End-to-end Dense Video Captioning as Sequence Generation

Wanrong Zhu[1], Bo Pang[2], Ashish V. Thapliyal[2], William Yang Wang[1], Radu Soricut[2]

[1]UC Santa Barbara, [2]Google Research

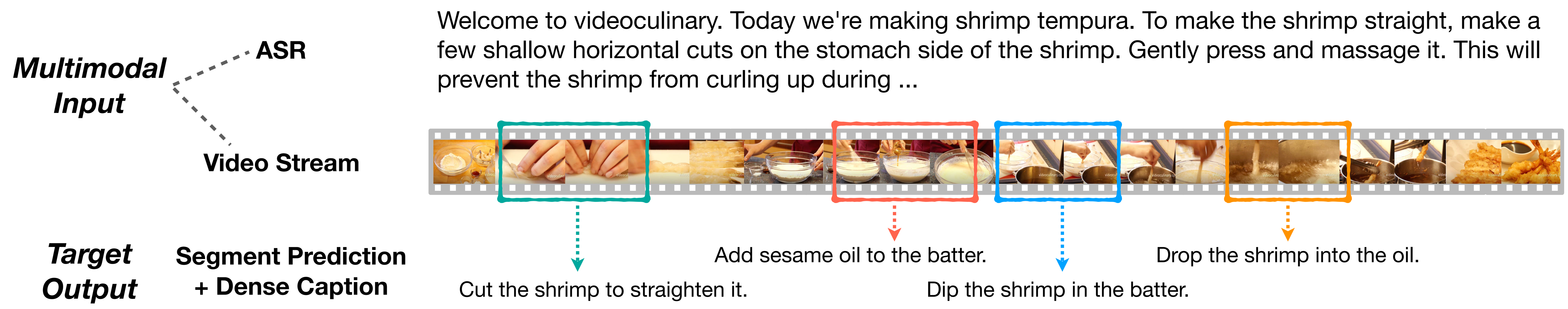

Figure 1: An example of the input video and output segmentations and captions for the dense video captioning task from the YouCook2 dataset.

Motivation

Dense video captioning (DVC) aims to identify the events of interest in an input video, and generate descriptive captions for each event. Figure 1 shows an example. Previous approaches usually follow a two-stage generative process, which first proposes a segment for each event, then renders a caption for each identified segment. In this work, we show how to model the two subtasks of dense video captioning jointly as *one* sequence generation task, and simultaneously predict the events and the corresponding descriptions.

Input Formulation For Multimodal Signals

We provide the multimodal input (video stream and ASR tokens) to the encoder in two ways:

- **Simple Concatenation**: concatenate the sequence of ASR token embeddings and the sequence of projected visual features.
- **Temporal Embedding (+Emb$_{\text{Time}}$)**: express the temporal alignment more explicitly in the input by adding learned temporal embeddings to both ASR tokens and visual frames.

Target String Formulations

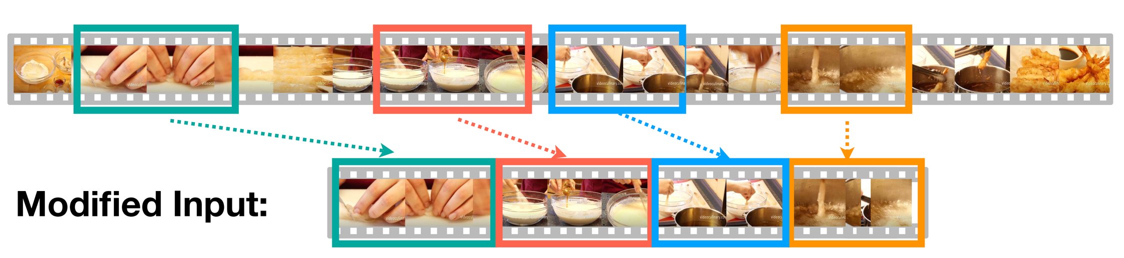

Figure 2: Modified dense video captioning: a simplified setting where the segments are concatenated to form the modified input with gaps removed.

Results

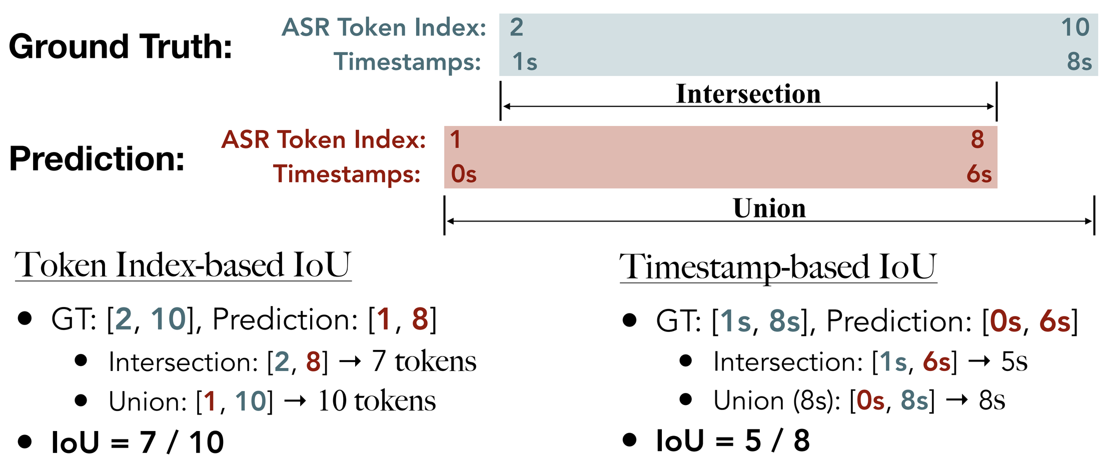

Figure 4: We use Intersection-over-Union (IoU) to measure segmentation performance. Here we compare the token index-based and timestamp-based IoU used in our study.

1. Target String Formulation

Target Formulation	Ckpt?	Seg-only mIoU	Seg-only F1	Seg+Cap mIoU	Seg+Cap BLEU-4
Random Partition		37.7	23.5	-	-
Tagging-based	-	33.6	24.5	19.7	0.1
	T5	12.1	2.8	6.7	0
Length-based	-	36.3	25.8	33.6	0.2
	T5	**42.7**	**31.2**	**42.8**	**1.8**

Table 1: In the modified setting for YouCook2, the length-based formulation achieves higher performance across the board when trained from scratch, and benefits from the T5 checkpoint.

2. Input Formulation

Dataset	Input Formulation	Seg-only mIoU	Seg-only F1	Seg+Cap mIoU	Seg+Cap BLEU-4
Youcook2	Random	20.6	10.5	-	-
	SimpleConcat	**27.8**	**16.9**	**30.3**	**3.0**
	+Emb$_{\text{Time}}$	26.5	15.8	28.7	2.6
ViTT	Random	21.9	12.5	-	-
	SimpleConcat	**41.9**	**31.3**	42.4	**1.3**
	+Emb$_{\text{Time}}$	41.6	30.8	**43.2**	1.2

Table 2: For the vanilla DVC tasks on YouCook2 and ViTT, results using SimpleConcat compared to their counterparts using Emb$_{\text{Time}}$ are mixed, both outperform the baseline of random partition with non-trivial improvements.

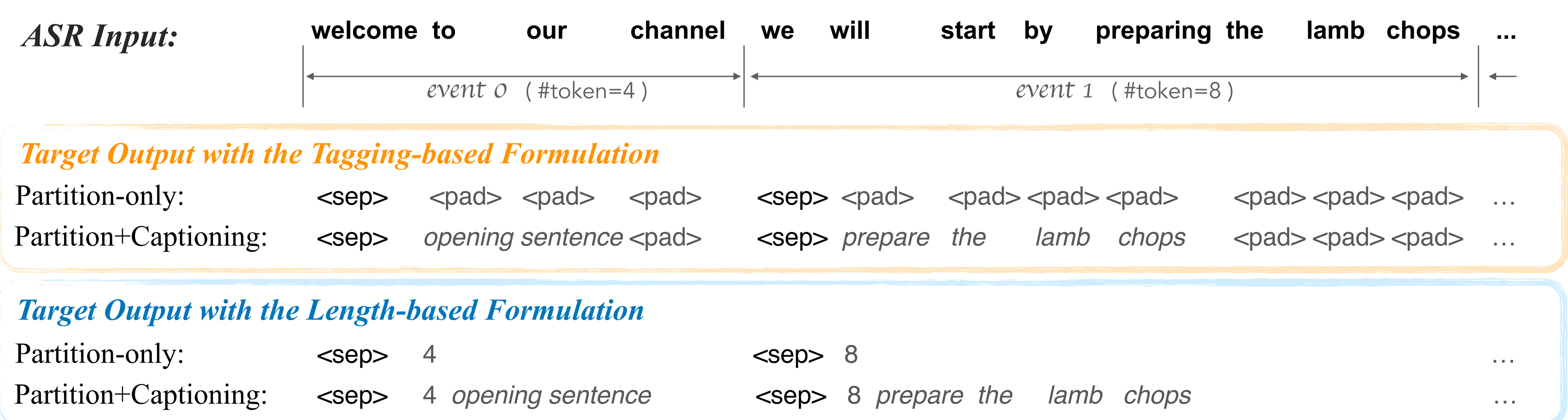

Figure 3: To jointly model segmentation and captioning subtasks as one single sequence generation task, we use the **tagging-based** and the **length-based** target formulations to encode both segmentation and captioning predictions in the target string. Here we show examples for the modified dense video captioning.

3. Effects of Pretraining

Dataset	Ckpt?	Seg-only mIoU	Seg-only F1	Seg+Cap mIoU	Seg+Cap BLEU-4
Youcook2	-	13.0	9.4	16.5	0.2
	T5	24.1	14.1	24.2	0.9
	WikiHow	22.6	13.3	23.3	0.7
	WikiHow T5	**27.8**	**16.9**	**30.3**	**3.0**
ViTT	-	33.9	23.0	32.7	0.1
	T5	37.9	27.2	38.1	0.6
	WikiHow	38.2	26.9	37.8	0.4
	WikiHow T5	**41.9**	**31.3**	**42.4**	**1.3**

Table 3: Performance reported with the SimpleConcat setting. For both YouCook2 and ViTT datasets, there are significant performance improvements from utilizing pretrained checkpoints in terms of both segmentation metrics and captioning metrics.

4. Effects of Joint Modeling

We observe a general trend in Table 2 where the Seg+Cap model outperforms the Seg-only model on the mIoU score. This indicates that with the right formulation, the segmentation subtask can indeed benefit from joint learning with a related captioning subtask.

Showcase

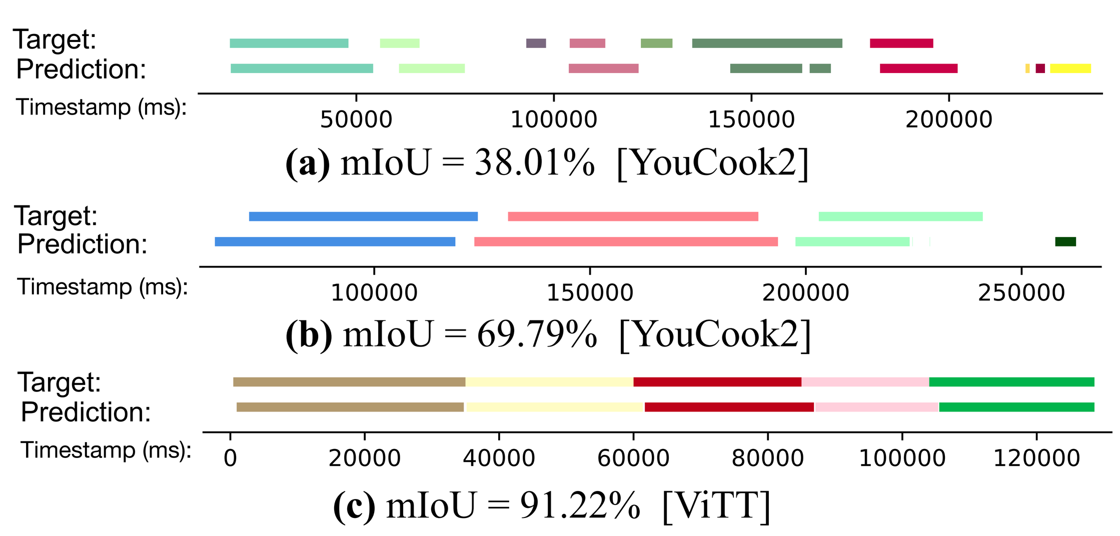

Figure 5: Example segmentation predictions corresponding to different mIoU scores.

	IoU	Segment Border (ms)	Caption
Tgt.	90.0%	[58000.0, 77000.0]	whisk eggs and season with salt
Pred.		[57309.0, 78429.5]	whisk the eggs in the deep plate
Tgt.	99.3%	[28000.0, 45000.0]	chop up the garlic in the food processer
Pred.		[28005.0, 44894.0]	chop garlic and place in the food processor
Tgt.	94.7%	[64199.0, 98080.0]	Preparining remaining ingredients
Pred.		[65710.0, 98380.0]	Chopping the remaining ingredients
Tgt.	97.0%	[65100.0, 124729.0]	Blow-drying the roots
Pred.		[63239.5, 124714.5]	Blow-drying hair

Table 4: Example caption predictions where the IoU ≥ 90% between the target (Tgt.) and the predicted (Pred.) segments. The first two examples are from YouCook2, the last two examples are from ViTT.

UC SANTA BARBARA
Google Research