Contextual Self-Paced Learning for Weakly-Supervised Spatio-Temporal Video Grounding

Akash Kumar[1] Zsolt Kira[2] Yogesh Singh Rawat[1]

[1]University of Central Florida [2]Georgia Institute of Technology

Open-World Free-form Grounding

- Open-world: Adapt to any scene (seen/unseen)
- Free-form: User can input any query
- Spatio-Temporal Video Grounding (STVG)

The man in blue shirt walks to car, speaks to man who gets out of car and walks behind him.

STVG (Qualitative Examples)

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.

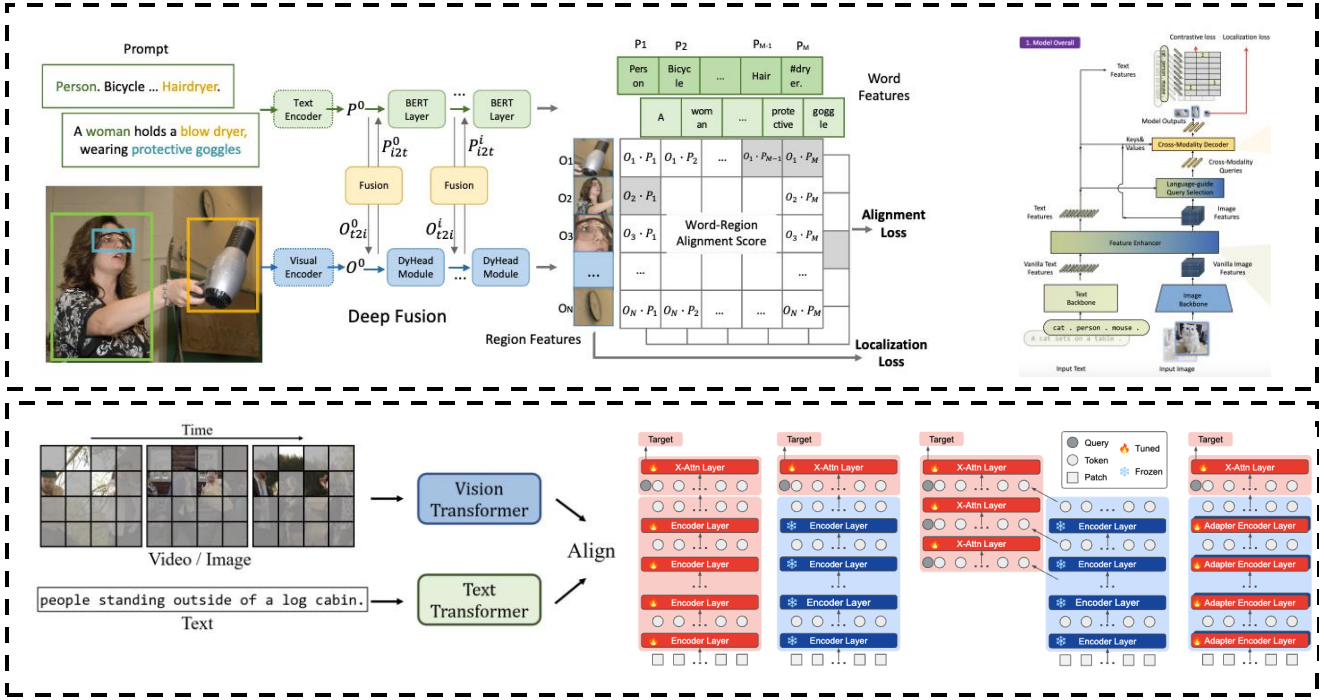
Can Vision Language Models (VLMs) solve this?

🤔 🤔 🤔

- It comprises of just two things:
  - Object Detection (man in white shirt)
  - Action Recognition (man in blue shirt walks and speaks)

Let’s see where VLMs stand

- Vision Language models
  - Image-based
    - Designed for Dense Tasks
    - Superior performance
    - Not designed for videos 😟
  - Video-based
    - Strong Trivial tasks
    - Can’t solve dense task 😟



[1] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European Conference on Computer Vision. Springer, Cham, 2025..

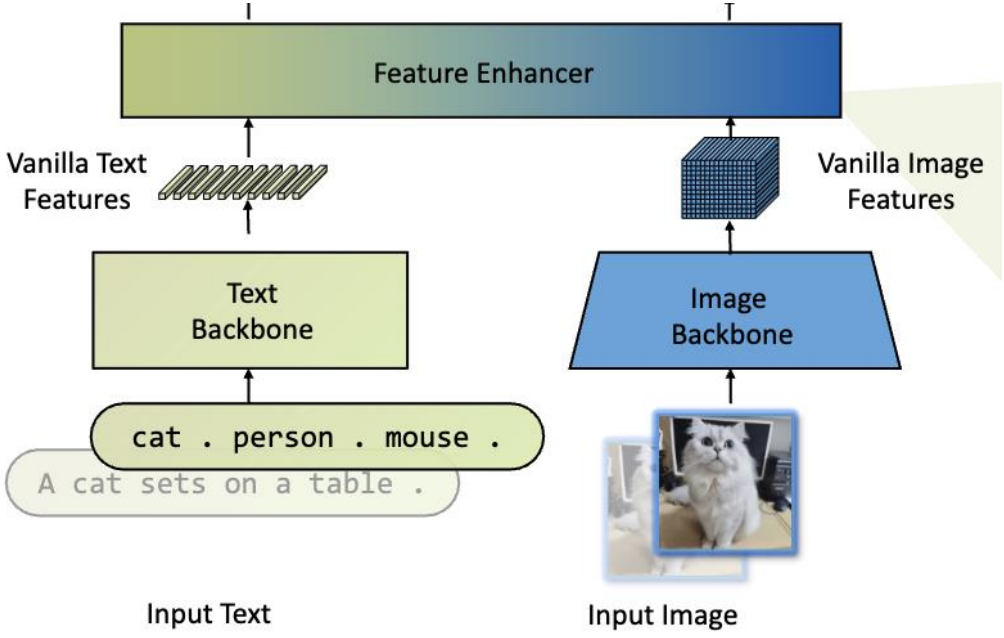
[2] Li, Liunian Harold, et al. "Grounded language-image pre-training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

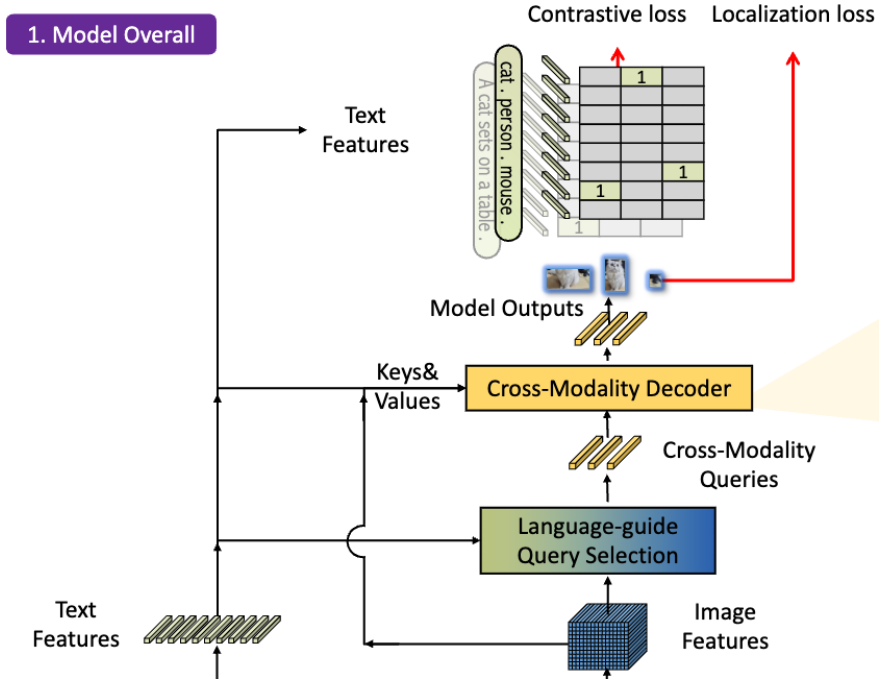
[3] Wang, Yi, et al. "Internvid: A large-scale video-text dataset for multimodal understanding and generation." arXiv preprint arXiv:2307.06942 (2023).

[4] Yuan, Liangzhe, et al. "Videoglue: Video general understanding evaluation of foundation models." arXiv preprint arXiv:2307.03166 (2023).

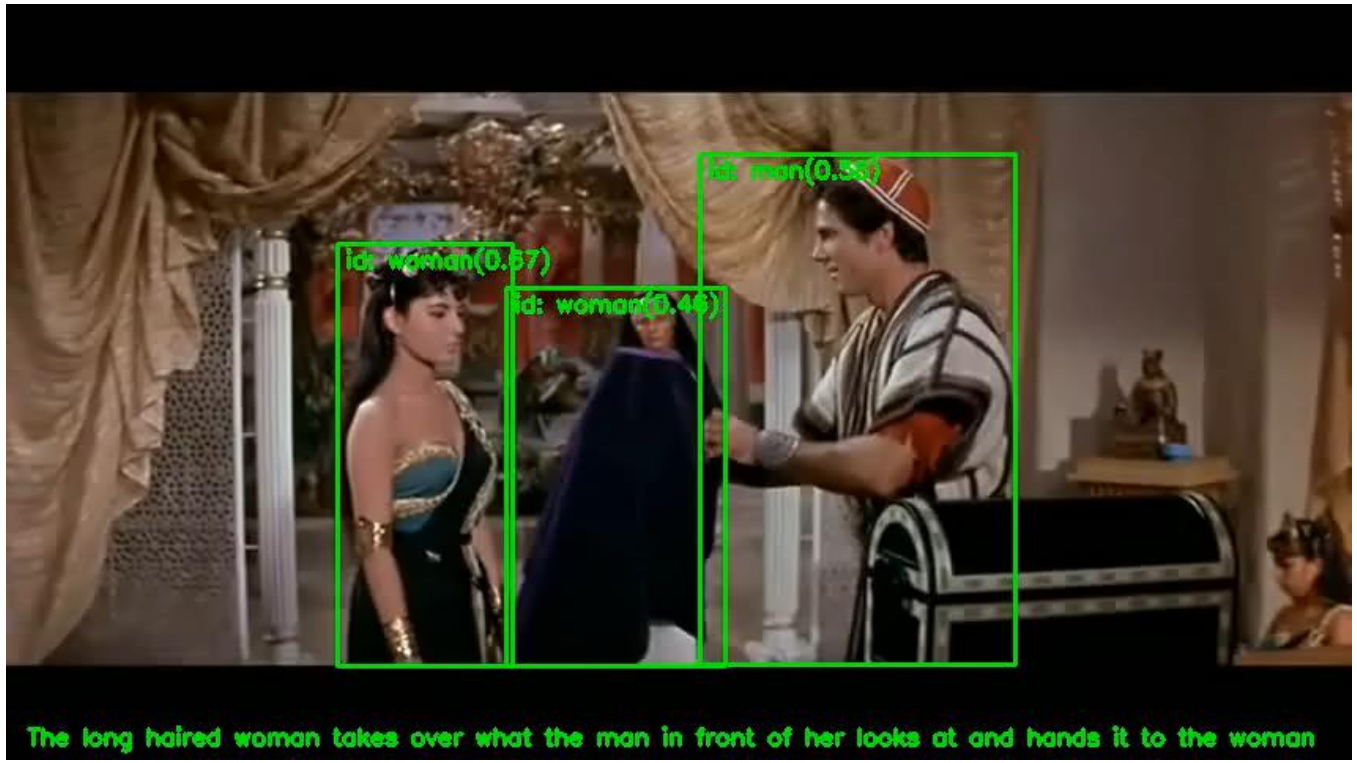
Evaluating VLMs…

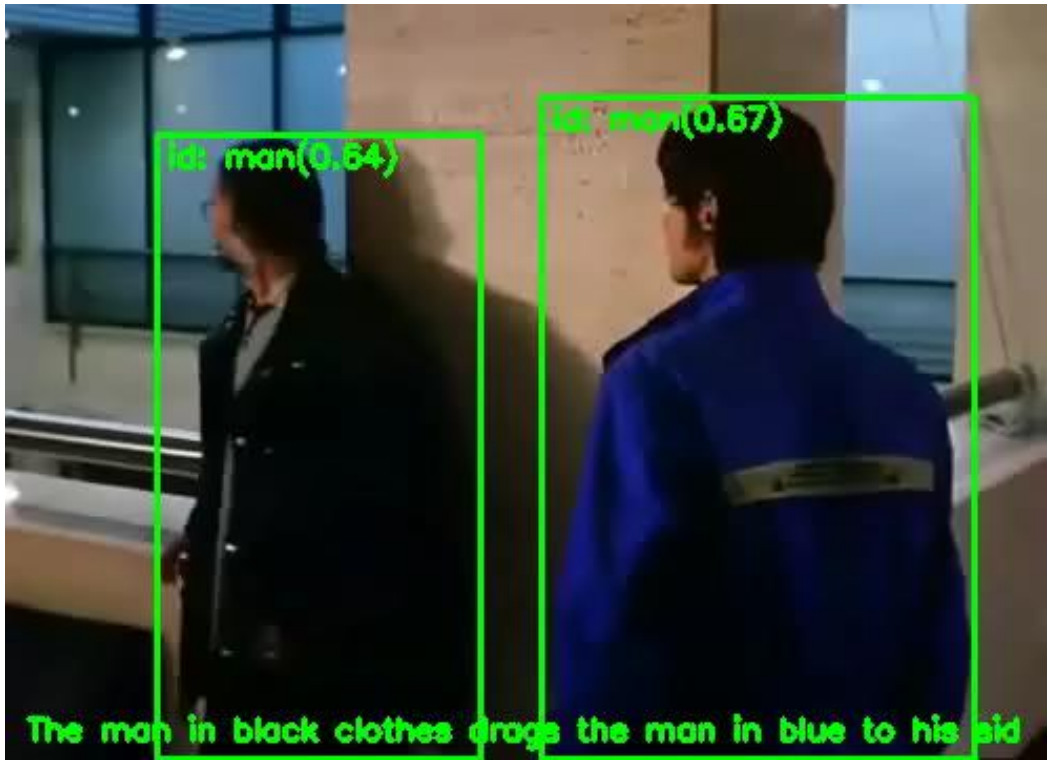
- Q: Why not CLIP/GPT-4v? A: No trivial extension exists!!!
- Q: Proposed Solution? A: Detection Model – Grounding DINO [ECCV'24]



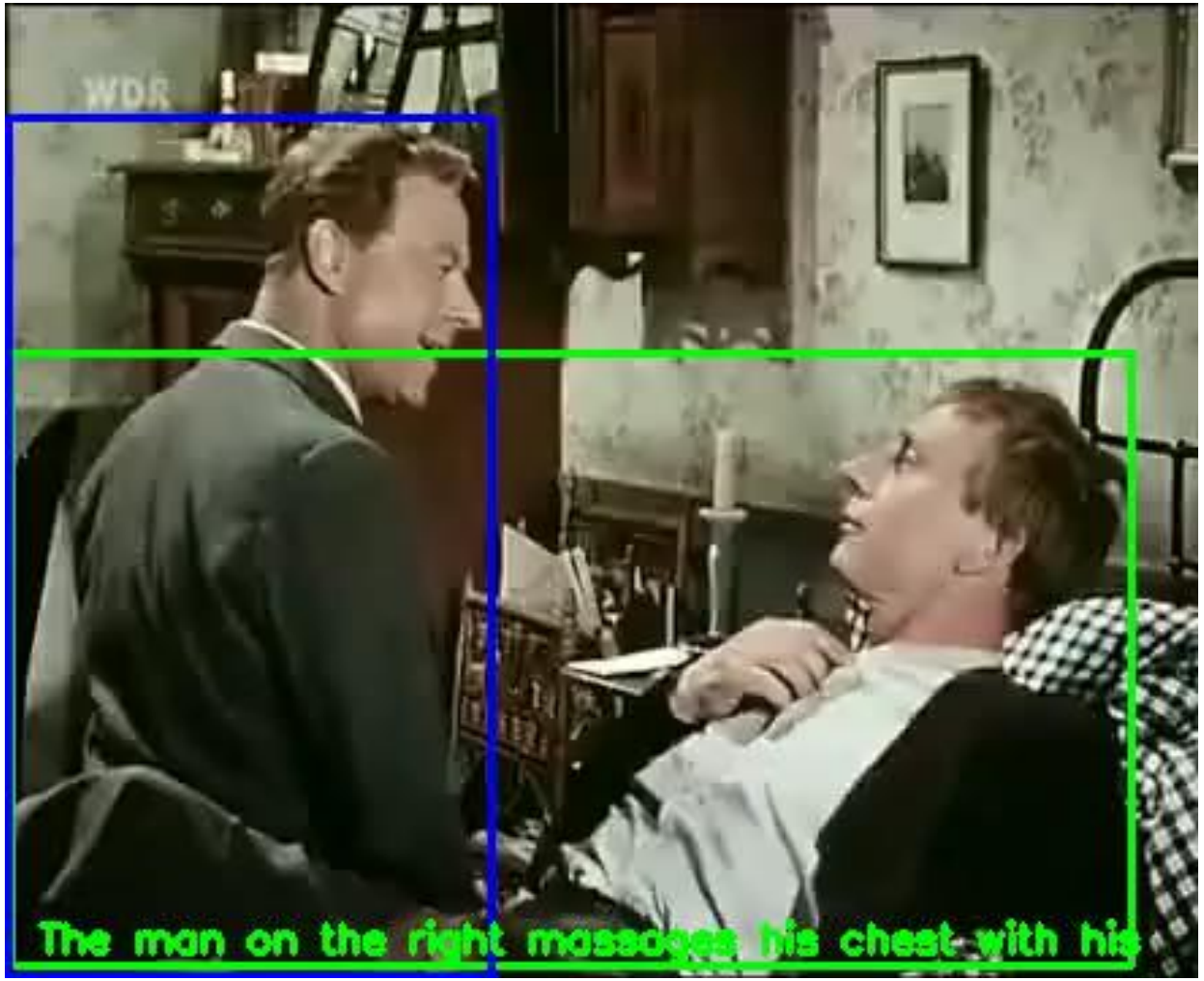


VLMs robust at object detection…





Fails when it comes to ground exact query.



Evaluating zero-shot capability of VLMs…

- Studied/Researched 60+ VLMs for dense image tasks
  - Grounding DINO [ECCV'24]
    - SOTA on image referral/phrase grounding/object detection tasks
  - Extended for STVG
    - Detector (Grounding DINO) + Tracker

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

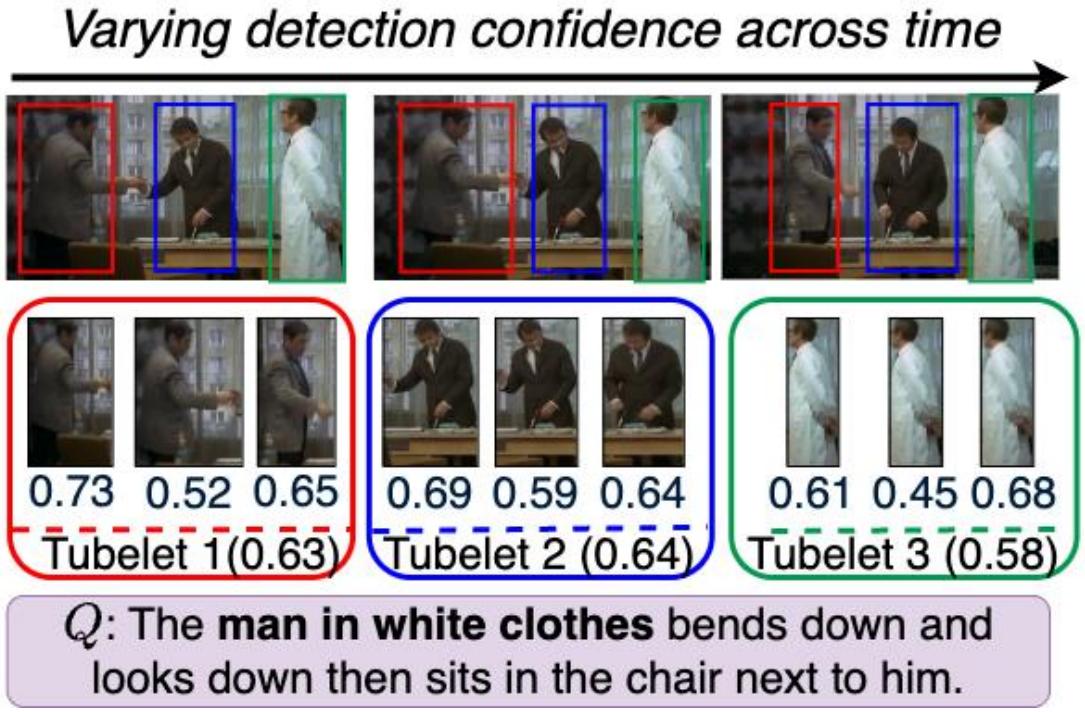
Proof: Trivial adaptation fails. VLMs are not designed for dense video tasks.

Why weakly supervised?

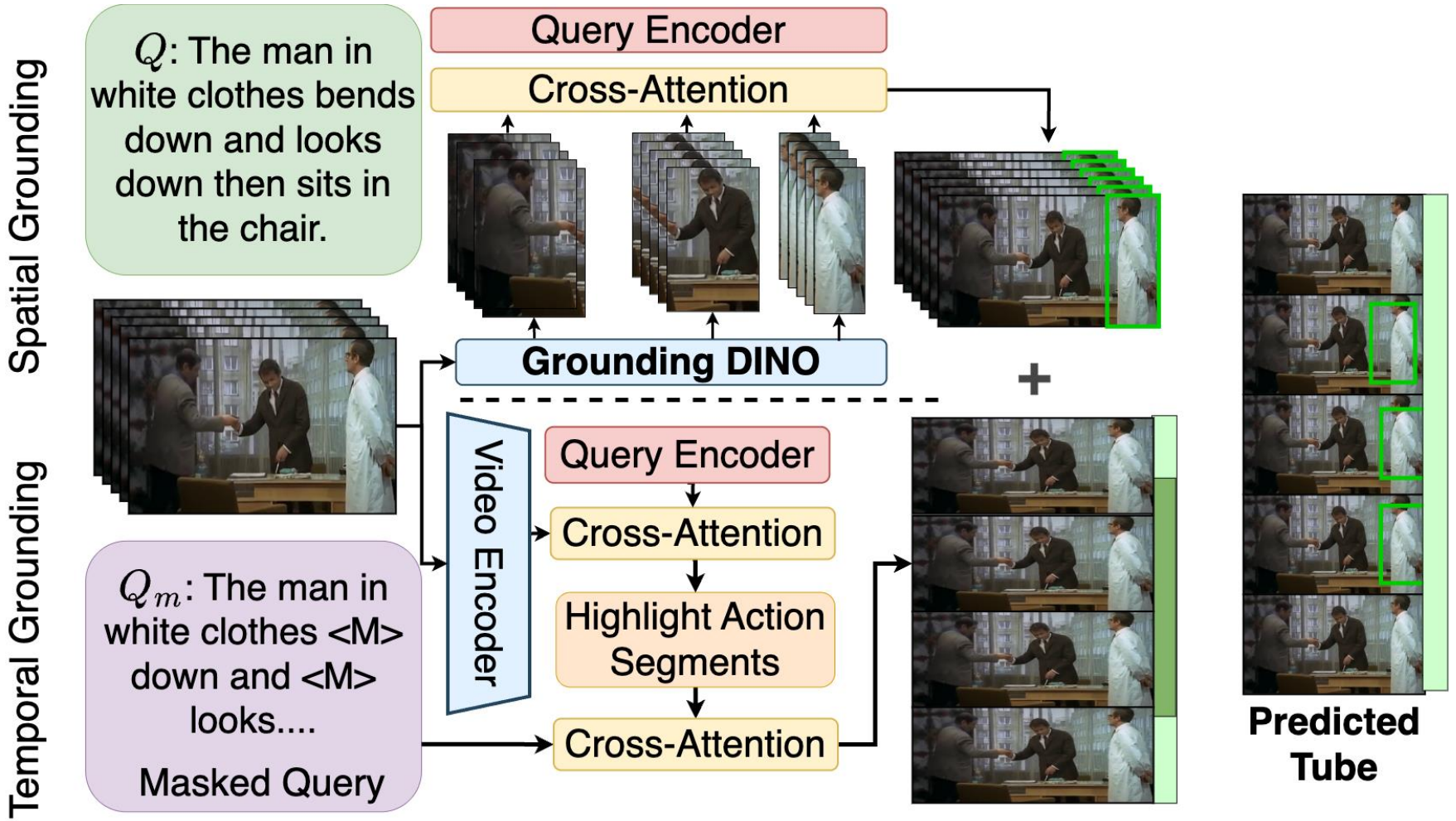
- VLMs - Not designed for end-to-end video
  - Zero-shot – Good Potential
- Expensive annotations dense task
  - Cost-effective data labels
- More Biasness - Finetune small datasets
  - Less biasness – Pre-trained on HUGE datasets
- Fully-Supervised – Scalability vs Compute Issues
  - Weakly Supervised - Scalable to large scale datasets
    - For e.g.: # Videos: 4.5k -> 100k, # Train Time: 4-5hr -> 10-12 hr (single GPU)

Let's start with looking into issues of VLM!!!

- **Aim:** Adapt VLMs for dense tasks
- **Why**: Unreliable Temporal Predictions



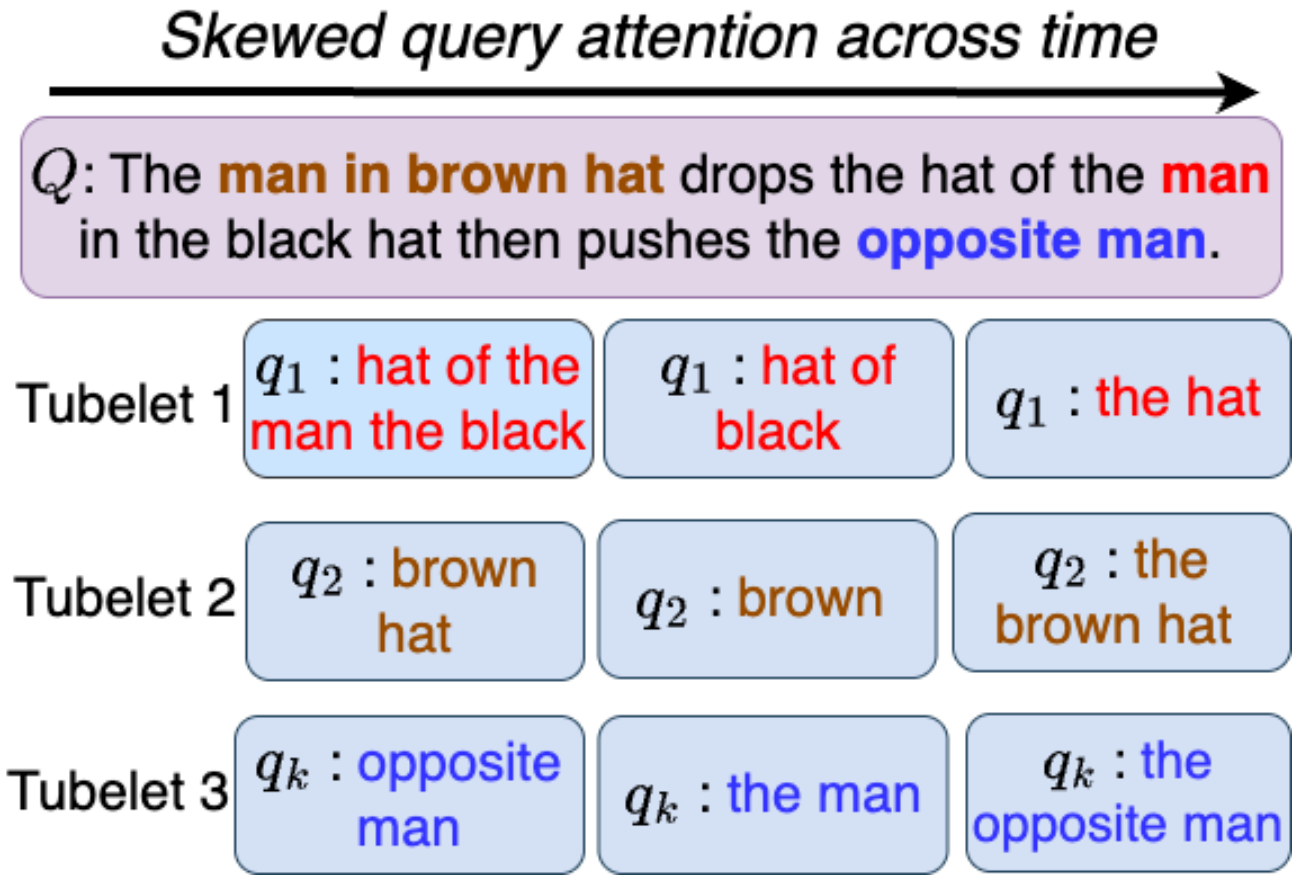
Tubelet Phrase Grounding (TPG)



First step towards adaptation without using any labels.

Can we refine the query for VLMs?

- **Aim:** Extract referential information from free-text query
- **Why:** Imbalanced Query Attention



Contextual Referral Grounding (CRG)

- Proposed Solution
  - Utilize LLMs for original query
  - Three sub-parts:
    - Spatial
      - Referral tubelet + attributes
    - Temporal
      - Referral tubelet action verb
    - Background Information
      - Dump it

Contextual Referral Grounding (CRG)

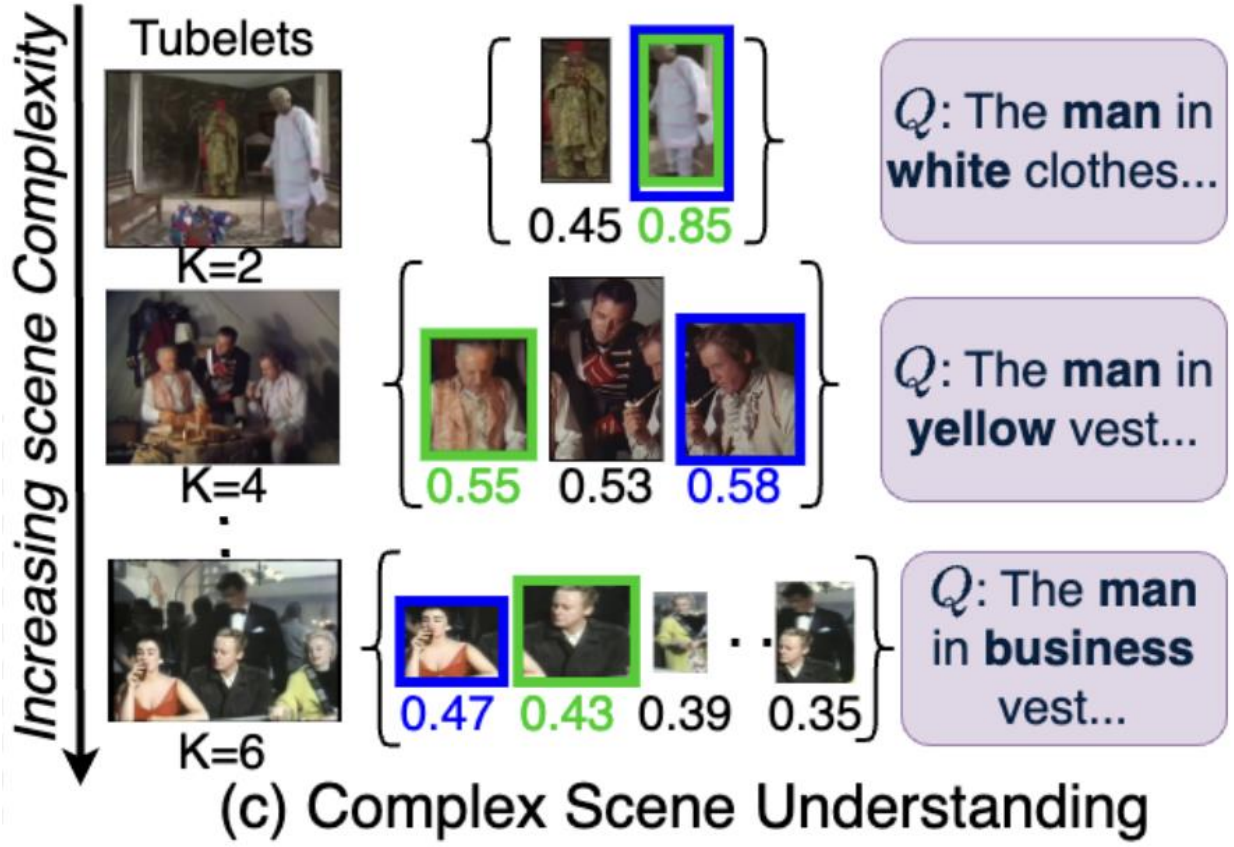
Noun	Adjective	Verb
Man	white	bends, looks, sits
Global Query		

Q: The man in white clothes bends down...

Local Query		
Attribute	**Verbs**	**Background**
The man in white clothes	He bends down. He looks down. He sits.	next to him
The man in white clothes bends down The man in white cloth looks down The man in white cloth sits		

Self-paced Scene Understanding(SPS)

- **Aim:** Coarse-to-fine scene understanding
- Why?
  - VLMs doesn't generalize well to complex scenes
    - Loose attention



Datasets Stats & Evaluation Metrics

- Datasets
  - HCSTVG-v1 – 5.5k videos
  - HCSTVG-v2 – 16k+ videos
  - VidSTG – 100k videos
    - Declarative
    - Interrogative
  - Both Datasets are difficult in their own ways
    - <2% improvement in 2 years on fully supervised.
- Evaluation Metrics:
  - mean vIoU: Spatio-Temporal Overlap
  - mean tIoU: Temporal Overlap

Comparison with Fully/Weakly-Supervised

Methods	HCSTVG - v1 tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	HCSTVG - v2 tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Fully-Supervised								
STGVT [TCSVT20] (Tang et al., 2020)	-	18.2	26.8	9.5	-	-	-	-
STVGBert [ICCV21] (Su et al., 2021)	-	20.4	29.4	11.3	-	-	-	-
TubeDETR [CVPR22] (Yang et al., 2022)	43.7	32.4	49.8	23.5	53.9	36.4	58.8	30.6
STCAT [NeurIPS22] (Jin et al., 2022)	49.4	35.1	57.7	30.1	-	-	-	-
CSDVL [CVPR23] (Lin et al., 2023)	-	36.9	62.2	34.8	58.1	38.7	65.5	33.8
CG-STVG [CVPR24] (Gu et al., 2024)	52.8	38.4	61.5	36.3	60.0	39.5	64.5	36.3
VGDINO [CVPR24] (Wasim et al., 2024)	-	38.3	62.5	36.1	-	39.9	67.1	34.5
Weakly-Supervised								
AWGU [ACMMM20] (Chen et al., 2020)	-	8.2	4.5	0.8	-	-	-	-
Vis-CTX [CVPR19] (Shi et al., 2019)	-	9.8	6.8	1.0	-	-	-	-
WINNER [CVPR23] (Li et al., 2023)	-	14.2	17.2	6.1	-	-	-	-
W-GDINO (Ours-Baseline)	18.0	9.0	11.6	4.6	23.3	9.9	13.3	5.6
CoSPaL (Ours)	**41.2**	**22.1**	**31.8**	**19.6**	**48.6**	**22.2**	**31.4**	**18.9**
	(+23.2)	(+7.9)	(+14.6)	(+13.5)	(+25.3)	(+12.3)	(+18.1)	(+13.3)

Scaling to large-scale dataset (100k+ videos)

Methods	Declarative Sentences			Interrogative Sentences		
	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
Two-stage pipelines						
GroundeR[ECCV16] [40]+LCNet [IEEE17] [58]	7.85	7.96	3.02	6.43	6.58	2.92
MATN[CVPR18] [62]+LCNet [IEEE17] [58]	8.16	8.03	3.59	6.97	6.64	3.05
GroundeR[ECCV16] [40]+CPL[CVPR22] [64]	8.28	8.35	3.68	7.16	7.28	3.23
RAIR [CVPR21] [33]+CPL [CVPR22] [64]	8.67	8.72	4.01	7.68	7.71	3.58
Single-stage pipelines						
WSSTG [ACL19] [9]	8.85	8.52	3.87	7.12	6.87	2.96
AWGU [ACMM20] [5]	8.96	7.86	3.10	8.57	6.84	2.88
Vis-CTX [CVPR19] [42]	9.34	7.32	3.34	8.69	7.18	2.91
WINNER [CVPR23] [23]	11.62	14.12	7.40	10.23	11.96	5.46
VCMA [ECCV24] [16]	<u>14.45</u>	<u>18.57</u>	<u>8.76</u>	<u>13.25</u>	<u>16.74</u>	<u>7.66</u>
W-GDINO (Ours-Baseline)	10.69	13.02	7.83	9.87	12.16	6.71
CoSPaL (Ours)	**16.04**	**20.12**	**13.16**	**13.53**	**16.42**	**10.27**

Ablations – Impact of Phrase Grounding & Curriculum Learning

S	TSA	T	tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
✓			26.2	13.5	17.7	7.3
✓	✓		27.3	13.9	18.6	6.9
✓		✓	35.2	18.0	26.3	14.1
✓	✓	✓	37.6	19.2	28.8	15.3

Stages	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
I	34.1	17.7	26.0	14.4
II	36.2	18.5	27.0	14.8
III	38.2	20.1	28.5	17.6

Main Ablation study

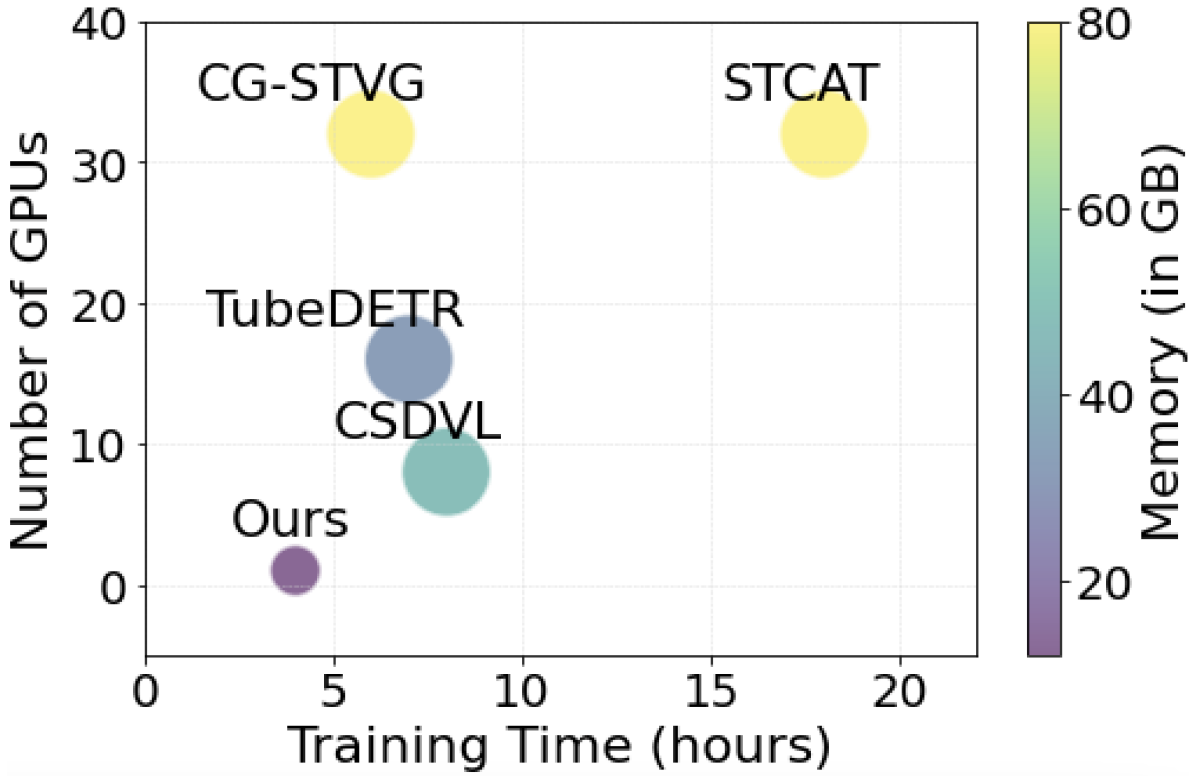
TPG	CRG	SPS	tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
			18.0	9.0	11.6	4.6
✓			37.6	19.2	28.8	15.3
	✓		35.8	20.2	30.5	17.6
✓	✓		37.8	21.0	31.7	16.8
✓		✓	38.2	20.1	28.5	17.6
	✓	✓	38.1	21.1	30.7	18.4
✓	✓	✓	41.2	22.1	31.8	19.6
			(+23.2)	(+13.1)	(+20.2)	(+15.0)

Discussion #1: Impact of Detector backbone

Methods	Detector	m_vIoU	vIoU@0.3	vIoU@0.5
WINNER	Faster-RCNN	11.6	14.1	7.4
CoSPaL	Faster-RCNN	16.4	23.7	11.1
CoSPaL	DETR	**22.1**	**31.8**	**19.6**

Outperform previous SOTA w/ similar backbone.

Discussion #2: Why weakly sup. matters?

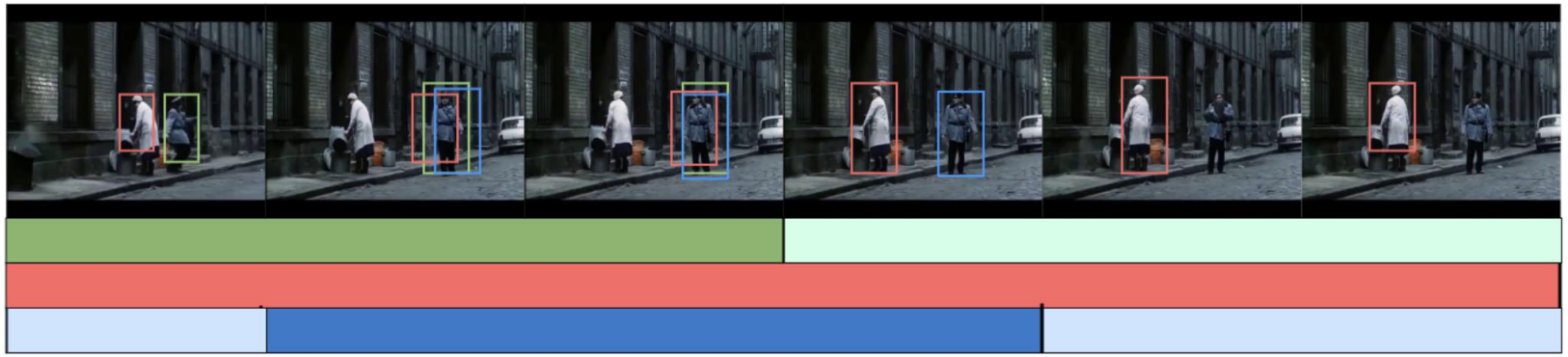


Weakly-sup. is efficient since all backbones are frozen.

Qualitative Analysis

The woman in blue clothes takes something on the table and wraps it around her wrists a few times.

The man in blue clothes runs to the side of the road, stops, turns around and speaks to the person in white , and then turns again.



Summary

- Contributions:
  - Developed first VLM for dense multimodal video detection task without any labels.
  - Make VLMs context aware via CoSPaL.
- Scalable to large datasets (minimal computation).

Visit our webpage!!!