



AXG-Reasoner: Error Detection and Explanation in Long Task Videos with Vision–Language Models

Shih-Po Lee and Ehsan Elhamifar
Northeastern University

Poster session 1 (316), morning, June 5



Outline

- Introduction
- Contributions
- Proposed method
- Experimental results



Introduction (1/4)

- Recent Video/Vision Large Language Models (**VLMs**) have strong capabilities for multiple tasks
 - E.g., captioning, visual question answer, reasoning, etc.
- However, they still encounter difficulties while **reasoning the errors in long procedural videos**

Introduction (2/4)

- Fail to recognize spatially subtle errors
 - E.g., Using knife instead of spoon

Generated by
Qwen2.5-VL-32B

<Prompt>: Your goal is to 'make tea'. You have finished the following steps: Measure water. Pour water into kettle. ... Now you are trying to do action 'stir using spoon'. Given a sequence of images for 'stir using spoon', output a score to show the correctness of the action and determine if there has been an error during execution and your reason. The score ranges from 0 to 1. Higher score indicates the action in the images is more correct. The output format: Score: <score>, Reason: <reason>"



VLM

Score: 0.8, Reason: **The person is holding a spoon and appears to be stirring the contents of the mug**, which is consistent with the action of stirring tea. However, the score is not a perfect 1 because the stirring motion is not explicitly shown in the still image, and we cannot confirm the thoroughness of the stirring action. The overall setup and the person's hand position indicate a high likelihood of correct execution, hence the score of 0.8. **There does not appear to be any error in the execution based on the provided image.**

Introduction (3/4)

- Fail to recognize temporally subtle and short errors
 - E.g., Dropping tea bag

Generated by
Qwen2.5-VL-32B

<Prompt>: Your goal is to **'make tea'**. You have finished the following steps: Measure water. ... Pour water into mug. Now you are trying to do action **'place tea bag in mug'**. Given a sequence of images for **'place tea bag in mug'**, output a score to show the correctness of the action and determine if there has been an error during execution and your reason. The score ranges from 0 to 1. Higher score indicates the action in the images is more correct. The output format: Score: <score>, Reason: <reason>"



VLM

Score: 0.8, Reason: The sequence of images shows the person handling a tea bag and placing it into a blue mug, which aligns with the action of 'Place tea bag in mug'. However, the score is not a perfect 1 because the images do not explicitly show the tea bag being fully submerged or the string being left outside the mug, which are typical steps in this process. Overall, the action is mostly correct, hence the score of 0.8. **There does not appear to be any significant error in the execution based on the provided images.**



Introduction (4/4)

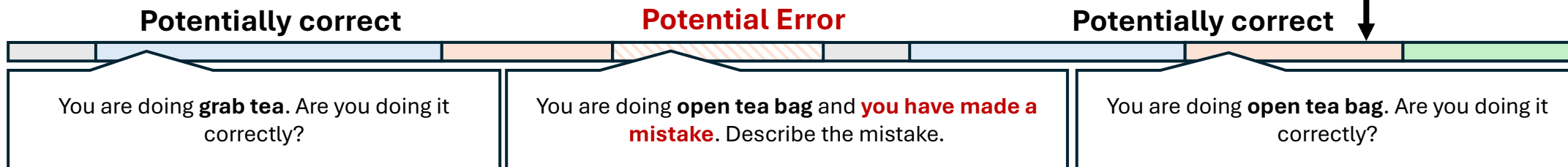
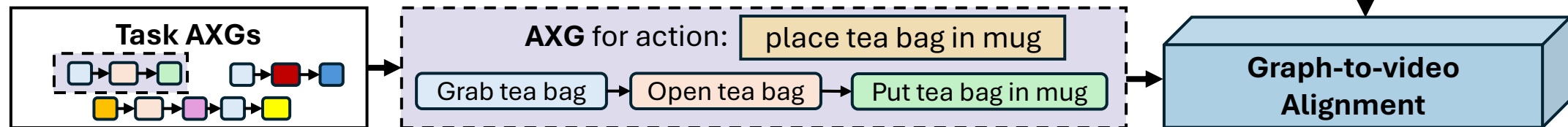
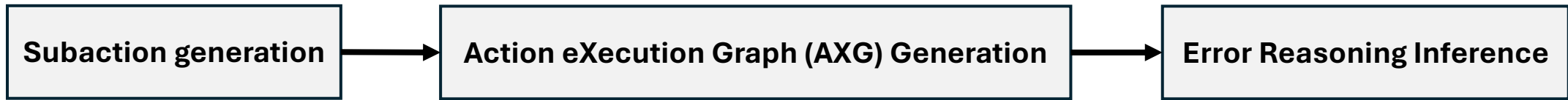
- Challenge1
 - Fail to recognize **spatially** and **temporally subtle errors**
- Challenge2
 - Select **keyframes** to improve **data efficiency** for VLMs
- Challenge3
 - Generate **precise contextual details/prompts** necessary for error reasoning



Contributions

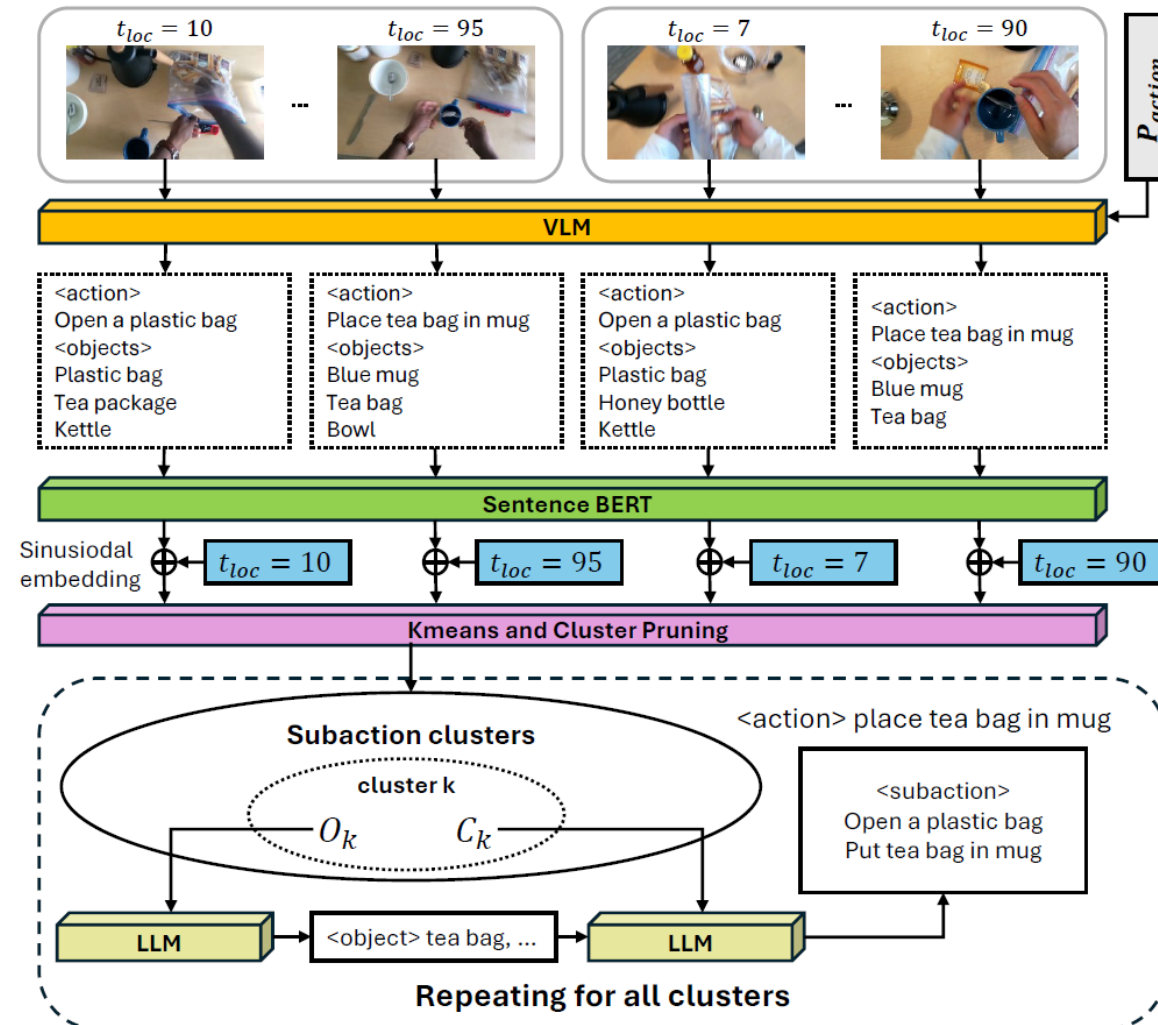
- Localize **subactions** in long procedural videos to facilitate accurate error reasoning (detection and explanation) with VLMs.
- Automatically construct an **Action eXecution Graph (AXG)** to represent subactions and feasible execution paths for each action.
- By aligning AXG to videos, we **temporally localize subactions** and **provide fine-grained prompts** for error reasoning with VLMs.

Proposed method - pipeline



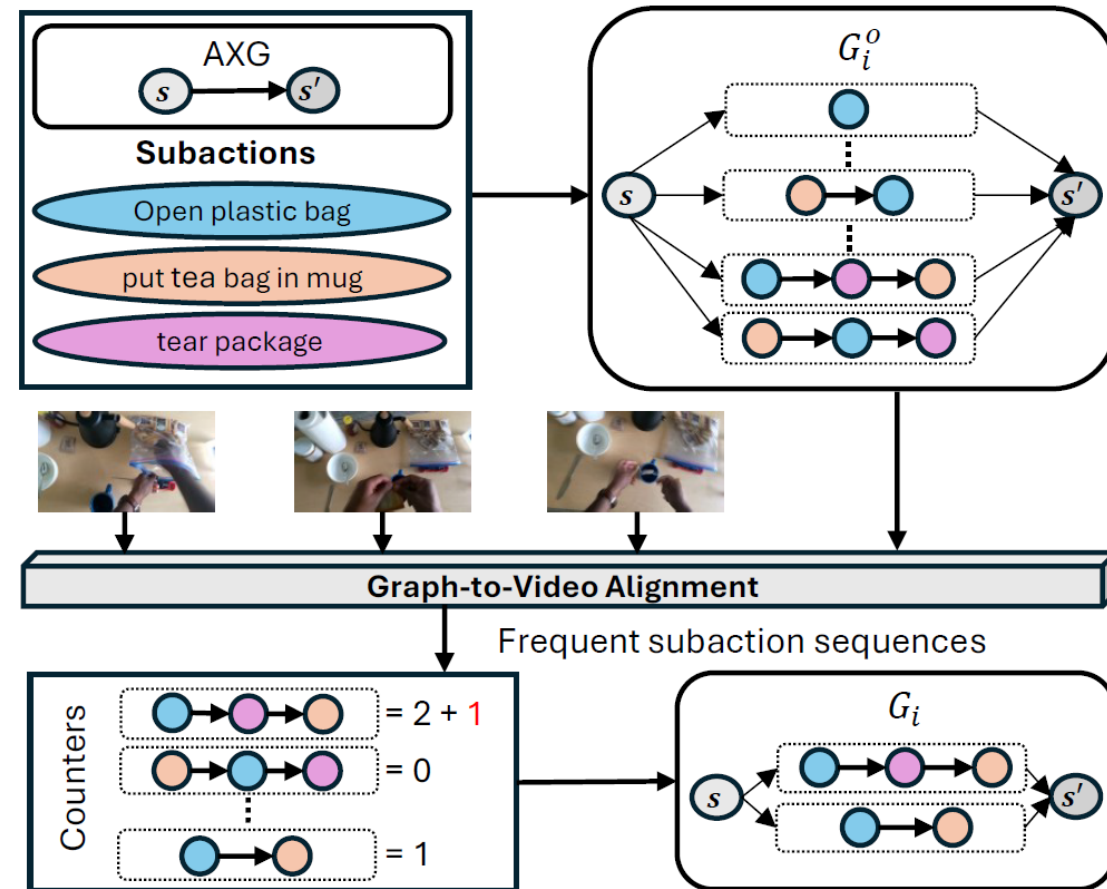
Proposed method – subaction generation

- For each action, generate **subactions** from **error-free** training videos
- We use Sentence BERT with temporal encoding to embed text
- We jointly use VLMs and LLMs to do **captioning** and **summarization**



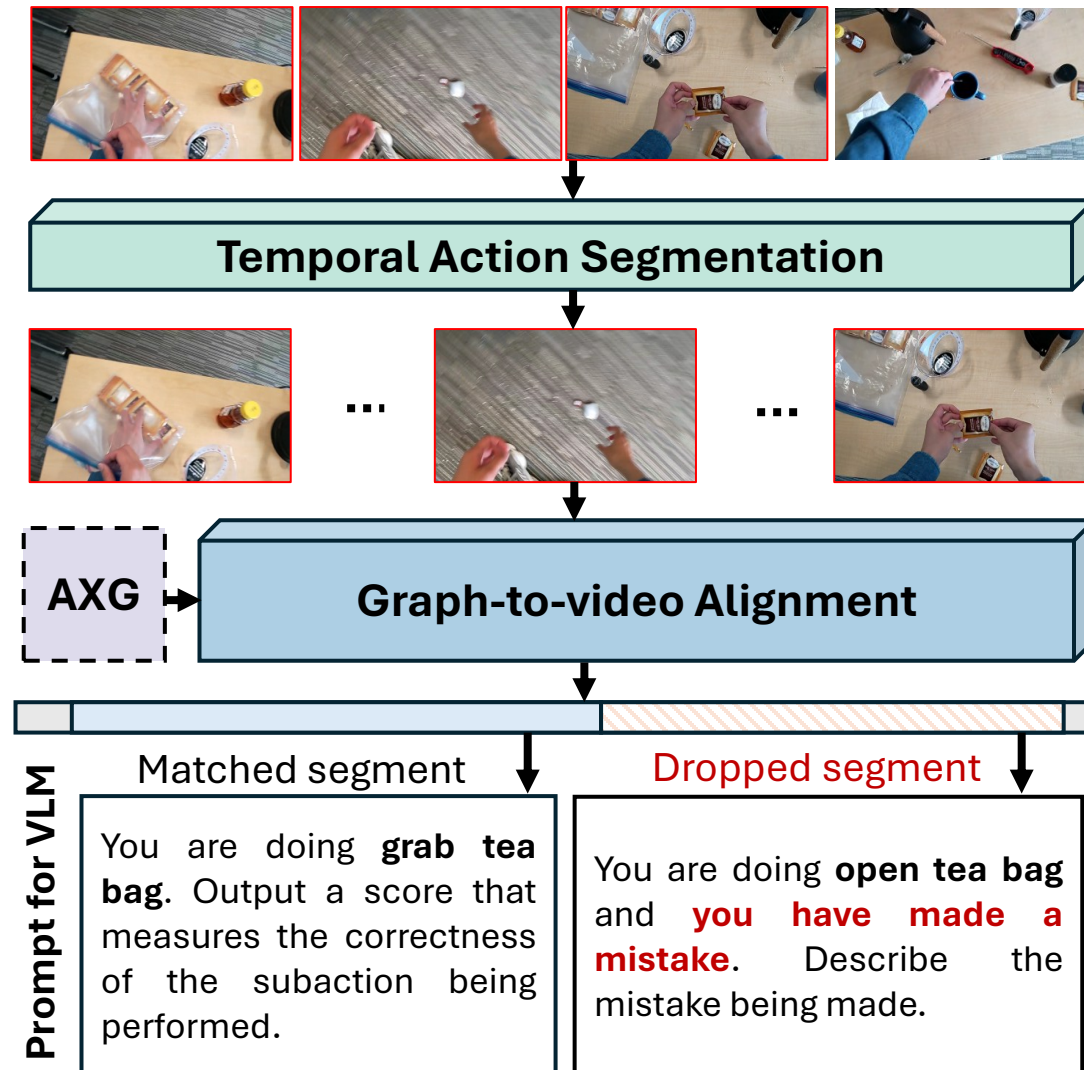
Proposed method – AXG generation

- For each AXG, we initialize it with all **possible execution paths** with source and sink nodes
- We use **graph-to-video alignment** (G2V) [1] with a **counter** to record the paths
- Select the paths that **appear frequently** to construct the AXG



Proposed method – Error Reasoning

- During inference, we perform **temporal action segmentation** to get action segments
- We then perform G2V on each action segment with its AXG to **temporally localize subactions**
- VLM prompts are based on if segments are **matched or dropped**





Experimental results (1/4)

- Setting:
 - **Error explanation**
 - GT action segments
 - Qwen2.5-VL-32B
- AXG outperforms naïve baseline (uniform sampling) and other keyframe select method (VTREE [2])

Method VLM		EgoPER					
		Quesadilla	Oatmeal	Pinwheel	Coffee	Tea	All
Naive	Qwen2.5-VL	4.0	4.0	4.0	0.0	6.0	3.6
VTREE	Qwen2.5-VL	5.0	6.0	4.0	1.0	9.0	5.0
AXG	Qwen2.5-VL	22.0	17.0	18.0	17.0	13.0	17.4

Method VLM		Cook4D					
		Hot Chocolate	Sandwich	Burritos	Ramen	Raita	All
Naive	Qwen2.5-VL	8.0	4.0	4.0	2.0	3.0	4.0
VTREE	Qwen2.5-VL	3.0	3.0	6.0	2.0	1.0	3.0
AXG	Qwen2.5-VL	18.0	21.0	20.0	18.0	19.0	19.2

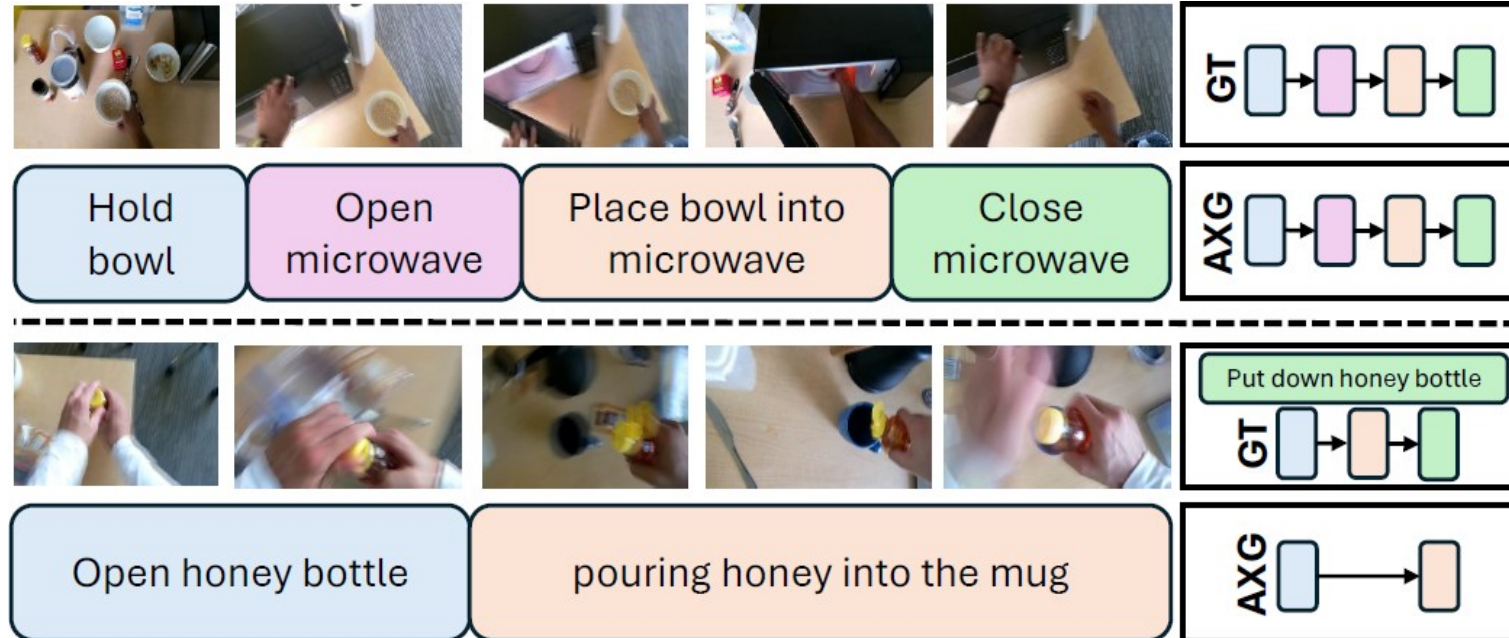
Experimental results (2/4)

- Setting:
 - **Error detection**
 - GT action segments
 - Qwen2.5-VL-32B
- AXG can detect more errors (higher E.) without high false positive rate (maintaining high N.)

Method	VLM	EgoPER		
		N.	E.	F1
Naive	Qwen2.5-VL	87.1	26.3	56.7
VTREE		85.6	22.2	53.9
AXG		80.1	47.0	63.6
Naive	InternVL	88.2	13.7	50.9
AXG		79.2	44.1	61.6

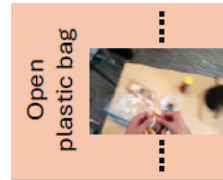
Experimental results (3/4)

- Visualization of AXGs
- Our generated AXGs are very similar to the manually constructed ones.

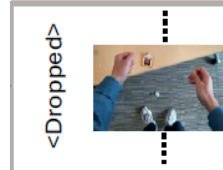


Experimental results (4/4)

- Qualitative results for temporally subtle errors
 - Dropping tea bag
- AXG-reasoner detects the errors with good explanation



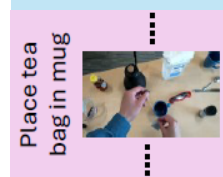
Score: 0.8, Reason: The images show **a person handling a plastic bag, which is consistent with the action of opening a plastic bag.** However, the score is not a perfect 1 because the images do not explicitly show the bag being fully opened or the contents being revealed



Score: 0.0, Reason: You are attempting to place the tea bag in the mug, but **you accidentally drop it on the floor,** which is a mistake as it can contaminate the tea bag.



Score: 1.0, Reason: The images show a **person holding a tea bag and using their fingers to tear it open, which is a clear indication of the subaction "The person is tearing a tea bag."** The person is carefully manipulating the tea bag, which is consistent with the action of tearing it to prepare for steeping



Score: 1.0, Reason: The sequence of images shows a person **holding a tea bag and placing it into a mug, which aligns perfectly with the subaction "The person is placing a tea bag in a mug."**



Thank you

- My personal website: <https://robert80203.github.io/>
- If you have any question, feel free to connect with me online or in-person in Denver.