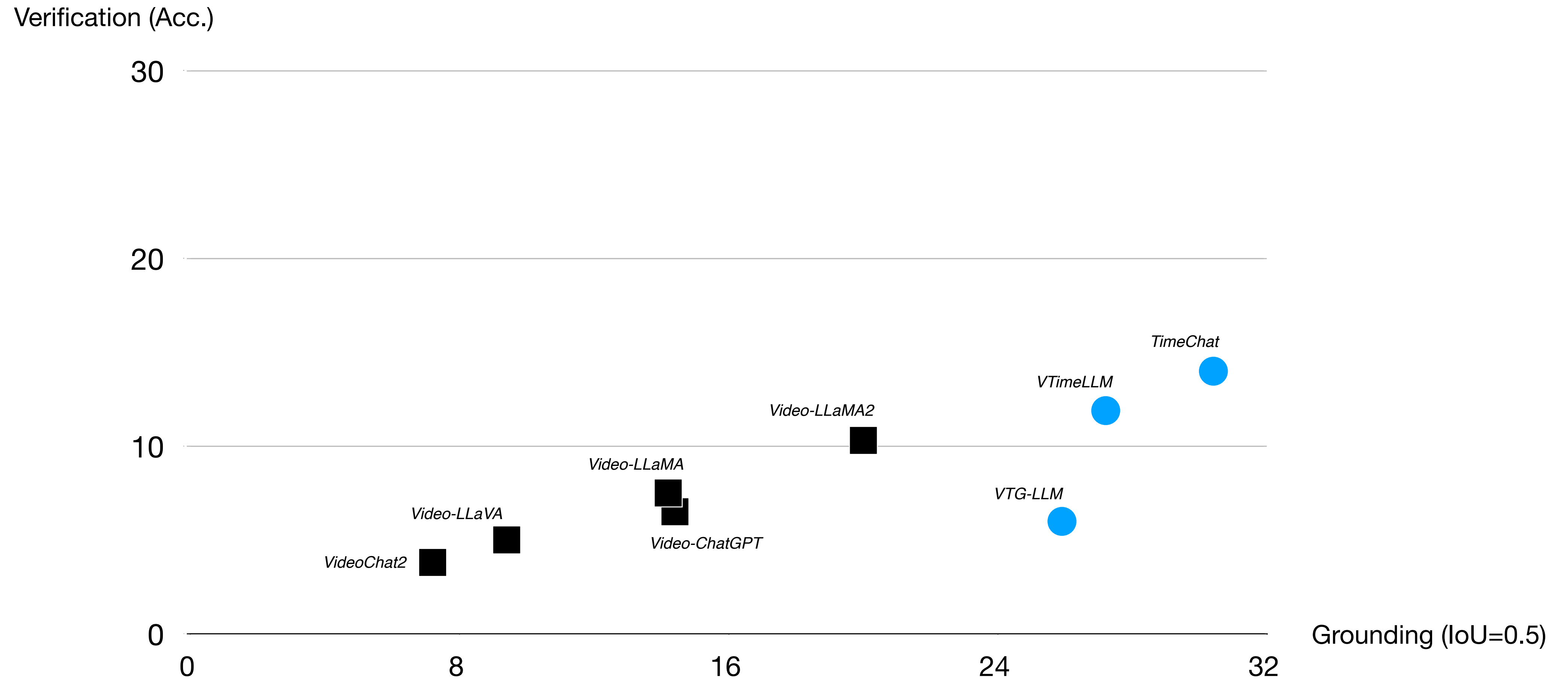
**A:** No, they are not heard singing.



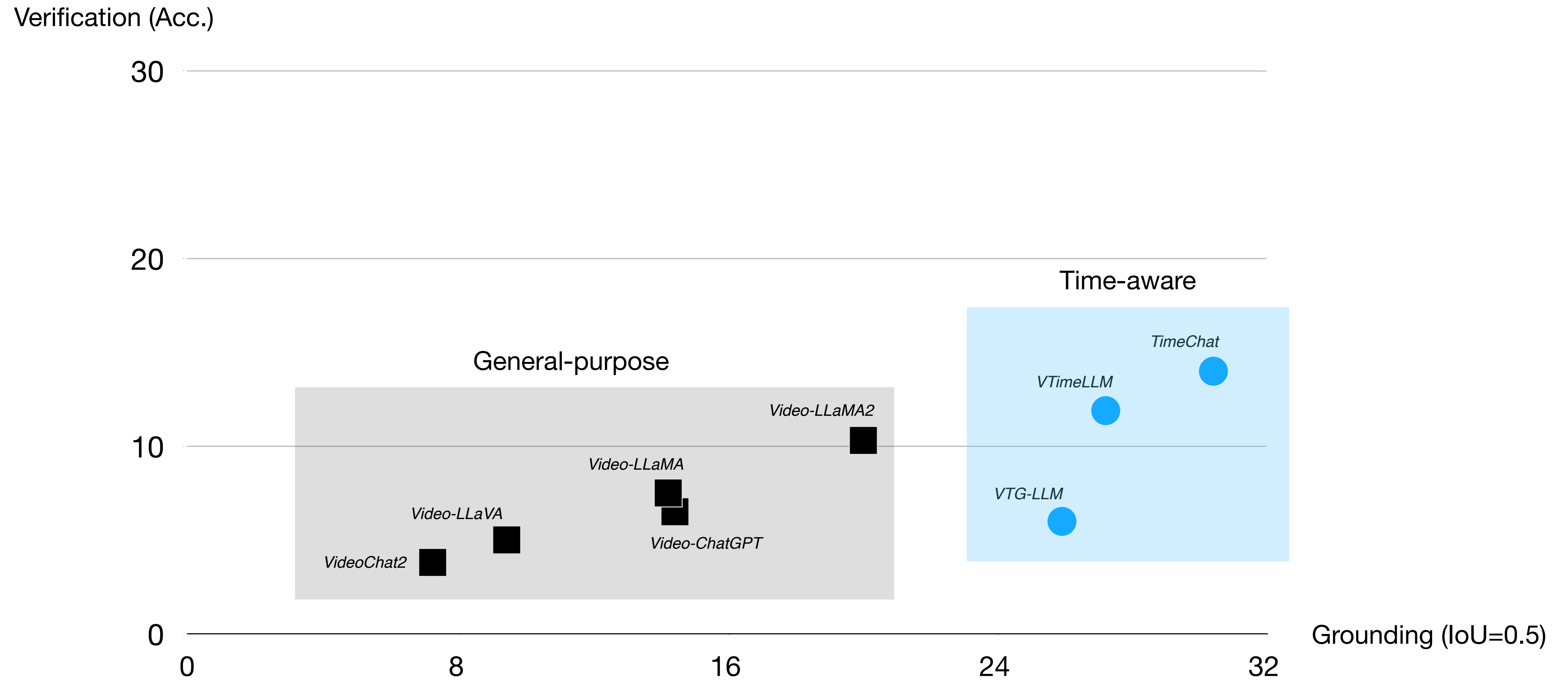
GPT-4o

**A:** No, they are playing instruments.

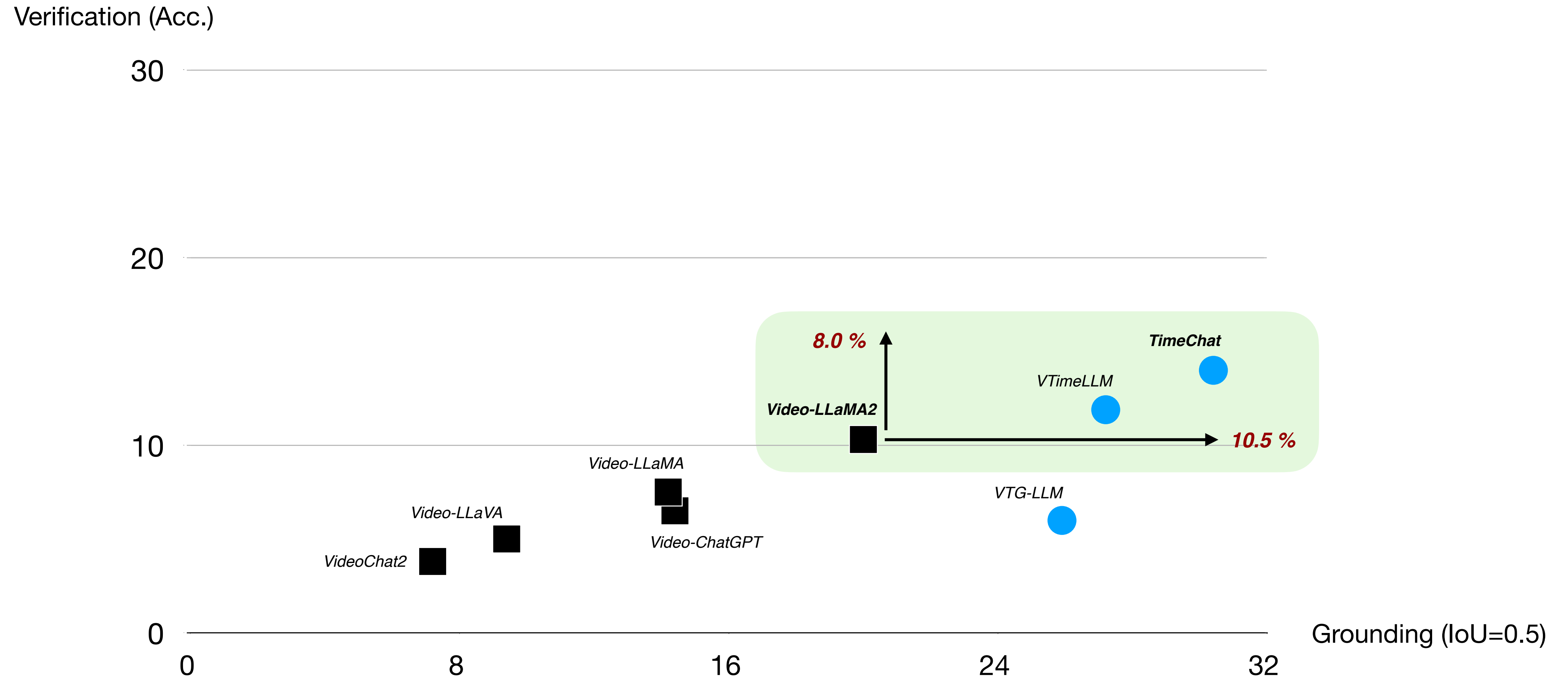
# Experiments



# Experiments

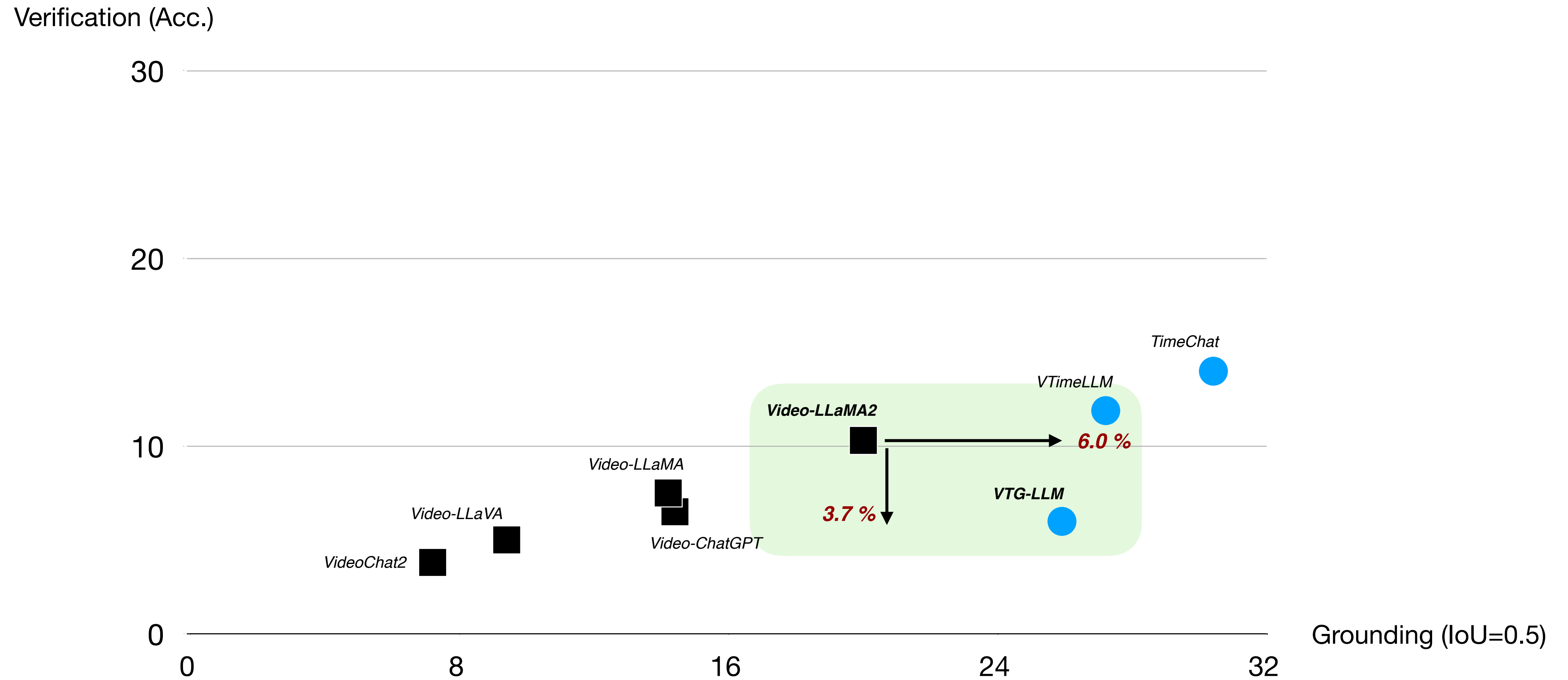


# Experiments



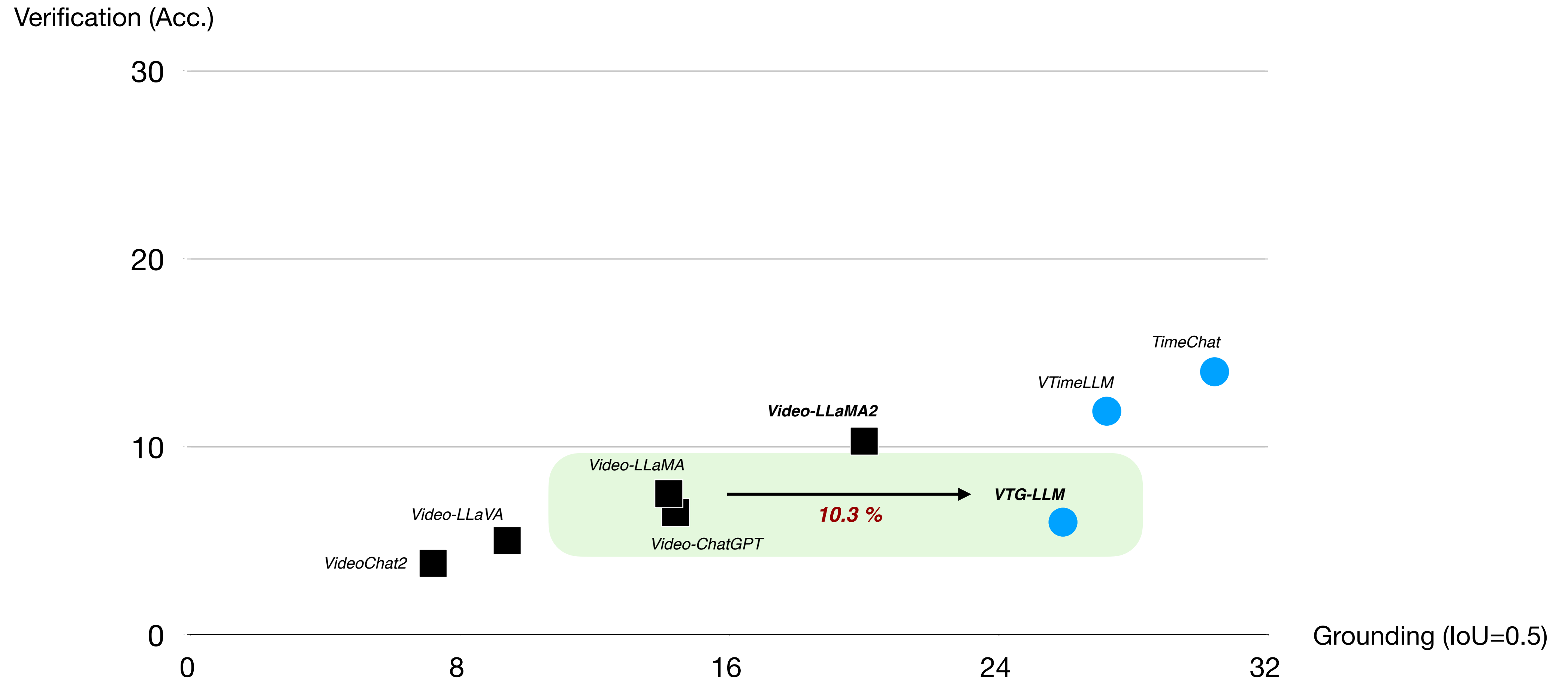
***Time-aware models generally outperform general-purpose ones for the grounding task.***

# Experiments



***However, improved grounding capabilities does not necessarily represent robustness or reliability.***

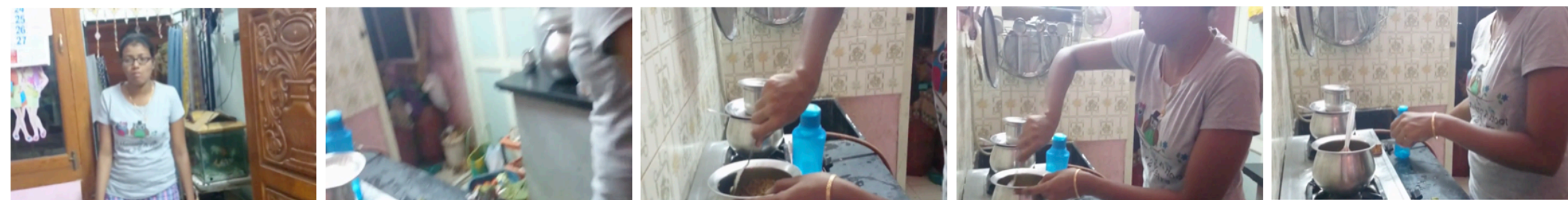
# Experiments



***However, improved grounding capabilities does not necessarily represent robustness or reliability.***

# Method

**Query:** A person starts cooking.



**Instruction Tuning (IT)**

12.4s

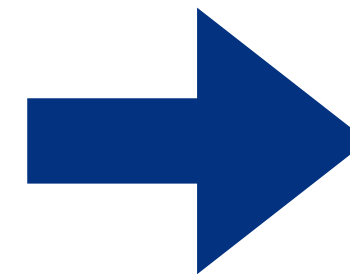
Ground-truth moment

27.0s

*Grounding Query*

**Q:** Localize the visual content described by the given textual query 'A person starts cooking' in the video, and output the start and end timestamps.

**A:** The given query happens in 12.4 to 27.0 seconds.



**Event Temporal Verification Tuning (VTune)**

*Event Verification Query*

**Q:** Is the event '**The individual is preparing a dish.**' present from 12.4 to 27.0 seconds in the video?

**A:** **Yes**, we can see the event 'The individual is preparing a dish.' from 12.4 to 27.0 seconds.

**Q:** Does the event '**The person is eating dinner at a table.**' not happen from 12.4 to 27.0 seconds in the video?"

**A:** Yes, the event 'The person is eating dinner at a table.' does not happen from 12.4 to 27.0 seconds in the video. **We cannot see a person eating dinner; the individual is cooking.**

*Temporal Verification Query*

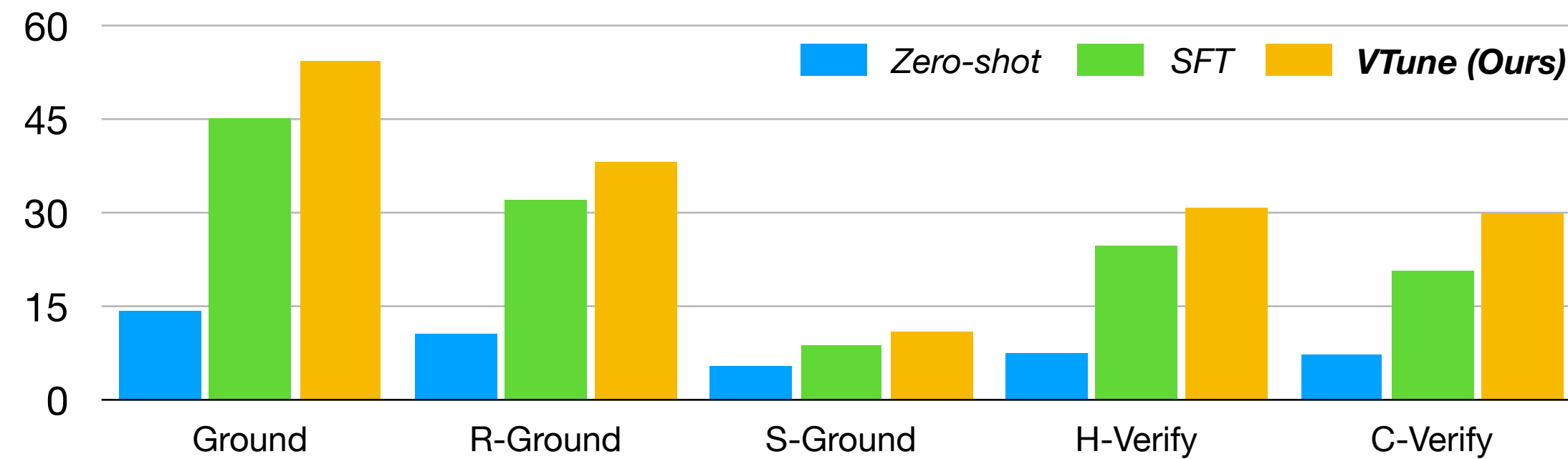
**Q:** Is the event 'A person starts cooking' present from **0 to 10 seconds** in the video?

**A:** **No**, we can see the event 'A person starts cooking' from **12.4 to 27.0 seconds**.

***We encourage the model to verify whether the given questions are both semantically and temporally correct.***

# Experiments

**Consistent Improvements**



**State-of-the-art Performance**

Method	Charades-STA		ActivityNet-Captions	
	R@1, 0.5	R@1, 0.7	R@1, 0.5	R@1, 0.7
<i>Task-Specific Models</i>				
BM-DETR [12]	59.4	38.3	49.6	30.6
Mr.BLIP [2]	69.3	49.2	53.9	35.5
<i>Video-LLMs</i>				
HawkEye [33]	58.3	28.8	34.7	17.7
VTG-LLM [8]	57.2	33.4	-	-
Video-LLaMA [41]	35.0	18.6	25.2	14.4
<b>Video-LLaMA-VT (Ours)</b>	37.1	20.1	34.3	<u>19.1</u>
TimeChat [31]	46.7	23.7	28.0	15.8
<b>TimeChat-VT (Ours)</b>	<b>58.4</b>	<b>34.7</b>	<b>41.0</b>	<b>23.7</b>

***Encouraging reasoning process inherently benefit the model's temporal understanding capabilities.***

# EgoExo-Con: Exploring View-Invariant Video Temporal Understanding



Minjoon Jung  
Seoul National University



Junbin Xiao  
National University of Singapore



Junhyun Kim  
Seoul National University



Byoung Tak Zhang  
Seoul National University



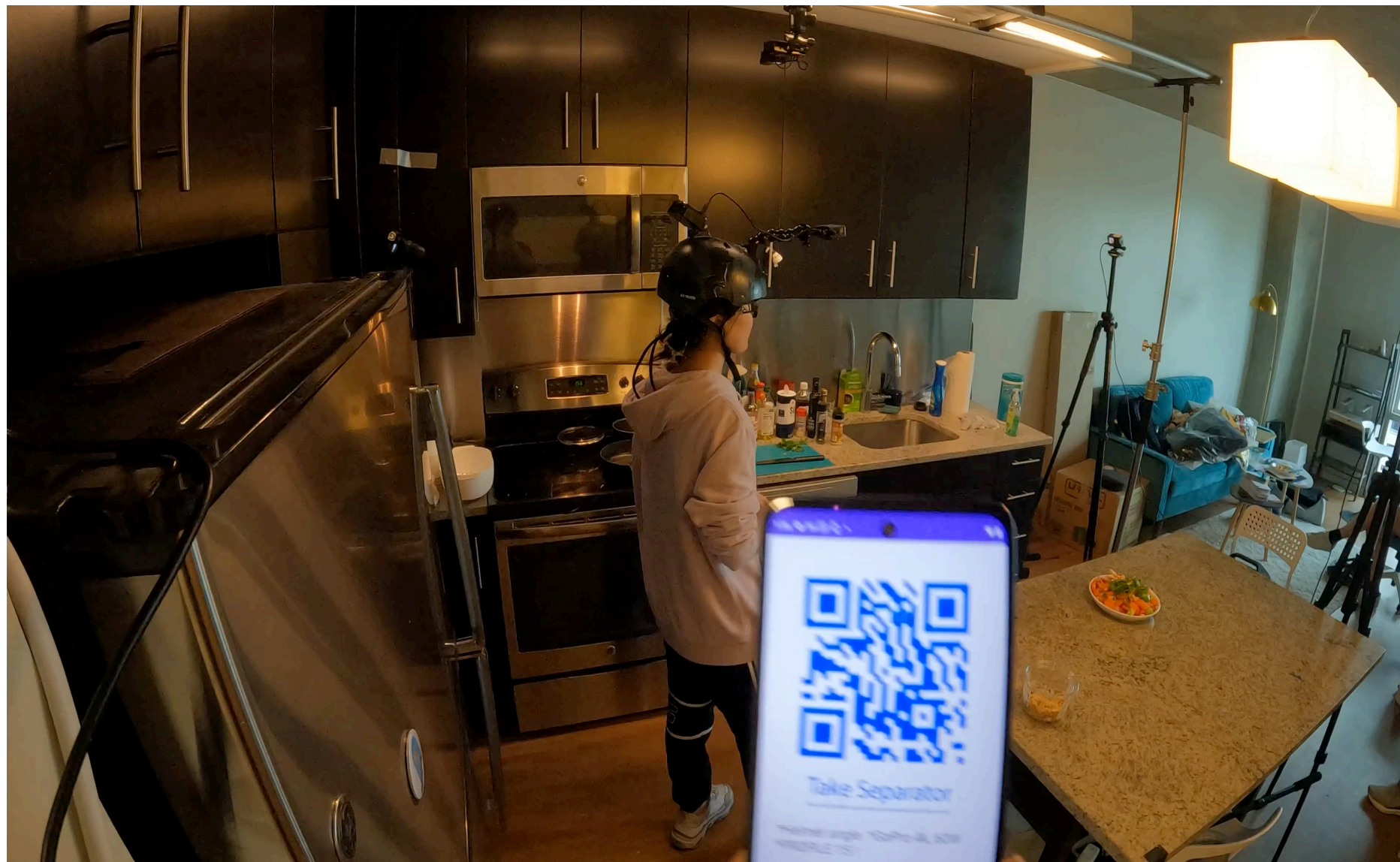
Angela Yao  
National University of Singapore



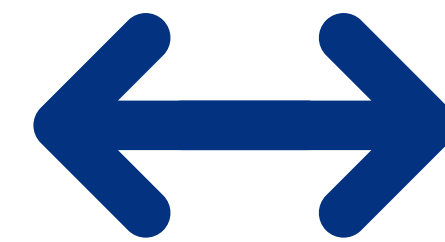
# Introduction

- Does a model generalize its temporal understanding across different viewpoints?

Exocentric (third-person view)



Egocentric (first-person view)



Synchronized!

# Introduction

- Unfortunately, existing benchmarks are insufficient to answer this question.

(a) CharadesEgo

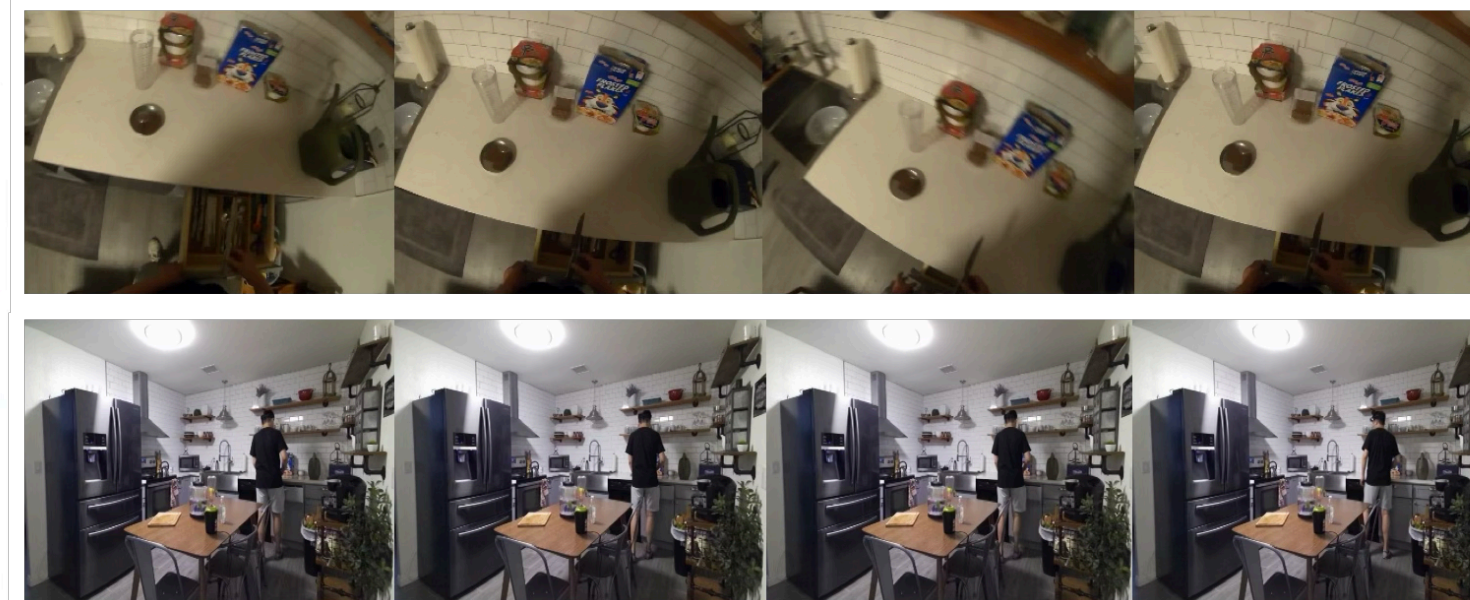
Query: A person is smiling.



*The egocentric view primarily captures hand–object interactions, making it difficult to infer a person’s appearance or facial expressions.*

(b) LEMMA

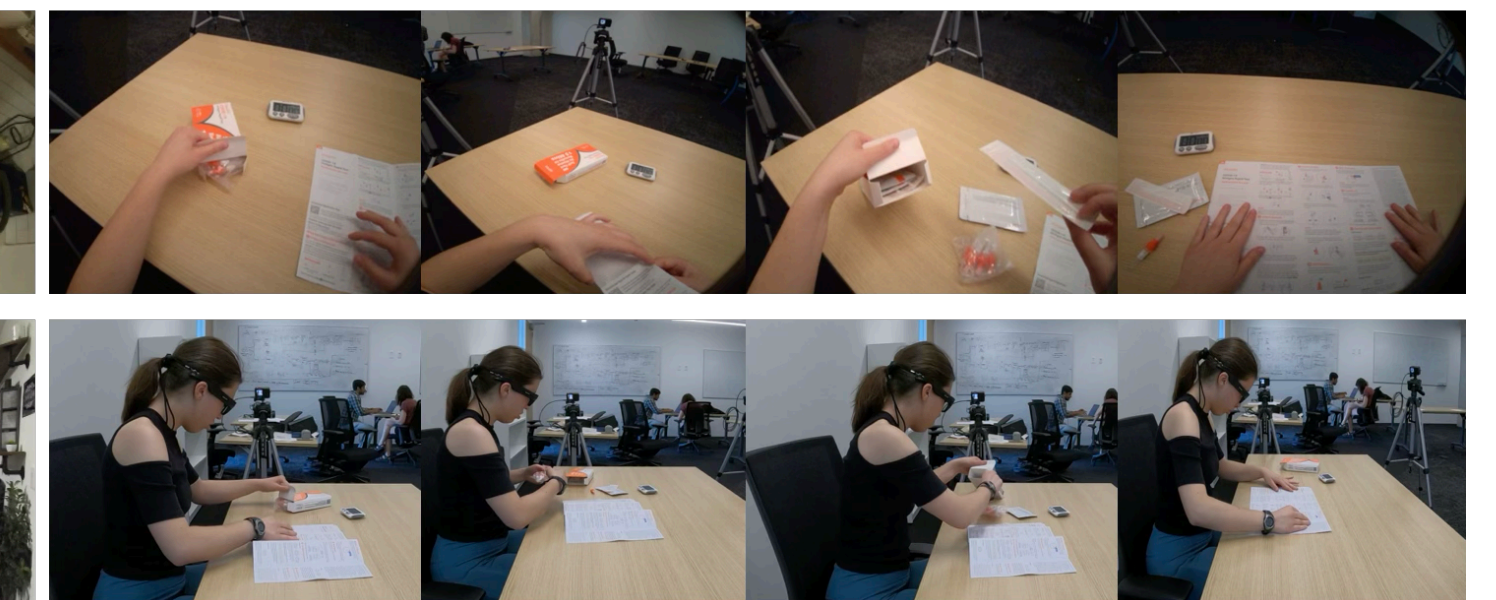
Query: Put a knife.



*Although a knife is clearly identifiable in the egocentric view, the object is ambiguous in the exocentric view due to the person's back being turned.*

(c) Ego-Exo4D

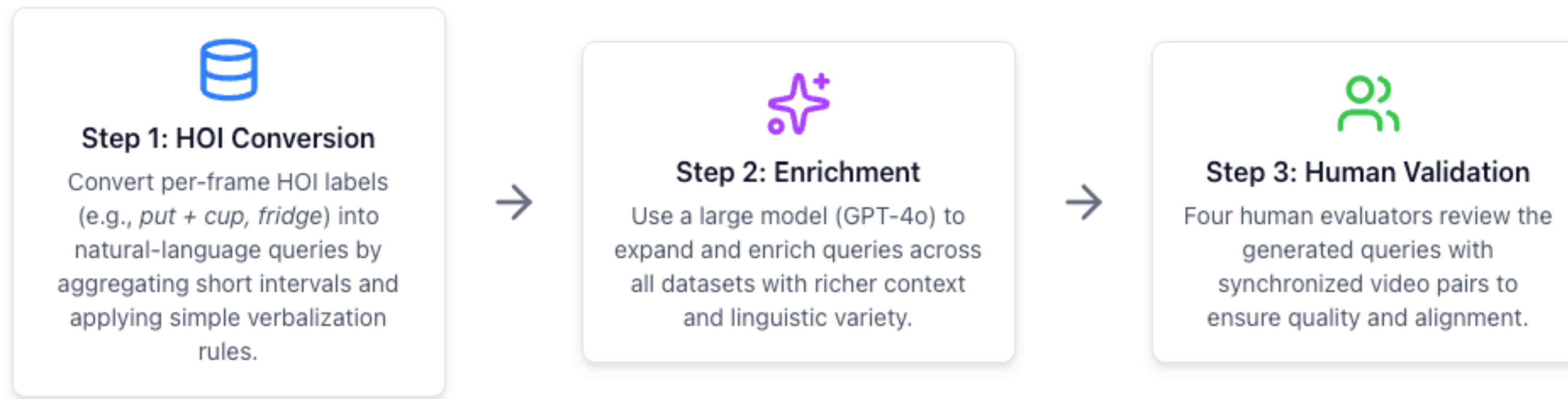
Query: Remove the test kits from the test pack.



*Both the egocentric and exocentric views provide complementary information.*

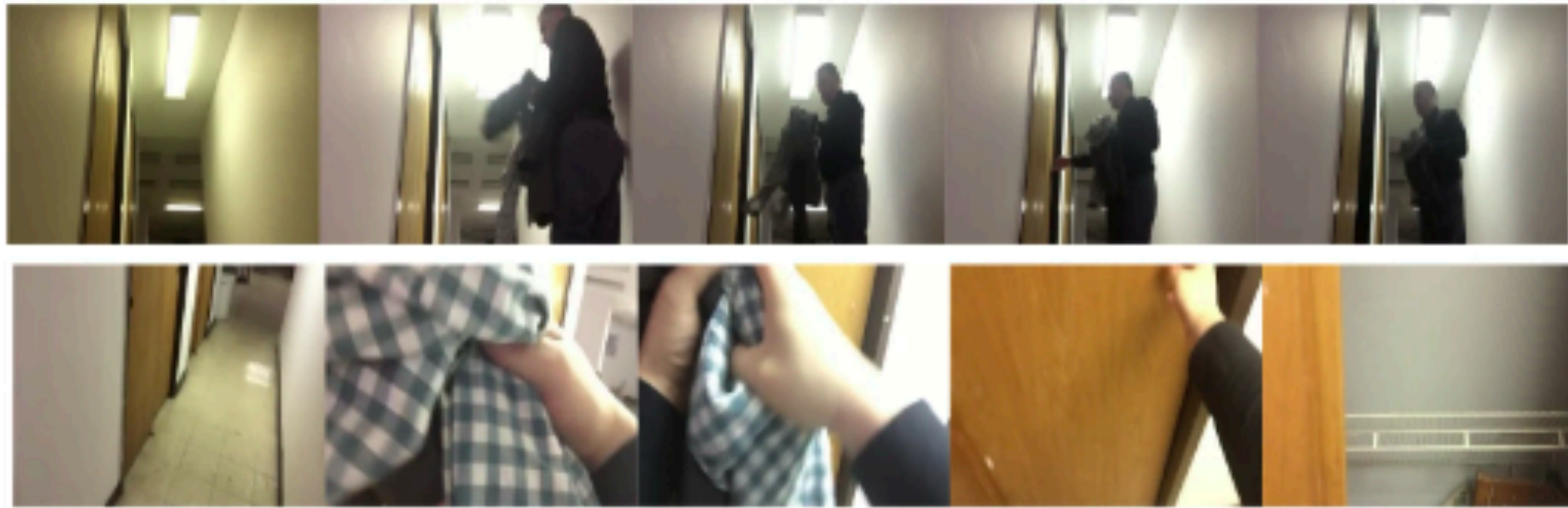
# Dataset

- To address this, we reformulate queries in multiple stages. And we perform human validation for all samples. Four human evaluators review the generated queries alongside associated synchronized video pairs.



# Experiments

Exocentric Video 
/ Egocentric Video



6.4 s ← ----- → 12.4 s

**Original Query**  
Closing a closet/cabinet.

**Refined Query**  
A person is holding clothes while opening a door in a hallway.

**Misaligned Query**  
A person is standing in a hallway, holding a suitcase, and is opening a cabinet door.


(V): Does the event <Refined Query> happen from 6.4 to 12.4 seconds in the video?

	A: Yes.	A: Yes. ✔ Consistent!
	A: Yes.	A: Yes. ✔ Consistent!

(V): Does the event <Misaligned Query> happen from 6.4 to 12.4 seconds in the video?

	A: Yes.	A: No. ✘ Inconsistent!
	A: No.	A: No. ✔ Consistent!

Exocentric Video 
/ Egocentric Video



11.1 s ← ----- → 16.9 s

**Original Query**  
Remove the test kits from the test pack.

**Refined Query**  
A person is sitting at a table, taking a test kit out of a box while looking at instructions.

**Misaligned Query**  
A person is standing at a table and reading a book.

(V): Does the event <Refined Query> happen from 11.1 to 16.9 seconds in the video?

	A: No.	A: Yes. ✘ Inconsistent!
	A: Yes.	A: Yes. ✔ Consistent!

(G): Please answer when the event <Refined Query> occurs in the video.

	A: 0.8 -8.0 (IoU = 0).	A: 0.0 -13.0 (IoU = 0.11). ✘ Inconsistent!
	A: 10.0 -15.0 (IoU = 0.55).	A: 10.0 -15.0 (IoU = 0.55). ✔ Consistent!

# Experiments

Methods	# F	Ego	EgoExo-Con					
			V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Human	-	-	92.1	91.3	89.4	72.4	73.0	67.3
<i>Closed-source</i>								
GPT-5 (OpenAI, 2025)	32	-	60.5	61.3	52.5	34.5	32.8	20.1
Gemini-2.5 Flash (Comanici et al., 2025)	1 fps	-	70.4	70.1	52.3	42.0	45.9	20.8
Random	-	-	50.0	50.0	50.0	12.5	12.5	12.5
<i>General-purpose</i>								
VideoChat2 (Li et al., 2024a)	16	✓	46.0	45.1	23.4	5.6	5.3	4.0
Qwen2.5-VL (Bai et al., 2025)	1 fps	✗	54.3	56.3	33.0	14.2	11.4	6.9
Video-LLaMA2 (Cheng et al., 2024b)	8	✓	53.3	52.1	27.9	12.0	11.5	7.5
Video-LLaMA3 (Zhang et al., 2025a)	1 fps	✓	56.7	54.6	36.6	27.7	<b>28.0</b>	16.2
<i>Time-aware</i>								
VTimeLLM (Huang et al., 2024)	100	✗	48.9	48.5	23.5	12.6	11.1	6.5
TimeChat (Ren et al., 2024)	96	✗	48.9	48.4	25.1	21.3	20.5	12.8
TimeSuite (Zeng et al., 2024)	128	✓	47.4	48.5	25.6	<b>28.2</b>	27.3	<b>18.7</b>
TimeChat-VT (Jung et al., 2025b)	96	✗	<b>62.1</b>	<b>61.4</b>	<b>42.1</b>	27.8	26.2	16.3

# Experiments

## ● Single view vs. Cross view

Methods	# F	Ego	EgoExo-Con					
			V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Human	-	-	92.1	91.3	89.4	72.4	73.0	67.3
<i>Closed-source</i>								
GPT-5 (OpenAI, 2025)	32	-	60.5	61.3	52.5	34.5	32.8	20.1
Gemini-2.5 Flash (Comanici et al., 2025)	1 fps	-	70.4	70.1	52.3	42.0	45.9	20.8
Random	-	-	50.0	50.0	50.0	12.5	12.5	12.5
<i>General-purpose</i>								
VideoChat2 (Li et al., 2024a)	16	✓	46.0	45.1	23.4	5.6	5.3	4.0
Qwen2.5-VL (Bai et al., 2025)	1 fps	✗	54.3	56.3	33.0	14.2	11.4	6.9
Video-LLaMA2 (Cheng et al., 2024b)	8	✓	53.3	52.1	27.9	12.0	11.5	7.5
Video-LLaMA3 (Zhang et al., 2025a)	1 fps	✓	56.7	54.6	36.6	27.7	<b>28.0</b>	16.2
<i>Time-aware</i>								
VTimeLLM (Huang et al., 2024)	100	✗	48.9	48.5	23.5	12.6	11.1	6.5
TimeChat (Ren et al., 2024)	96	✗	48.9	48.4	25.1	21.3	20.5	12.8
TimeSuite (Zeng et al., 2024)	128	✓	47.4	48.5	25.6	<b>28.2</b>	27.3	<b>18.7</b>
TimeChat-VT (Jung et al., 2025b)	96	✗	<b>62.1</b>	<b>61.4</b>	<b>42.1</b>	27.8	26.2	16.3

While a modest gap between individual viewpoints, Video-LLMs in particular achieve barely half of their single-view performance.

# Experiments

## ● Human vs. Closed-sourced

Methods	# F	Ego	EgoExo-Con					
			V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Human	-	-	92.1	91.3	89.4	72.4	73.0	67.3
<i>Closed-source</i>					↓ ~37%			↓ ~47%
GPT-5 (OpenAI, 2025)	32	-	60.5	61.3	52.5	34.5	32.8	20.1
Gemini-2.5 Flash (Comanici et al., 2025)	1 fps	-	70.4	70.1	52.3	42.0	45.9	20.8
Random	-	-	50.0	50.0	50.0	12.5	12.5	12.5
<i>General-purpose</i>								
VideoChat2 (Li et al., 2024a)	16	✓	46.0	45.1	23.4	5.6	5.3	4.0
Qwen2.5-VL (Bai et al., 2025)	1 fps	✗	54.3	56.3	33.0	14.2	11.4	6.9
Video-LLaMA2 (Cheng et al., 2024b)	8	✓	53.3	52.1	27.9	12.0	11.5	7.5
Video-LLaMA3 (Zhang et al., 2025a)	1 fps	✓	56.7	54.6	36.6	27.7	<b>28.0</b>	16.2
<i>Time-aware</i>								
VTimeLLM (Huang et al., 2024)	100	✗	48.9	48.5	23.5	12.6	11.1	6.5
TimeChat (Ren et al., 2024)	96	✗	48.9	48.4	25.1	21.3	20.5	12.8
TimeSuite (Zeng et al., 2024)	128	✓	47.4	48.5	25.6	<b>28.2</b>	27.3	<b>18.7</b>
TimeChat-VT (Jung et al., 2025b)	96	✗	<b>62.1</b>	<b>61.4</b>	<b>42.1</b>	27.8	26.2	16.3

*Yet a substantial gap in cross-view consistency remains compared to humans.*

# Experiment

## ● Limitation of ego-view training

Methods	# F	Ego	EgoExo-Con					
			V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Human	-	-	92.1	91.3	89.4	72.4	73.0	67.3
<i>Closed-source</i>								
GPT-5 (OpenAI, 2025)	32	-	60.5	61.3	52.5	34.5	32.8	20.1
Gemini-2.5 Flash (Comanici et al., 2025)	1 fps	-	70.4	70.1	52.3	42.0	45.9	20.8
Random	-	-	50.0	50.0	50.0	12.5	12.5	12.5
<i>General-purpose</i>								
VideoChat2 (Li et al., 2024a)	16	✓	46.0	45.1	23.4	5.6	5.3	4.0
Qwen2.5-VL (Bai et al., 2025)	1 fps	✗	54.3	56.3	33.0	14.2	11.4	6.9
Video-LLaMA2 (Cheng et al., 2024b)	8	✓	53.3	52.1	27.9	12.0	11.5	7.5
Video-LLaMA3 (Zhang et al., 2025a)	1 fps	✓	56.7	54.6	36.6	27.7	<b>28.0</b>	16.2
<i>Time-aware</i>								
VTimeLLM (Huang et al., 2024)	100	✗	48.9	48.5	23.5	12.6	11.1	6.5
TimeChat (Ren et al., 2024)	96	✗	48.9	48.4	25.1	21.3	20.5	12.8
TimeSuite (Zeng et al., 2024)	128	✓	47.4	48.5	25.6	<b>28.2</b>	27.3	<b>18.7</b>
TimeChat-VT (Jung et al., 2025b)	96	✗	<b>62.1</b>	<b>61.4</b>	<b>42.1</b>	27.8	26.2	16.3

*Models including egocentric videos for training do not consistently outperform others trained on exocentric videos alone.*

# Experiments

- Despite utilizing twice the data, training on both viewpoints (EgoExo) yields only marginal gains and often underperforms the models trained on a single view (Ego or Exo).

Methods	CharadesEgo			EgoExo-4D		
	V-Exo	V-Ego	V-ExoEgo	V-Exo	V-Ego	V-ExoEgo
VideoChat2	46.3	44.4	22.2	41.4	41.5	19.8
+ Ego	56.2	<b>59.2</b>	<b>36.4</b>	46.5	49.2	29.7
+ Exo	<b>56.6</b>	57.5	35.5	46.1	47.3	29.4
+ EgoExo	56.4	57.1	34.7	<b>48.3</b>	<b>50.1</b>	<b>30.1</b>
Video-LLaMA2	54.0	52.4	28.2	51.5	52.8	28.1
+ Ego	57.0	<b>58.2</b>	<b>31.7</b>	<b>60.2</b>	<b>60.3</b>	38.8
+ Exo	57.6	56.1	31.4	59.8	60.1	<b>39.2</b>
+ EgoExo	<b>58.5</b>	57.3	31.0	57.5	59.6	39.1

Methods	CharadesEgo			EgoExo-4D		
	G-Exo	G-Ego	G-ExoEgo	G-Exo	G-Ego	G-ExoEgo
TimeChat	44.9	46.1	30.1	4.9	6.3	3.3
+ Ego	<b>62.0</b>	<b>62.1</b>	<b>48.3</b>	9.7	<b>13.1</b>	<b>4.9</b>
+ Exo	58.8	60.0	47.1	10.4	10.7	4.2
+ EgoExo	60.3	61.8	46.5	<b>10.6</b>	11.5	4.3
TimeSuite	63.4	56.2	44.8	5.8	8.3	2.3
+ Ego	61.0	61.5	54.0	6.7	6.2	4.8
+ Exo	<b>74.6</b>	<b>68.7</b>	<b>59.5</b>	<b>10.5</b>	<b>9.9</b>	<b>5.7</b>
+ EgoExo	67.8	61.1	51.3	9.6	8.9	5.3

*Naively blending perspectives is insufficient for view-invariant temporal understanding!*

# Method

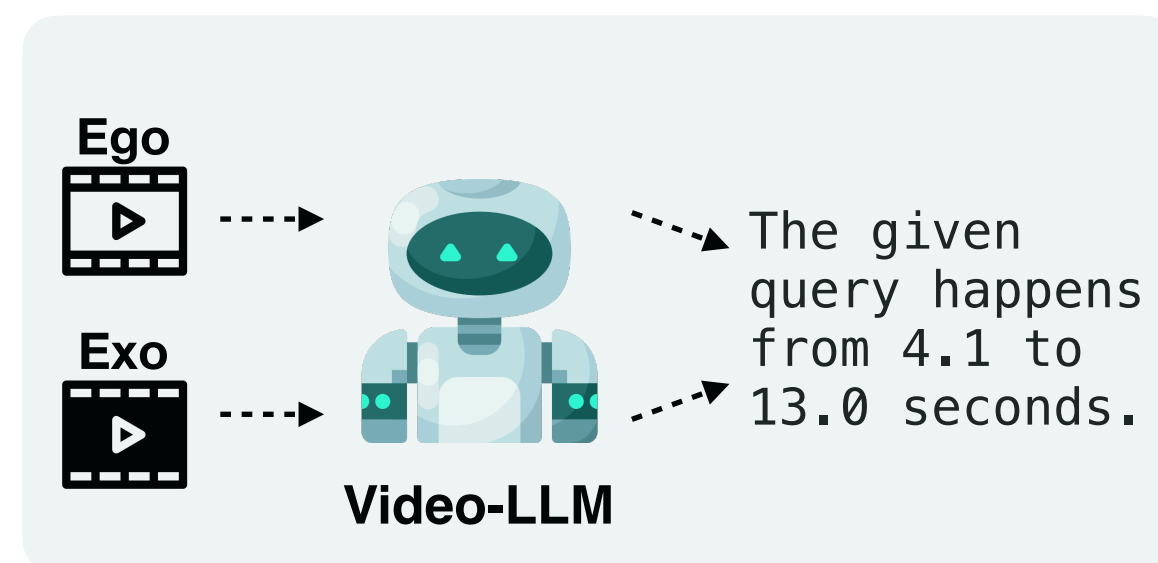
- We propose View-GRPO that guides models toward developing viewpoint-specific reasoning while encouraging shared consistency.

Query: 'Tidying some clothes'

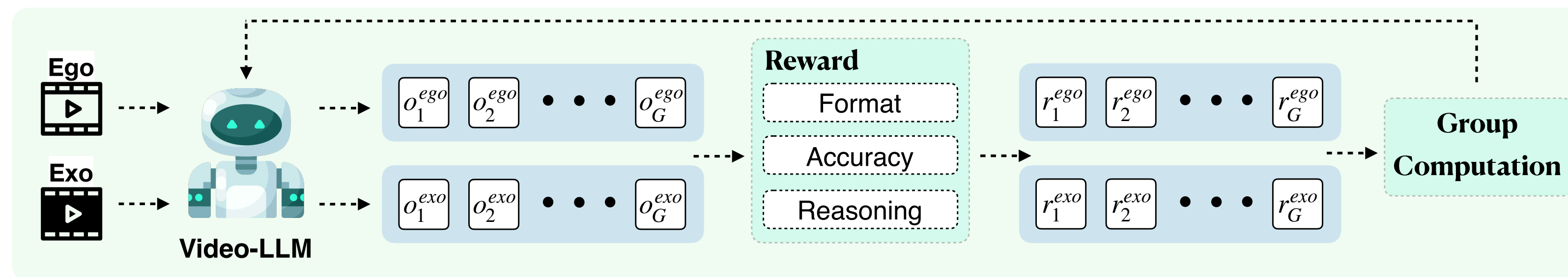


GPT-5: `<think>` The video captures an interior bedroom scene featuring a woman lying on a bed with a patterned blanket. The main activity of interest, "tidying clothes," begins as she sits up. The action is visually defined by a sequence of continuous hand motions, including picking up, folding, and arranging garments into an organized pile. This event occurs from approximately 4.1 seconds to 13.0 seconds. `</think>` `<time>`[4.1, 13.0]`</time>`

GPT-5: `<think>` Overall scene:The video shows a person "tidying clothes" from an egocentric/first-person perspective. The activity is defined by a series of hand movements—including lifting, grasping, folding, and arranging—that occur continuously between 4.1 and 13.0 seconds. The scene is a small room with a bed and household items, and the person is wearing a purple-and-white striped garment. `</think>` `<time>`[4.1, 13.0]`</time>`



(a) Supervised Fine-Tuning



(b) View-GRPO

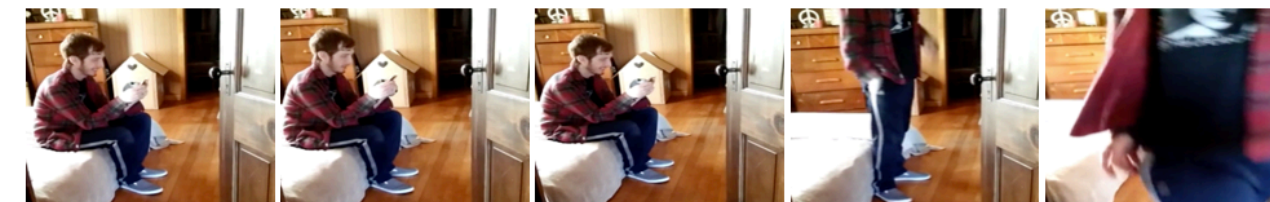
*Enhancing viewpoint-specific reasoning while aligning the final answers!*

# Experiments

## ● Effectiveness of View-GRPO

Methods	EgoExo-Con					
	V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Qwen2.5-VL-3B	51.0	52.5	28.1	10.1	9.9	7.9
+ SFT	<u>52.7</u>	51.2	<u>30.6</u>	<u>16.3</u>	<u>16.6</u>	<u>12.9</u>
+ GRPO	52.5	<u>52.9</u>	30.3	15.4	16.3	10.3
+ <b>View-GRPO</b>	<b>54.0 ↑3.0</b>	<b>54.1 ↑1.6</b>	<b>33.9 ↑5.8</b>	<b>18.6 ↑8.5</b>	<b>17.9 ↑8.0</b>	<b>14.8 ↑6.9</b>
Qwen2.5-VL-7B	54.3	56.3	33.0	14.2	11.4	6.9
+ SFT	<u>57.6</u>	<u>58.0</u>	<u>41.4</u>	18.3	<u>17.8</u>	<u>14.9</u>
+ GRPO	55.2	57.6	39.8	<u>18.6</u>	16.1	14.3
+ <b>View-GRPO</b>	<b>58.3 ↑4.0</b>	<b>58.1 ↑1.8</b>	<b>44.7 ↑11.7</b>	<b>21.5 ↑7.3</b>	<b>21.0 ↑9.6</b>	<b>18.3 ↑11.4</b>

Query: Holding a phone/camera



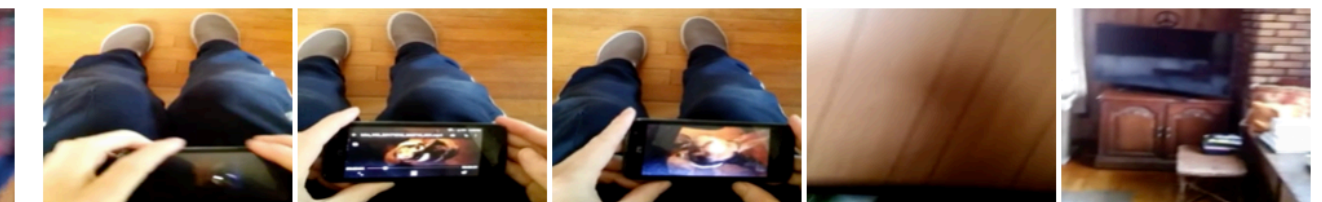
**View-GRPO:** <think> Given the event "holding a phone/camera," the person in the video is indeed holding a phone or camera. The primary focus is on the person using their phone or a camera. To accurately pinpoint the time periods, we need to consider the duration of the activity: 1. The person is actively engaged with her device throughout the video. ... 3. There is no indication of a break or transition in the person's activity with their device....</think> <answer> 0.00 to 23.56 </answer>

**Ground-Truth:** <think> The shot is exocentric (third-person viewpoint) and consistently shows a handheld smartphone in the lower-right...1. 0.0~8.0s - Establishing view: the person and room are visible but the foreground hand and phone are either out of...5.~26.7s - End of sustained hold: the hand either withdraws the phone away from the lower-right view or lowers it enough that the phone is no longer clearly visible...Based on these observable cues, the best-aligned interval is the segment where the phone is steadily visible and held...</think> <answer> 0.0 to 26.5 </answer>

$r_{sim} = 0.6$  (by Qwen2.5-3B)

$r_{sim} = 0.7$  (by Qwen2.5-0.5B)

IoU = 0.89



**View-GRPO:** <think> The person is holding a phone/camera throughout the video segment. Analysis of key frames reveals consistent activity with this object in the person's hand. This is supported by the observation of the phone/camera being prominently displayed and the person's hand consistently interacting with it. The absence of any breaks or interruptions in the holding position of the phone/camera aligns with the continuous observation pattern.</think> <answer> 0.00 to 30.63 </answer>

**Ground-Truth:** <think> A person is seated with their legs and shoes visible, holding a smartphone horizontally over their lap while watching a video. The room and a TV screen appear in the background as the person shifts the phone and later pans the camera upward...0.0~0.5s - Initial establishing view: both hands are visible holding a phone horizontally above the lap;0.5~20.0s - Continuous holding...After ~26.5s - The scene shows ceiling, doorway, and other room features with no handheld phone visible.</think> <answer> 0.0 to 26.5 </answer>

$r_{sim} = 0.6$  (by Qwen2.5-3B)

$r_{sim} = 0.8$  (by Qwen2.5-0.5B)

IoU = 0.87

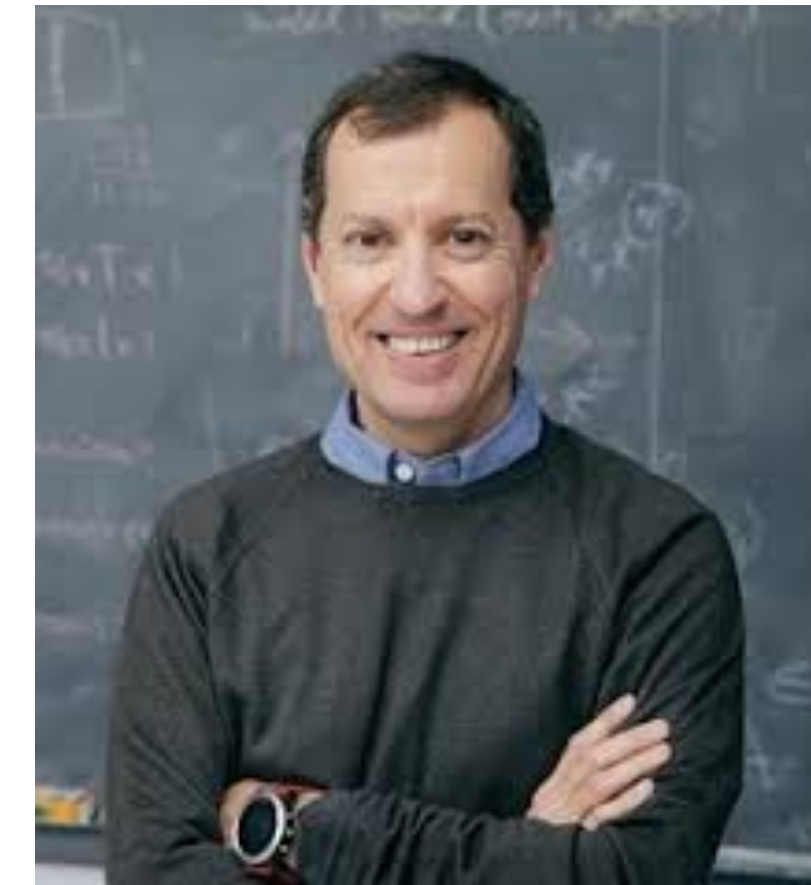
# EvoGround: Self-Evolving Video Agents for Video Temporal Grounding



Minjoon Jung  
Seoul National University



Byoung Tak Zhang  
Seoul National University

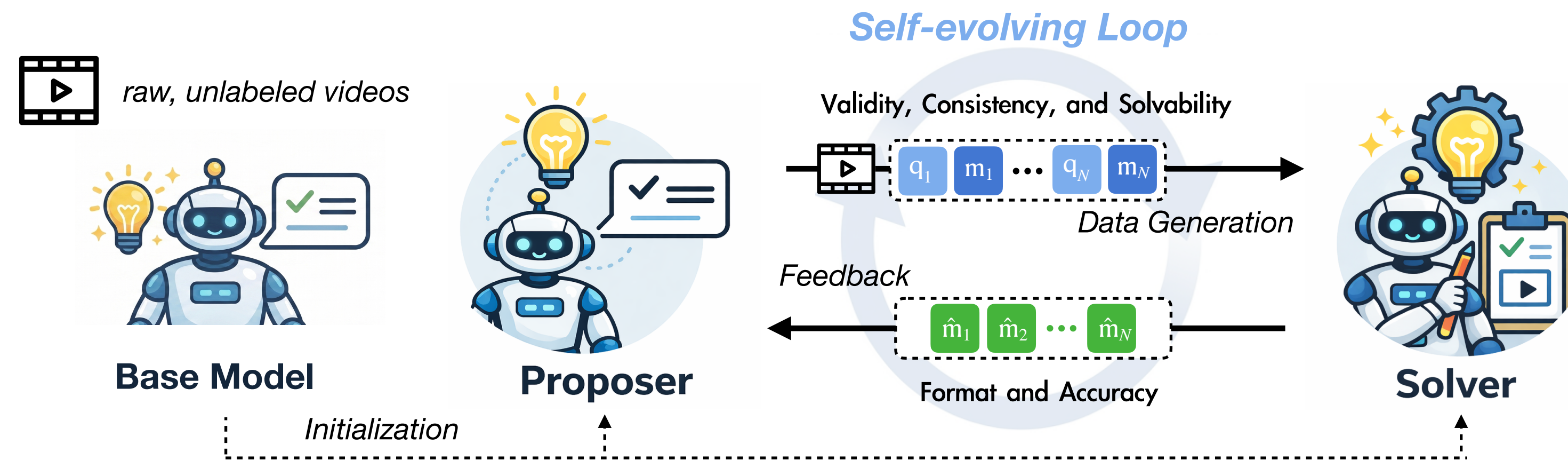


Lorenzo Torresani  
Northeastern University



# Introduction

- We propose EvoGround, a self-evolving video agents that learn temporal information from raw videos without any manual annotation.

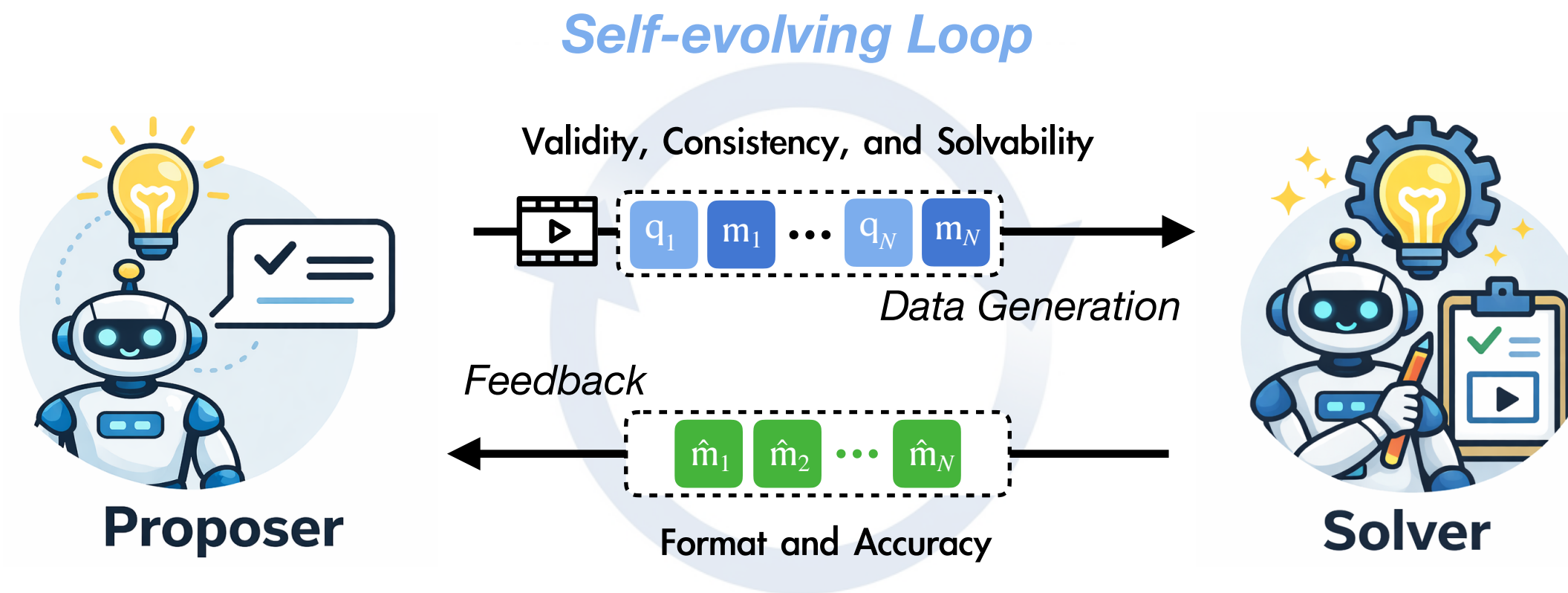


# Method

- **Reward designs for the proposer:**
  - Format reward (Validity): Are generated timestamps well-formed and within the video?
  - Consistency reward (Consistency): Does a query describe its corresponding video moment?
  - Feedback reward (Solvability): Can the solver actually localize this data?

# Method

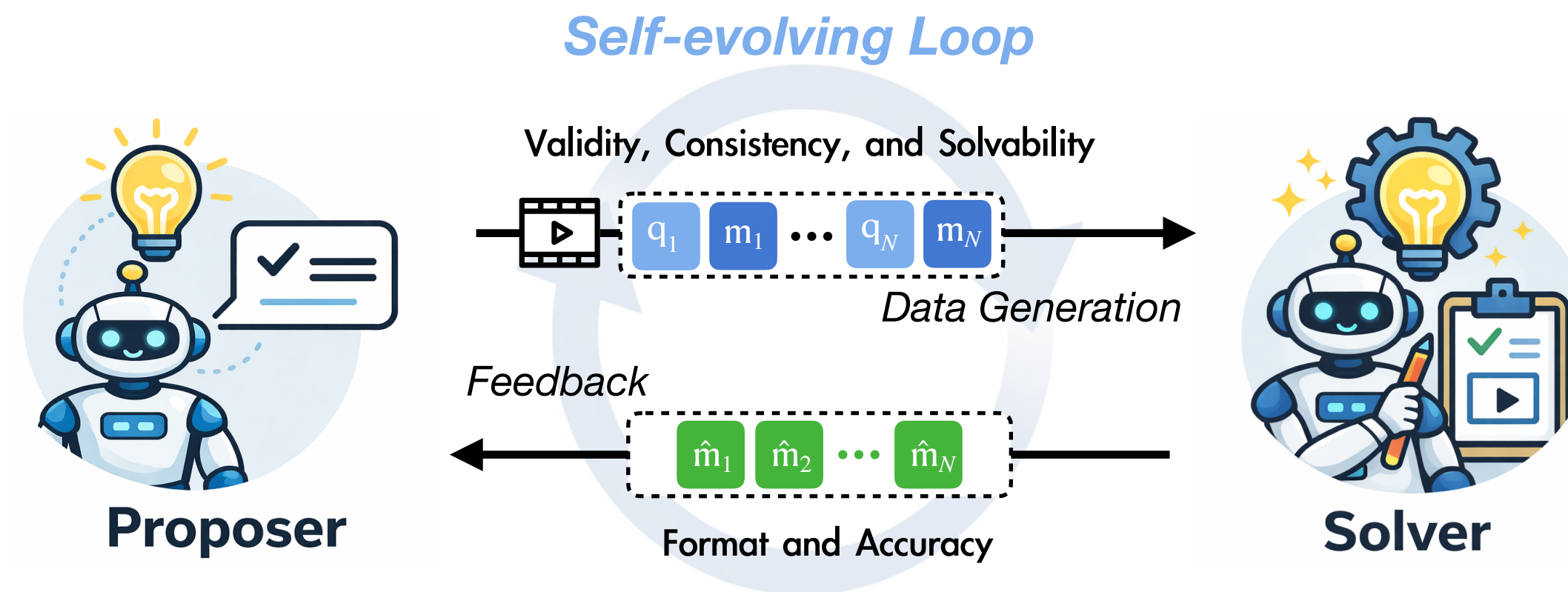
- **Format reward confirms whether the proposer generates valid outputs. Let the proposer generate  $N$  query-moment pairs  $\{(q_n, m_n)\}_{n=1}^N$ , where  $q_n$  denotes the  $n$ -th query and  $m_n = (s_n, e_n)$  represents the corresponding start and end time.**



$$R_{\text{format}}(\mathbf{m}_i) = \begin{cases} \frac{1}{N} \sum_{n=1}^N \mathbb{1}((s_n, e_n) \in \mathcal{V}), & \text{if the proposer follows the template,} \\ 0, & \text{otherwise} \end{cases}$$

# Method

- Beyond enforcing valid formatting, we define the conditioned format reward based on the solver's accuracy.



$$\mathbb{I}((s_n, e_n) \in \mathcal{V} \wedge \text{tIoU}(m_n, \hat{m}_n) \geq \delta)$$

↑ Solver's prediction

↓ Threshold

# Method

- The consistency reward measures two complementary properties of each query-moment pair.

$$R_{\text{consistency}} = \frac{1}{N} \sum_{n=1}^N \underbrace{\exp\left(-\gamma \sqrt{\frac{1}{|m_n|} \sum_{t=1}^{|m_n|} (S_{n,t} - \mu_n)^2}\right)}_{\text{Intra-consistency}} \cdot \underbrace{\frac{\exp(\mu_n/\tau)}{\sum_{j=1}^N \exp(\mu_j/\tau)}}_{\text{Inter-consistency}}$$

# Method

- The intra-consistency score penalizes high standard deviation in  $S_{n,t}$ , rewarding pairs where the query aligns uniformly with frames in the moment  $m_n$ .

$$R_{\text{consistency}} = \frac{1}{N} \sum_{n=1}^N \underbrace{\exp \left( -\gamma \sqrt{\frac{1}{|m_n|} \sum_{t=1}^{|m_n|} (S_{n,t} - \mu_n)^2} \right)}_{\text{Intra-consistency}} \cdot \underbrace{\frac{\exp(\mu_n/\tau)}{\sum_{j=1}^N \exp(\mu_j/\tau)}}_{\text{Inter-consistency}}$$

*Controlling sensitivity*     *Frame-query sim.*

*Avg sim.*

Intra-consistency: high when scores within the moment are consistent (low variance).

# Method

- The inter-consistency score addresses this by measuring how discriminatively the query matches its own moment relative to others, computed as a softmax over cross-moment similarities with temperature  $\tau$ .

$$R_{\text{consistency}} = \frac{1}{N} \sum_{n=1}^N \underbrace{\exp \left( -\gamma \sqrt{\frac{1}{|m_n|} \sum_{t=1}^{|m_n|} (S_{n,t} - \mu_n)^2} \right)}_{\text{Intra-consistency}} \cdot \underbrace{\frac{\exp(\mu_n/\tau)}{\sum_{j=1}^N \exp(\mu_j/\tau)}}_{\text{Inter-consistency}}$$

Inter-consistency: high when the moment is more relevant than other candidate moments.

# Method

- The feedback reward confirms whether the pair is solvable and is measured by the solver's accuracy.

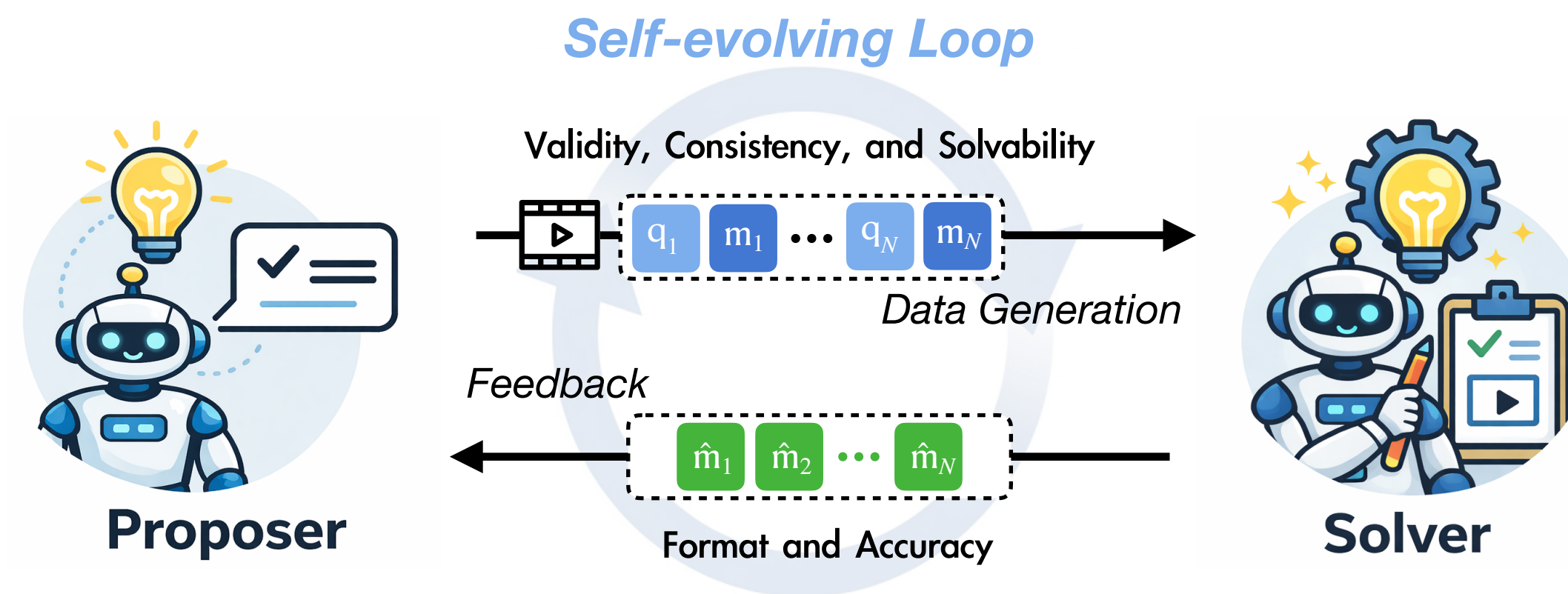
$$R_{\text{feedback}}(\mathbf{s}_i) = \frac{1}{N} \sum_{n=1}^N \text{tIoU}(m_n, \hat{m}_n).$$

Generated moment

Predicted moment

# Method

- For the solver, we keep the design simple following the previous RL-based works.

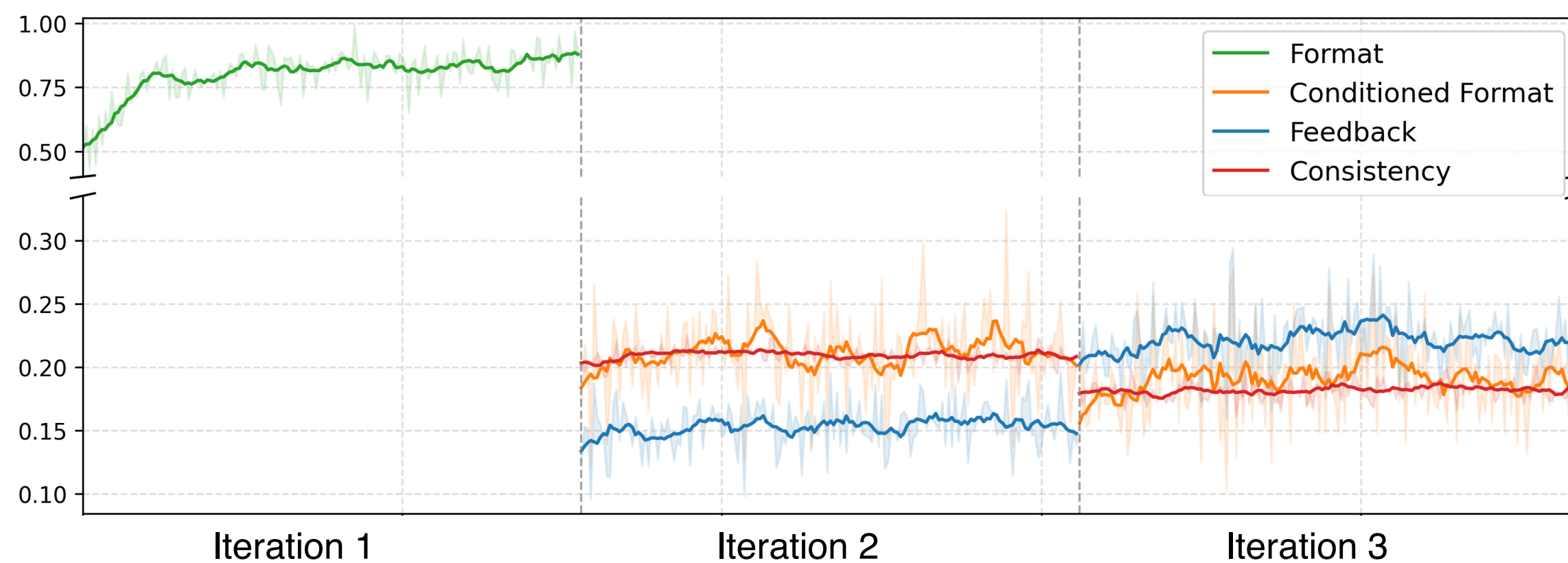


Solver's output :  $\underbrace{\langle think \rangle \dots \text{The model's intermediate reasoning} \dots \langle /think \rangle}_{\text{format}} \underbrace{\langle answer \rangle \hat{s} \text{ to } \hat{e} \text{ seconds} \dots \langle /answer \rangle}_{\text{accuracy}}$

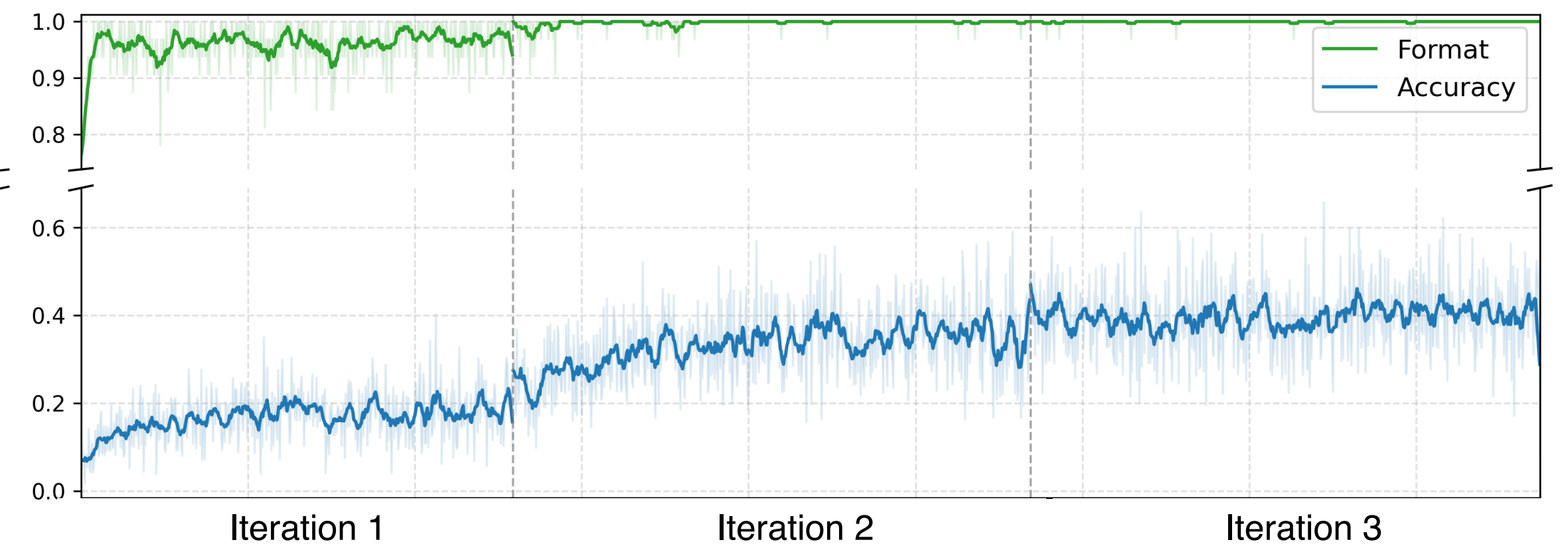
# Method

- In the first iteration, we only apply the format reward to the proposer and update the solver, since neither agent is yet tailored to the task.
- In subsequent iterations, we apply the consistent and feedback rewards to the proposer and progressively increase  $\delta$ .

(a) Proposer



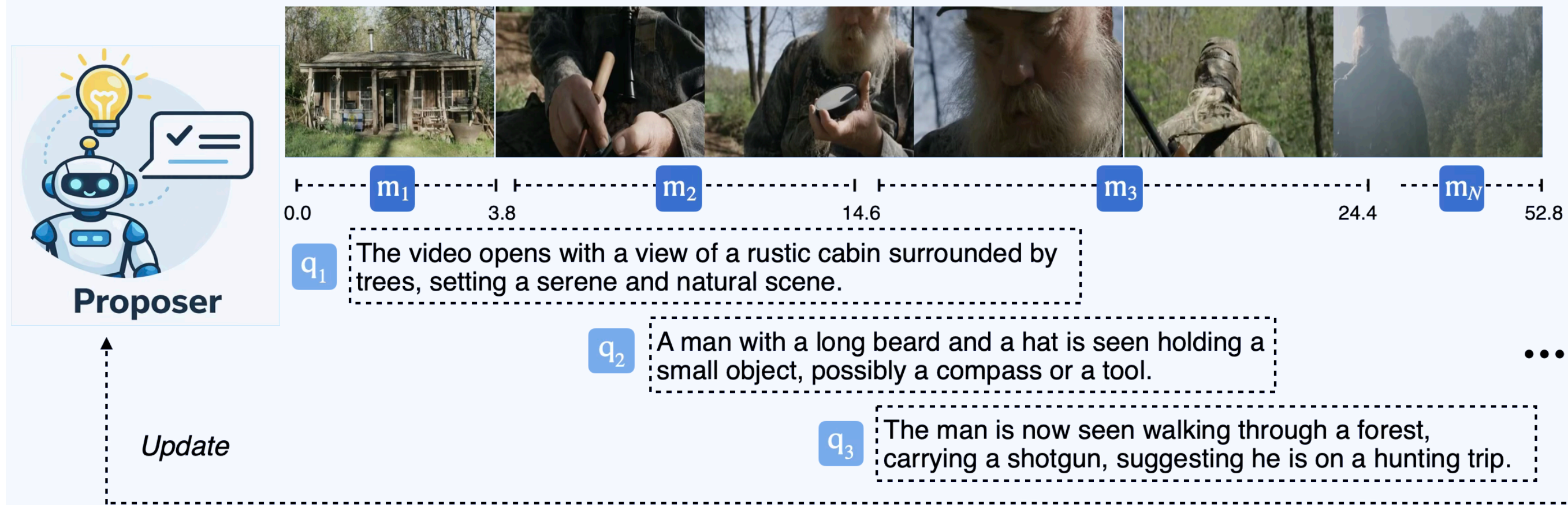
(b) Solver



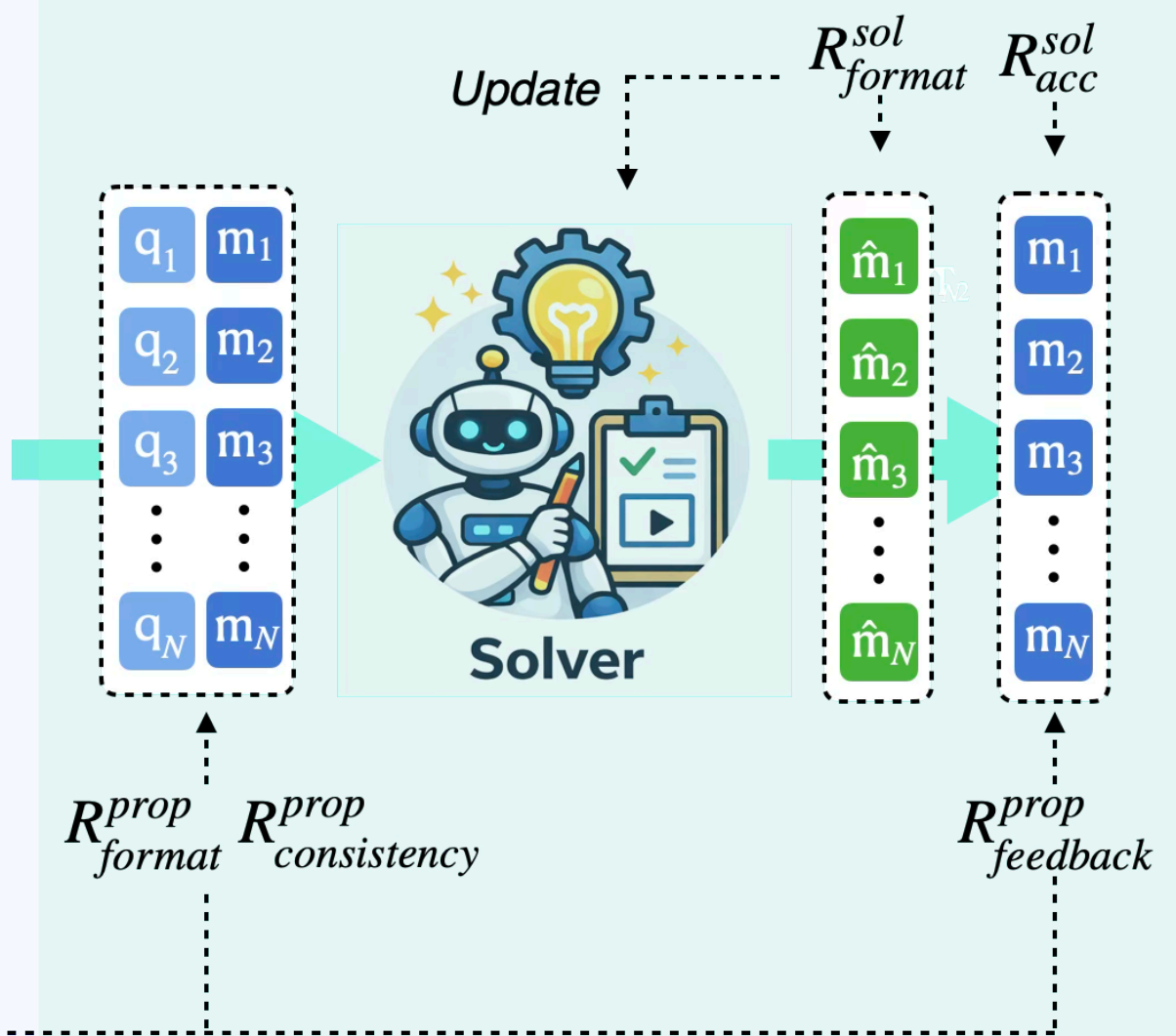
# Method

## Proposer Stage

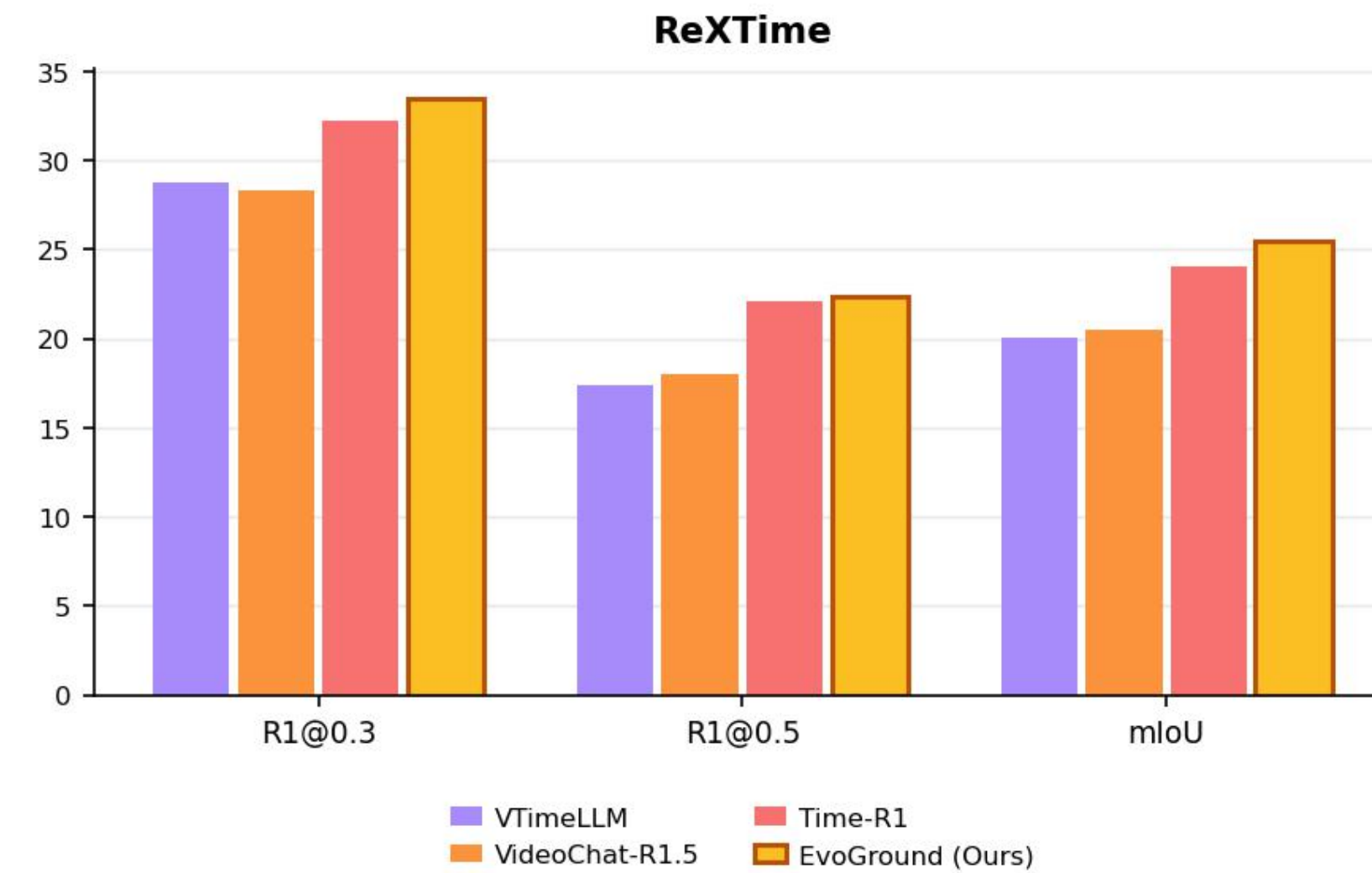
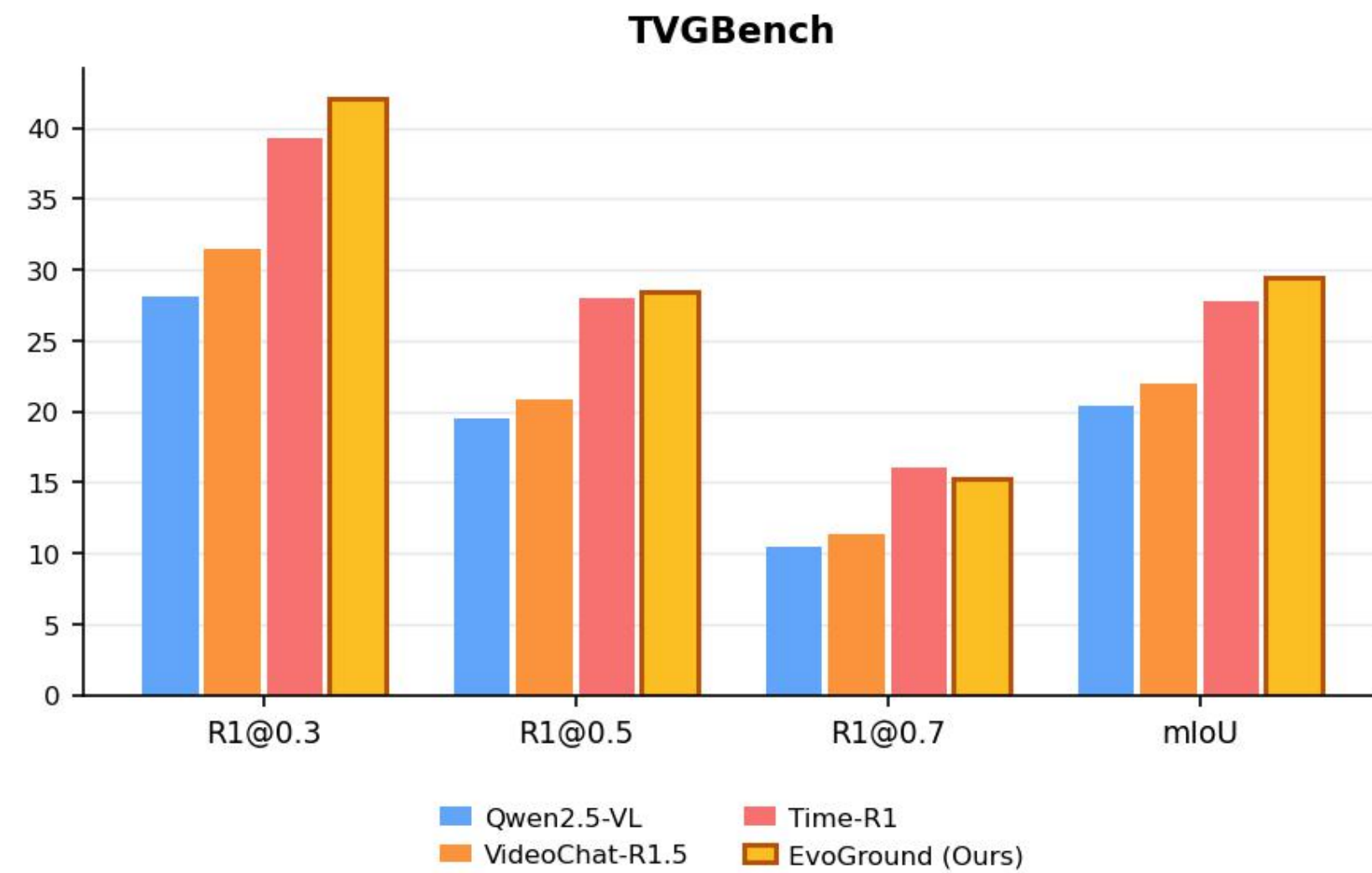
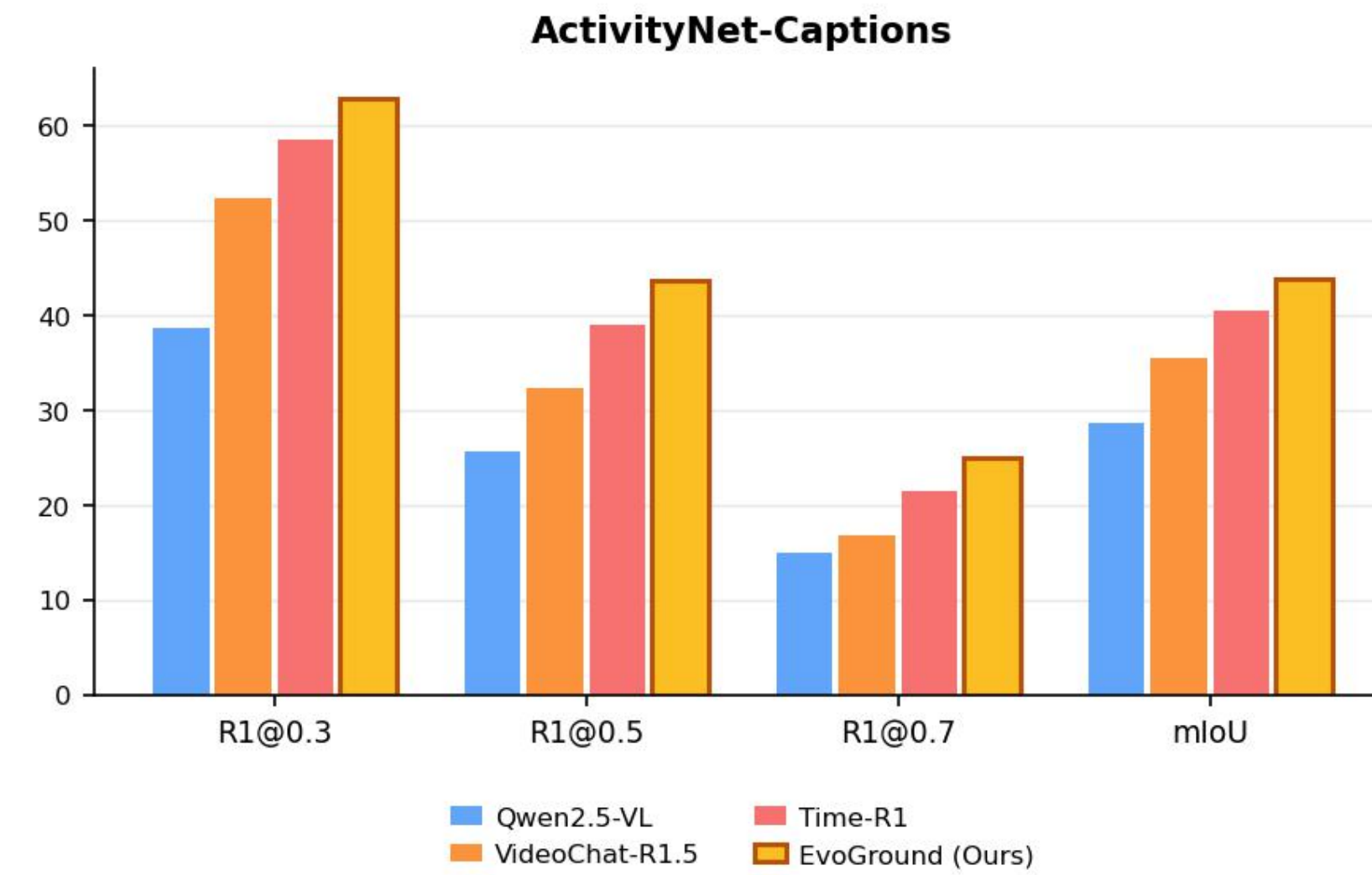
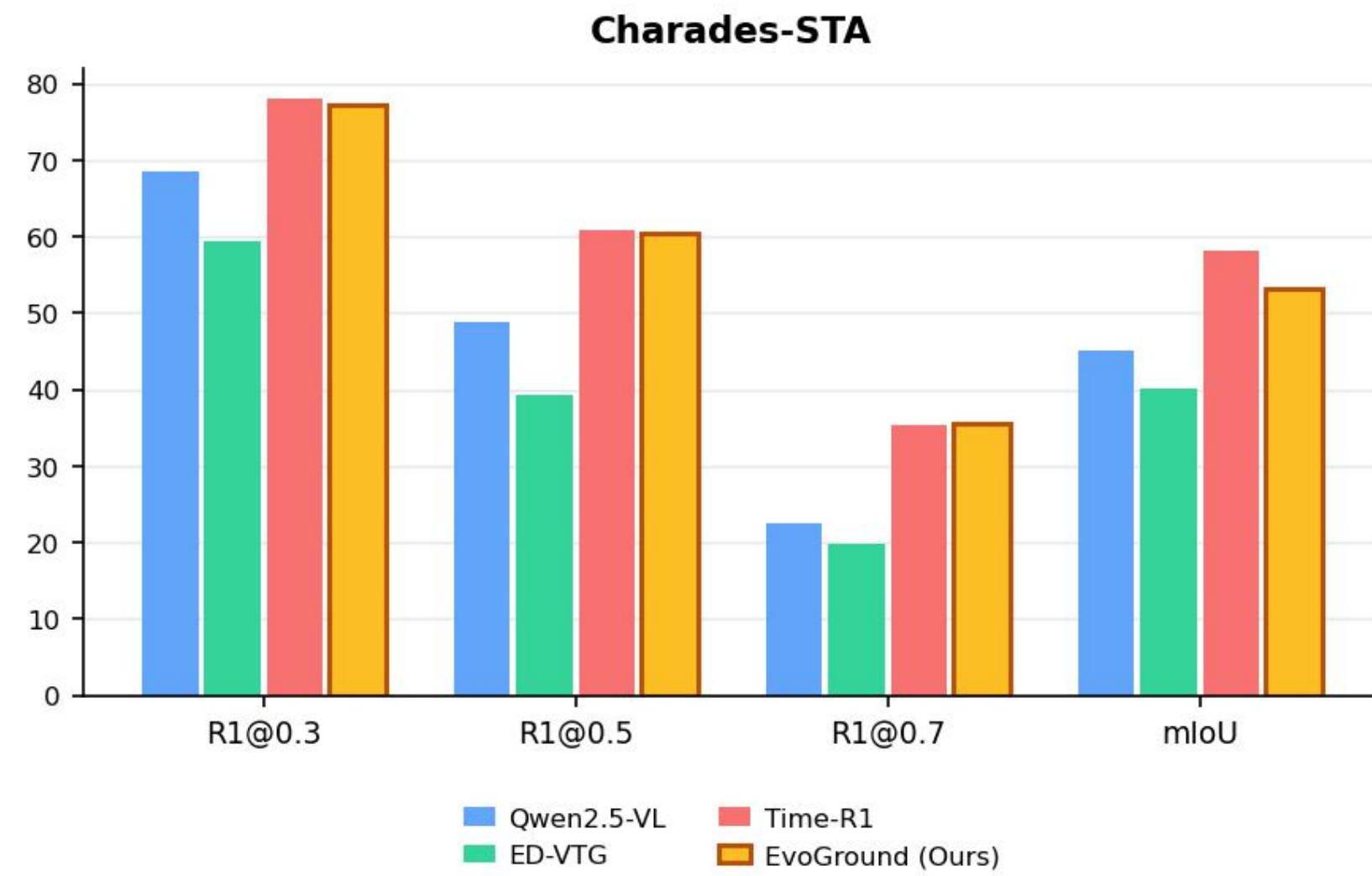
You are an expert video analyst. You are given a video with a total duration of 52.8 seconds. Your task is to identify distinct, non-overlapping temporal segments and provide a clear description of what happens in each segment.



## Solver Stage



# Experiments



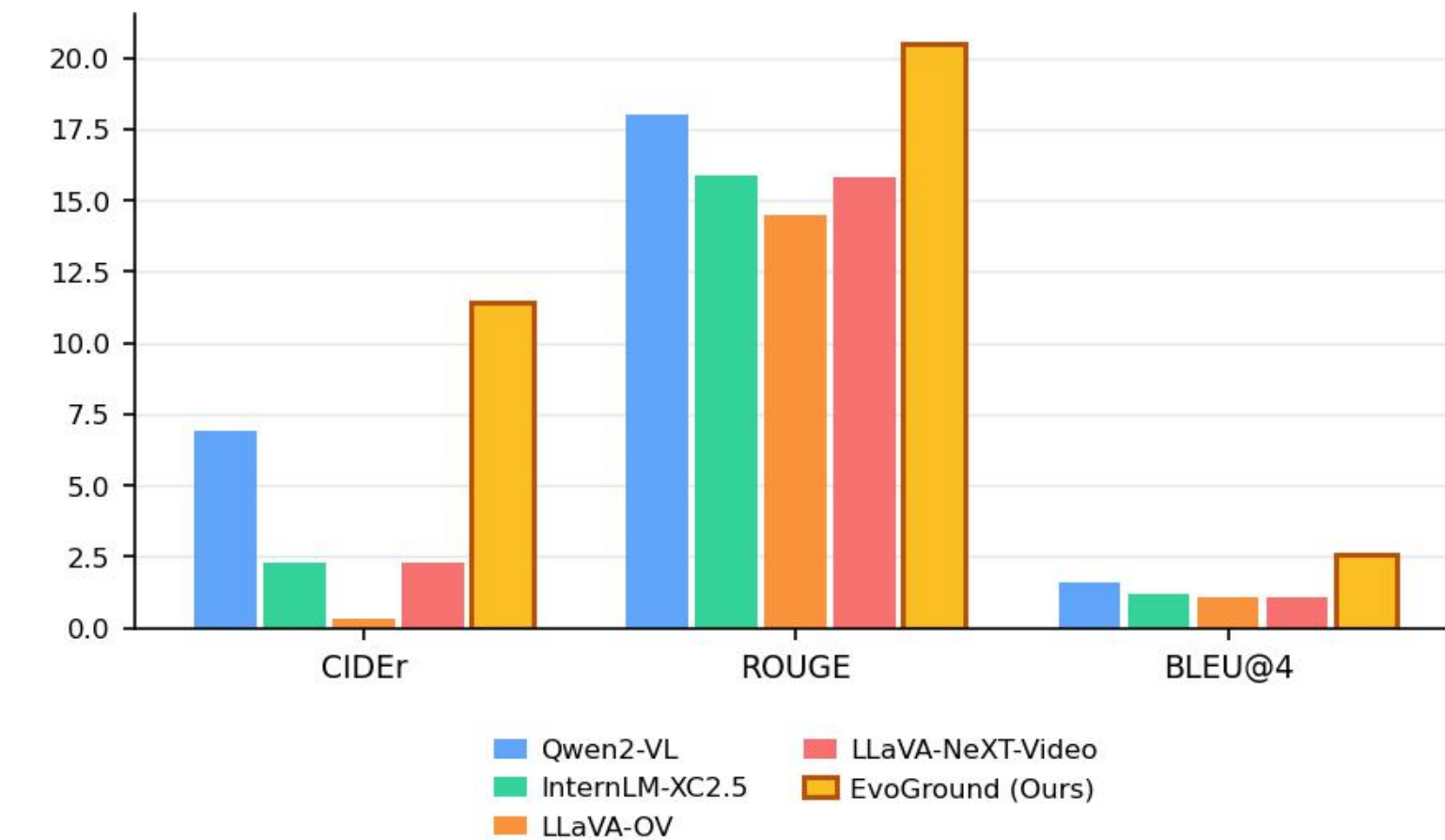
# Experiments



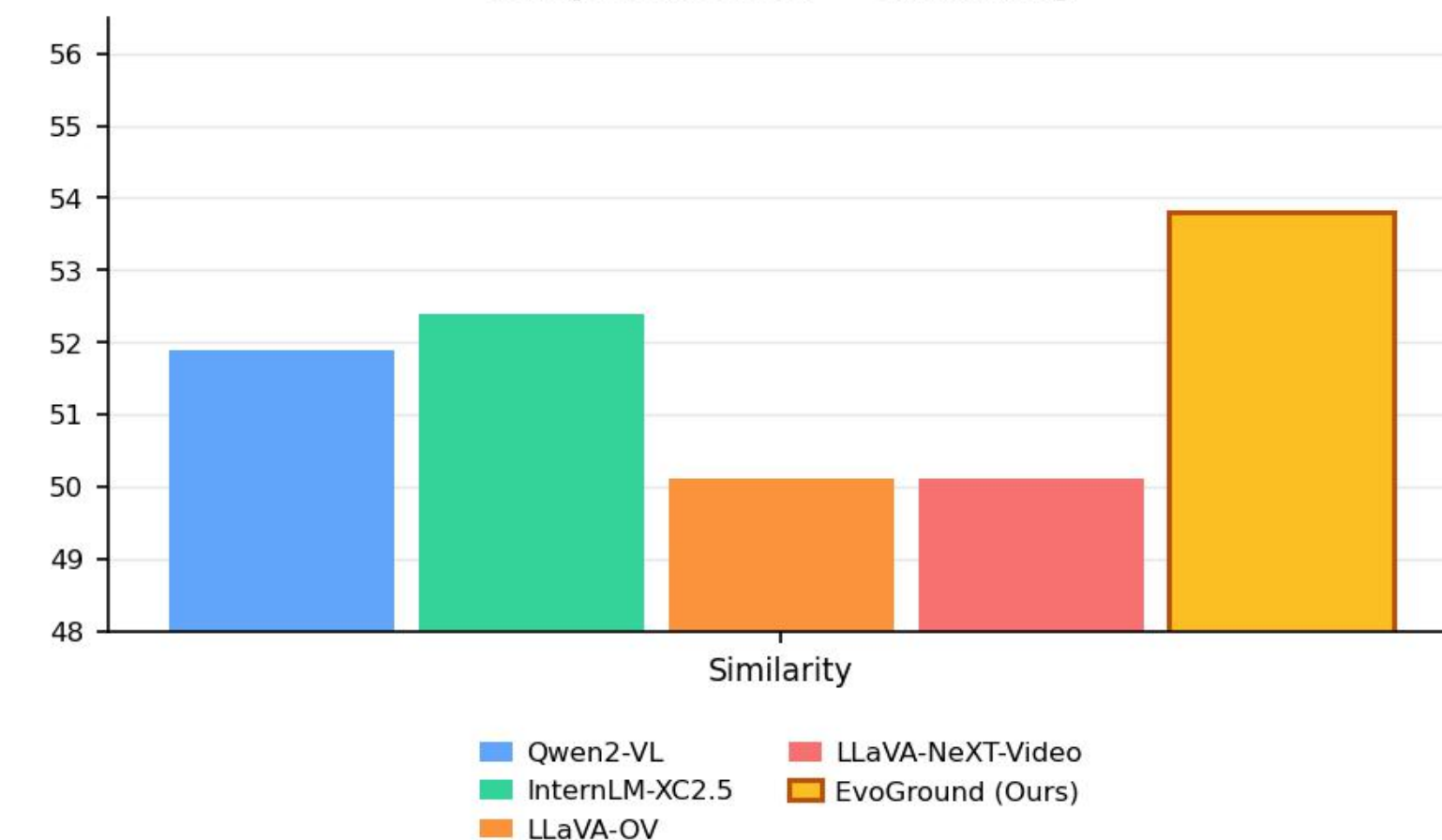
**Ground-truth:** Women dressed in blue uniforms and hats walk in two lines while holding hands. They are smiling and waving at the crowd of people who are cheering them on by clapping their hands on the side. A group of men wearing heavy dark jackets and hat walk through while shaking hands and smiling at the crowd.

**EvoGround:** Students in robes walk through a grand hall, greeted by applause from others. The group of students continues walking, with some students in uniform hats and robes. The students pass by a group of adults in robes, who are clapping and smiling. The students continue walking, with the crowd of students and adults still applauding and smiling.

TemporalBench — CIDEr / ROUGE / BLEU@4

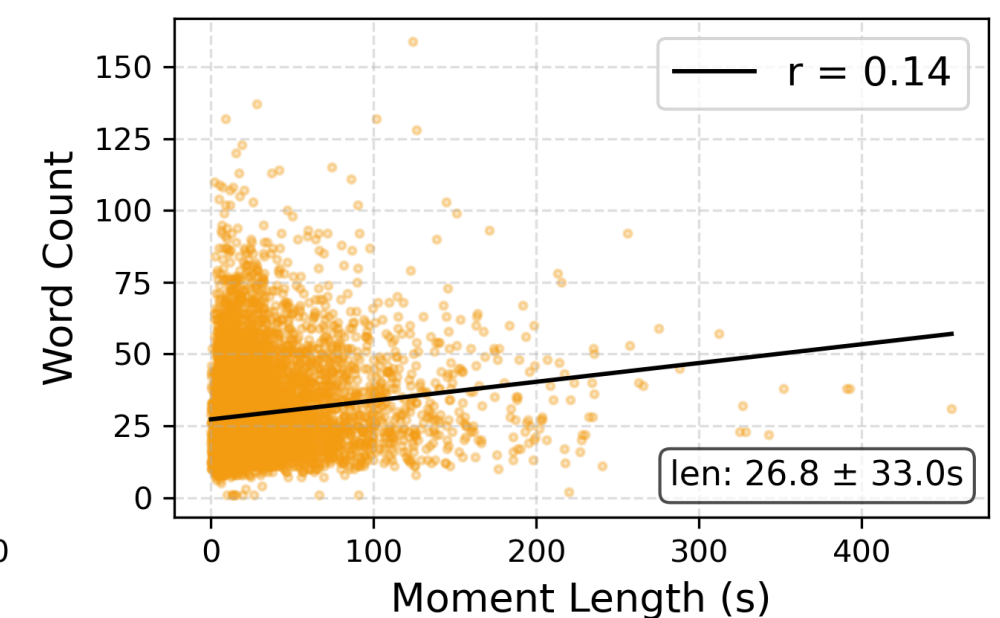
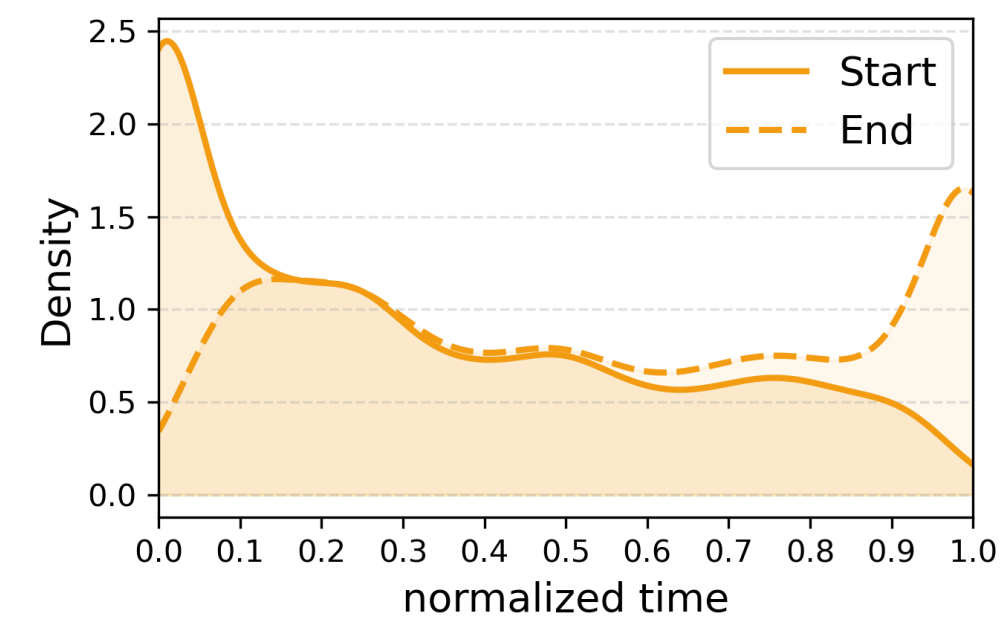
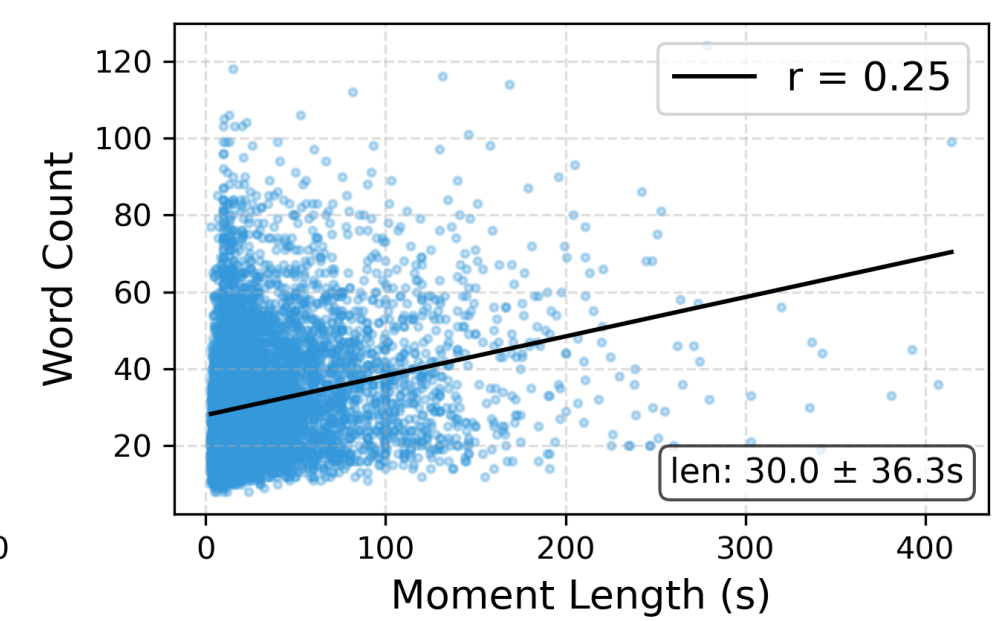
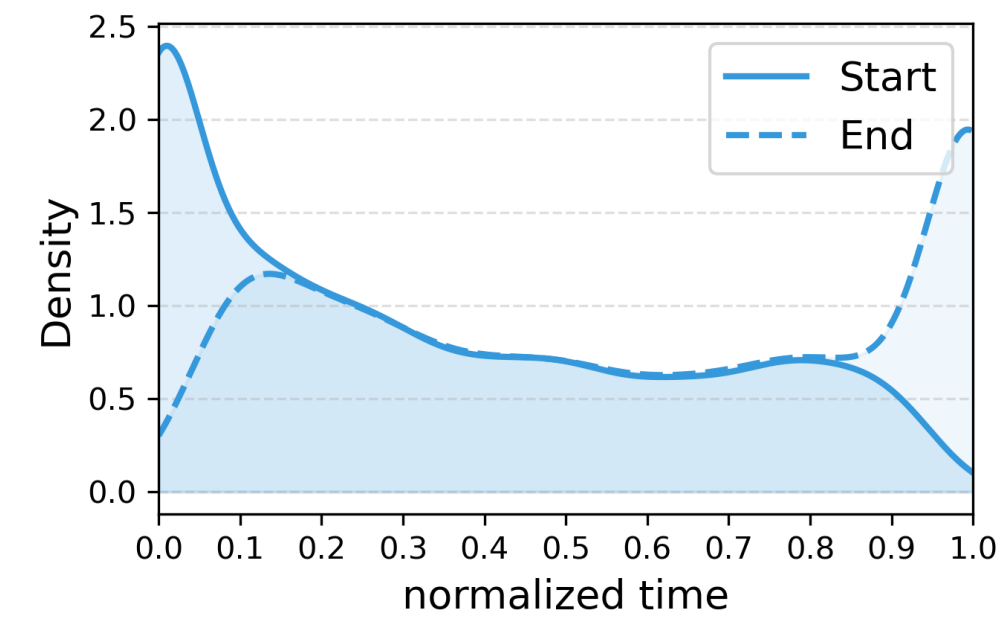
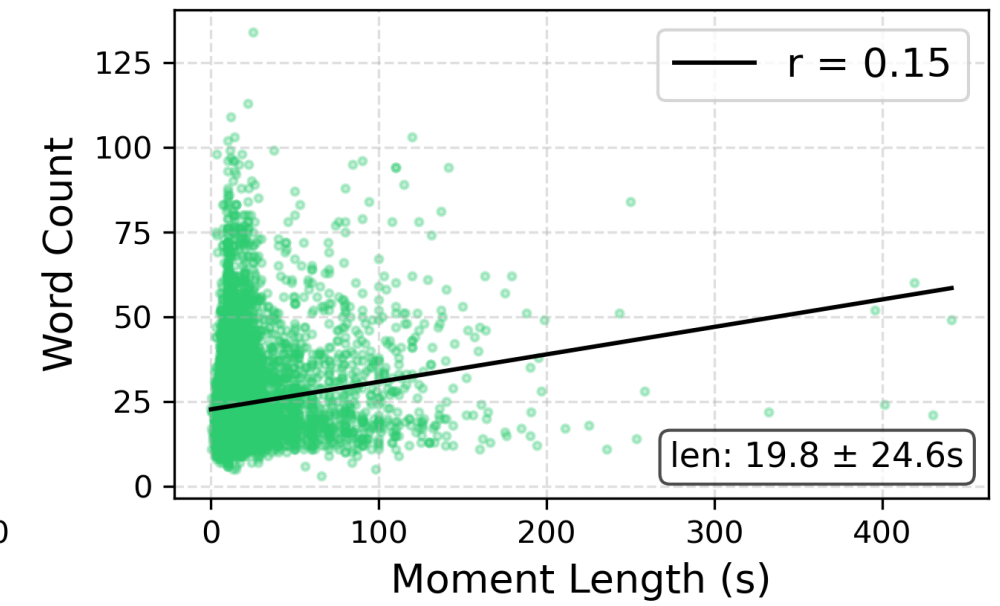
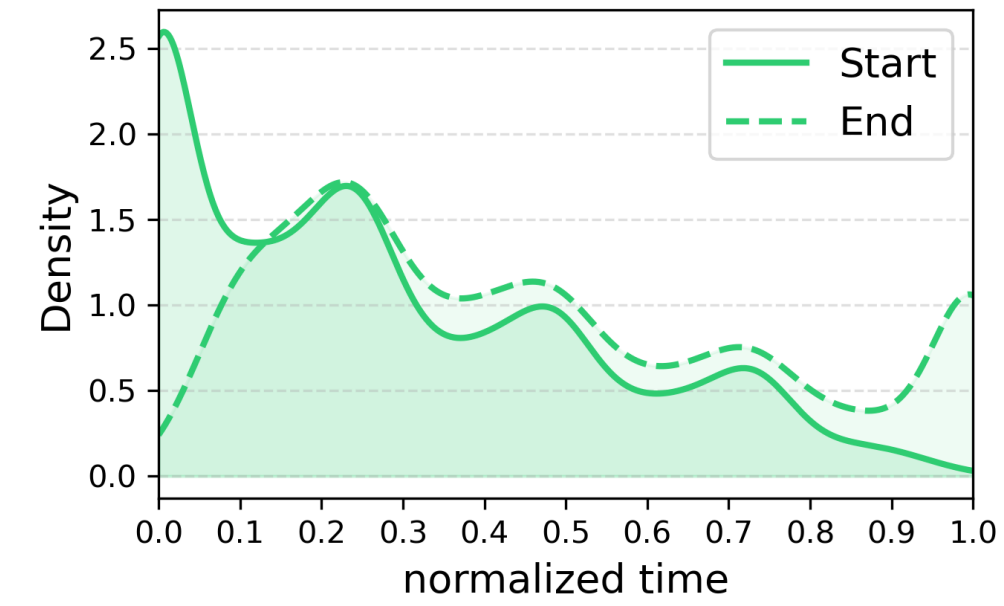
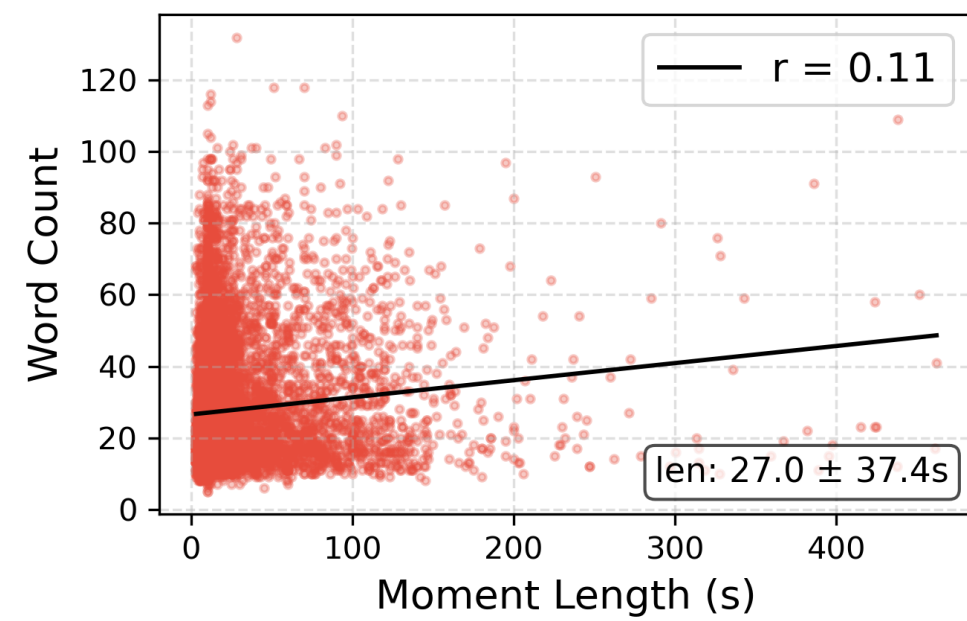
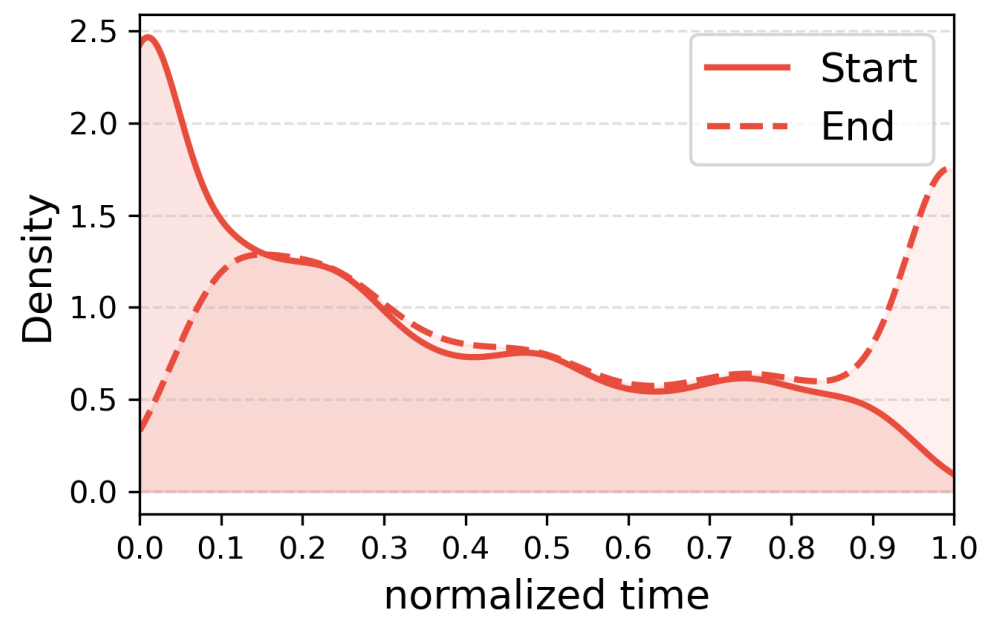


TemporalBench — Similarity



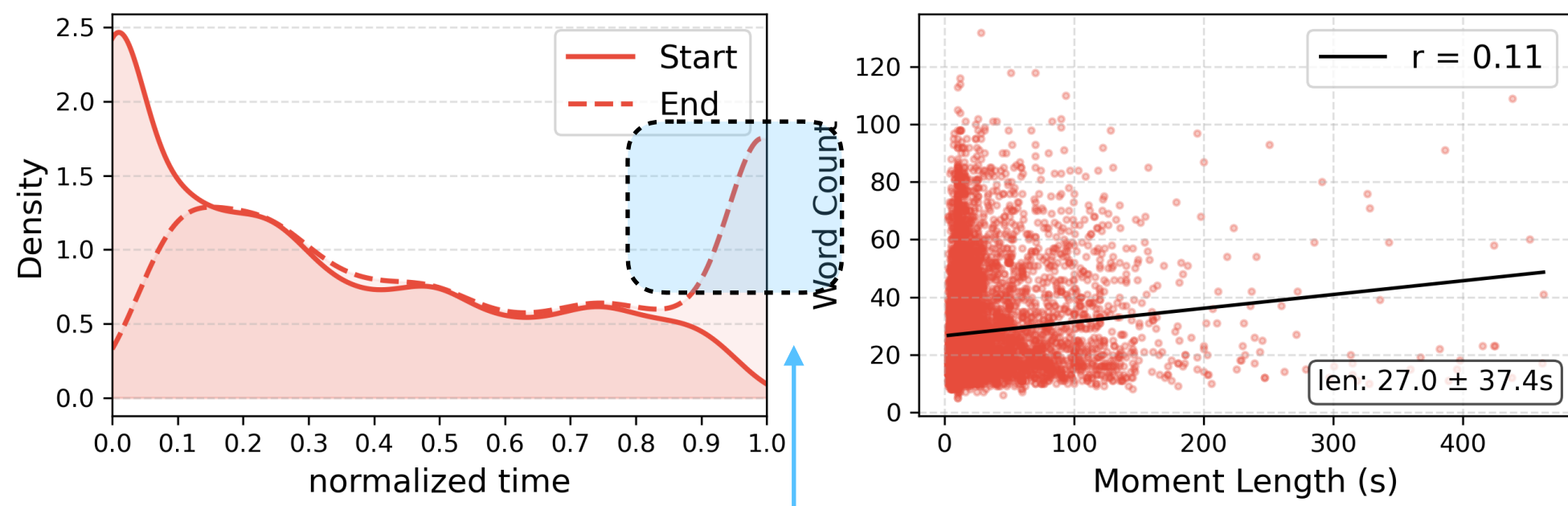
# Experiments

## Format Only



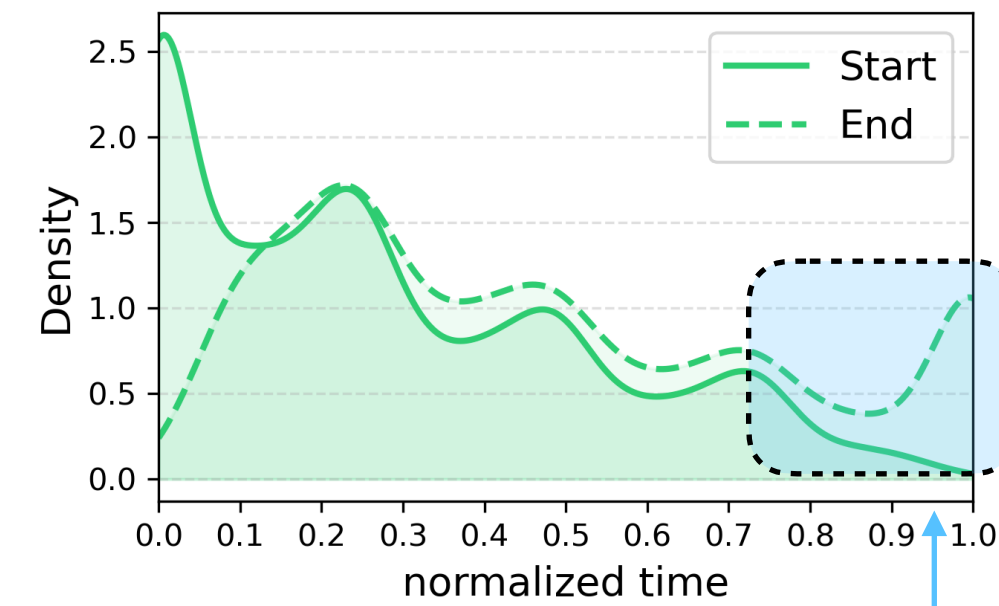
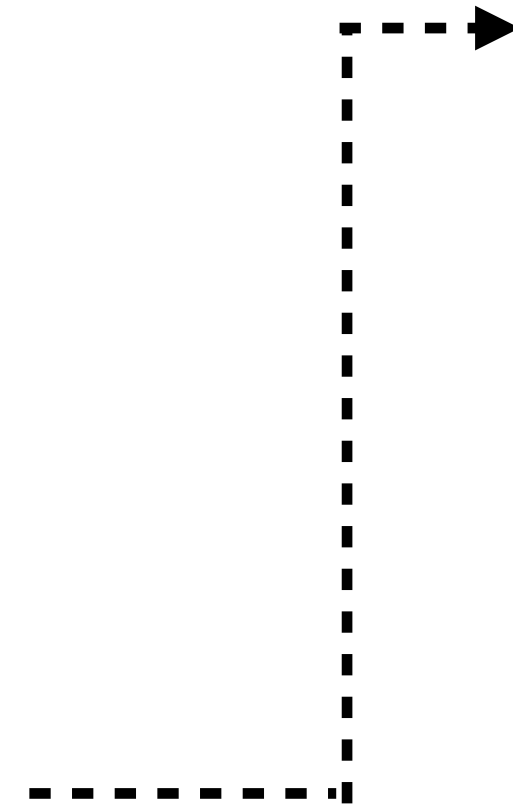
# Experiments

*Format Only*

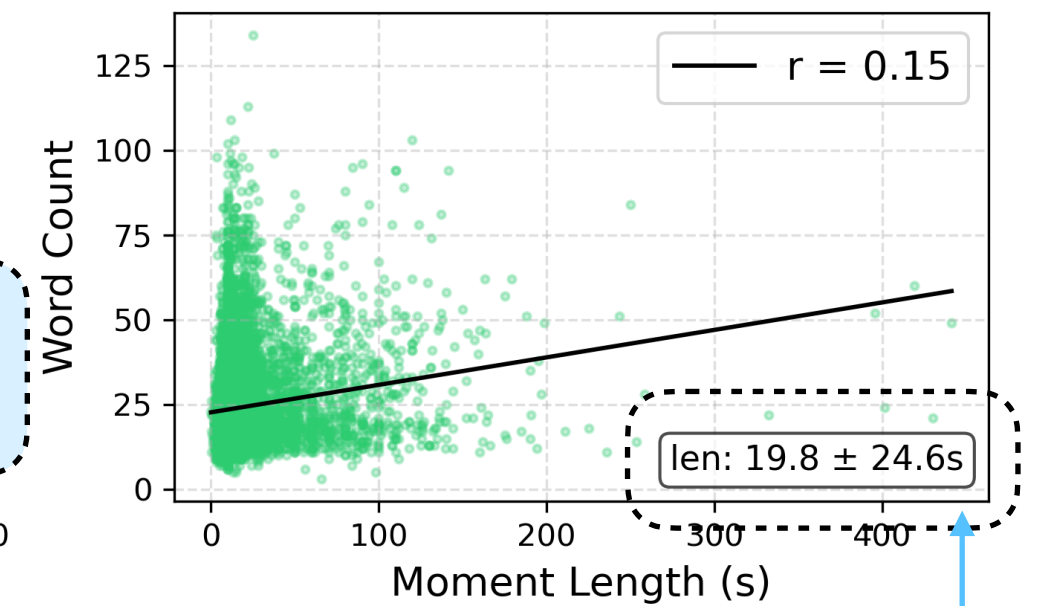


*Concentrated!*

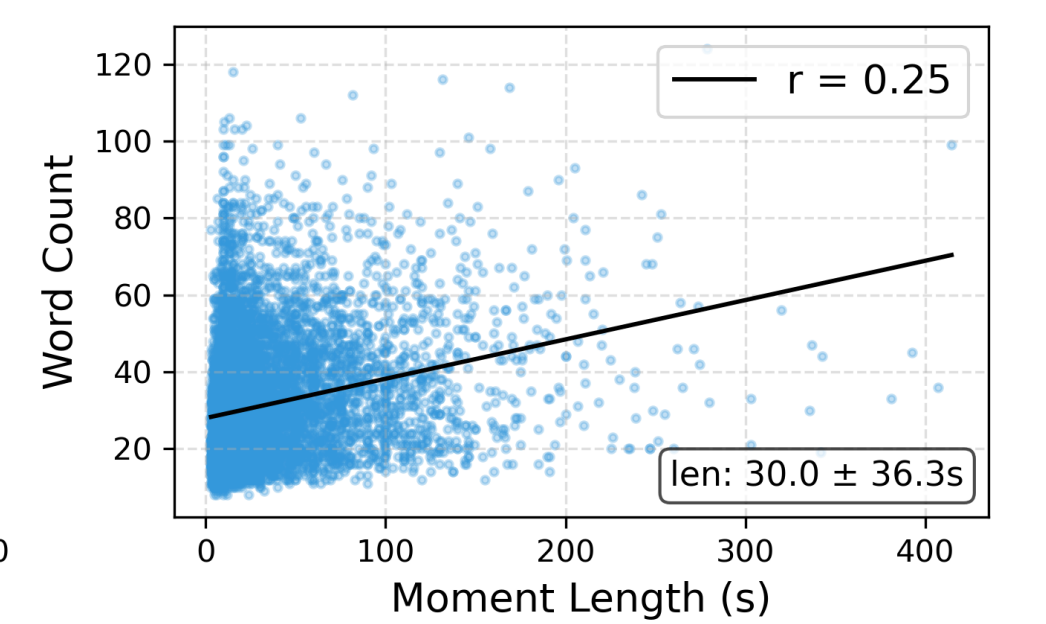
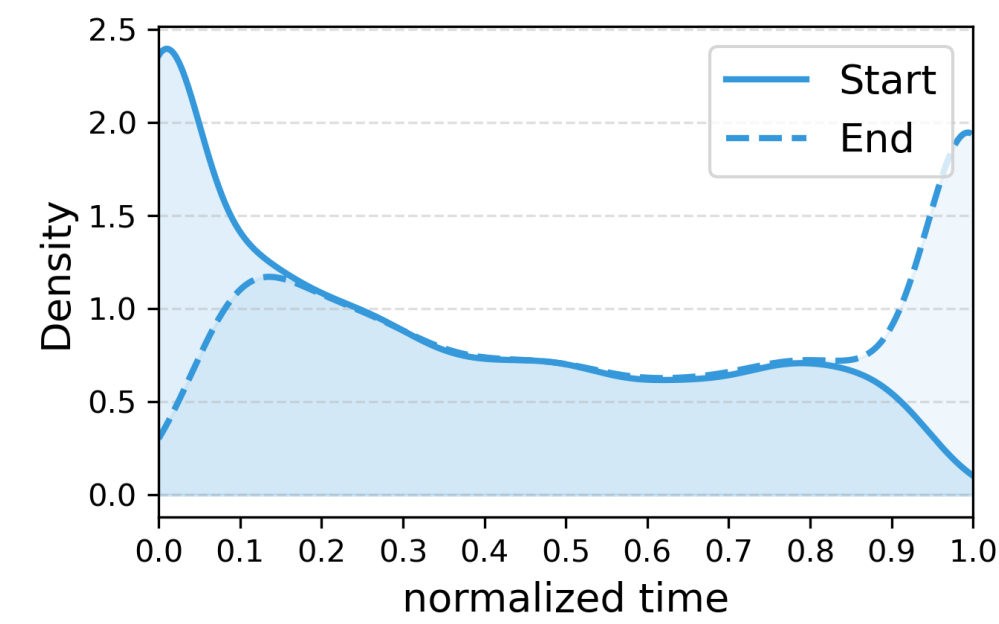
*Consistency*



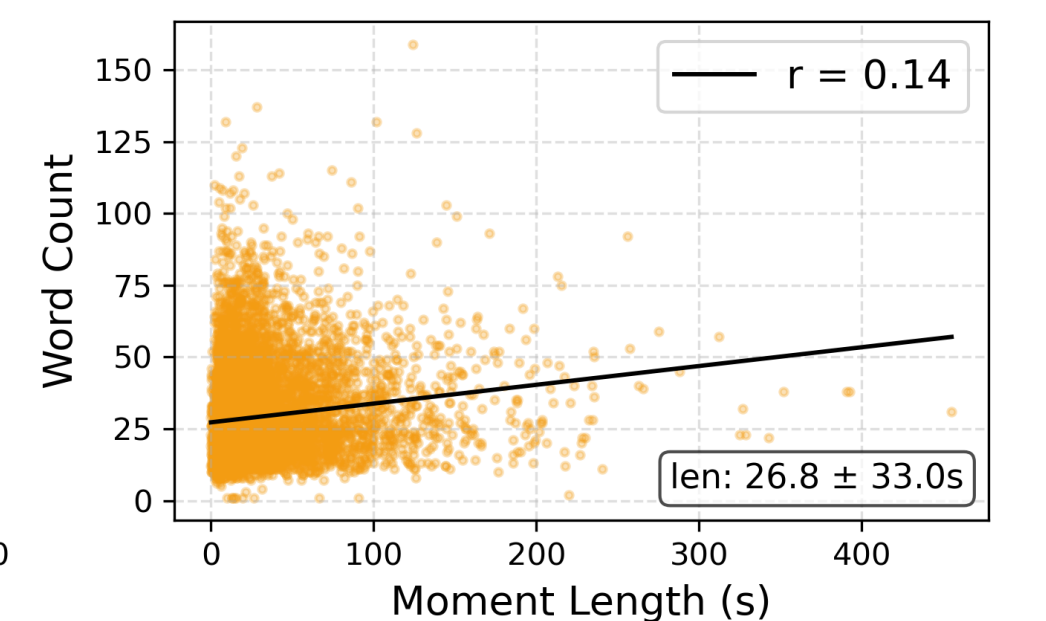
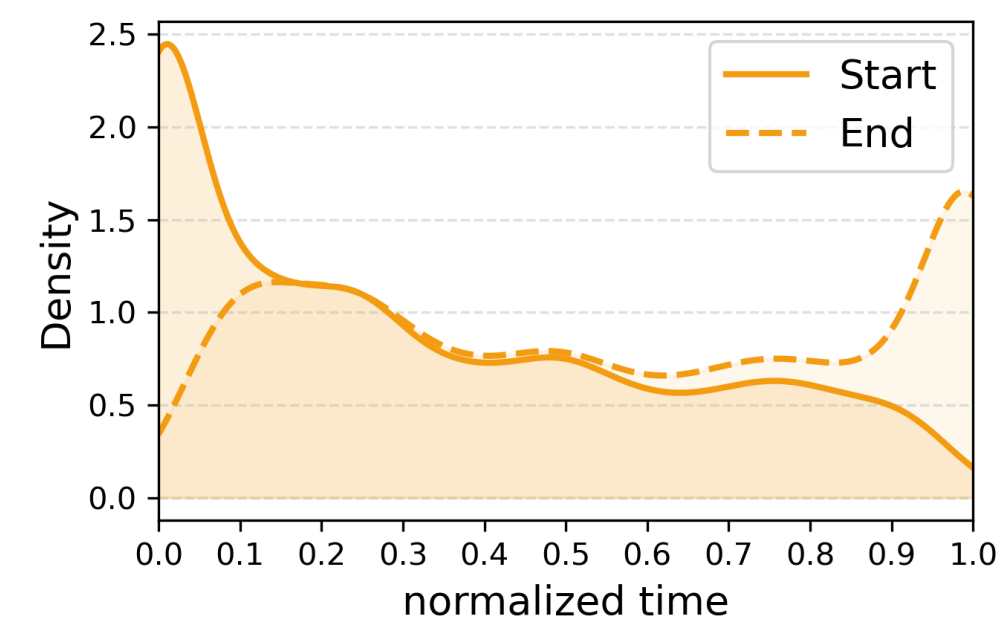
*Sparse!*



*Tight!*

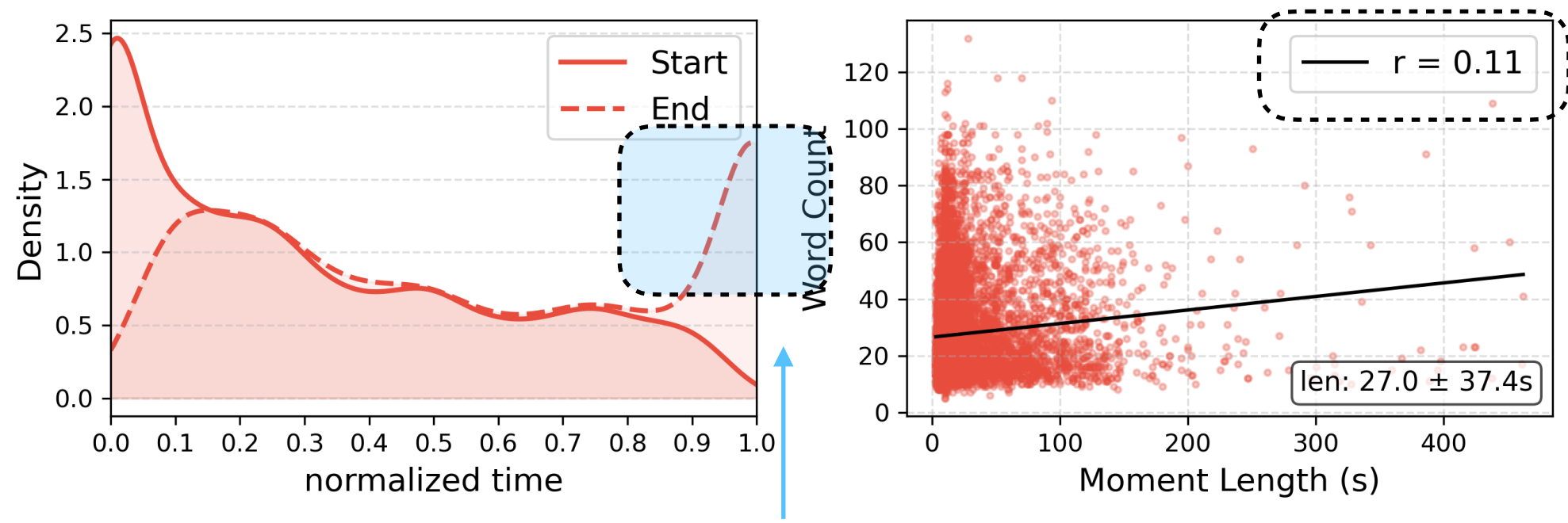


*The consistency reward sharpens the moment distribution!*



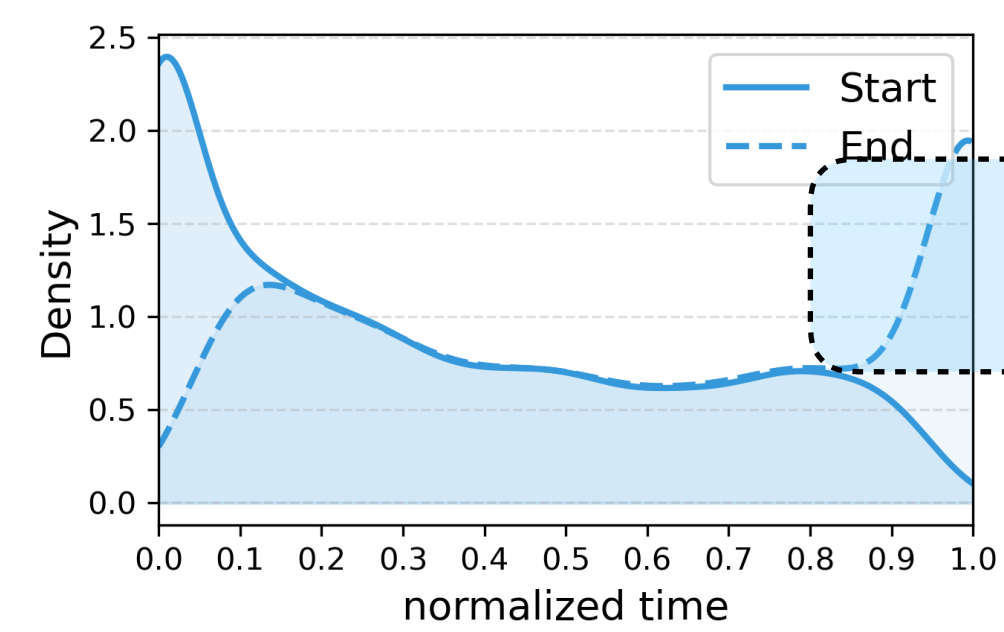
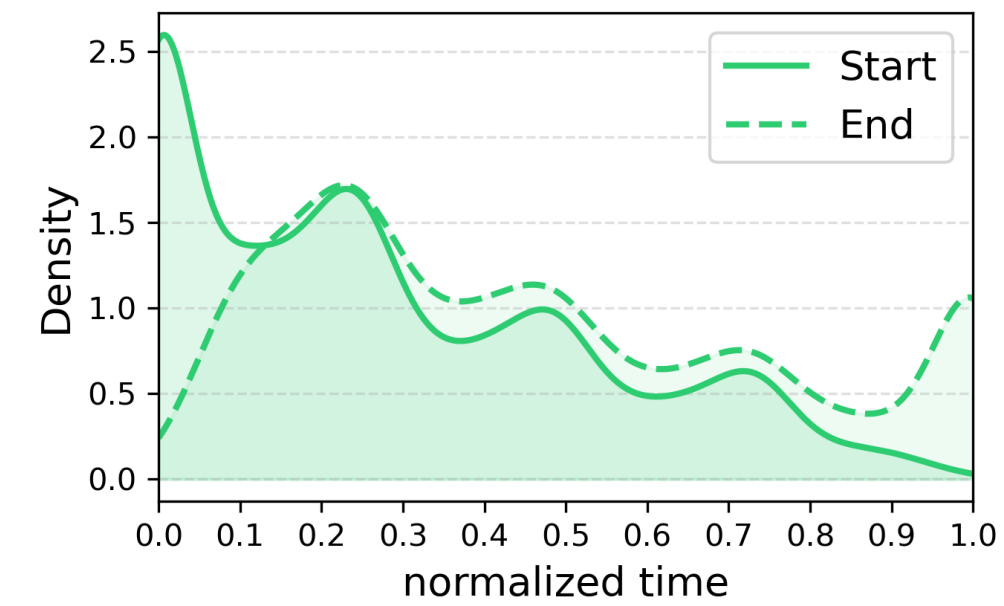
# Experiments

*Format Only*



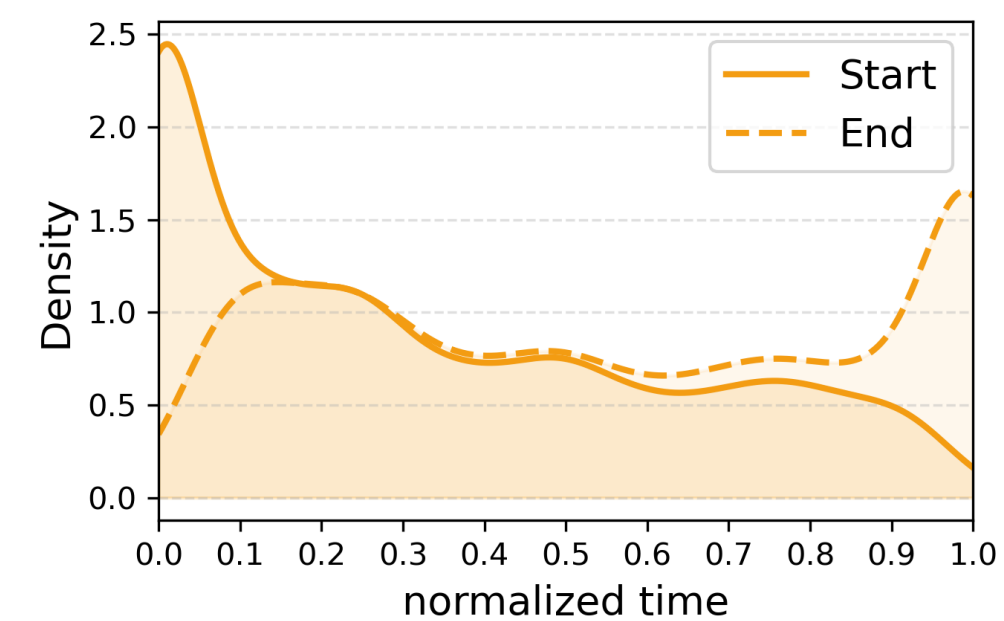
*Concentrated!*

*Feedback*



*Concentrated!*

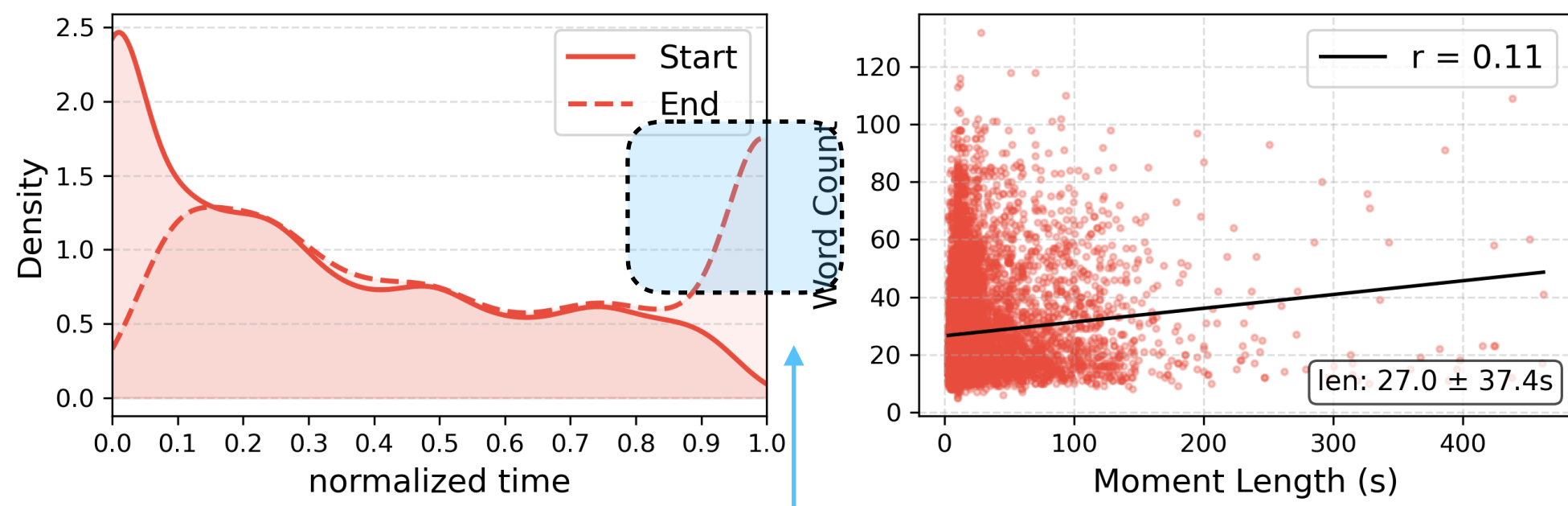
*Lengthy!*



*The feedback reward encourages the proposer to generate more descriptive queries!*

# Experiments

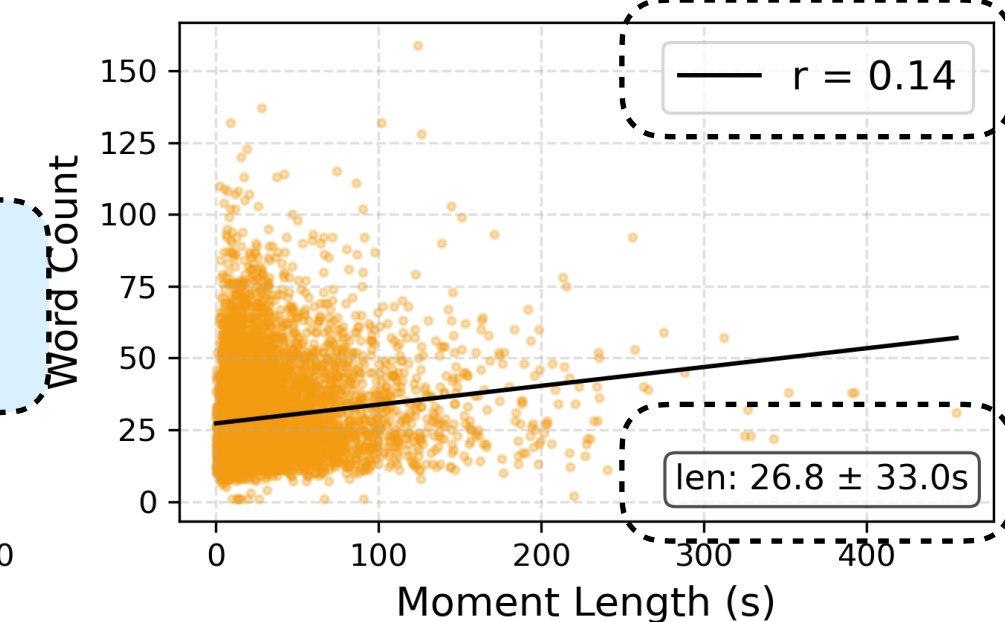
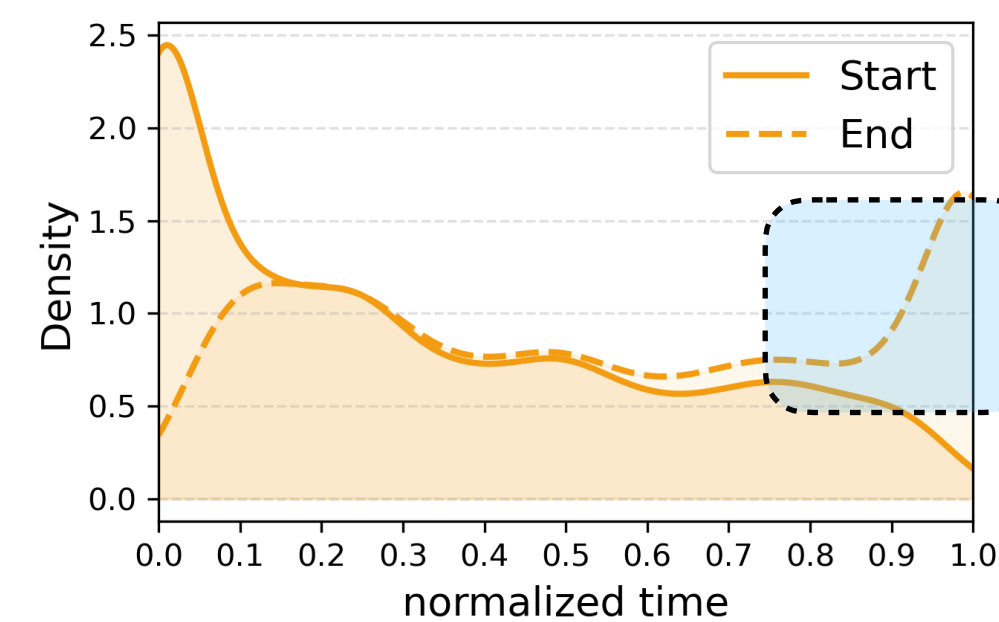
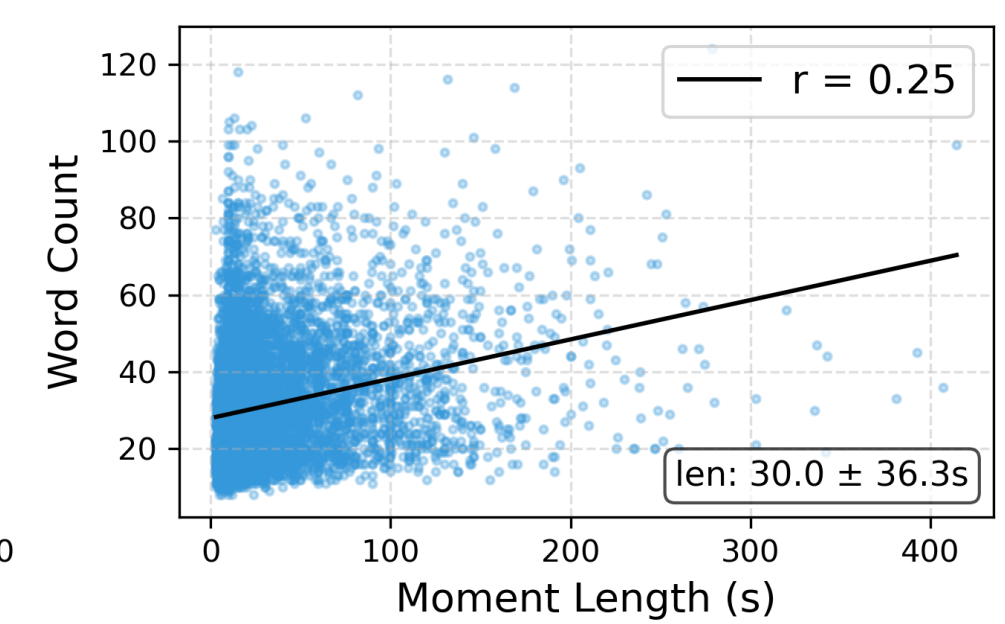
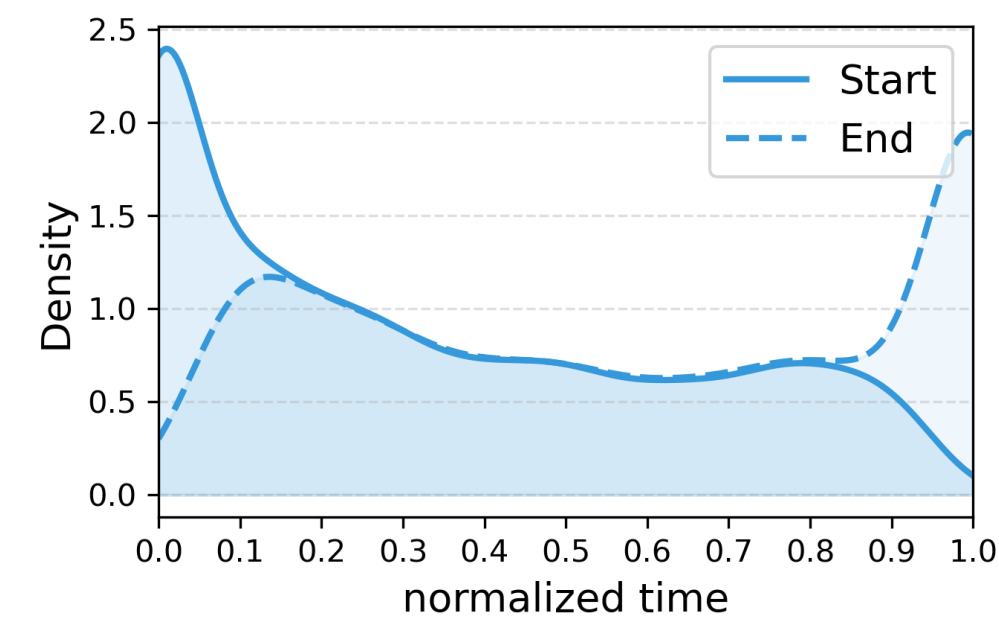
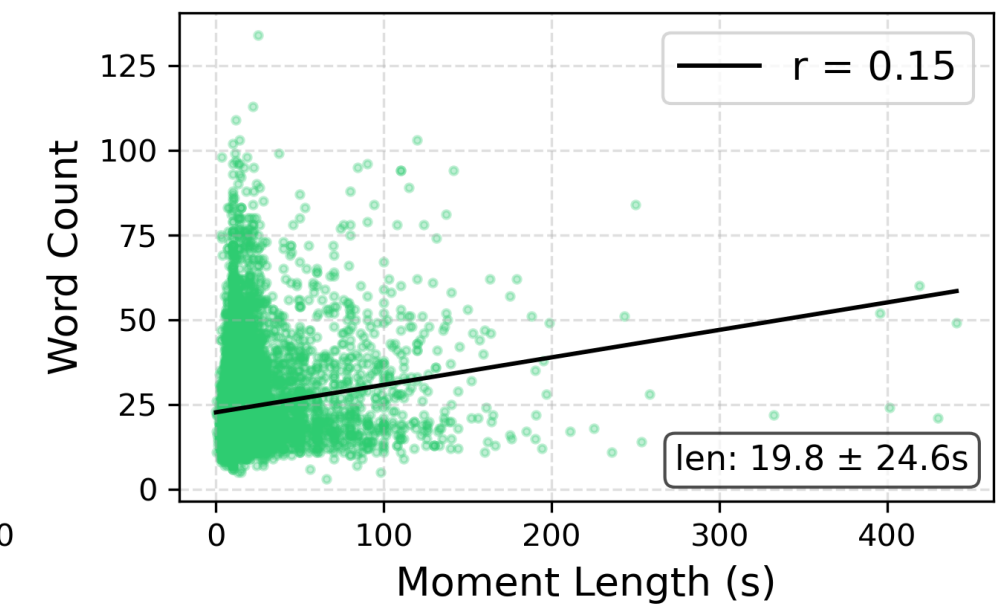
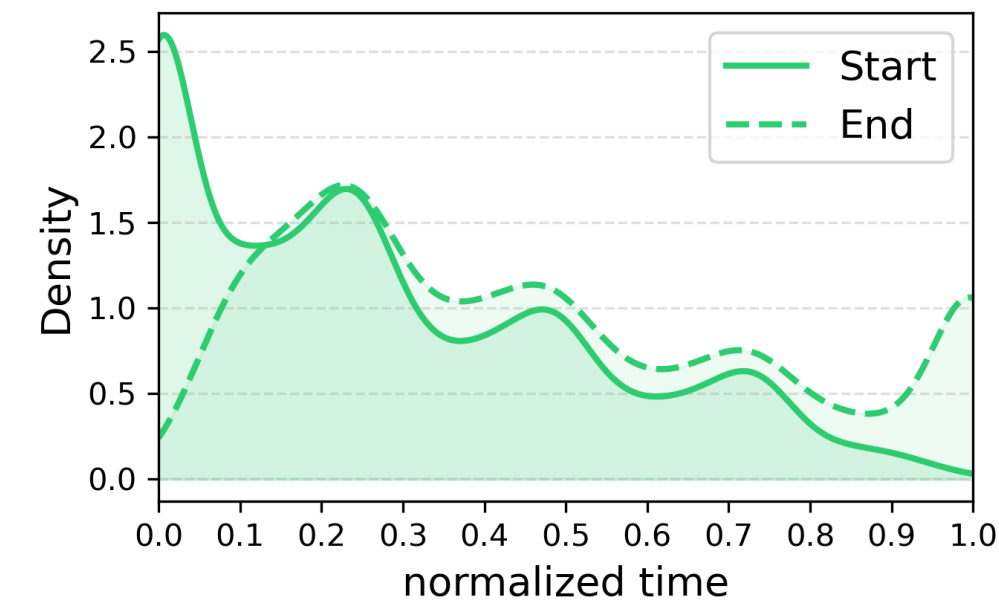
*Format Only*



*Concentrated!*

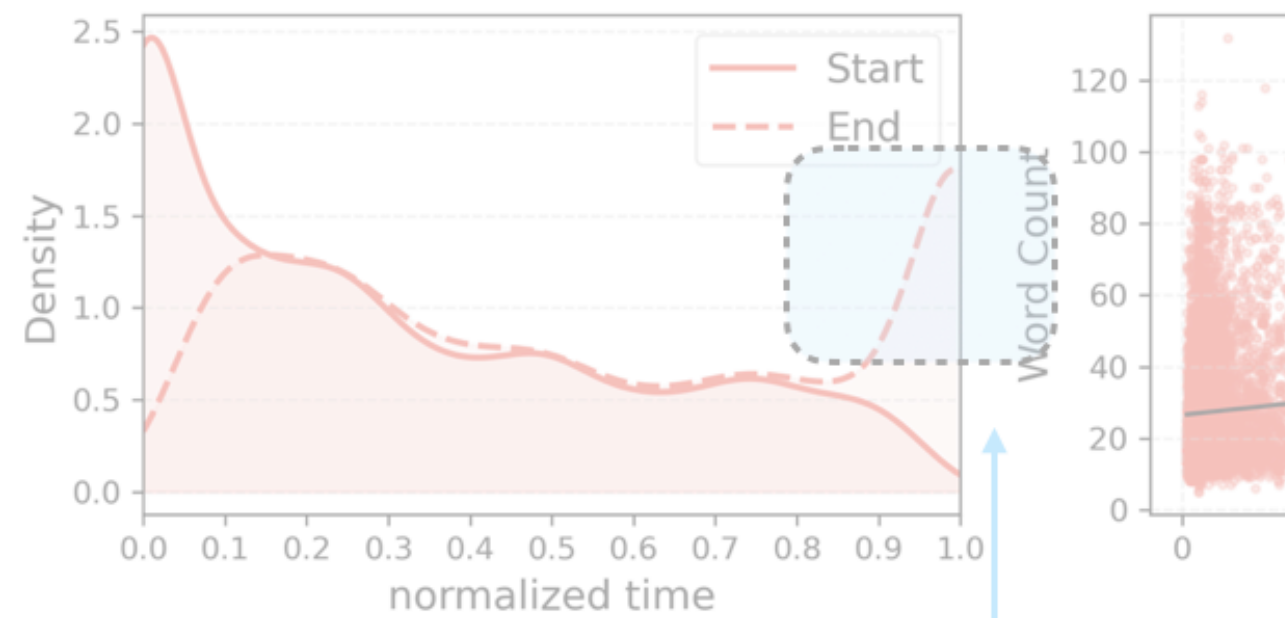
*The consistency reward sharpens the moment distribution!*

*All*



# Experiments

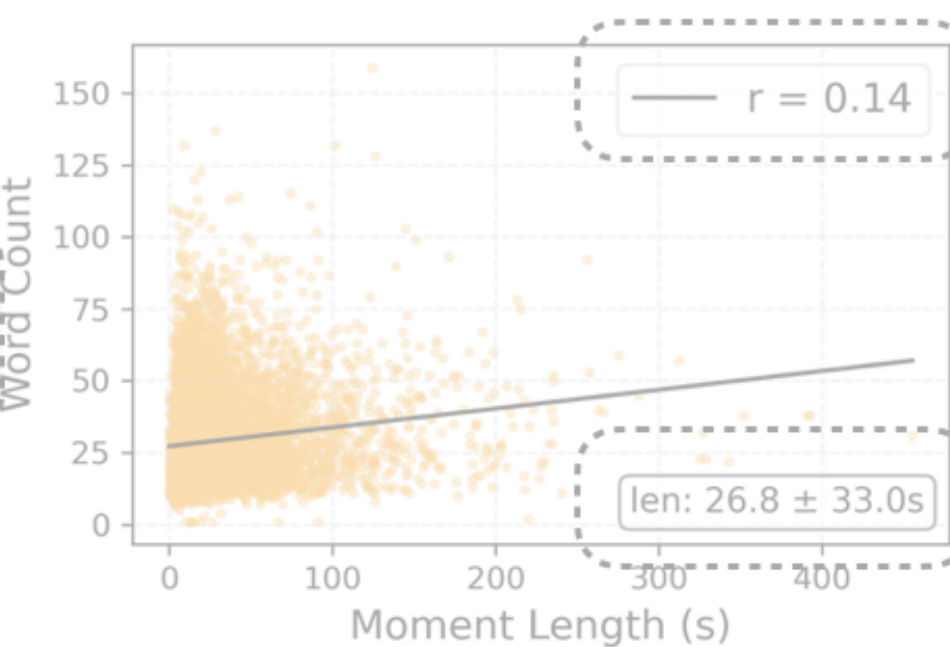
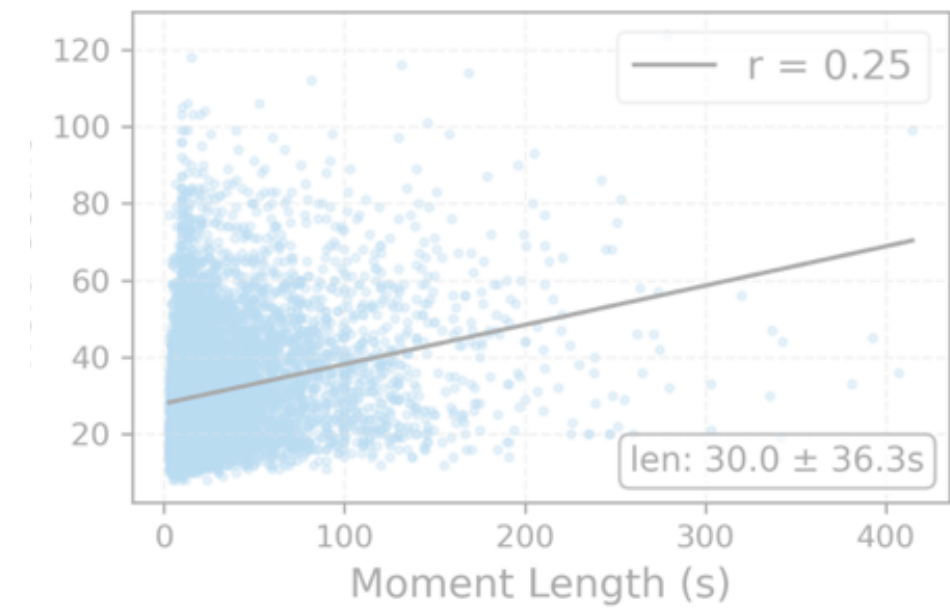
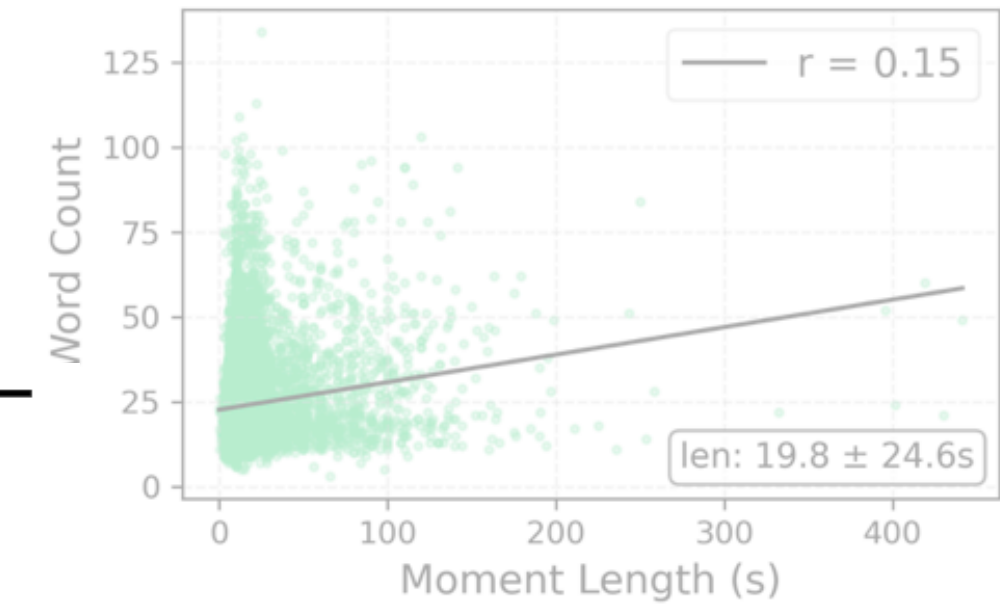
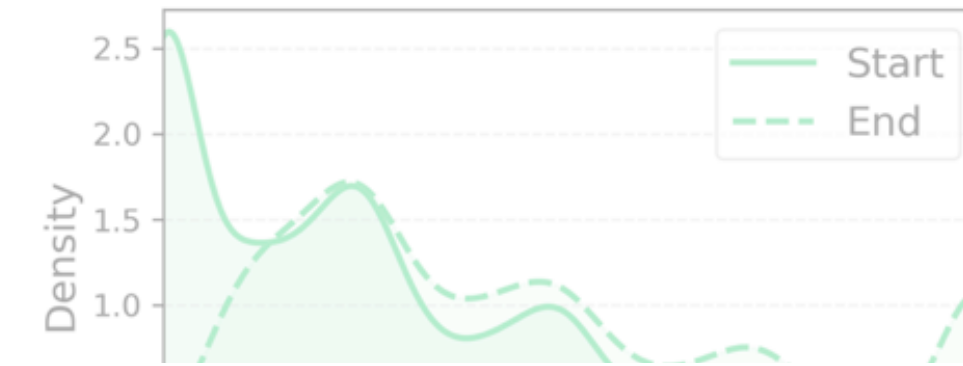
*Format Only*



*Concentrated!*

*The consistency reward sharpens the moment distribution!*

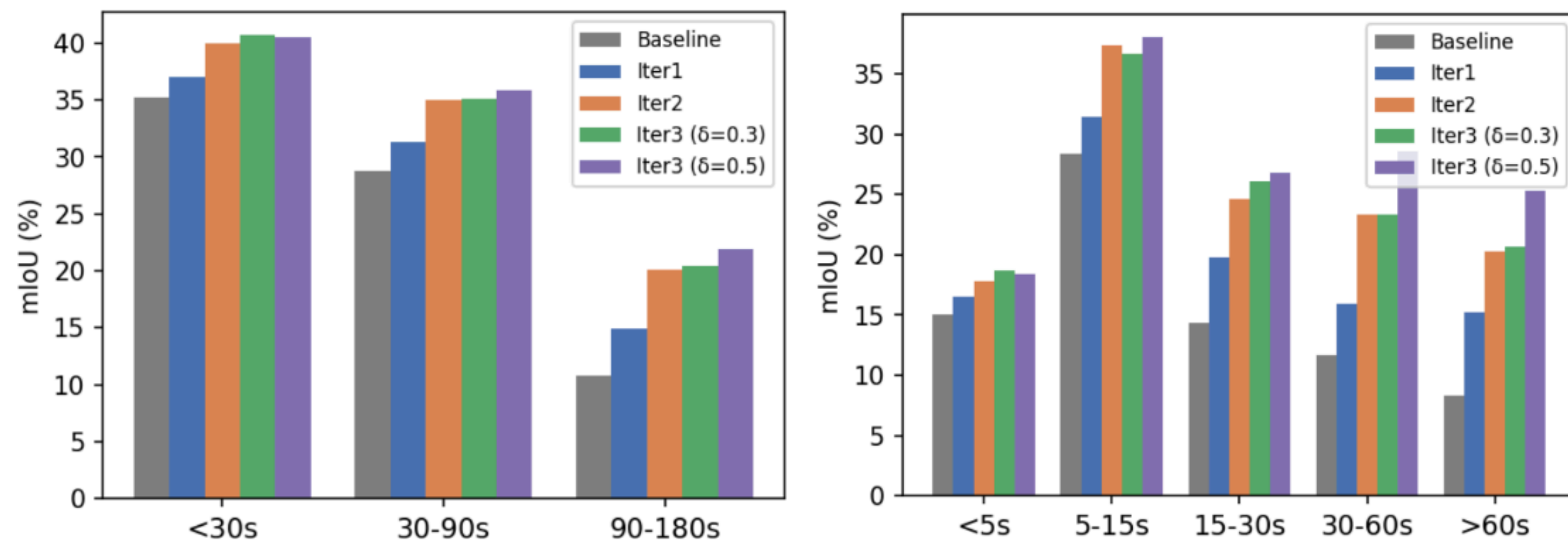
*All*



Iter	$R_{fmt}$	$R_{con}$	$R_{feed}$	TVGBench		
				R1@0.3	R1@0.5	R1@0.7
1	✓			34.0	21.2	12.0
2	✓			37.0	23.8	12.5
2	✓	✓		<u>38.7</u>	<u>26.2</u>	<u>14.6</u>
2	✓		✓	<u>38.1</u>	<u>26.0</u>	<u>14.4</u>
2	✓	✓	✓	<b>40.1</b>	<b>27.4</b>	<b>14.9</b>
3	✓			37.2	23.6	13.5
3	✓	✓		41.7	27.3	<u>15.0</u>
3	✓		✓	<u>41.8</u>	<u>27.4</u>	14.8
3	✓	✓	✓	<b>42.1</b>	<b>28.5</b>	<b>15.3</b>

# Experiments

*EvoGround demonstrates consistent Improvement across different duration and moment lengths*

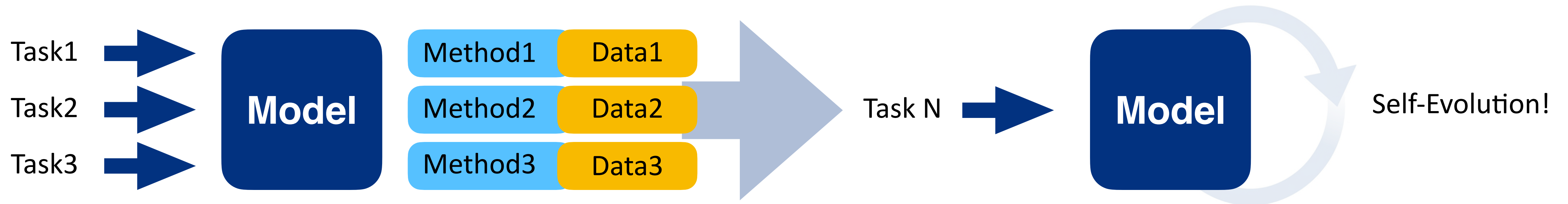


*Scaling data restarts the loop*

Method	FT	Charades-STA			
		R1@0.3	R1@0.5	R1@0.7	mIoU
Hawkeye [46]	✓	72.5	58.3	28.8	49.3
TimeChat [30]	✓	-	46.7	23.7	-
TimeChat-VT [12]	✓	-	58.4	34.7	-
ED-VTG [50]	✓	78.2	62.1	35.0	52.6
VideoChat-R1 [34]	✓	-	71.7	<b>50.2</b>	<u>60.8</u>
VideoChat-R1.5 [35]	✓	<b>82.8</b>	71.6	48.3	60.6
Time-R1 [14]	✓	<b>82.8</b>	<b>72.2</b>	<u>50.1</u>	<b>60.9</b>
<b>EvoGround (2.5K)</b>	✗	77.2	60.5	35.5	53.1
<b>EvoGround (7.5K)</b>	✗	<u>81.7</u>	67.7	45.6	60.3

# Summary

- Existing video large language models exhibit significant consistency failures across diverse scenarios.
- Such deficiencies can be improved through specialized objectives and curated data training. Yet, these solutions are not scalable.
- Enabling models to discover and address their own weaknesses is a promising path toward more robust and general video understanding.



**Thank You**