Supplementary Material to: Identity-Aware Multi-Sentence Video Description

Here, we provide some technical details as well as qualitative examples and analysis. Section A provides details regarding our local person ID re-labeling, data augmentation, and accuracy metric for the **Fill-in the Identity** task. Section B includes a qualitative comparison of our approach to **Fill-in the Identity** and two concurrent methods, Yu et al. and Brown et al. We also discuss some failure cases. In Section C we provide further insights into human performance on the **Fill-in the Identity** task. Finally, in Section D we include **Identity-Aware Video Description** scores on the validation set and more qualitative results for our approach to Identity-Aware Video Description task.

A Technical Details

First, we illustrate our local person ID re-labeling and training data augmentation in Figure 1. One can see the default segmentation (consecutive sets of 5 clips) and additional segmentations, which serve for data augmentation. Each individual set of 5 clips with associated local IDs serves as a unique data point for our model.

Next, we illustrate our proposed accuracy metric (introduced in Section 5.2 of the main paper) in Figure 2. We consider pairwise comparisons between predicted IDs (whether they are the same or not), and compare that to ground-truth pairs. Accuracies are then computed respectively as the number of correct pairs divided by the total number of ground truth pairs with the same IDs (Same-Acc), with the different IDs (Different-Acc), and with all the pairs (Instance-Acc). In the figure, there are 2 ground truth pairwise comparisons with same IDs, 8 with different IDs, and 10 total. Note that the Same-Acc and Different-Acc are used to calculate the Class-Acc in the main paper.

B Fill-in the Identity: Qualitative Results

In the qualitative examples, our model is able to recognize different characters and link the same characters across the video[1]. In Figure 3 (a), our model consistently links the character that appears from the second clip with the same ID. Likewise, we get the correct identities for video clip that involves more than two characters in Figure 3 (b). On the other hand, other state of the art models either tend to predict characters that are not present in the video e.g. predicting [PERSON5] for the last sentence in Figure 3 (a), or fail to correctly link

[1] Note, that we skip the clips/sentences with no blanks, therefore, sometimes resulting in less than 5 clips per set.

		LOCAL PERSON IDs		
CLIP	REFERENCE SENTENCE	DEFAULT SEGMENTATION	ADDITIONAL SEGMENTATIONS	
1	SOMEONE<NEVILLE> turns to the others.	[PERSON1]	…	…
2	He<NEVILLE> opens the gold-framed painting.	[PERSON1]	[PERSON1]	…
3	SOMEONE<NEVILLE> steps aside, revealing SOMEONE<HARRY>.	[PERSON1],[PERSON2]	[PERSON1],[PERSON2]	[PERSON1],[PERSON2]
4	SOMEONE<HARRY> sees a hoard of students in a large room strung with hammocks.	[PERSON2]	[PERSON2]	[PERSON2]
5	He<HARRY> steps down.	[PERSON2]	[PERSON2]	[PERSON2]
6	They take it in turn to hug him.	_	_	_
7	SOMEONE<NEVILLE> approaches a student, who runs to a radio.	[PERSON1]	[PERSON1]	[PERSON1]
8	SOMEONE<HARRY> shifts uncomfortably and turns to the expectant faces.	[PERSON2]	[PERSON2]	[PERSON1]
9	SOMEONE<GINNY WEASLEY> rushes into the room.	[PERSON3]	[PERSON3]	[PERSON2]
10	SOMEONE<GINNY> ignores SOMEONE<RON>.	[PERSON3],[PERSON4]	[PERSON3],[PERSON4]	[PERSON2],[PERSON3]
…	…	…	…	…

Fig. 1: Illustration of local person ID re-labeling and training data augmentation for the **Fill-in the Identity** task.

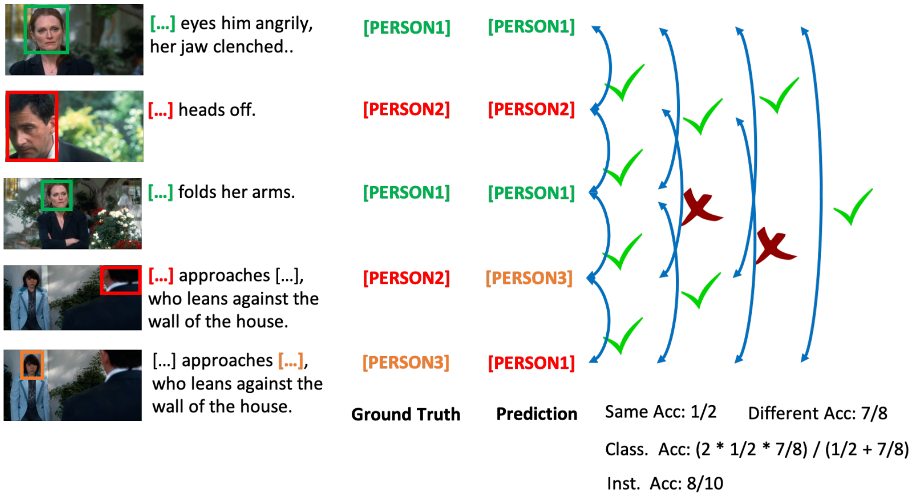

[...] eyes him angrily, her jaw clenched..
[...] heads off.
[...] folds her arms.
[...] approaches [...], who leans against the wall of the house.
[...] approaches [...], who leans against the wall of the house.
[PERSON1]
[PERSON2]
[PERSON1]
[PERSON2]
[PERSON3]
[PERSON1]
[PERSON2]
[PERSON1]
[PERSON3]
[PERSON1]
Ground Truth
Prediction
Same Acc: 1/2
Different Acc: 7/8
Class. Acc: (2 * 1/2 * 7/8) / (1/2 + 7/8)
Inst. Acc: 8/10

Fig. 2: Illustration of the accuracy metric for the **Fill-in the Identity** task. For each pair of blanks, we assign "correct" if the IDs are the same or different in *both* ground truth and predictions. See Section A for more details.

to previously seen characters e.g. mixing up [PERSON4] in Figure 3 (a) and [PERSON2] in Figure 3 (b). We also study if the model is possibly biased towards number of blanks in each clip. In particular, it is likely that sentence with more than one blank slot may also involve visual content with more than one character. In Figure 3 (c), we show a sequence of video clips involving only one character. While our prediction identifies all the blanks as the same character, the other models struggle to do so and predict different identities. This pattern is not surprising, as we've seen that these models tend to over-predict diverse IDs (2,3,...) based on Figure 4 of the main paper.

We show some failure cases in Figure 4. In the second clip and the last clip of Figure 4 (a), we fail to identify which character is holding the gun and who they are pointing at. This results in the swapped character IDs. Note, that we still limit our predictions to two characters, while other methods predict more

Method	**Inst-Acc**
Human w/o video (median)	70.0
Human (median)	87.0
Human w/o video (max)	85.1
Human (max)	96.0

Table 1: **Fill-in the Identity**: median and maximum human performance over 200 random sets of clips (Test set).

characters. Our model also struggles to correctly link previously seen characters in clips with a crowd of people. In Figure 4 (b), the model incorrectly links the characters within in the third clip due to a large crowd; however, it still links the last two IDs as the same person in the first clip, which matches the ground truth labels.

C Fill-in the Identity: Additional Analysis

In Section 5.3.2 of the main paper, we have presented human performance on the **Fill-in the Identity** task, measured as a median accuracy across three annotators. It is also worth looking at the maximum accuracy across three workers. We include that in Table 1 (here we only report the Instance Accuracy). As we see, the numbers are substantially higher, if we consider the upper-bound accuracy across the workers. When seeing the video, humans can get up to 96% accuracy. We analyze the cases when none of the three annotators were able to get the correct person IDs, and find that most of the time that happens (a) in more complex scenes (with multiple participants), (b) in symmetrical cases like "[...] and [...] walk in", (c) in ambiguous cases, such as "[...] gives [...] a look", where it may be hard to tell which of the two persons was meant, etc.

D Identity-Aware Video Description: Additional Results

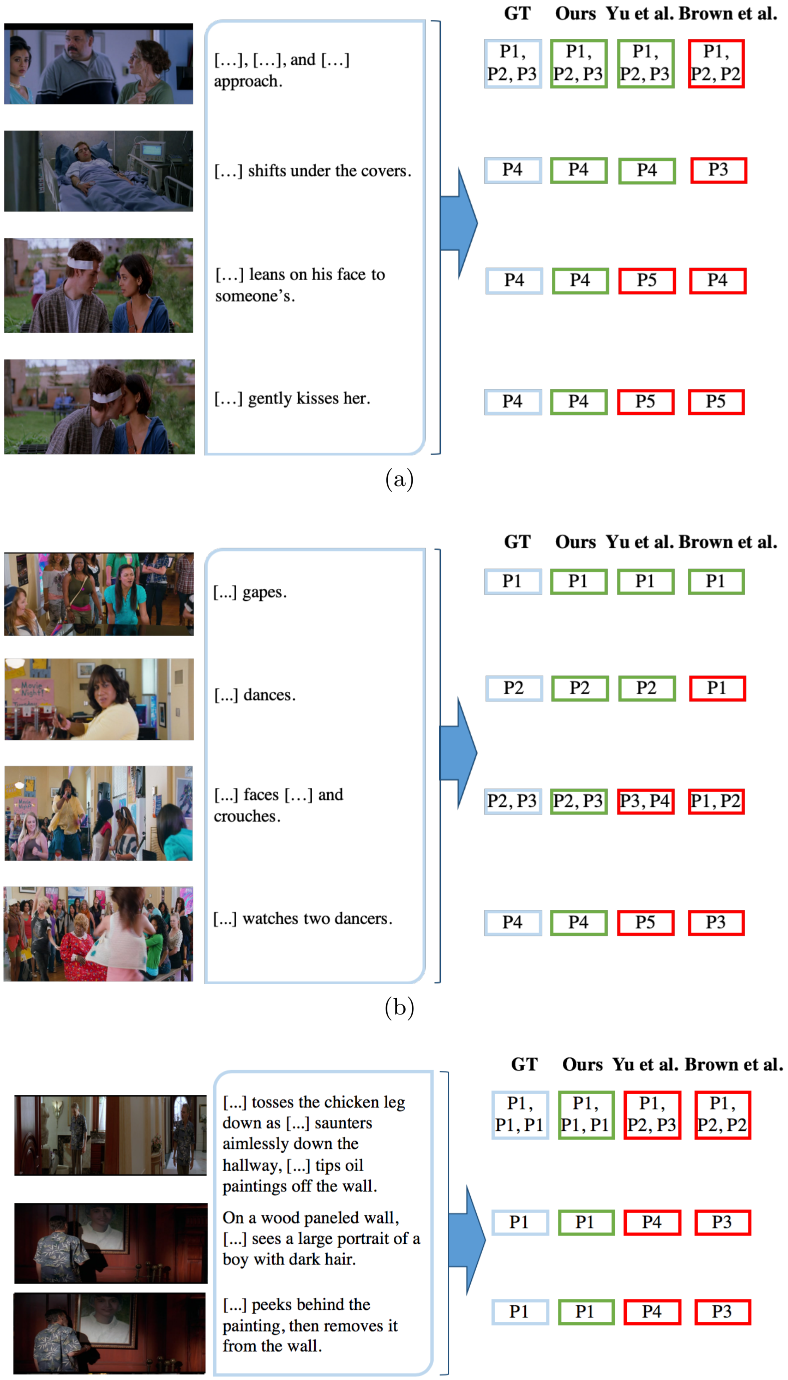
We show results from our model and other baselines on the LSMDC Validation set in Table 2. We use the same metrics as in Table 5 of the main paper.

Finally, in Figure 5, we show descriptions generated by our baseline model with predicted character IDs. Overall, our Fill-in the Identity model correctly links relevant activity to the character IDs. PERSON2 in Figure 5 (a) looks and walks away in the video, and PERSON2 is also predicted correctly as the woman kissing and smiling in the last two clips of Figure 5 (b). We observe that the model still performs reasonably well even for descriptions that may not be perfectly aligned with the video. For example, there is no person walking away in the third clip of Figure 5 (c), but the model recognizes that PERSON2 has her head away from the camera and predicts her as the one leaving. However, we do acknowledge that our two-stage pipeline approach is not perfect; our model does

	Per set, MAX score		
Method	**METEOR**	**BLEU@4**	**CIDEr-D**
Same ID	9.41	1.57	7.03
All different IDs	9.11	1.36	7.00
Ours Text-Only	10.53	1.77	7.73
Ours	10.68	1.80	7.77

Table 2: **Identity-Aware Video Description** scores for our method on the LSMDC validation set.

not identify all the characters in the video, e.g. there is no ID for the woman looking at PERSON2 in Figure 5 (a) nor the man hiding in the bushes in the first clip of Figure 5 (b). We leave improving the description quality with character identification to future work.

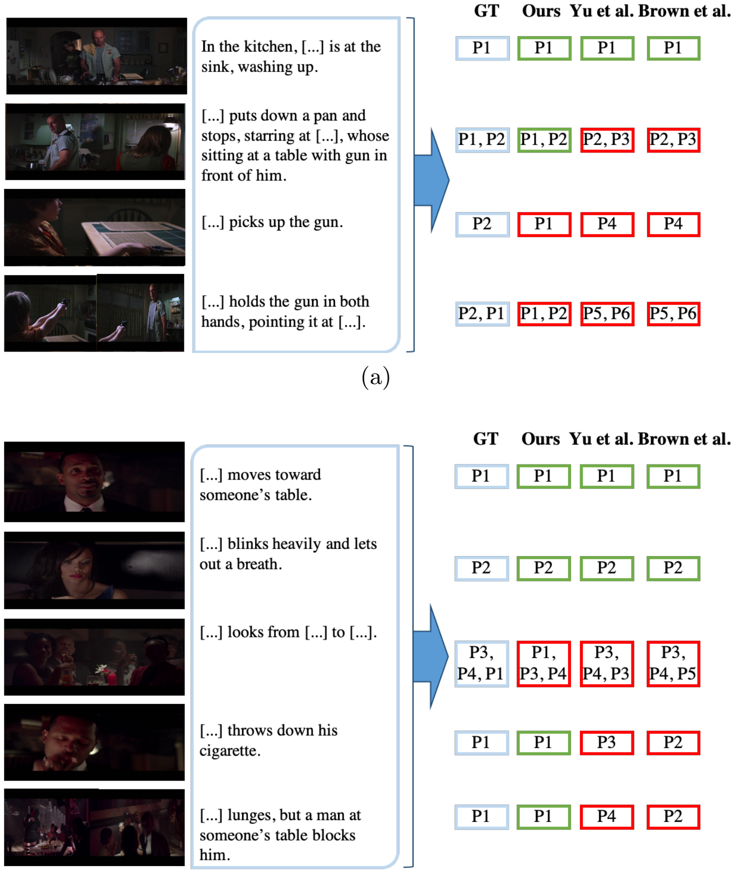

GT
Ours
Yu et al.
Brown et al.
[…], […], and […] approach.
P1, P2, P3
P1, P2, P3
P1, P2, P3
P1, P2, P2
[…] shifts under the covers.
P4
P4
P4
P3
[…] leans on his face to someone's.
P4
P4
P5
P4
[…] gently kisses her.
P4
P4
P5
P5
GT
Ours
Yu et al.
Brown et al.
[...] gapes.
P1
P1
P1
P1
[...] dances.
P2
P2
P2
P1
[...] faces […] and crouches.
P2, P3
P2, P3
P3, P4
P1, P2
[...] watches two dancers.
P4
P4
P5
P3
GT
Ours
Yu et al.
Brown et al.
[...] tosses the chicken leg down as [...] saunters aimlessly down the hallway, [...] tips oil paintings off the wall.
P1, P1, P1
P1, P1, P1
P1, P2, P3
P1, P2, P2
On a wood paneled wall, [...] sees a large portrait of a boy with dark hair.
P1
P1
P4
P3
[...] peeks behind the painting, then removes it from the wall.
P1
P1
P4
P3

(a)

(b)

(c)

Fig. 3: Qualitative examples for the **Fill-in the Identity** task, comparison between our approach and two concurrent methods. Correct/incorrect predictions are labeled with green/red, respectively.



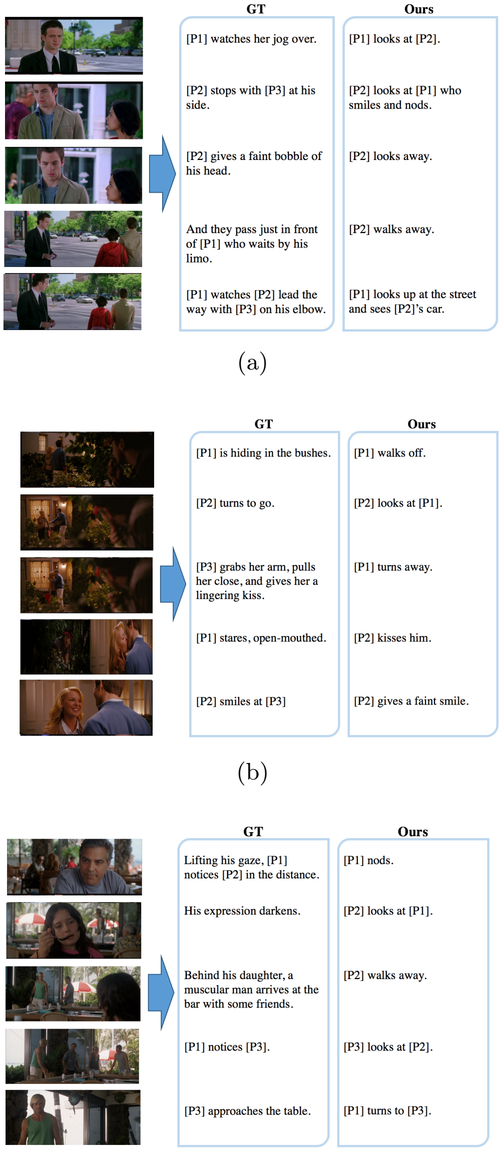
	GT	Ours	Yu et al.	Brown et al.
In the kitchen, [...] is at the sink, washing up.	P1	P1	P1	P1
[...] puts down a pan and stops, starring at [...], whose sitting at a table with gun in front of him.	P1, P2	P1, P2	P2, P3	P2, P3
[...] picks up the gun.	P2	P1	P4	P4
[...] holds the gun in both hands, pointing it at [...].	P2, P1	P1, P2	P5, P6	P5, P6

(a)

	GT	Ours	Yu et al.	Brown et al.
[...] moves toward someone's table.	P1	P1	P1	P1
[...] blinks heavily and lets out a breath.	P2	P2	P2	P2
[...] looks from [...] to [...].	P3, P4, P1	P1, P3, P4	P3, P4, P3	P3, P4, P5
[...] throws down his cigarette.	P1	P1	P3	P2
[...] lunges, but a man at someone's table blocks him.	P1	P1	P4	P2

(b)

Fig. 4: Failure examples for our model on the **Fill-in the Identity** task, comparison between our approach and two concurrent methods. Correct/incorrect predictions are labeled with green/red, respectively.



GT	Ours
[P1] watches her jog over.	[P1] looks at [P2].
[P2] stops with [P3] at his side.	[P2] looks at [P1] who smiles and nods.
[P2] gives a faint bobble of his head.	[P2] looks away.
And they pass just in front of [P1] who waits by his limo.	[P2] walks away.
[P1] watches [P2] lead the way with [P3] on his elbow.	[P1] looks up at the street and sees [P2]'s car.

(a)

GT	Ours
[P1] is hiding in the bushes.	[P1] walks off.
[P2] turns to go.	[P2] looks at [P1].
[P3] grabs her arm, pulls her close, and gives her a lingering kiss.	[P1] turns away.
[P1] stares, open-mouthed.	[P2] kisses him.
[P2] smiles at [P3]	[P2] gives a faint smile.

(b)

GT	Ours
Lifting his gaze, [P1] notices [P2] in the distance.	[P1] nods.
His expression darkens.	[P2] looks at [P1].
Behind his daughter, a muscular man arrives at the bar with some friends.	[P2] walks away.
[P1] notices [P3].	[P3] looks at [P2].
[P3] approaches the table.	[P1] turns to [P3].

(c)

Fig. 5: Qualitative examples of our approach on the **Identity-Aware Video Description** task.