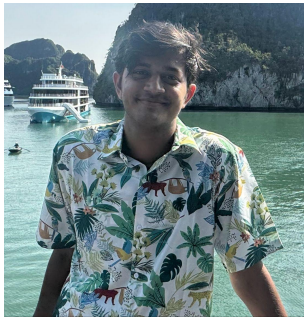
STPro: Spatial and Temporal Progressive Learning for Weakly Supervised Spatio-temporal Video Grounding



Aaryan Garg[1]

Akash Kumar[2]

Yogesh Singh Rawat[2]

[1]BITS Pilani

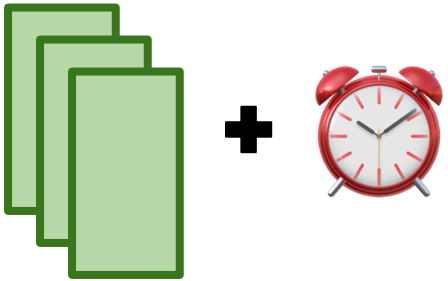
[2]University of Central Florida

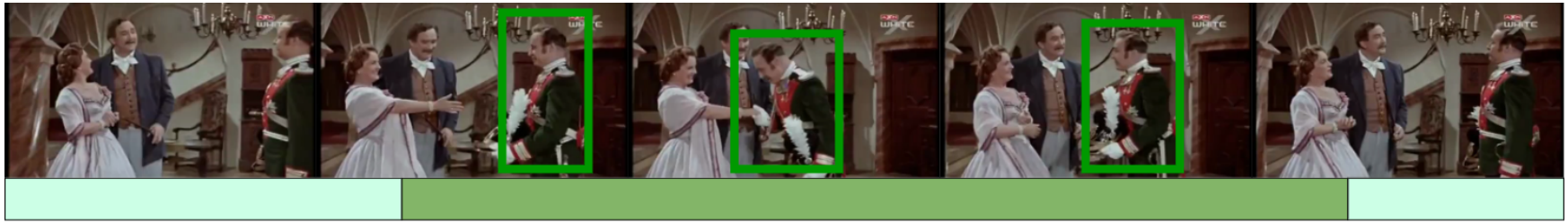
What is Open-World Free-form Grounding?

Untrimmed
Seen/Unseen

Any Caption

No fixed form
Has spatio-temporal reference





The ***man in red*** clothes ***puts*** the hat to his left hand, then ***pulls*** the womans hand and ***kisses***, then ***puts*** the hat to his right hand.

Qualitative Samples

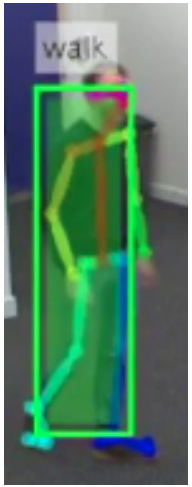
Query: A **dog** with a rope **walks** in front of a **man in white**.

Query: The **man in brown clothes** **pours** the contents of the bag into his hand, and then **takes out** a piece of paper from the bag and **opens it**.

Can’t Vision Language Models (VLMs) **easily** solve STVG? 🤔

The ***bottom man*** with his ***head up***

Referring Expression Grounding



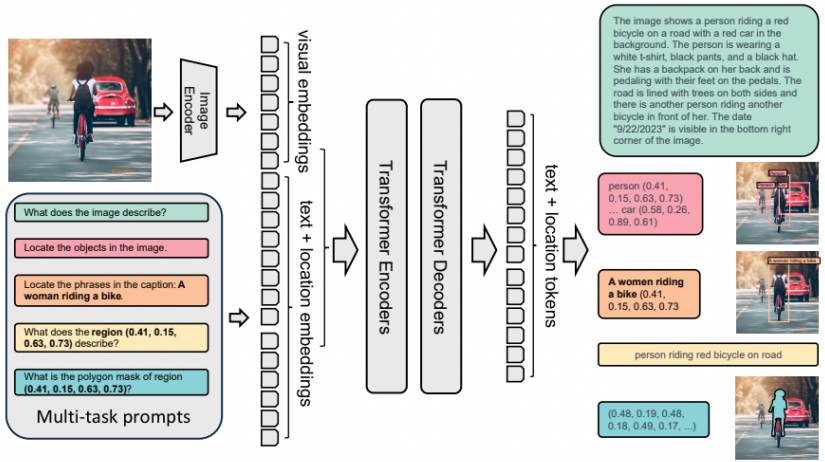
The **walking** man

Action Recognition

Which VLMs to start from?

Multi-modal LLMs (MLLMs)?

Florence-2 [CVPR'24] (Microsoft)

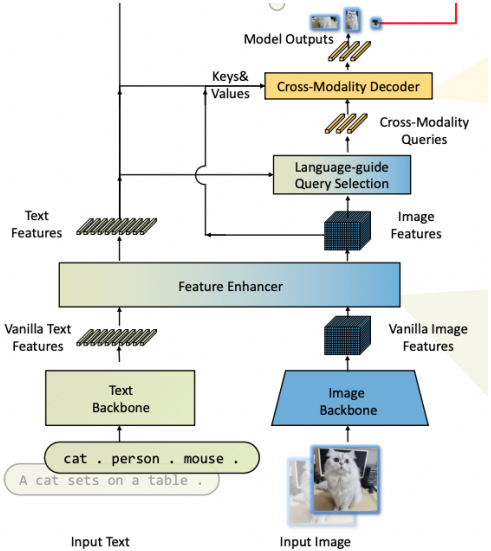


Bounding-Box ✅ Confidence ❌

Localized Features ❌

Task-specific VLMs?

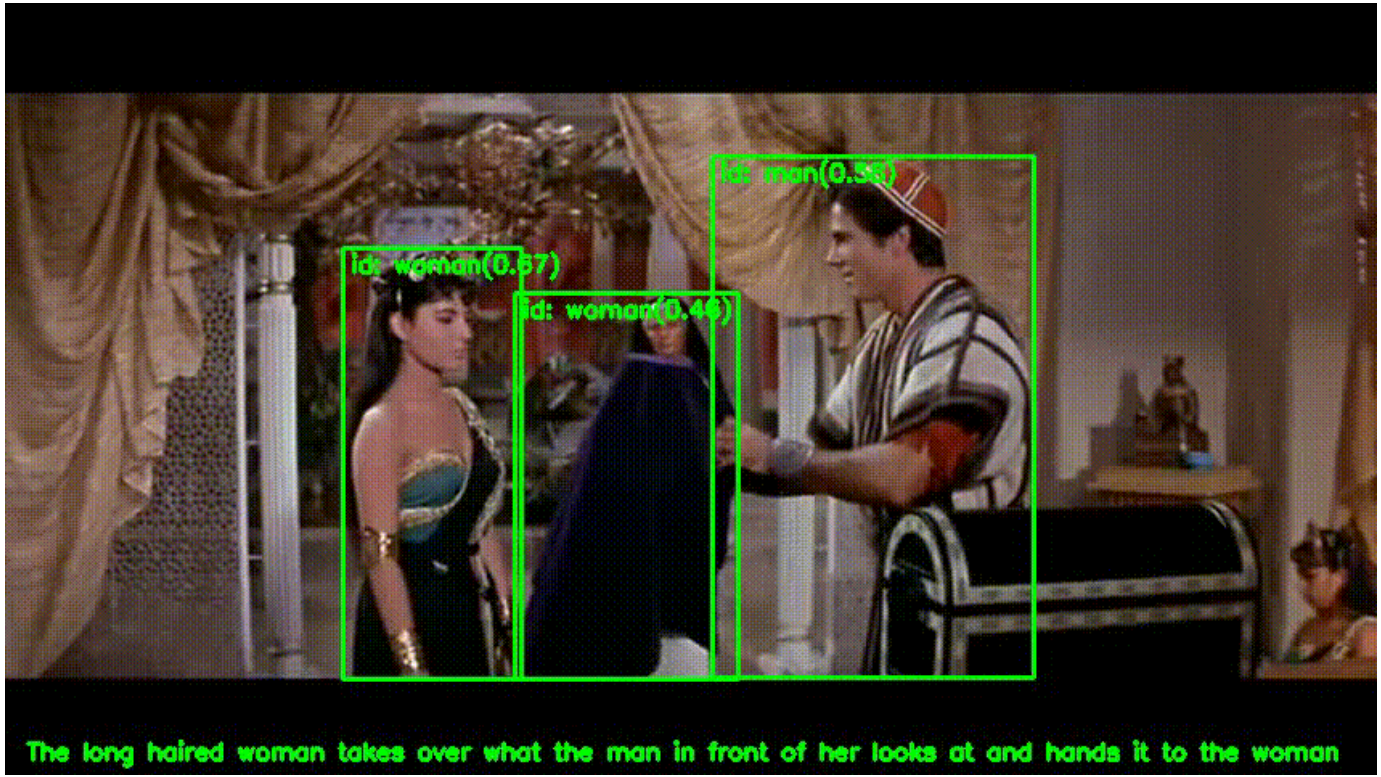
Grounding-DINO [ECCV'24]



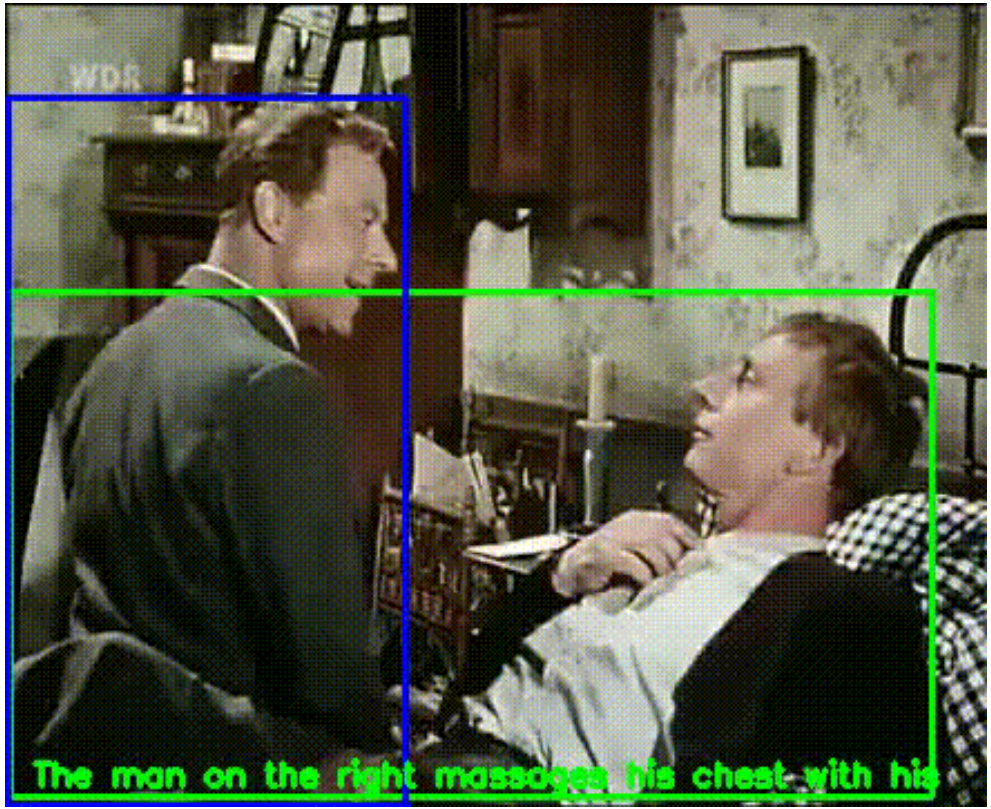
Bounding-Box ✅ Confidence ✅

Localized Features ✅

Multimodal DETR are **robust** at detection 🙂



Fails at exact query grounding ☹



Results? 🧐

Methods	HCSTVG-v1			VidSTG		
	mvIoU	vIoU@0.3	vIoU@0.5	mvIoU	vIoU@0.3	vIoU@0.5
AWGU [CVPR19]	8.2	4.5	0.8	8.8	7.4	3.0
Vis-CTX [CVPR19]	9.8	6.8	1.0	9.0	7.3	3.1
WINNER [CVPR23]	14.2	17.2	6.1	10.9	13.0	6.4
W-GDINO [ECCV24]	9.0	11.6	4.6	10.2	12.6	7.3

Proof: Trivial adaptation fails. VLMs are not designed for dense video tasks.

What's limiting W-GDINO's Performance?

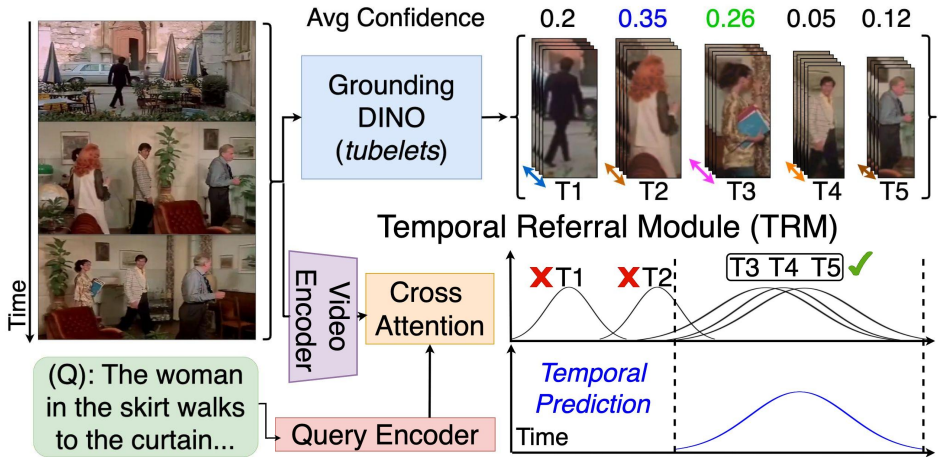
Actor's entire presence grounded, instead of specified temporal boundary

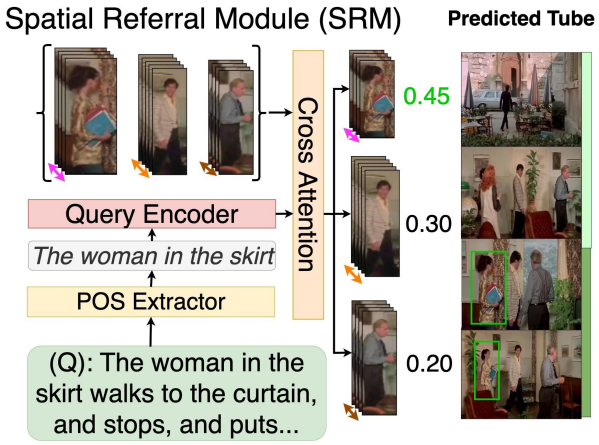
The **man in red** clothes **puts** the hat to his left hand, then **pulls** the womans hand and **kisses**, then **puts** the hat to his right hand.

Confidence score changes vastly within tubelets due to challenging scenes

The **man in black suit follows** the man, **chases** him and **run** towards the car

Weakly-Supervised Adaptation of DETR





Why weakly supervised? 🤨

Weakly Supervised

💵💵

Less Bias (❄️ Models)

Easy to Scale (large datasets)

Fully Supervised

💵💵💵💵💵💵

More Bias (🔥 Models)

Hard to Scale (large datasets)

Results? 🧐

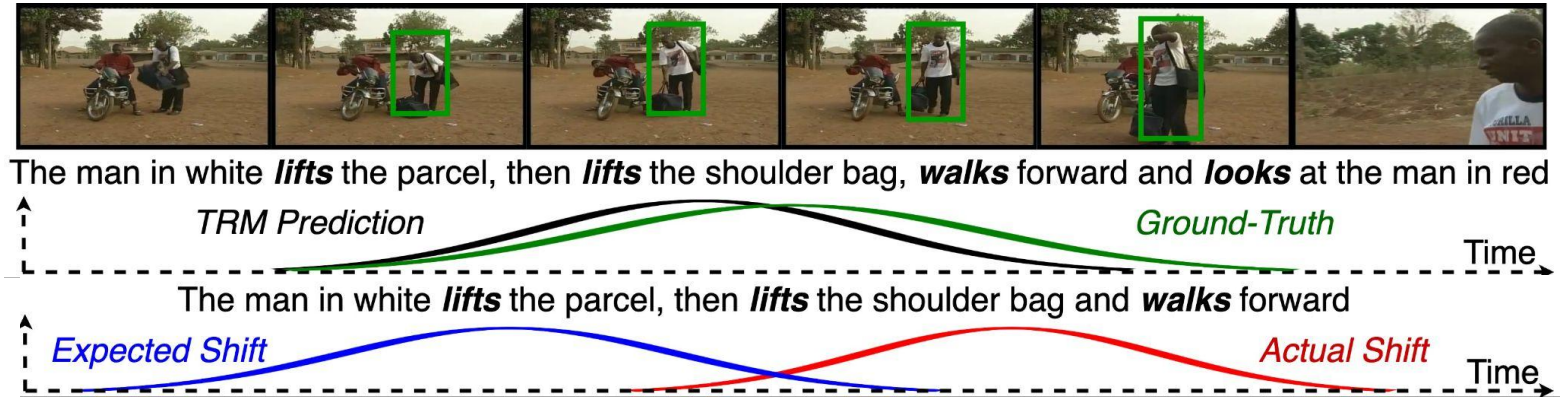
Methods	VidSTG-Declarative			HCSTVG-1		
	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
AWGU [5]	9.0	7.9	3.1	8.2	4.5	0.8
Vis-CTX [33]	9.3	7.3	3.3	9.8	6.8	1.0
WINNER [18]	11.6	14.1	7.4	14.2	17.2	6.1
W-GDINO [24]	10.6	13.0	7.8	9.0	11.6	4.5
SRM + TRM	**15.52**	**19.39**	**12.69**	**14.81**	**21.81**	**10.26**

Some improvement! But, we're still far away from what's achievable!

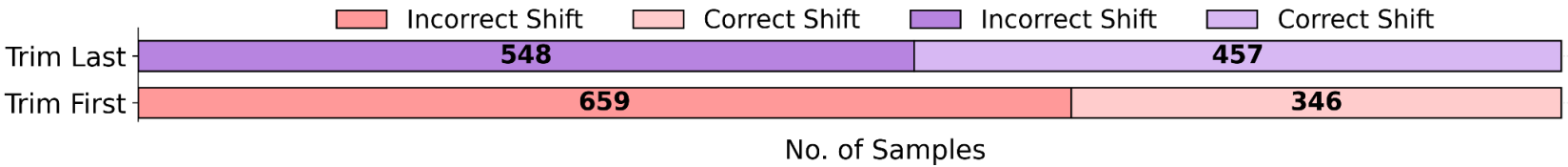
Dataset	m_tIoU	m_vIoU	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.5
	Detection			Post-Tracking		
HCSTVG-v1	92.1	69.3	86.2	83.6	65.0	76.3
HCSTVG-v2	95.1	68.8	87.0	81.5	60.0	68.0
VidSTG-D	83.8	52.7	60.9	72.5	45.4	50.6
VidSTG-I	83.7	48.8	54.7	72.7	41.6	44.6

So, what's limiting TRG? Where does it fail and why?

Lacks Action Composition Understanding!

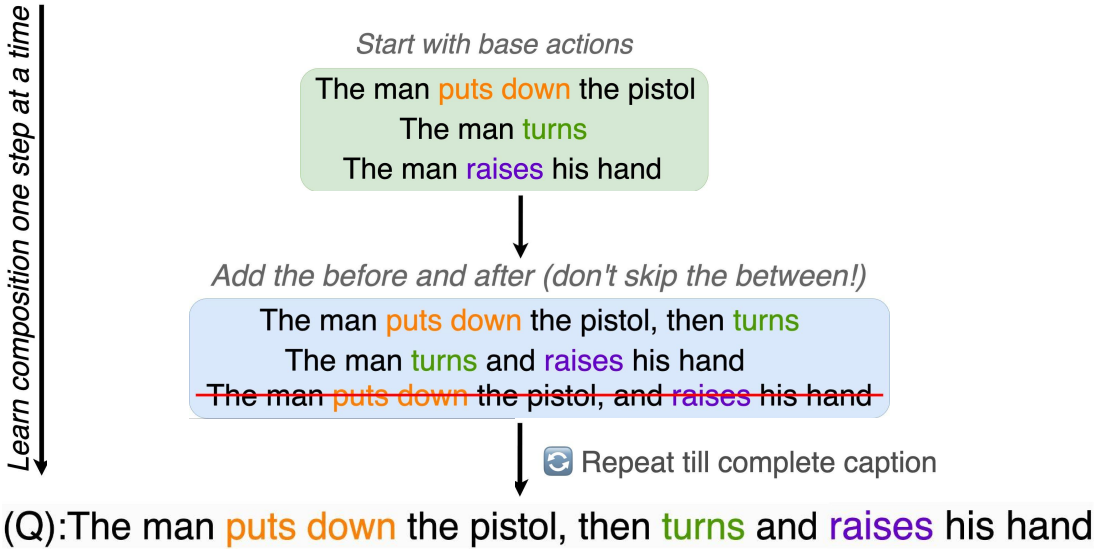


Not just some samples! We see consistent failure.

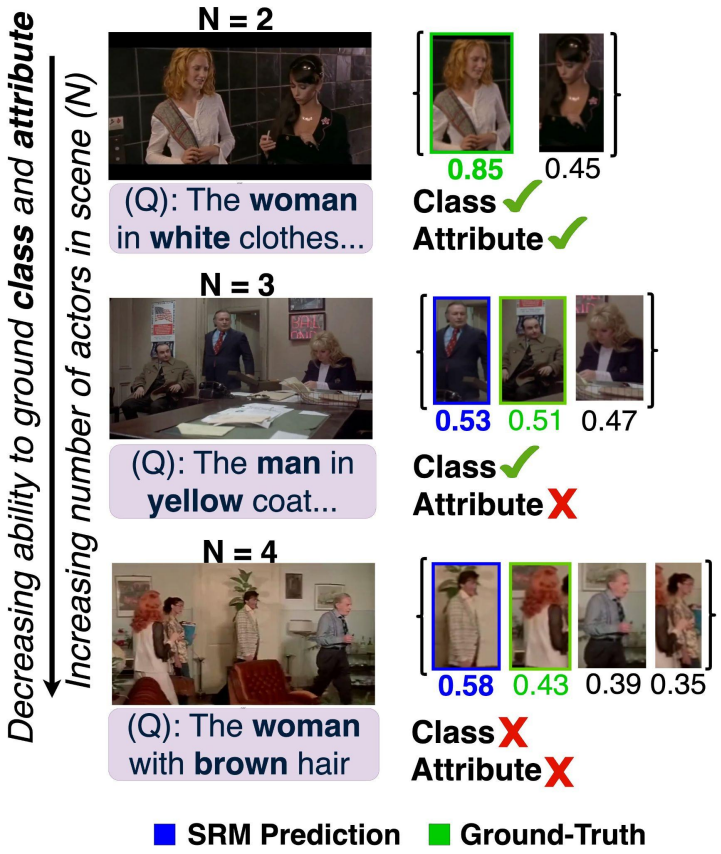


Solution? **Sub-Action** Curriculum

- **Aim:** Teach the model individual actions and how they stack
- **Why:** Compositional understanding enables new-scenario generalization



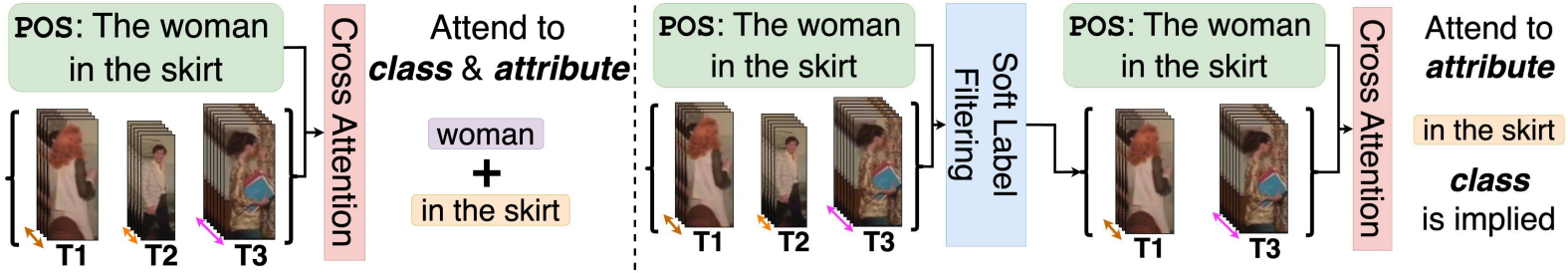
Deterioration of SRM On Dense Scenes



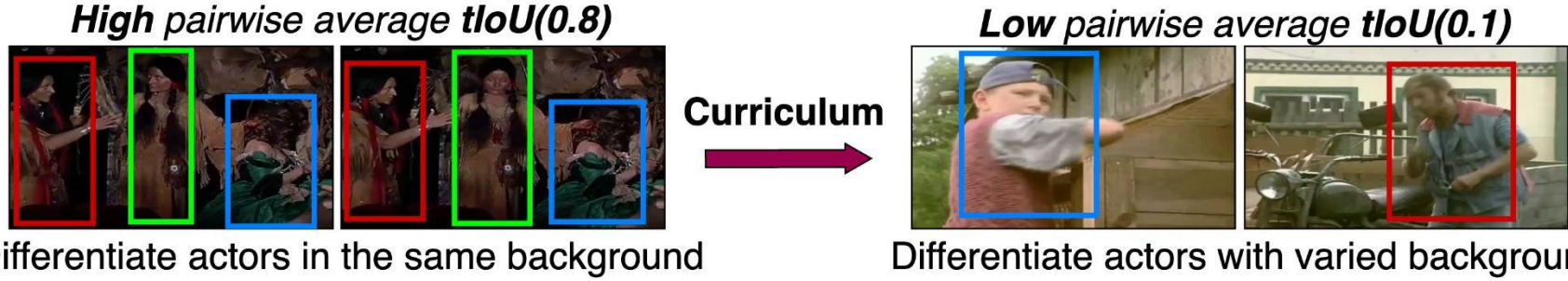
- **Challenge:** As actors increase, SRM fails to ground the right class and attribute.
- **Importance:** Independent class and attribute understanding leads to better new-scenario generalization.

Solution? **Spatial** Curriculum

Soft-Label Filtering (SLF) - Let SRM focus on attribute. Imply class

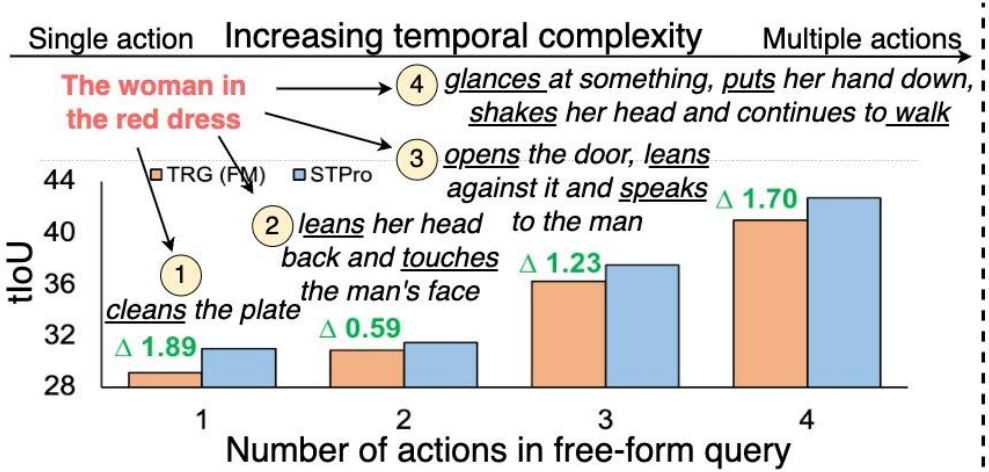


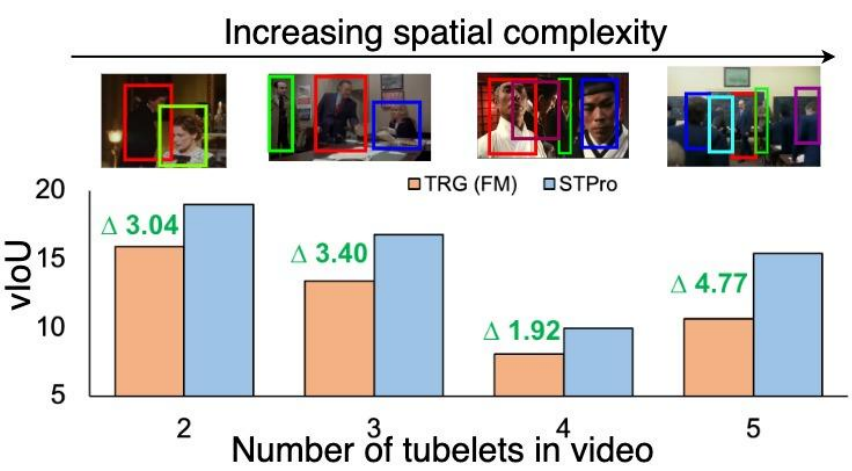
Congestion Guided Sampling (CGS) - Eliminate background distraction



Do The **Curriculums** Really Help? 🧐

Significant performance improvements on more complex spatial and temporal scenes!





Summary

- Contributions:
  - Novel simple adaptation of VLMs on dense multimodal videos.
  - **First** work to make VLMs **spatio-temporal** contextual aware via **curriculum** learning.
- **Scalable** to large-scale datasets (minimal computation)

Visit Poster #307, Exhibit Hall D on June 13!!!