Weakly Supervised Video Scene Graph Generation via Natural Language Supervision

Kibum Kim[1], Kanghoon Yoon[1], Yeonjun In[1], Jaehyeong Jeon[1], Jinyoung Moon[2], Donghyun Kim[3], Chanyoung Park[1]

KAIST[1], ETRI[2], Korea University[3]

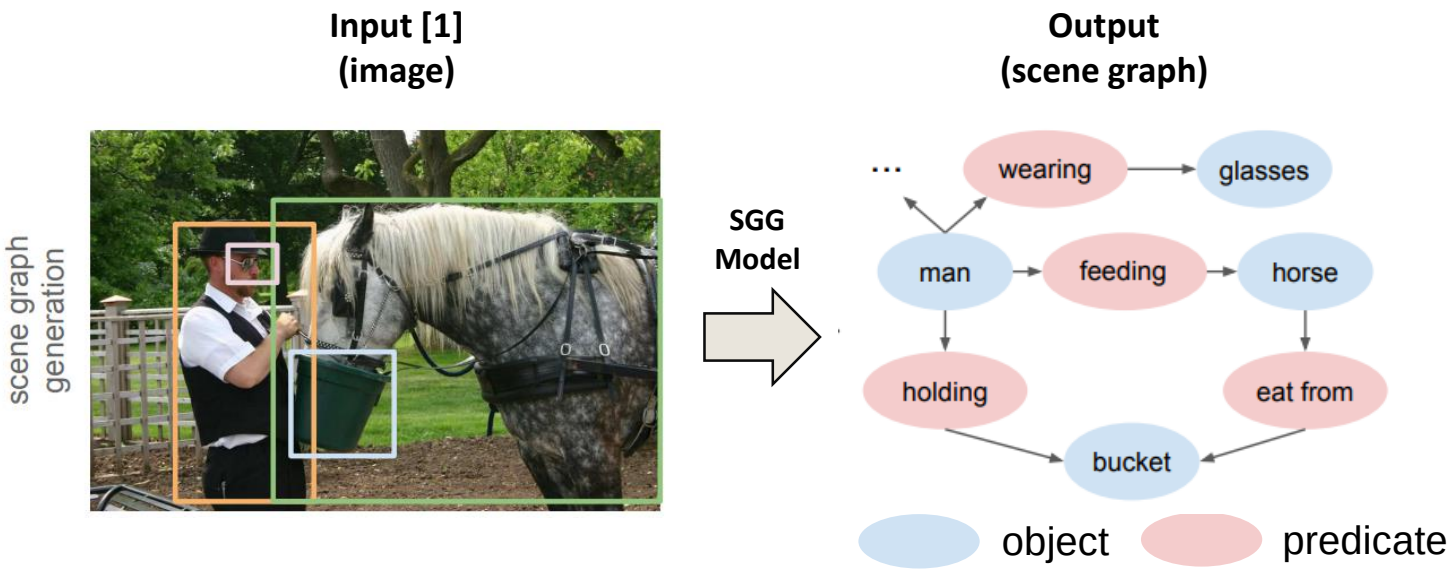
Presenter: Kibum Kim

CONTENT

- Introduction
  - Video Scene Graph Generation
- Method
- Experiment
- Summary

SCENE GRAPH GENERATION (SGG)

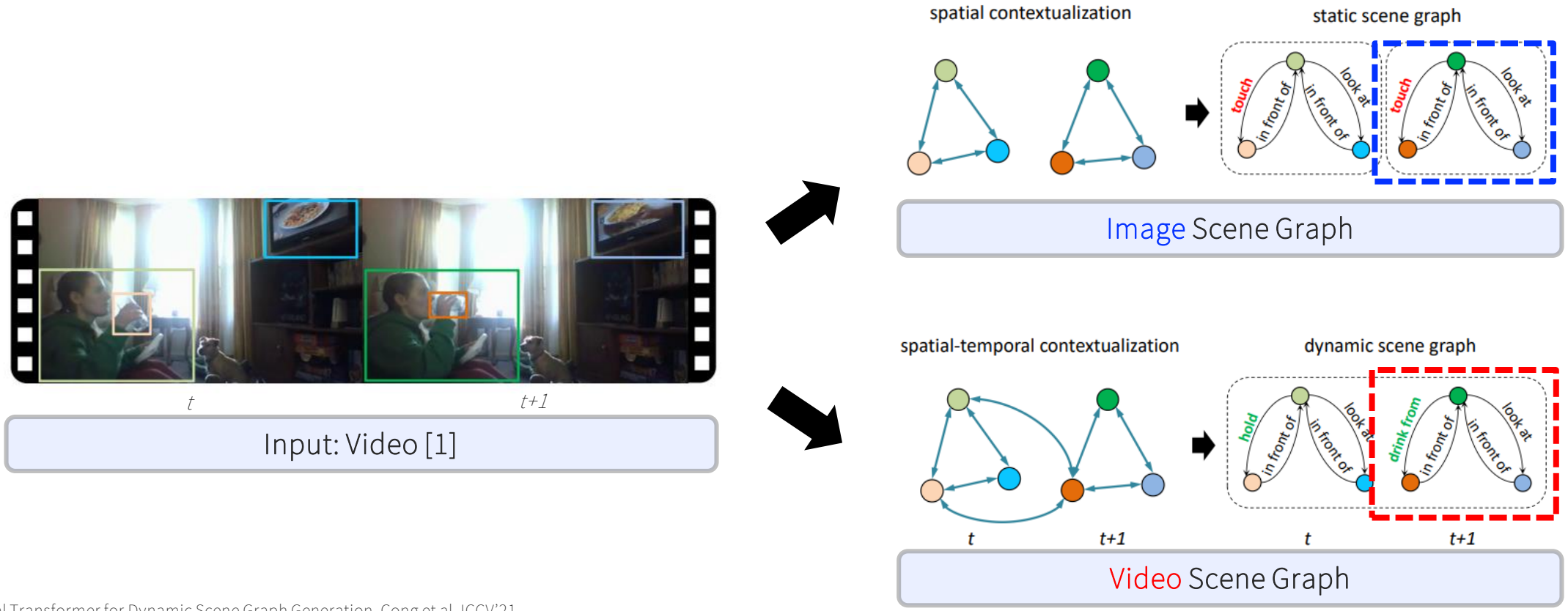
- SGG aims to represent observable knowledges in an image in the form of a graph
- The knowledge includes 1) object information and 2) their relation information, which is mapped to a scene graph
  - E.g., Object information: {*man, horse, glasses, bucket*}
  - E.g., Relationship information between objects: {*feeding, wearing, …, holding, eat from*}



[1] Scene Graph Generation by Iterative Message Passing. Danfei Xu et al. CVPR'2017

VIDEO SCENE GRAPH GENERATION

- Difference between Image scene graph and Video scene graph
  - Image SG: Considering only spatial context → Prediction of touch predicate
  - Video SG: Considering both spatial and temporal context → Prediction of drink from predicate

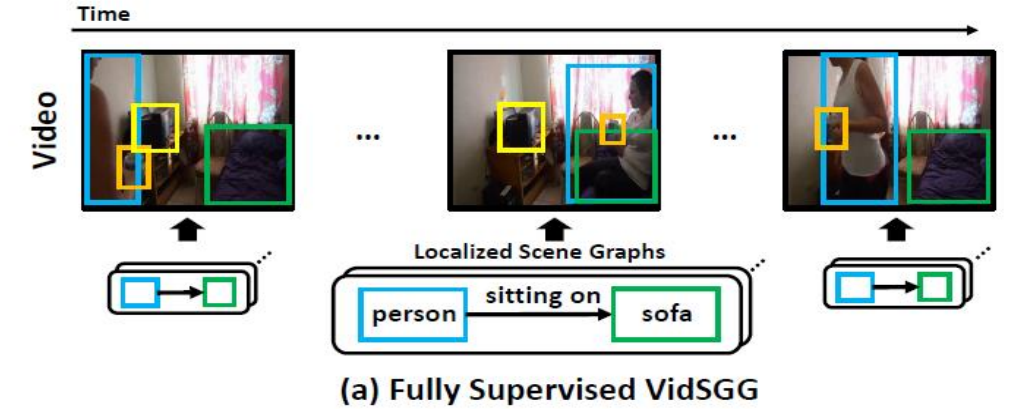


[1] Spatial-Temporal Transformer for Dynamic Scene Graph Generation. Cong et al. ICCV'21

Weakly Supervised Video Scene Graph Generation via Natural Language Supervision

MOTIVATION

- Limitation of Fully-supervised approach
  - Training the model relying on costly annotations → Challenges in extending large-scale datasets
  - Costly annotations (Figure a): each frame includes bounding boxes, the corresponding entity classes, and predicate information (i.e., localized SG)

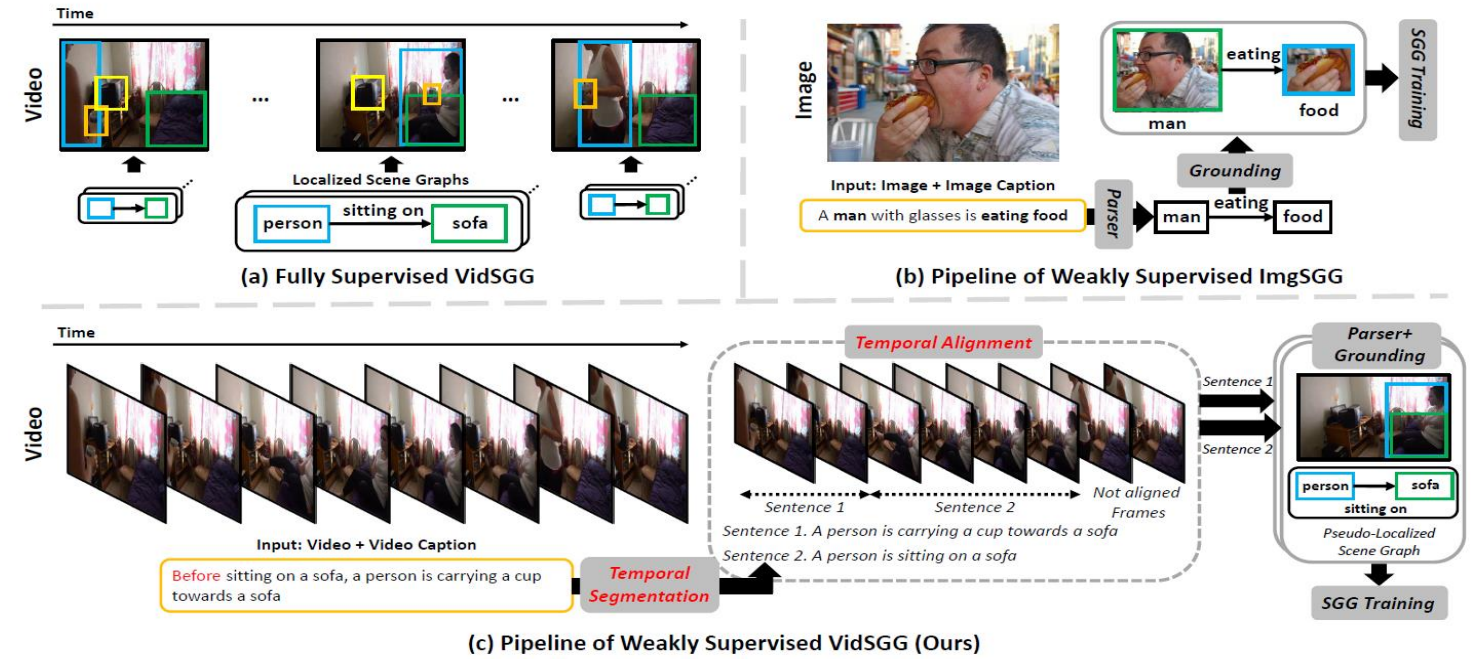


Costly annotations for Video SGG

- Goal: Alleviate reliance on costly annotations by training a Video SGG model without localized SG for each frame

MOTIVATION

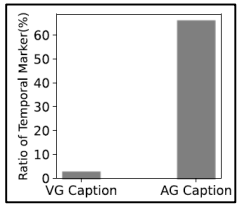
- Question: What if we apply a weakly supervised image SGG approach (Figure b) to the weakly supervised Video SGG?
- Naïve approach to adopt a weakly supervised image SGG
  - 1. Parsing the video caption into triplets
  - 2. Assigning triplets to frames within the same segment
    - <person, sitting on, sofa> : 1~4 frame
    - <person, carrying on, cup>: 5~8 frame



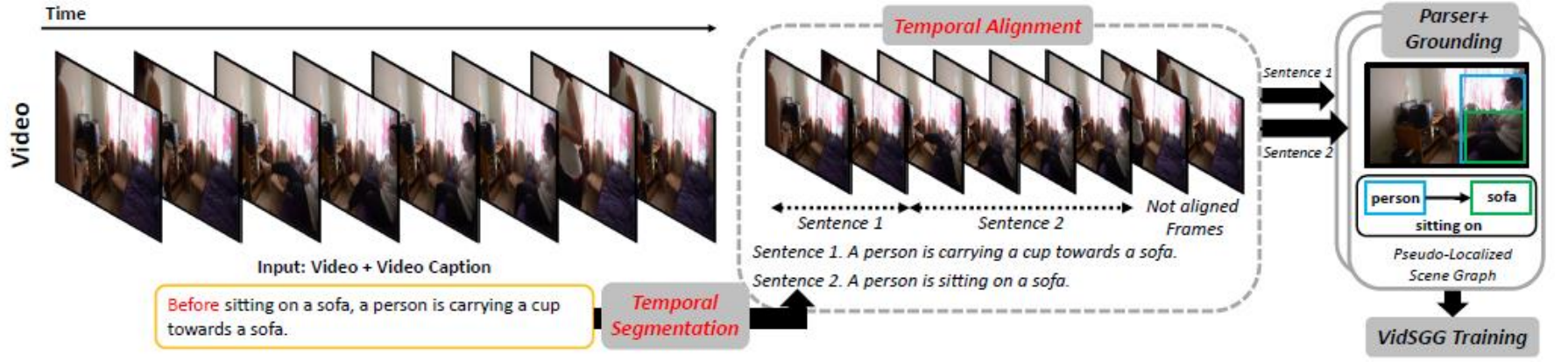
Motivation Figure

MOTIVATION

- Challenge in applying the weakly supervised image SGG approach
  - 1. Temporality within Video caption: unlike image captions, video captions include temporal marker (e.g., before, while)
    - Without considering this, the model may get incorrect supervision—such as mistakenly identifying the person as sitting on a sofa in the earlier frames
    - Actually, the video caption contains contain numerous temporal markers



  - 2. Variability in Action Duration: the various action described in a video caption do not occur within the same interval.
    - Without considering it, incorrect supervision may be provided —such as falsely indicating that the person is sitting in the last frame

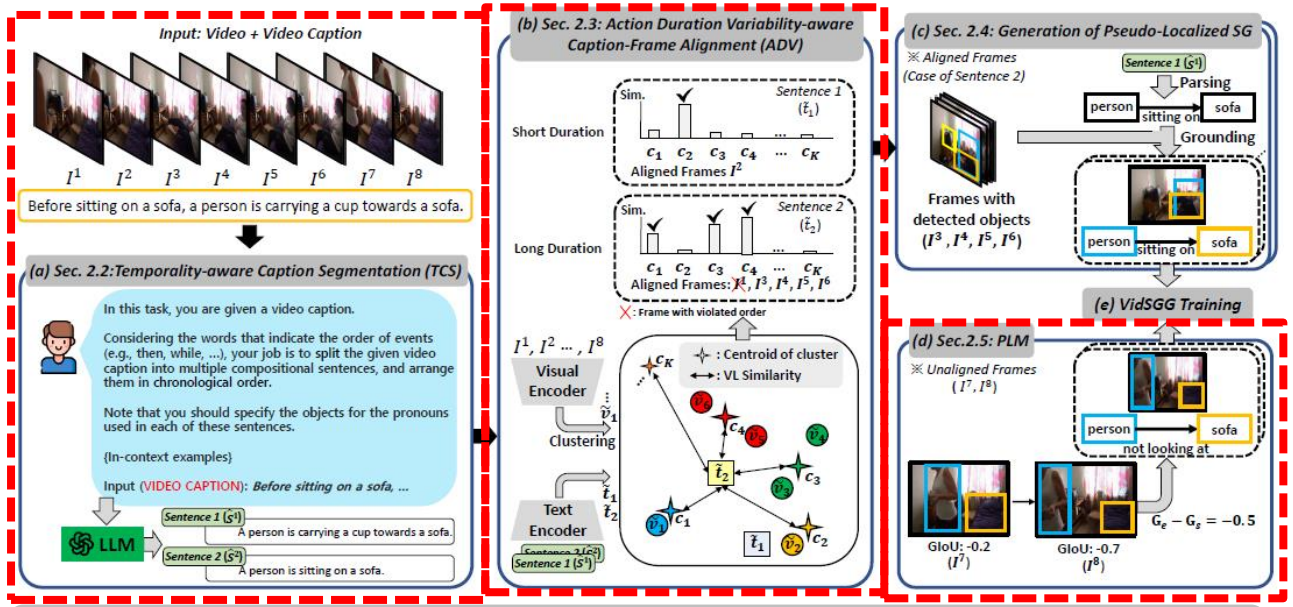


Overall framework of Weakly Supervised Video SGG

We propose a weakly supervised video SGG approach that uses caption-based supervision, addressing two key challenges

METHOD: OVERVIEW

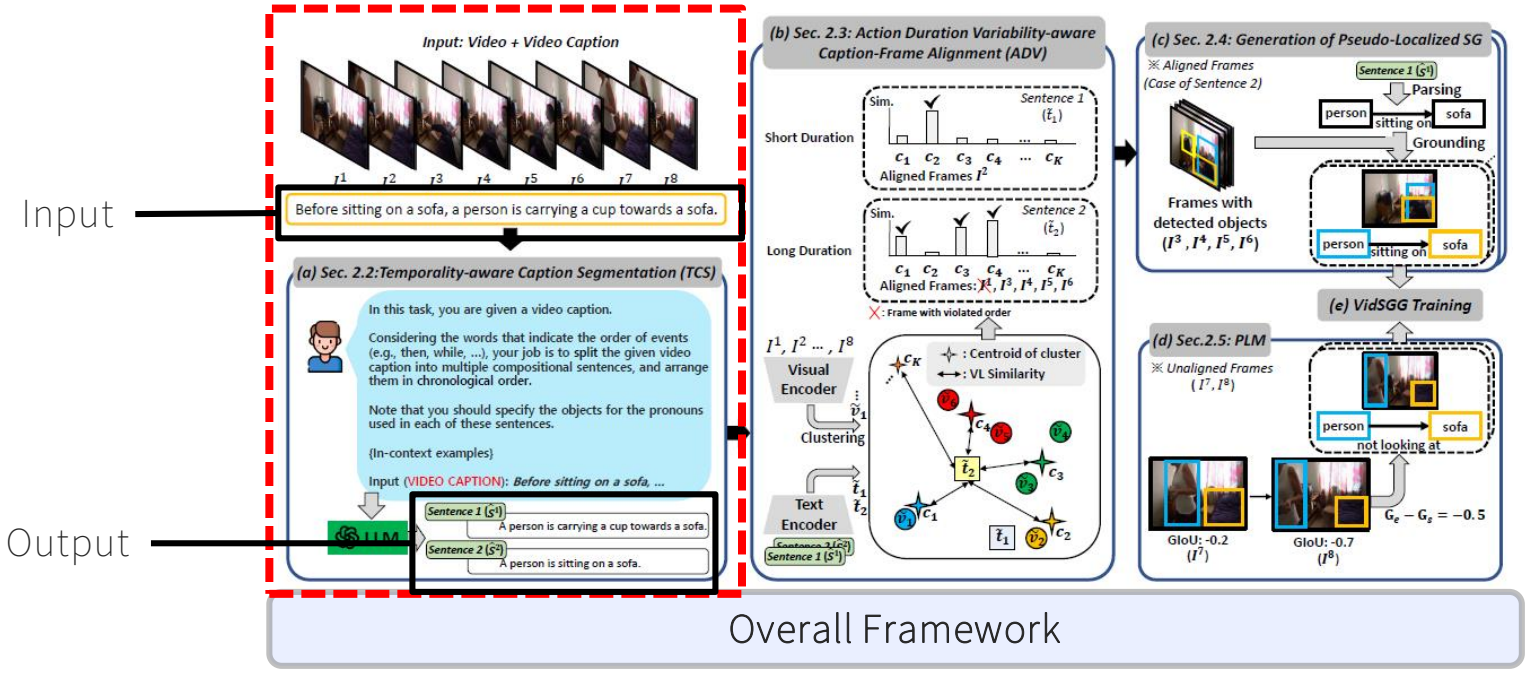
- 1. Temporal-aware Caption Segmentation (TCS): Segmenting video captions in temporal order by taking their inherent temporality into account
  - Leverage the LLM
- 2. Action Duration Variability-aware Caption-Frame Alignment (ADV): A module designed to account for action variability
  - Using the VL score from vision-language model
- 3. Pseudo-Labeling strategy based on Motion Cue (PLM)
  - A module for pseudo-labeling of negative classes
  - Captions typically do not contain negation



Overall Framework

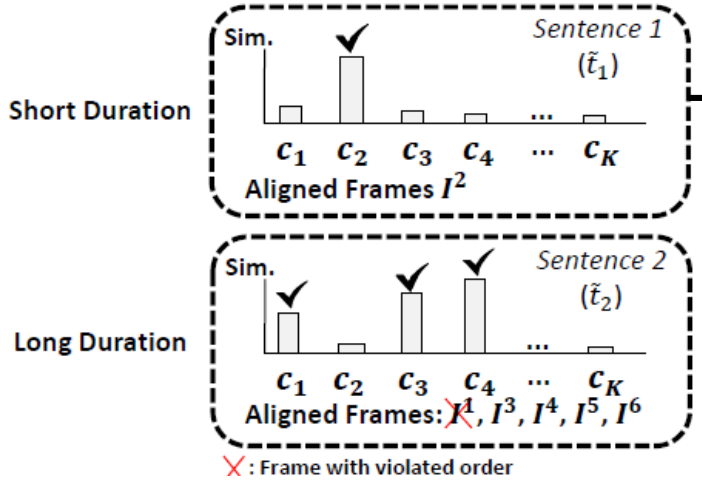
METHOD: TCS MODULE (1/4)

- 1. Temporal-aware Caption Segmentation (TCS)
- Segmenting the video caption into multiple sentences in temporal order, based on LLM's understanding of temporality
  - Explicitly state in the task description part of the prompt to consider the temporal marker.
  - Input: Video caption, Output: Segmented Sentence

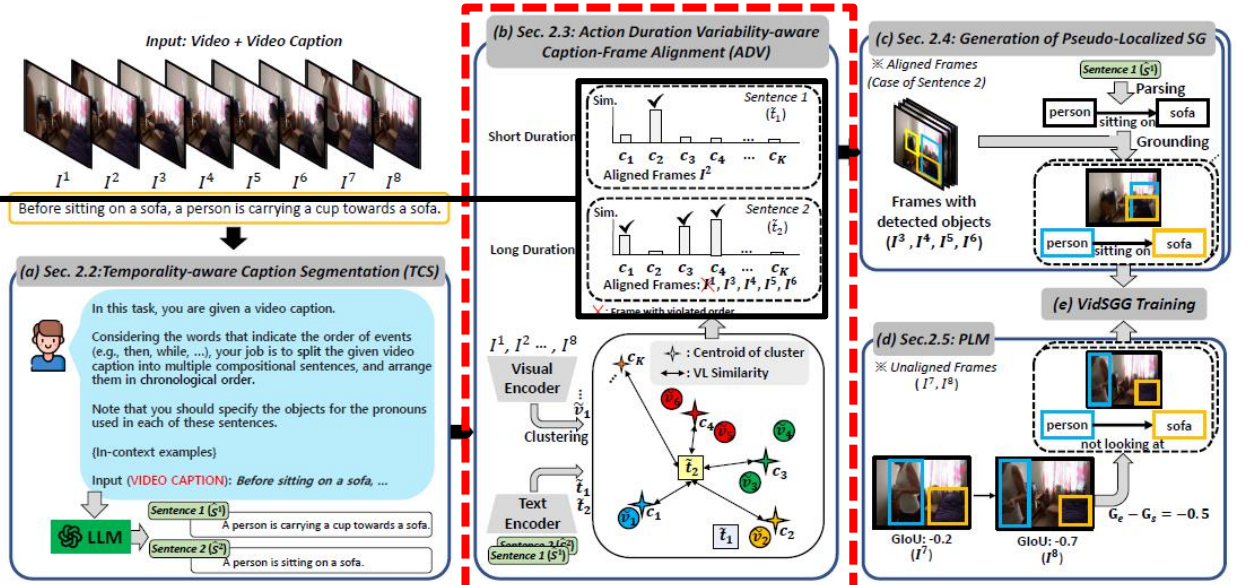


METHOD: ADV MODULE (2)

- 2. Action Duration Variability-aware Caption-Frame Alignment (ADV)
- Accounting for action variability using the VL score distribution from the Vision-Language (VL) model
  - 1) Generating proposals by applying *K*-means clustering to the visual features of each frame extracted by visual encoder
  - 2) Computing the similarity between each cluster's centroid and the text feature of the segmented sentence
  - 3) Assigning each segmented sentence to the corresponding cluster based on the distribution of VL scores (similarity scores)
    - Assigning based on the point with the largest drop in score



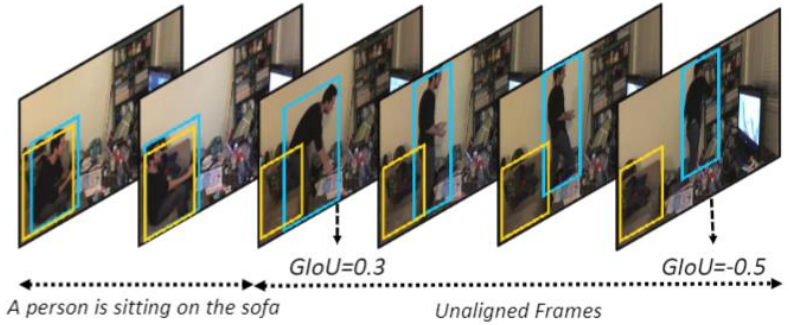
Frame Assignment based on the action variability



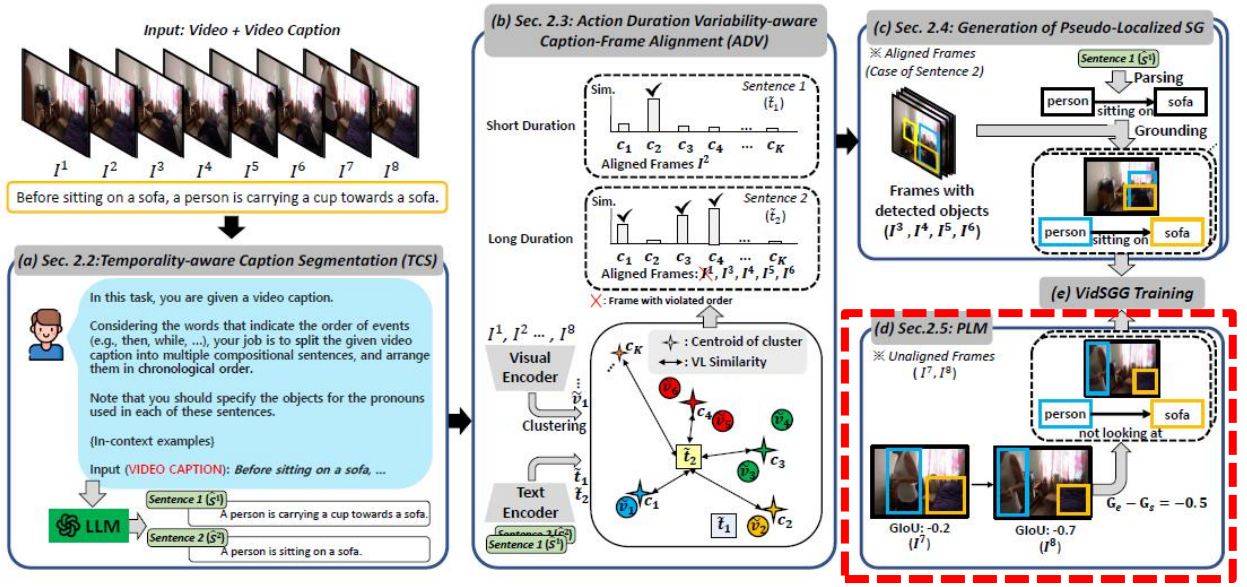
Overall Framework

METHOD: PLM MODULE (3/4)

- 3. Pseudo-Labeling strategy based on Motion Cue (PLM)
- Pseudo-labeling for negative classes using the GIoU metric
  - GIoU ↓: Indicates that the two bounding boxes are far apart, GIoU ↑: Indicates that the two bounding boxes are close to each other
- Key Idea: If a person quickly moves away from an object, it is likely that they are "not looking at" or "not contacting" the object
  - Assign negative classes only to unaligned frames where the GIoU difference is small



Motion Cue 를 이용한 Pseudo-labeling



Overall Framework

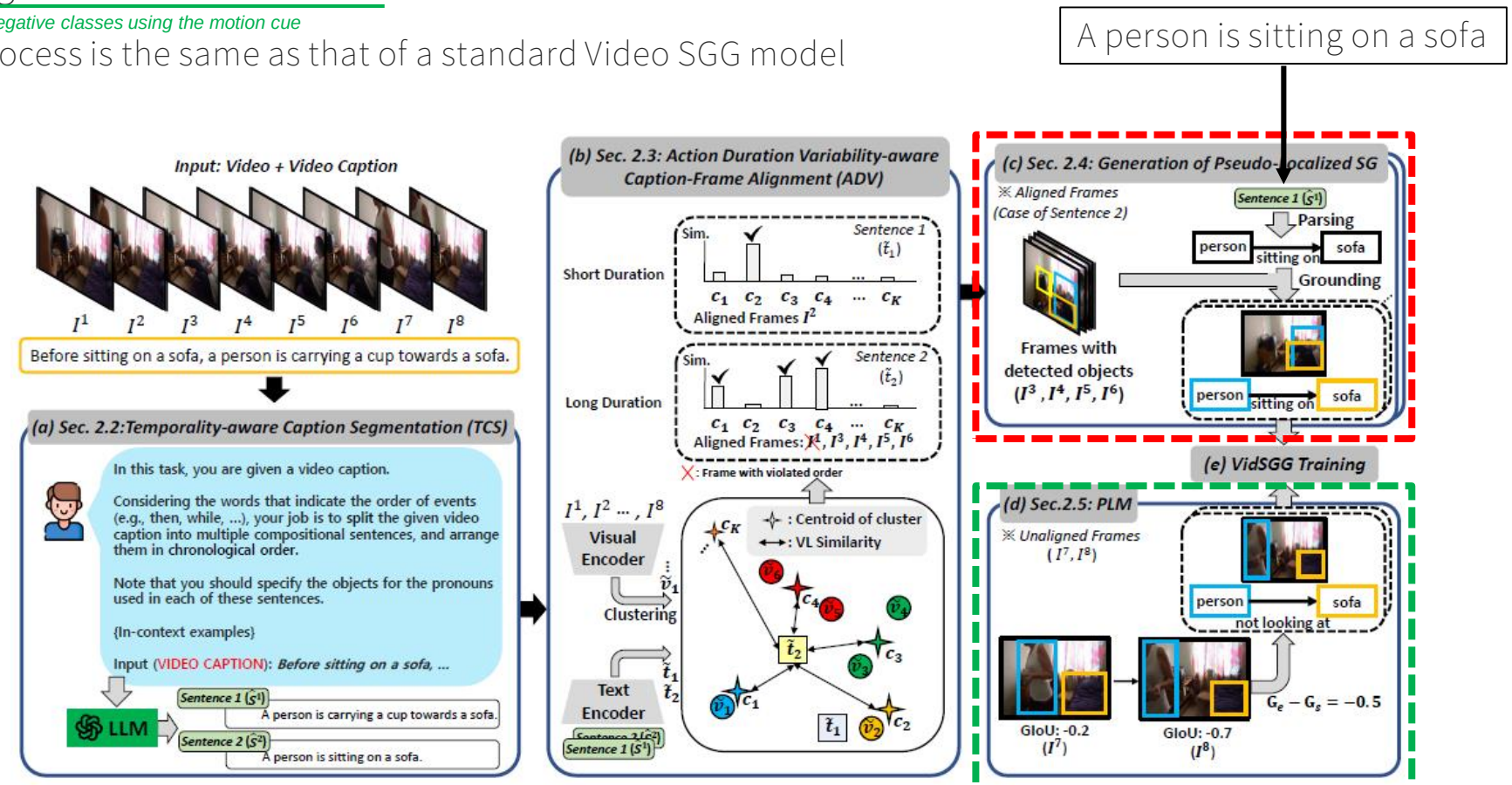
METHOD: MODEL TRAINING (4/4)

- Training the model using Pseudo-Localized Scene Graphs from the frames assigned to segmented sentences and Pseudo-Labeling from the PLM module

  Scene Graph extracted from the video captions

  Pseudo-labeling for negative classes using the motion cue

  - The training process is the same as that of a standard Video SGG model



Overall Framework

EXPERIMENT SETTING

- Evaluation Dataset: Action Genome
- Four type of Supervision
  - 1. Zero-shot: During evaluation, apply an image-based pretrained model (RLIP, RLIPv2) to each frame for inference
  - 2. Full: Training the model using the localized scene graphs (Ground-truth) in the Action Genome dataset
  - 3. Weak (GT Unlocalized SG): Using the GT unlocalized scene graphs of the middle frame (PLA [1] approach)
  - 4. Weaker (Natural Language): Only using the video caption for model training (Ours)

[1] Video Scene Graph Generation from Single-Frame Weak Supervision. Chen et al. ICLR'23

EXPERIMENTAL RESULTS (1/3)

- Performance comparison with baselines
  - 1. WS-ImgSGG vs. NL-VSGG: Our approach (NL-VSGG) outperforms the baseline of applying a weakly supervised image scene graph approach
  - 2. PLA_{cap} vs. NL-VSGG: Our approach outperforms the baseline that adjusts the same experimental settings
  - 3. PLA_{cap} vs. $\mathrm{PLA}_{simp.}$: The baseline heavily relies on the costly annotations of the middle frame
    - $\mathrm{PLA}_{simp.}$: Training the model without introducing any modules proposed by PLA

Backbone	Method	Supervision	With Constraint		No Constraint	
			R@20	R@50	R@20	R@50
RLIP	ImgSGG	Zero-shot	7.93	9.16	9.70	13.80
RLIPv2			8.37	10.05	14.60	21.42
STTran	Vanilla	Full	33.98	36.93	36.20	48.88
	+PLA	Weak (GT Unlocalized SG)	20.94	25.79	22.34	31.69
	+$\mathrm{PLA}_{simp.}$		20.42	25.43	21.72	30.87
	+WS-ImgSGG	Weaker (Natural Language)	10.01	12.83	9.02	14.05
	+PLA_{cap}		10.40	13.26	10.64	15.13
	+NL-VSGG		**15.61**	**19.60**	**15.92**	**22.56**
DSG-DETR	Vanilla	Full	34.80	36.10	40.90	48.30
	+PLA	Weak (GT Unlocalized SG)	21.30	25.90	22.70	31.90
	+$\mathrm{PLA}_{simp.}$		20.78	25.79	22.31	31.69
	+WS-ImgSGG	Weaker (Natural Language)	10.05	12.96	10.29	14.77
	+PLA_{cap}		10.36	13.53	10.57	15.41
	+NL-VSGG		**15.75**	**20.40**	**16.11**	**23.21**

Performance Comparison

EXPERIMENTAL RESULTS (2/3)

- Ablation studies
  - Each module demonstrates the effectiveness

Row	TCS	ADV	PLM	With Constraint R@20	With Constraint R@50	No Constraint R@20	No Constraint R@50	Mean R@20	Mean R@50
(a)				10.01	12.83	9.02	14.05	9.52	13.44
(b)	✓			11.09	14.66	11.34	16.70	11.22	15.68
(c)	✓	✓		11.98	15.58	11.93	17.36	11.96	16.47
(d)	✓	✓	✓	**15.61**	**19.60**	**15.92**	**22.56**	**15.77**	**21.08**

Ablation study

- Performance comparison over various video length
  - Our proposed method shows the effectiveness on the videos with various length
  - Our method can use readily available video caption dataset for model training (Scalability)

Training Dataset (Caption)	Method	Avg. Video Length	With Constraint R@20	With Constraint R@50	No Constraint R@20	No Constraint R@50	Mean
Action Genome	WS-ImgSGG	29.9 seconds	10.01	12.83	9.02	14.05	11.48
	NL-VSGG		**15.61**	**19.60**	**15.92**	**22.56**	**18.42**
MSVD	WS-ImgSGG	9.5 seconds	6.22	8.03	7.69	12.31	8.56
	NL-VSGG		**9.05**	**11.31**	**10.22**	**16.60**	**11.80**
ActivityNet	WS-ImgSGG	117.3 seconds	10.86	14.47	10.07	15.80	12.80
	NL-VSGG		**13.46**	**17.58**	**13.94**	**21.41**	**16.60**

Performance over various video length

EXPERIMENTAL RESULTS (3/3)

- Combining weakly supervised and fully supervised datasets
  - The performance of the model fine-tuned to the AG dataset outperforms that of the model initially trained on the AG dataset
  - It indicates that our proposed approach can synergize with the fully supervised approach

Training dataset	With Constraint		No Constraint		Mean
	R@20	R@50	R@20	R@50	
AG (Full)	33.98	36.93	36.20	48.88	39.00
AG+MSVD (Weak) → AG (Full)	**34.84**	**37.64**	**38.40**	**49.55**	**40.11**

Combining weakly and fully supervised approach

Qualitative results for broader range of action classes

- Qualitative results for broader range of action classes
  - It demonstrates that incorporating various action classes in video captions allows for predicting a broader range of action classes
  - E.g., standing on vs. walk into // holding vs. close, open

SUMMARY

- We propose the weakly supervised video SGG approach that trains the model using only the video captions
  - Therefore, it alleviates the reliance on expensive annotations for model training

- We identify the two key challenges of weakly supervised Video SGG: 1) Temporality within caption, 2) Variability in Action Duration

- To address two challenges, we propose TCS and ADV modules

- We demonstrates the effectiveness of each module on the Action Genome dataset. Furthermore, our approach highlights the potential to broaden the range of action classes and use the readily available video caption datasets.

Thank you