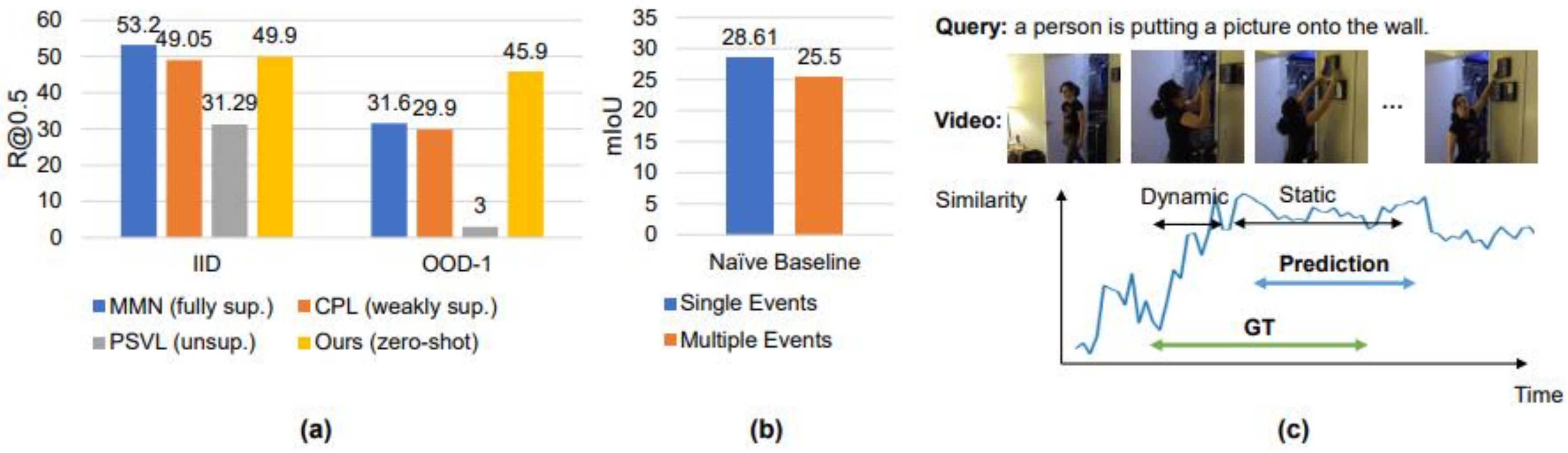
Training-free Video Temporal Grounding using Large-scale Pre-trained Models

Minghang Zheng[1], Xinhao Cai[1], Qingchao Chen[2], Yuxin Peng[1], and Yang Liu[1,3⋆]

[1] Wangxuan Institute of Computer Technology, Peking University
[2] National Institute of Health Data Science, Peking University
[3] State Key Laboratory of General Artificial Intelligence, Peking University
{minghang,qingchao.chen,pengyuxin,yangliu}@pku.edu.cn
xinhao.cai@stu.pku.edu.cn

Presenter : 김진용

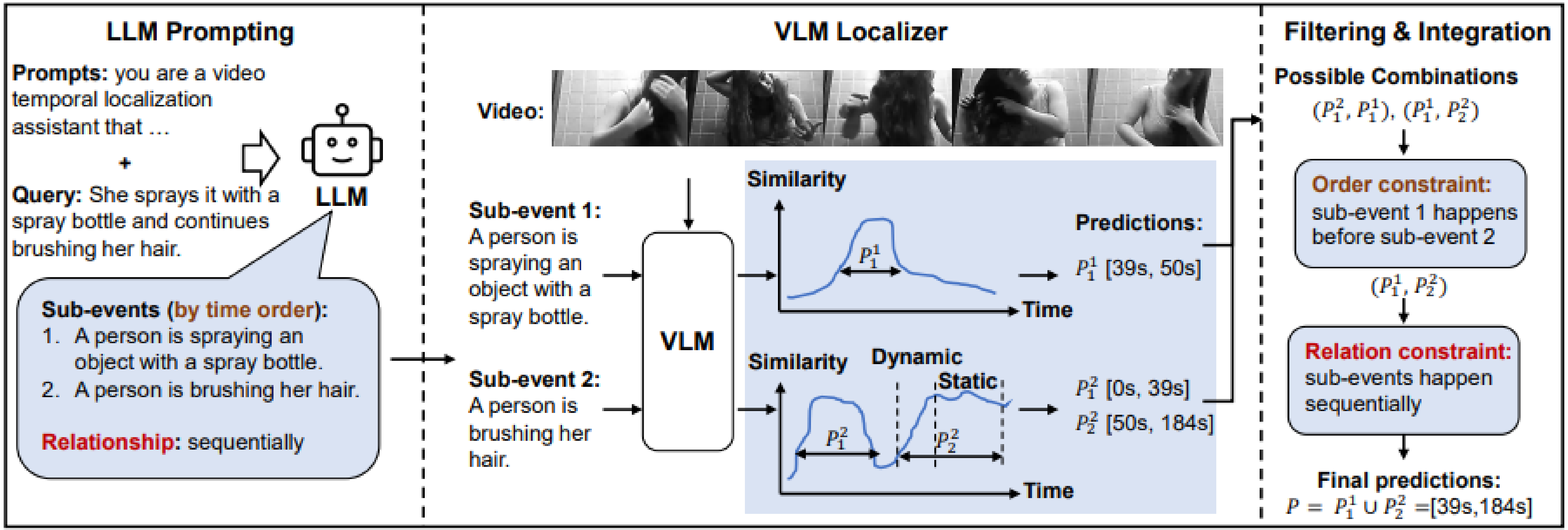
Introduction



Contribution

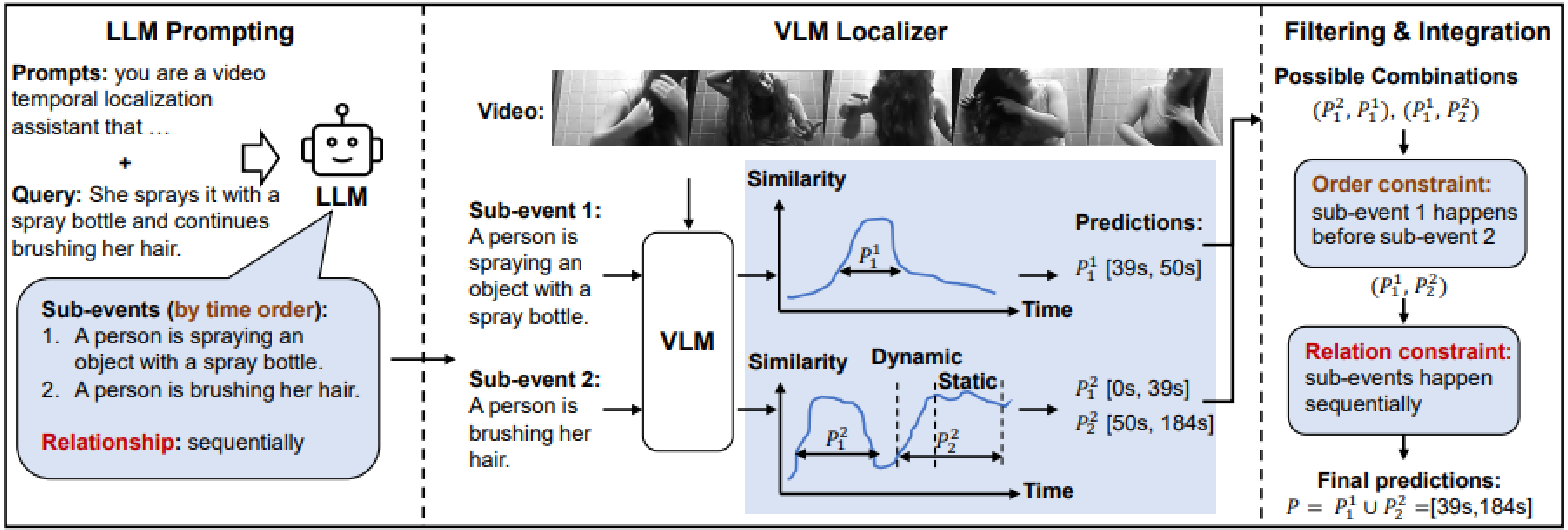
- We propose a training free pipeline for video temporal grounding using LLM and VLMs.
- To help VLM better understand the dynamic transitions in the video, we divide the events into dynamic and static parts and model them separately.
- Our method achieves the best performance on zero-shot temporal grounding on both the Charades-STA and Activity Net Captions datasets and has a greater advantage in cross-dataset and OOD settings.

Method



LLM Prompting
Prompts: you are a video temporal localization assistant that …
+
Query: She sprays it with a spray bottle and continues brushing her hair.
LLM
Sub-events (by time order):
1. A person is spraying an object with a spray bottle.
2. A person is brushing her hair.
Relationship: sequentially
VLM Localizer
Video:
Sub-event 1: A person is spraying an object with a spray bottle.
Sub-event 2: A person is brushing her hair.
VLM
Similarity
Time
Similarity
Dynamic
Static
Time
Predictions:
P_1^1 [39s, 50s]
P_1^2 [0s, 39s]
P_2^2 [50s, 184s]
Filtering & Integration
Possible Combinations
(P_1^2, P_1^1), (P_1^1, P_2^2)
Order constraint: sub-event 1 happens before sub-event 2
(P_1^1, P_2^2)
Relation constraint: sub-events happen sequentially
Final predictions:
$P = P_1^1 \cup P_2^2$ =[39s,184s]

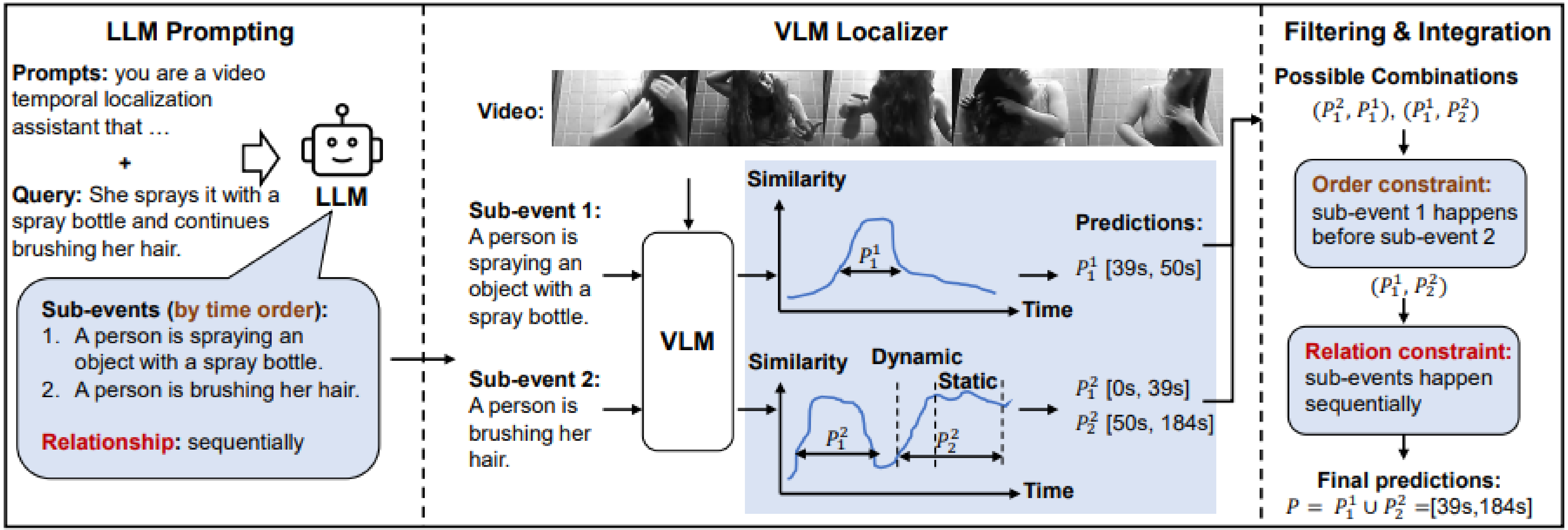
Method



What does LLM do for VTG?

- LLM analyze events in user's query and divide the events into sub-events.
- Infer order and relationships of the sub-events.
- Example)
  - Order : A->B
  - Relationships : single, sequentially, simultaneously

Method



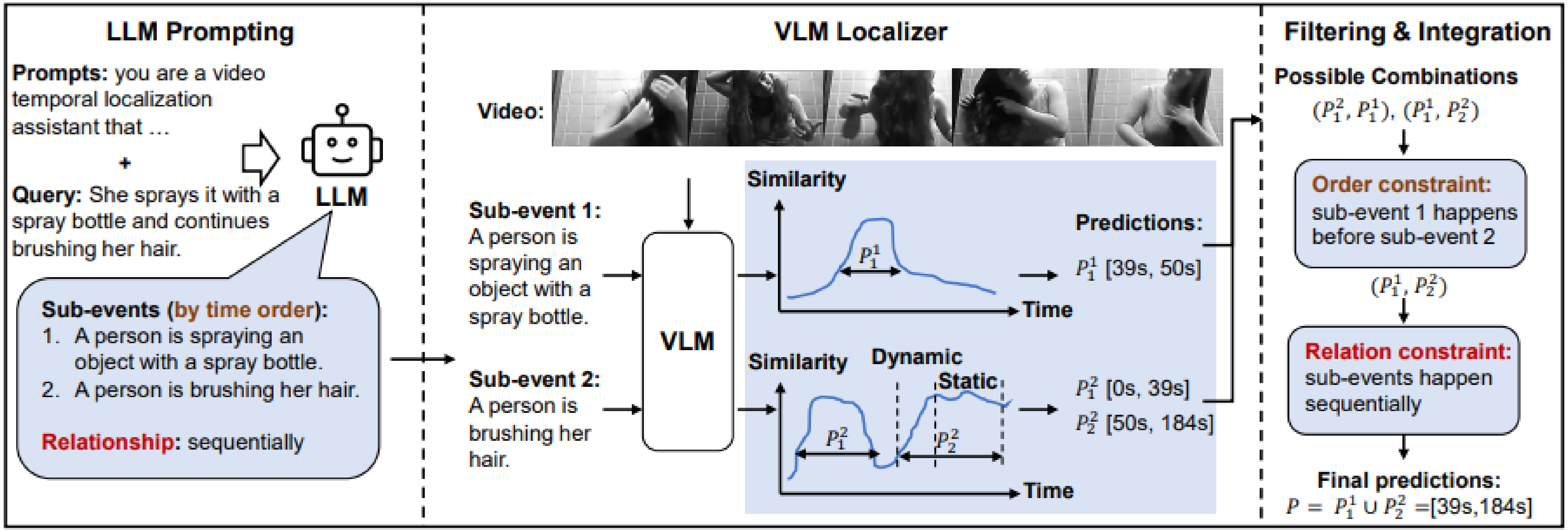
What does VLM Localizer do for VTG?

- **Calculate (cosine) similarity score between video frames and descriptions of sub-events.**

$$S = \frac{f^c F^{v\top}}{\|f^c\| \|F^v\|} \in \mathbb{R}^N$$

- f^c: text features of BLIP2 Q former, f^c: vision features of BLIP2 Q former, N: the number of frames

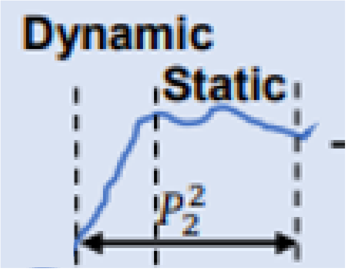
Method



What does VLM Localizer do for VTG?

- **Calculate (cosine) similarity score between video frames and descriptions of sub-events.**
- **Classify Static and Dynamic segments.**
  - **Localizer can't response sensitively during dynamic transitions.**
  - **For example, in given query "A person sits down", localizer tends to predict segments where the person is already seated on the chair rather than the process of the person gradually from standing up to sitting down.**

Method



Dynamic : increasing
Static : high score average

$$S_{i,k}^{dynamic} = \begin{cases} \sum_{l=i}^{k} D_l, & D_l > \delta, \forall l \in [i, k] \\ 0, & otherwise \end{cases}$$

- $D = S_i - S_{i-1}$, k: end stamptime of dynamic, i: start stamptime of dynamic
- **If score differential is over than threshold, the range is designated to dynamic segment.**
- **Dynamic score is defined to sum of score differential values in dynamic segement.**

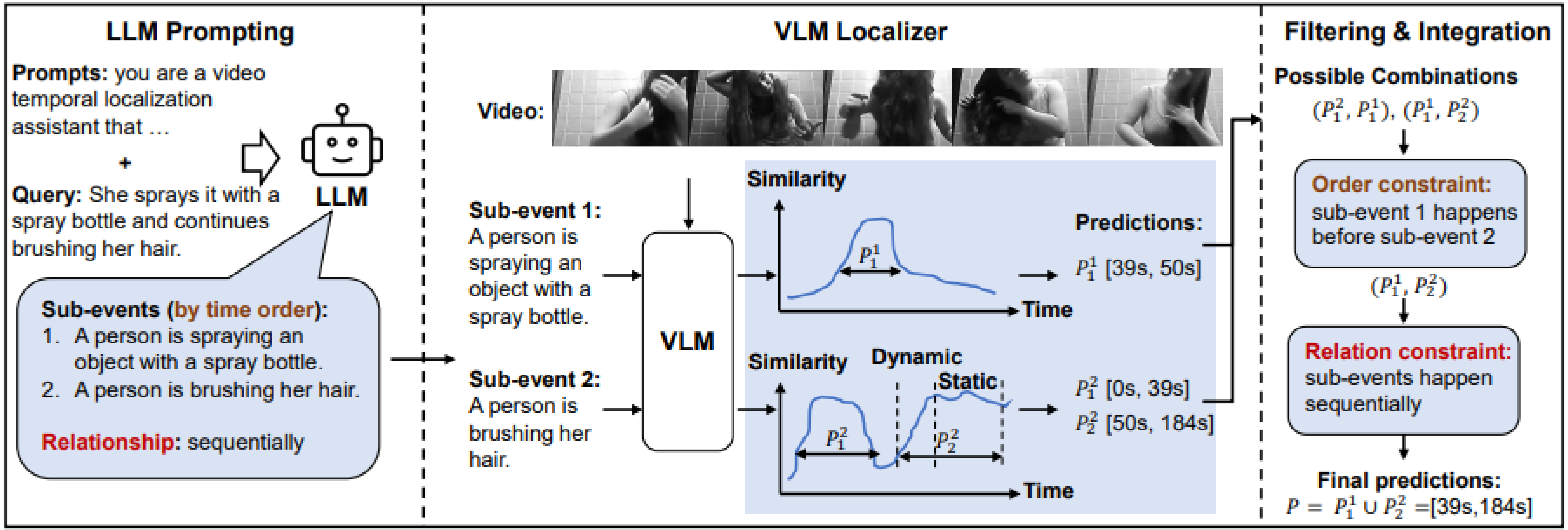
$$S_{k,j}^{static} = \frac{1}{j-k} \sum_{l \in [k,j]} S_l - \frac{1}{N-(j-k)} \sum_{l \notin [k,j]} S_l$$

- k: end stamptime of dynamic, i: start stamptime of dynamic, j: end stamptime of static
- **Static score is defined to subtraction of sum of score differential values in static segment and sum of score differential values not in static segment.**

$$S_{i,j}^{final} = \max_{k=i}^{j}(S_{i,k}^{dynamic} + S_{k,j}^{static})$$

- **“j” is determined by maximizing sum of dynamic score and static score.**

Method



What does Filtering & Integration do for VTG?

- **Filtering**
  - **Order constraint:** $\mathrm{P}^{1:\,\text{sub-event1}}_{1:\text{first segment}} \rightarrow \mathrm{P}^{2:\text{sub-event2}}_{2:\text{second segment}}$ **sub event1 happens before sub-event2**
- **Integration**

$$P^{final} = \begin{cases} P_1 \cap P_2 \cap \ldots \cap P_m, & \text{relation is `simultaneously'} \\ P_1 \cup P_2 \cup \ldots \cup P_m, & \text{relation is `sequentially'} \end{cases}$$

Experiments

Datasets
- **Activity Net Captions**
- **Charades-STA**

Implementation Details
- **VLM : BLIP2 Q-Former**
    - **3 FPS of videos**
    - $\delta = 5 \times 10^{-4}$
- **LLM : GPT4-Turbo**

Experiments

Method	Setting	VLM	LLM	Charades-STA R@0.3	R@0.5	R@0.7	mIoU	ActivityNet Captions R@0.3	R@0.5	R@0.7	mIoU
2D-TAN [57]	fully	✗	✗	-	39.81	23.25	-	58.75	44.05	27.38	-
EMB [11]				**72.50**	58.33	39.25	**53.09**	**64.13**	44.81	26.07	**45.59**
MGSL-Net [25]				-	63.98	41.03	-	-	51.87	31.42	-
EaTR [14]				-	**68.47**	**44.92**	-	-	**58.18**	**37.64**	-
CRM [12]	weakly	✗	✗	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM [60]				60.39	35.43	15.45	-	55.68	33.33	-	-
CPL [61]				66.40	49.24	22.39	-	55.73	31.37	-	-
Huang et al. [13]				**69.16**	**52.18**	**23.94**	**45.20**	**58.07**	**36.91**	-	**41.02**
Gao et al. [8]	unsup. [5]	✓	✗	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL [33]				46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR [39]				46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al. [15]				52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
SPL [59]				**60.73**	**40.70**	**19.62**	**40.47**	**50.24**	27.24	15.03	**35.44**
GroundingGPT [24]	fully [6]	✓	✓	-	29.6	11.9	-	-	-	-	-
VTimeLLM-13B [10]				**55.3**	**34.3**	**14.7**	**34.6**	**44.8**	**29.5**	**14.2**	**31.4**
VideoChat-7B [22]	zero-shot	✓	✓	9.0	3.3	1.3	6.5	8.8	3.7	1.5	7.2
VideoLLaMA-7B [55]		✓	✓	10.4	3.8	0.9	7.1	6.9	2.1	0.8	6.5
VideoChatGPT-7B [30]		✓	✓	20.0	7.7	1.7	13.7	26.4	13.6	6.1	18.9
Luo et al. [28]		✓	✗	56.77	42.93	20.13	37.92	48.28	27.90	11.57	32.37
VTG-GPT [46]		✓	✓	59.48	43.68	**25.94**	39.81	47.13	28.25	12.84	30.49
Ours w/o LLM		✓	✗	65.46	48.01	22.07	43.37	48.84	26.64	13.10	33.61
Ours		✓	✓	**67.04**	**49.97**	24.32	**44.51**	**49.34**	**27.02**	**13.39**	**34.10**

Table 1: Evaluation Results on the Charades-STA and ActivityNet Captions Datasets.

Experiments

Method	Setting	Charades-STA						ActivityNet-Captions					
		OOD-1			OOD-2			OOD-1			OOD-2		
		R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
LGI [32]	fully	42.1	18.6	41.2	35.8	13.5	37.1	16.3	6.2	22.2	11.0	3.9	17.3
2D-TAN [57]		27.1	13.1	25.7	21.1	8.8	22.5	16.4	6.6	23.2	11.5	3.9	19.4
MMN [44]		31.6	13.4	33.4	27.0	9.3	30.3	20.3	7.1	26.2	14.1	**5.2**	20.6
VDI [27]		25.9	11.9	26.7	20.8	8.7	22.0	**20.9**	7.1	**27.6**	**14.3**	**5.2**	**23.7**
DCM [50]		**44.4**	**19.7**	**42.3**	**38.5**	**15.4**	**39.0**	18.2	**7.9**	24.4	12.9	4.8	20.7
CNM [60]	weakly	9.9	1.7	21.6	6.1	0.5	16.6	**6.1**	**0.4**	21.0	**2.5**	0.1	16.8
CPL [61]		**29.9**	**8.5**	**32.2**	**24.9**	**6.3**	**30.5**	4.7	**0.4**	**21.1**	2.1	**0.2**	**17.7**
PSVL [33]	unsup.	**3.0**	0.7	8.2	**2.2**	0.4	6.8	-	-	-	-	-	-
PZVMR [39]		-	**8.6**	**25.1**	-	**6.5**	**28.5**	-	**4.4**	**28.3**	-	**2.6**	**19.1**
Luo et al. [28]	zero-shot	40.3	18.2	38.2	38.9	17.0	37.8	18.4	6.8	21.1	**18.6**	7.4	20.6
Ours		**45.9**	**20.8**	**43.0**	**43.8**	**20.0**	**42.6**	**20.4**	**11.2**	**31.7**	18.5	**10.0**	**30.3**

Table 2: Results under OOD setting on the Charades and ActivityNet Dataset.

Method	Setting	Charades-CD			Charades-CG					
		test-ood			novel-composition			novel-word		
		R@0.3	R@0.5	R@0.7	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
2D-TAN [57]	fully	43.45	30.77	11.75	30.91	12.23	29.75	29.36	13.21	28.47
TSP-PRL [45]		31.93	19.37	6.20	16.30	2.04	13.52	14.83	2.61	14.03
SCDM [54]		**52.38**	**41.60**	**22.22**	27.73	12.25	30.84	-	-	-
VISA [19]		-	-	-	45.41	**22.71**	**42.03**	**42.35**	**20.88**	**40.18**
DeCo [49]		-	-	-	**47.39**	21.06	40.70	-	-	-
WSSL [6]	weakly	**35.86**	**23.67**	**8.27**	3.61	1.21	8.26	2.79	0.73	**7.92**
CPL [61]		-	-	-	**39.11**	**15.60**	**35.53**	**45.90**	**22.88**	-
SPL [59]	unsup.	62.96	38.25	15.53	-	-	-	-	-	-
Luo et al. [28]	zero-shot	-	-	-	40.27	16.27	-	45.04	21.44	-
Ours		**65.07**	**49.24**	**23.05**	**43.84**	**18.68**	**40.19**	**56.26**	**28.49**	**46.90**

Table 3: Results under OOD setting on the Charades-CD and Charades-CG Dataset.

Inserting a segment of random generated video at the beginning of test videos.

Experiments

Ablation Study

Method	R@1 R@0.5	R@1 R@0.7	R@5 R@0.5	R@5 R@0.7
SCDM [54]	15.91	6.19	54.04	30.39
2D-TAN [57]	15.81	6.30	59.06	31.53
Debias-TLL [3]	21.45	10.38	62.34	32.90
Ours	**49.97**	**24.32**	**83.5**	**42.2**

Table 4: Cross-dataset performance when training on ActivityNet captions and evaluate on Charades-STA.

	LLM prompting	VLM localizer	Filtering & Integration	R@0.5	R@0.7	mIoU
1	✗	✗	✗	42.32	18.91	31.61
2	✓			43.17	18.56	32.14
3	✓		✓	44.12	19.21	33.07
4		✓		48.01	22.07	43.37
5	✓	✓		48.41	21.94	42.76
6	✓	✓	✓	**49.97**	**24.32**	**44.51**

Table 5: Ablations on each component.

Dynamic Scoring	Static Scoring	R@0.5	R@0.7	mIoU
✗	✗	42.32	18.91	31.61
✓		47.63	20.13	41.68
	✓	45.48	22.02	41.81
✓	✓	**48.01**	**22.07**	**43.37**

Table 6: Ablations on VLM localizer.

Order Constraint	Relation Constraint	R@0.5	R@0.7	mIoU
✗	✗	42.32	18.91	31.61
✓		43.01	19.03	31.73
	✓	43.97	19.11	32.76
✓	✓	**44.12**	**19.21**	**33.07**

Table 7: Ablations on LLM prompting.

Experiments

Ablation Study

VLMs	Type	R@0.5	R@0.7	mIoU
CLIP [34]	Image	42.68	18.92	38.89
BLIP-2 [20]		**48.01**	**22.07**	**43.37**
InterVideo [43]	Video	44.60	20.51	40.72
ViCLIP [42]		44.01	20.48	40.25

Table 8: Ablations on the VLMs.

LLMs	R@0.5	R@0.7	mIoU
None	48.01	22.07	43.37
Gemini-1.0-Pro [36]	48.97	22.76	44.12
GPT-3.5 Turbo	49.23	23.11	**44.69**
GPT-4 Turbo	**49.97**	**24.32**	**44.51**

Table 9: Ablations on the LLMs.