

Contributions

1. Challenging new dataset
2. 2SEAL - Action Localization multimodal method
3. Extensive Analysis on action localization in vlogs

Dataset

Videos	171					
Video hours	20					
Transcript words	302,316					
Clips	1,246	#actions	Vis. (%)	#videos	#clips	
Actions	13,380	Train	4,939	35.1	110	680
Visible actions	3,131	Val	1,264	35.9	26	187
Non-visible actions	10,249	Test	3,456	25.7	35	275

Table 1: Data Statistics.

Table 2: Statistics for the experimental data split. "Vis." is the % of visible actions

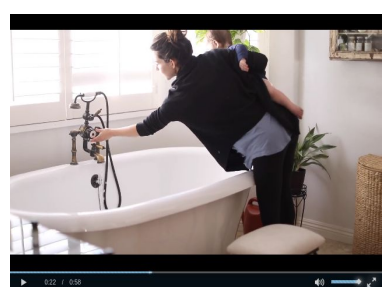
	Transcript Actions	Actions	Timestamp
	"clean up"	"clean up"	[1.4, 19.0]
+	"add their toys"	"add their toys"	[31.0, 40.0]
	"add bubble bath"	"add bubble bath"	[47.0, 55.0]

Figure 2: Action temporal localization annotation.

Multimodal Model

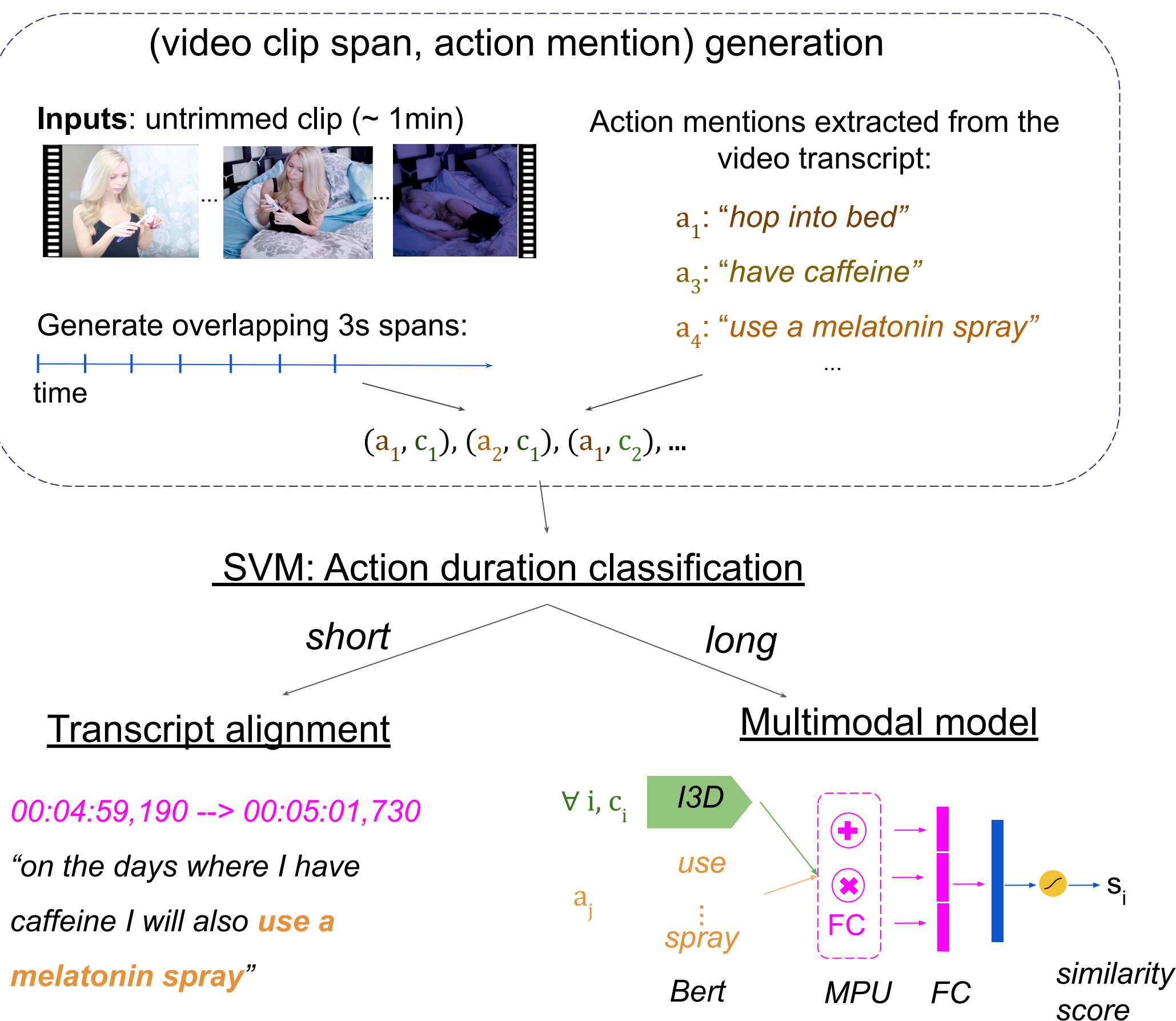
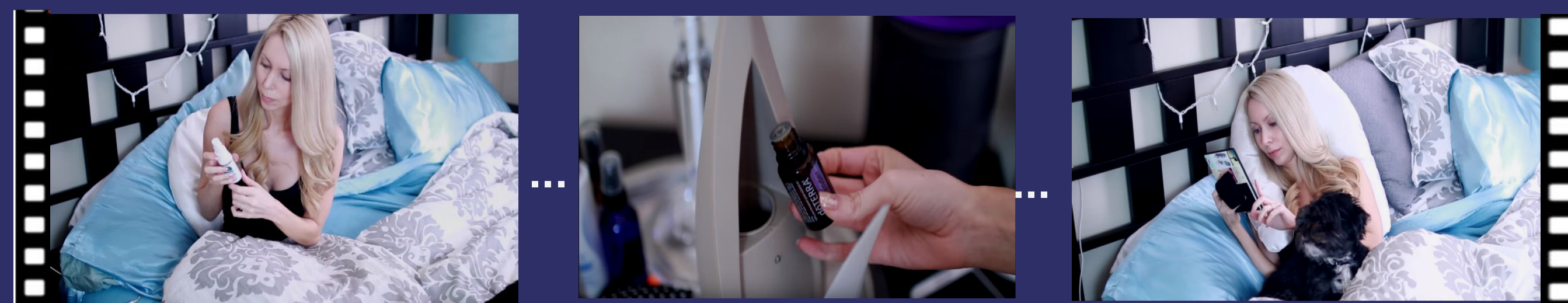


Figure 3: 2SEAL method architecture. The depicted MPU-based multimodal model can be replaced with any multimodal model.



Duration-informed Temporal Localization of Narrated Actions in Vlogs

Action duration greatly influences the performance of Action Temporal Localization.



"on days where I drink coffee I will use a melatonin spray to help me fall asleep"

"using lavender essential oil in my diffuser"

"then I grab my Kindle to do some reading I'm currently reading a book called the girl on the train"

time

Figure 1: Overview of the dataset (Ignat et al., 2019): distinguishing between actions that are narrated by the vlogger but not visible in the video and actions that are both narrated and visible in the video (underlined), with a highlight on visible actions that represent the same activity (same color). The arrows represent the temporal alignment between when the visible action is narrated as well as the time it occurs in the video.

Evaluation

Method	VA	Recall				
		IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
All visible	25.7	67.4	23.6	8.3	4.1	21.6
All non-visible	74.3	0.0	0.0	0.0	0.0	0.0
Transcript Alignment (ours)	25.7	73.3	47.3	22.2	7.2	30.8
MPU	75.5	57.9	27.0	12.4	6.2	21.4
2SEAL (ours) + MPU	79.0	74.6	48.7	22.8	8.6	31.9
MIL-NCE	26.1	62.9	22.2	8.0	4.2	20.5
2SEAL (ours) + MIL-NCE	34.4	74.4	47.8	21.7	7.9	31.4
SCA	24.2	49.9	17.0	6.0	3.4	15.9
2SEAL (ours) + SCA	26.1	72.2	46.7	21.4	7.6	30.5
Human	n/a	83.5	71.8	52.0	35.0	50.3

Table 6: Results on the test set. "VA" stands for Visibility Accuracy.

Recall	0-15s		16-35s		36-60s	
	MPU	Align	MPU	Align	MPU	Align
IoU=0.1	49.5	71.6	90.7	76.6	95.2	83.3
IoU=0.3	5.4	49.0	73.4	51.4	81.0	0.0
IoU=0.5	2.0	25.0	22.0	17.8	78.6	0.0
IoU=0.7	0.8	9.4	5.6	1.9	66.7	0.0
mIoU	12.0	32.0	38.9	29.9	71.7	16.5

Table 7: Breakdown by action duration (time span): MPU performance increases with the increase of action time span, while transcript alignment (Align) performance decreases.

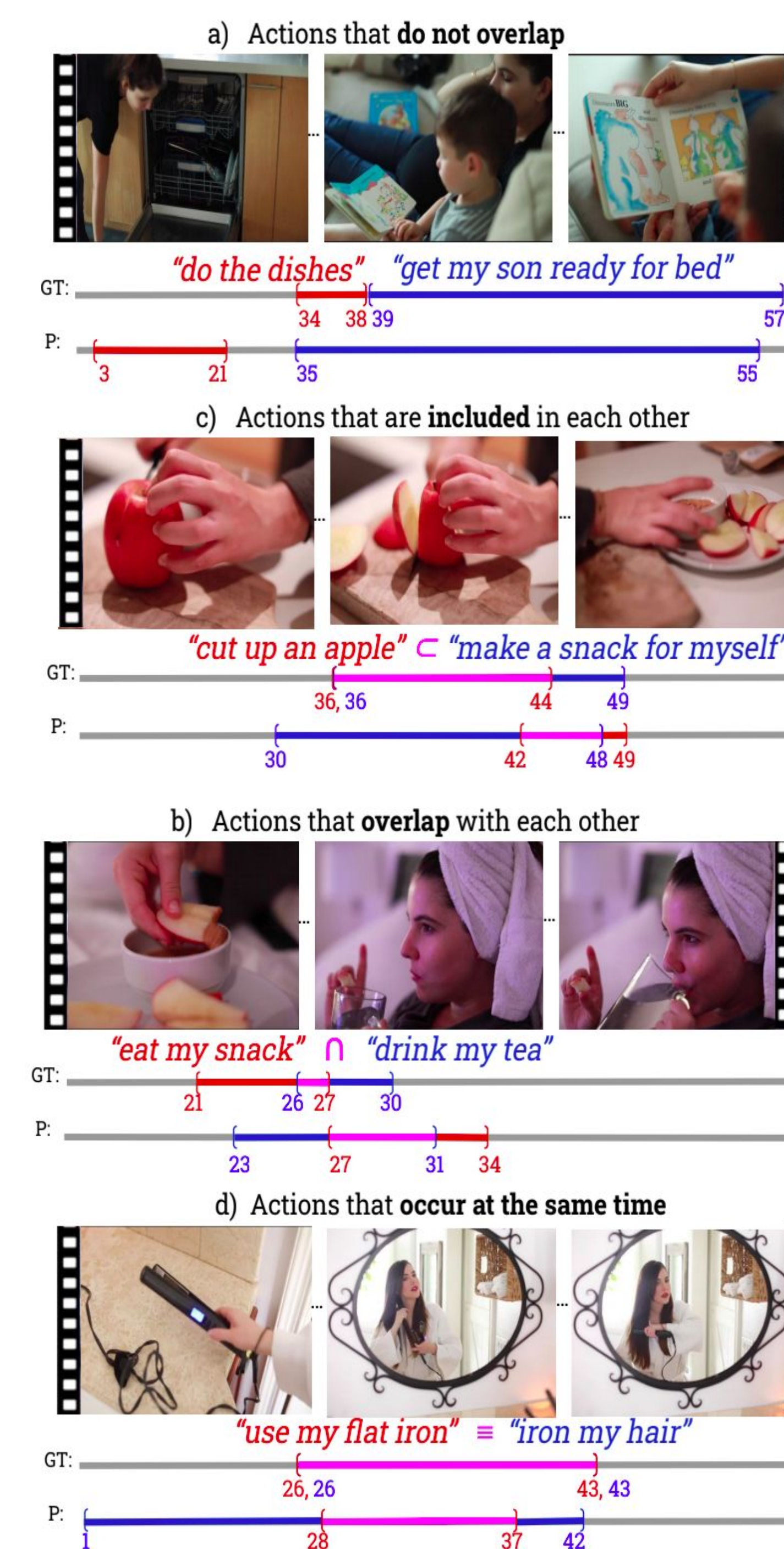


Figure 5: Localization results for different cases of action overlapping.