Fine-Grained Captioning of Long Videos through Scene Graph Consolidation

ComputerVisionLab Seoul National University

Sanghyeok Chu, Seonguk Seo, Bohyung Han

Computer Vision Lab, Seoul National University

ICML International Conference On Machine Learning

Motivation

Recent advances in VLMs have significantly improved captioning for images and short videos.
However, captioning longer videos remains challenging:

- Limited temporal receptive fields restrict holistic contextual understanding of long videos.
- Train models on long videos require large training datasets and substantial computational resources.
- LLM-based methods avoid training but have high inference costs and produce suboptimal results.

Key Idea

Our intuitions are straightforward:

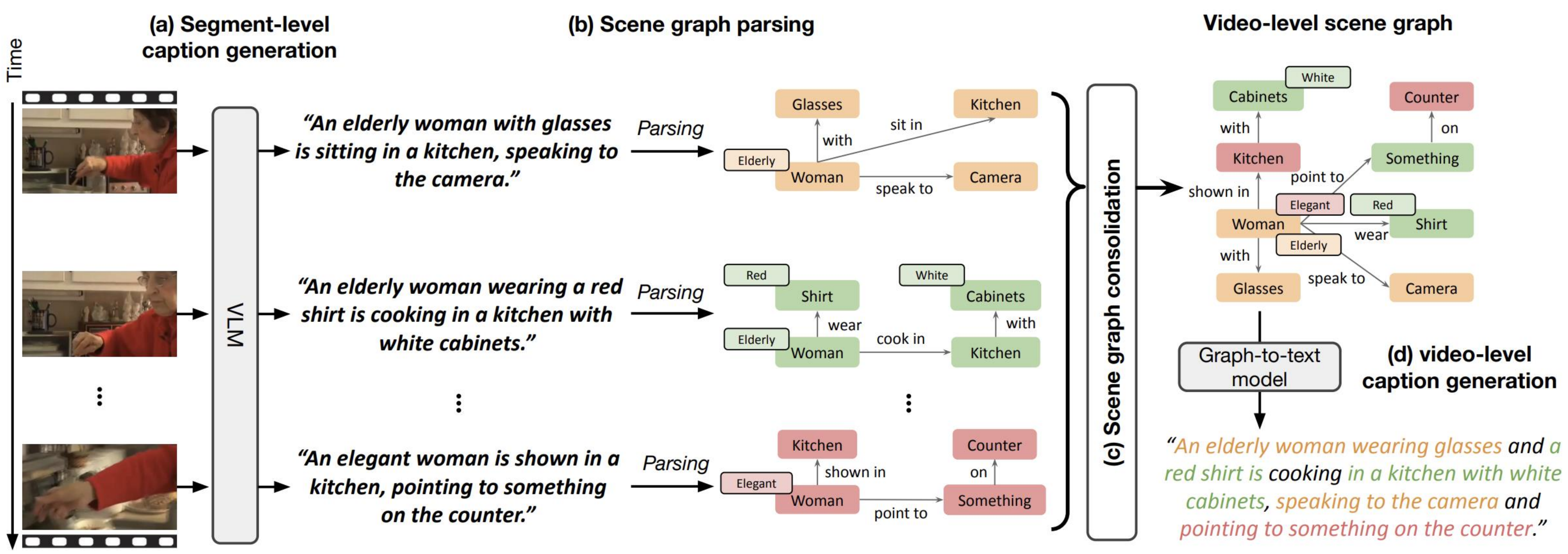
- Captioning short segments and long videos share common goals.
- High-quality information for each segments can be extracted by leveraging existing models.
- Video contain closely related contexts, motivating effective consolidation of local information.

  → Consolidate segment-level information using a graph structure!

Proposed Framework

Our framework: 1) is capable of generating fine-grained captions for long videos, 2) does not require any target dataset annotations, and 3) avoids high inference costs.

1. Segment-level caption generation: Generate captions for short segments using off-the-self VLMs.
2. Scene graph parsing: Convert segment captions into scene graphs using a textual scene graph parser.
3. Graph consolidation: Perform Hungarian Matching between two sets of object nodes from each graph.
4. Graph-to-text generation: Translate consolidated graph into a video caption using graph-to-text model.

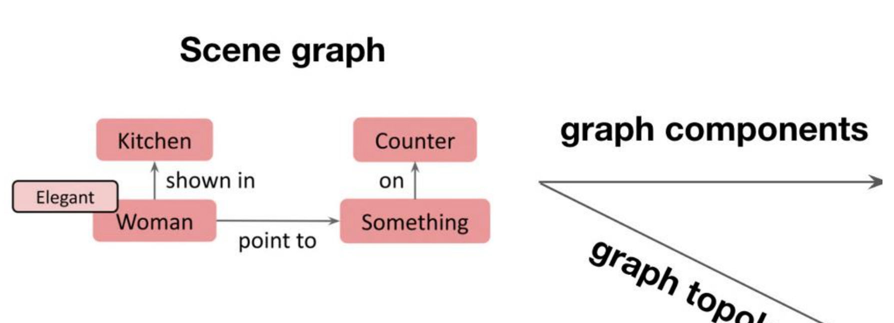


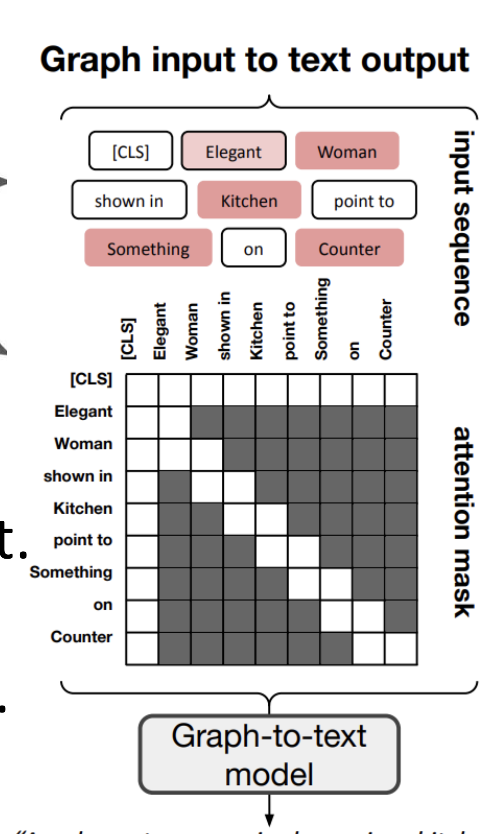
Algorithm 1 Scene graph consolidation

Input:
  $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$: set of scene graphs
  $\phi(\cdot)$: a graph encoder
  $\psi_i(\cdot)$: a function returning the i^{th} object in a graph
  π: a permutation function
  τ: a threshold
Output: G_{video}: a video-level scene graph
while $|\mathcal{G}| > 1$ **do**
  Retrieve the most similar pair $\{G^s, G^t\}$ from $\mathcal{G}$
  $G^s = (\mathcal{O}^s, \mathcal{E}^s)$, $G^t = (\mathcal{O}^t, \mathcal{E}^t)$
  $G^m = (\mathcal{O}^m, \mathcal{E}^m) \leftarrow (\mathcal{O}^s \cup \mathcal{O}^t, \mathcal{E}^s \cup \mathcal{E}^t)$
  $\pi^* \leftarrow \arg\max_{\pi \in \Pi} \sum_i \frac{\psi_i(\phi(G^s))}{\|\psi_i(\phi(G^s))\|} \cdot \frac{\psi_i(\phi(G^t_\pi))}{\|\psi_i(\phi(G^t_\pi))\|}$
  for $(p, q) \in \mathcal{M}$ such that $s_{p,q} > \tau$ **do**
    Set the class label of the merged object, $\hat{c}$
    $\hat{o}_m \leftarrow (\hat{c}, \mathcal{A}^s_p \cup \mathcal{A}^t_q)$
    $\mathcal{O}^m \leftarrow \{\hat{o}_m\} \cup \left(\mathcal{O}^m \setminus \{o^s_p, o^t_q\}\right)$
    Update $\mathcal{E}^m$: $e_{m,*} \leftarrow e_{p,*}$ and $e_{*,m} \leftarrow e_{*,q}$
  end for
  $\mathcal{G} \leftarrow \{G^m\} \cup (\mathcal{G} \setminus \{G^s, G^t\})$
end while
$G_{\text{video}} \leftarrow \text{extract}(\mathcal{G})$
return G_{video}

↑ **Graph Consolidation through Hungarian Matching**

Graph-to-Text Model





- Architecture: Graph encoder + Text decoder.
- Training dataset: 2.5M graph-text pair.
  - Constructed by parsing scene graph from given text.
  - Texts are collected from various sources, including image caption datasets and model-generated ones.
- Objective function: $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log P_\theta(t_i \mid t_{1:i-1}, G)$

Experiments

Zero-shot video captioning results

Table 1. Zero-shot video captioning results on the MSR-VTT (Xu et al., 2016) and MSVD (Chen & Dolan, 2011) test sets, comparing our method (SGVC) with LLM-based video understanding methods. † indicates that the method utilizes reference captions from the target dataset to construct few-shot exemplar prompts. Bold numbers indicate the highest scores among methods not using reference captions.

Dataset	Method	Backbone VLM	B@4	METEOR	CIDEr	P_{BERT}	R_{BERT}	F_{BERT}
MSR-VTT	VidIL (Wang et al., 2022b)	BLIP+CLIP	3.2	14.8	3.1	0.134	0.354	0.225
	VidIL† (Wang et al., 2022b)		13.6	20.0	20.2	0.461	0.552	0.490
	Video ChatCaptioner (Chen et al., 2023)	BLIP2	13.2	22.0	16.5	0.396	0.510	0.436
	SGVC (Ours)	BLIP	17.7	22.5	24.0	**0.476**	0.539	**0.490**
		BLIP2	**18.4**	**23.1**	**26.1**	0.467	**0.542**	0.487
MSVD	VidIL (Wang et al., 2022b)	BLIP+CLIP	2.5	16.5	2.3	0.124	0.404	0.238
	VidIL† (Wang et al., 2022b)		30.7	32.0	60.3	0.656	0.726	0.674
	Video ChatCaptioner (Chen et al., 2023)	BLIP2	22.7	31.8	35.8	0.496	0.651	0.550
	SGVC (Ours)	BLIP	22.6	30.2	50.2	**0.575**	0.646	0.589
		BLIP2	**25.3**	**32.0**	**53.3**	0.571	**0.669**	**0.597**

Table 2. Zero-shot video captioning results on the MSR-VTT (Xu et al., 2016) and MSVD (Chen & Dolan, 2011) test sets, comparing SGVC with the LLM summarization baseline. Bold numbers indicate the highest scores.

Dataset	Method	Backbone VLM	B@4	METEOR	CIDEr	P_{BERT}	R_{BERT}	F_{BERT}
MSR-VTT	Summarization w/ Mistral-7B	BLIP	9.6	21.6	10.8	0.313	0.516	0.395
		BLIP2	11.5	**23.1**	15.4	0.308	0.528	0.397
	SGVC (Ours)	BLIP	17.7	22.5	24.0	**0.476**	0.539	**0.490**
		BLIP2	**18.4**	**23.1**	**26.1**	0.467	**0.542**	0.487
MSVD	Summarization w/ Mistral-7B	BLIP	15.2	28.3	30.3	0.477	0.623	0.527
		BLIP2	22.5	31.9	41.6	0.500	0.664	0.558
	SGVC (Ours)	BLIP	22.6	30.2	50.2	**0.575**	0.646	0.589
		BLIP2	**25.3**	**32.0**	**53.3**	0.571	**0.669**	**0.597**

- (left) SGVC outperforms LLM-based video understanding when using the same VLM backbone.
- (right) Given the same set of captions, graph consolidation outperforms LLM summarization.

Zero-shot video paragraph captioning

Table 3. Zero-shot video paragraph captioning results on the ActivityNet Captions (Krishna et al., 2017a) *ae-val* set, comparing our method (SGVC) with LLM-based video understanding methods. † indicates that the method utilizes reference captions from the target dataset to construct few-shot exemplar prompts. Bold numbers indicate the highest scores among methods not using reference captions.

Method	Backbone VLM	B@4	METEOR	CIDEr	P_{BERT}	R_{BERT}	F_{BERT}
VidIL (Wang et al., 2022b)	BLIP+CLIP	1.0	5.8	4.6	0.122	0.135	0.125
VidIL† (Wang et al., 2022b)		2.9	7.6	3.3	0.414	0.243	0.323
Video ChatCaptioner (Chen et al., 2023)	BLIP2	2.4	8.9	1.6	0.207	0.202	0.200
SGVC (Ours)	BLIP	6.7	11.6	16.6	**0.367**	0.285	0.322
	BLIP2	**7.4**	**12.4**	**20.9**	**0.367**	**0.304**	**0.331**

Table 4. Zero-shot video paragraph captioning results on the ActivityNet Captions (Krishna et al., 2017a) *ae-val* set, comparing SGVC with the LLM summarization baselines. Bold numbers indicate the highest scores.

Method	Backbone VLM	B@4	METEOR	CIDEr	P_{BERT}	R_{BERT}	F_{BERT}
Summarization w/ Mistral-7B	BLIP	3.4	9.4	7.5	0.292	0.268	0.276
	BLIP2	4.1	10.4	9.6	0.307	0.293	0.295
	InternVL2.5	4.5	10.8	11.6	0.333	0.318	0.319
Summarization w/ GPT-4o mini	BLIP	4.6	10.2	10.3	0.325	0.284	0.300
	BLIP2	5.0	10.6	12.1	0.343	0.301	0.317
	InternVL2.5	5.8	11.4	15.3	0.352	**0.332**	0.336
SGVC (Ours)	BLIP	6.7	11.6	16.6	**0.367**	0.285	0.322
	BLIP2	7.4	12.4	20.9	**0.367**	0.304	0.331
	InternVL2.5	**8.0**	**13.2**	**24.1**	0.359	0.326	**0.338**

- (left) Effectiveness of SGVC becomes more evident when captioning longer and complex videos.
- (right) SGVC even outperforms stronger LLM summarization baselines using GPT-4o mini.

Efficiency comparison

Table 5. Comparison of computational costs between SGVC and LLM-based methods on the MSR-VTT test set.

Method	VLM Backbone	Params. (B)	GPU (GB)	Time (s)	CIDEr	Using reference	Using GPT API
VidIL	BLIP+CLIP	0.67	3.57	1.32	20.2	✓	✓
Video ChatCaptioner	BLIP2	3.75	14.53	3.65	16.5	-	✓
Summarization w/ Mistral-7B	BLIP	7.50	14.50	1.27	10.8	-	-
	BLIP2	11.00	28.20	1.51	15.4	–	–
SGVC (Ours)	BLIP	0.74	5.07	1.14	24.0	-	-
	BLIP2	4.24	18.40	1.37	26.1	–	–

[Ground-truth] *Two men are at a gym to demonstrate proper form for the exercise. The man in the black shorts gets on one knee as the instructor gives instructions on what to do. The man in black shorts lifts a bar from the kneeling position. After a few reps, the two men conclude the video.*

[LLM summ.] Two men working out in a gym, performing various activities such as weightlifting, martial arts, and stretching.

[VidIL] A group of men and women are seen working out in a gym, doing various exercises such as flipping tires, punching bags, and using a mesh sled.

[Video ChatCaptioner] The video features a man wearing a black shirt standing on a ledge in front of a red wall indoors. He appears to be leaning forward and looking at the camera with a nervous expression.

[Ours] *Two young men are standing in a gym, practicing martial arts. One of the men is holding a baseball. The other man is wearing a gray shirt. The man is standing behind the man. The man is holding a weight. The man is standing with his arms raised.*

[Ground-truth] *A track runner is preparing to run a race.*

[LLM summ.] *A group of runners, including females, stretch, crouch at the starting line, and.*

[VidIL] *A group of athletes competing in various track and field events.*

[Video ChatCaptioner] *The video shows a woman participating in a track and field event, wearing a red shirt and shorts.*

[Ours] *A group of runners crouching down a line on a track competing in a race.*

[Ground-truth] *A mom and daughter are walking around around town.*

[LLM summ.] *A woman and her daughter, accompanied by two other women, are walking down a street.*

[VidIL] *A group of people are walking down a street in Japan.*

[Video ChatCaptioner] *The video shows a girl wearing a white shirt walking down a street with a bag. The color of the bag is not known.*

[Ours] *A woman and her daughter walk down a street with a bicycle in the background.*