MICap: A Unified Model for Identity-aware Movie Descriptions

SUPPLEMENTARY MATERIAL

Haran Raajesh*[1] Naveen Reddy Desanur*[1] Zeeshan Khan[2] Makarand Tapaswi[1]
[1]CVIT, IIIT Hyderabad, India
[2]Inria Paris and Département d'informatique de l'ENS, CNRS, PSL Research University
https://katha-ai.github.io/projects/micap/
* denotes equal contribution

We present additional insights and results in the supplementary material. In Appendix A, we highlight how our auto-regressive Transformer decoder attends to various memory features. For the id-aware captioning task, we show the relative importance of the 3 visual features, while for the Fill-in-the-blanks (FITB) task, we highlight how our model attends to correct face clusters. Next, in Appendix B, we show qualitative results for both tasks, FITB and id-aware captioning. We also illustrate how our new identity-aware metric, iSPICE, is calculated on some examples. Finally, we end with discussion of some limitations in Appendix C.

A. Analyzing Model Attention

In this section, we visualize and discuss the attention scores from MICap's auto-regressive Transformer decoder. In particular, we focus on the cross-attention scores of the last layer as they reveal interesting insights about the features that the captioning model uses. Throughout this section, we analyze MICap trained jointly on id-aware captioning and FITB. All attention scores are obtained in inference mode.

A.1. Attention Patterns in Id-aware Captioning

In id-aware full captioning, for a particular videoset $\mathcal{N} = \{V_i\}_{i=1}^{N}$, we first encode the videos to obtain memory tokens M and pass them through a Transformer decoder auto-regressively to generate one token (word) at a time. If we consider that the number of tokens in the predicted captionset is L, we can compute a matrix of cross-attention scores $\alpha = L \times |M|$, where $|M|$ is the number of tokens in the decoder memory. Note, while we use multi-head attention, scores over the heads are averaged obtain α.

We split the L tokens into 2 groups: (i) one group consists of person id label predictions or *person tokens* (PT); and (ii) the other group consists of all other tokens referred to as *caption tokens* (CT). For visualization, we sum over the attention scores for each of the token types (id labels and text) and convert our attention map to a matrix of $2 \times |M|$.

Next, we also group the memory tokens into 3 types of visual features used in our work: action (I3D), face (Arcface), and semantic features (CLIP). Thus, we obtain a 2×3 matrix of cross-attention scores for each sample.

Results. We compute attention scores over all samples of the validation set and plot them as a probability density function in Fig. 1. PT (red) and CT (green) represent the person and caption tokens respectively. We observe that: (i) The model relies on CLIP features to predict captions (depicted by the overall high attention scores from 0.5-0.7). (ii) When predicting person tokens (PT) of the identity-aware captions, the model tends to look at face features (0.1-0.6) more than when predicting caption tokens (0-0.4). (iii) Finally, while action features are useful for captioning, they are less useful for predicting person-id labels. This is expected as action recognition is an identity-agnostic task.

A.2. Attention Patterns in FITB

For the FITB task, we analyze how the person id predictions attend to *face features* from the decoder memory. For a videoset $\mathcal{N} = \{V_i\}_{i=1}^{N}$ and its corresponding captionset with blanks $\hat{\mathcal{C}}$ we obtain a cross-attention map of $\alpha = |\mathcal{B}| \times F$, where $|\mathcal{B}|$ is the number of blanks in the captionset, and F is the number of face detections across the videoset. Each row of this matrix is normalized to sum to 1.

The attention scores and captionsets with blanks are presented in Fig. 2. In the next paragraphs, we will analyze the 3 types (columns) of the presented scores.

Cross-attention scores for face detections. In the left column of Fig. 2, we visualize the attention scores directly for each face detection. In the plot, x-axis spans time across different videos. Our model tends to show a diagonal pattern indicating that person id label predictions tend to look at faces in the same video (facilitated through the $\mathbf{E}^{\text{vid}}$ embeddings). However, as seen in captionset 5, left, row 1, the

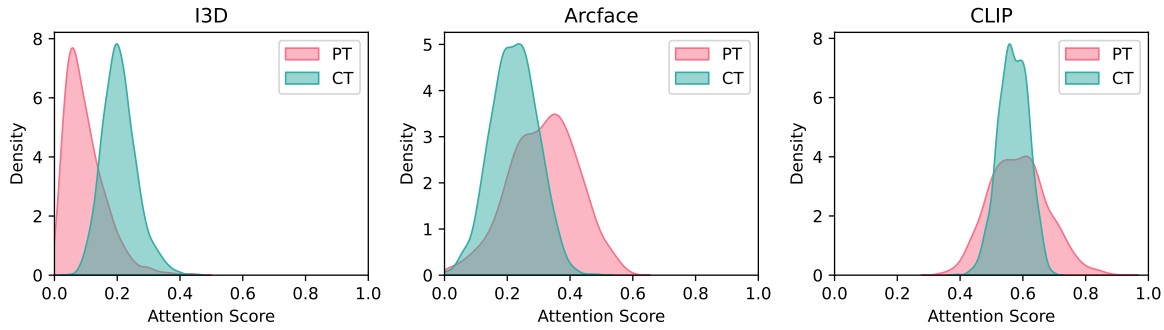


Figure 1. Cross-attention scores density plots for the id-aware captioning task. We group decoder output tokens into two types: person id label tokens (PT), and caption tokens that represent other words (CT). Attention scores are grouped across the three input visual features capturing actions (I3D, left), faces (Arcface, middle), and semantic content (CLIP, right). Please refer to Appendix A.1 for a discussion.

model may also attend to other face detections of the same person across videos. This highlights that being able to attend to faces across videos is useful (compared to [1] that only looks at faces within the same video).

Cross-attention scores for face clusters grouped by video index. Shown in the middle column of Fig. 2, we group the F face detections into clusters, but split them based on video index in the videoset. For example, in captionset 1, we see that faces in cluster 1 appears across videos 1, 2, 4 (C1/V1, C1/V2, C1/V4). This allows us to explain some of the predictions made by our model.

Please note that the face cluster index and person id labels need not match numerically. That is, cluster 2 could be assigned the label P1 and cluster 1 the label P2. These changes are acceptable as we only consider person id labels in a local videoset.

In cationset 3, we see that cluster 2 corresponds to the prediction P1 (first two rows) and cluster 4 (C4/V3) corresponds to person id label P2 (bottom two rows). In the last row of captionset 3, we see that our model predicts P2 for the video id 4 correctly, while looking at cluster 4 in video 3 (C4/V3). Previous work [1] is unable to use such cross-video information.

Cross-attention scores for clusters. In the right columns of Fig. 2, we show attention scores directly grouped by cluster ids. Here, the original attention map of $|\mathcal{B}| \times F$ is grouped to $|\mathcal{B}| \times |\mathcal{G}|$, where $|\mathcal{G}|$ is the number of face clusters obtained after performing DBSCAN on the F face detections.

Captionset 2 is an example with multiple blanks and 4 characters. We observe that some confusion in attention scores leads to errors in the predicted person id labels. In captionset 4, we also see 6 blanks, now with 3 characters. In the last row, while the model wrongly predicts P1, the model does look at cluster 3 (corresponding to P3) correctly. Captionset 1 and 2 are examples of perfect attention scores and clusters. P1 and C1, and P2 and C2 go together strongly in these examples.

Impact of number of clusters on FITB. Fig. 3 shows the results on FITB class-accuracy for varying the DBSCAN epsilon parameter. These results indicate the importance of clustering across videos and choosing an appropriate number of clusters. Qualitatively, we adopt 0.75 and it is unlikely to merge characters incorrectly.

B. Qualitative Results

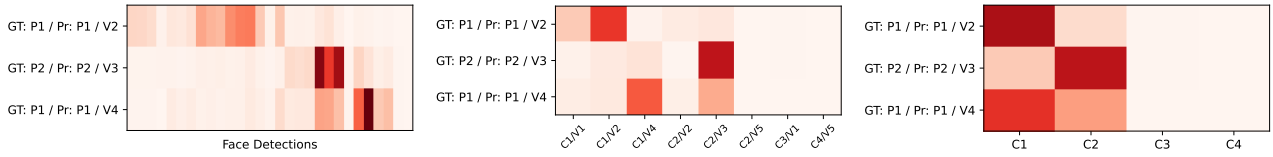
iSPICE validation examples. To validate our new metric, we propose an experiment that measures similarity between captions when identity names are added, removed, or replaced (Sec. 4 of the main paper). While the quantitative results favor iSPICE, as seen in Tab. 1 of the main paper, we illustrate with examples the process of metric computation in Fig. 4. We observe that the small difference in identity names is captured correctly by iSPICE, due to the focus on tuples containing identities, while other metrics do not show this sensitivity.

FITB examples. While Fig. 2 clearly shows the importance of cross-attention scores of detected faces and computed clusters, the challenging visual scenarios are not evident. We pick two examples (captionset 3 and 4) from Fig. 2 and pair them together with one frame from each video of the videosets. Fig. 5 shows the challenging nature of the videos where characters are often not looking at the camera (example 1 video 1, 3), the scene is dark, or the face may not even be visible (example 1 video 4 or example 2 video 3). MICap leverages the ability to look at faces and clusters across videos to improve results on the FITB task.

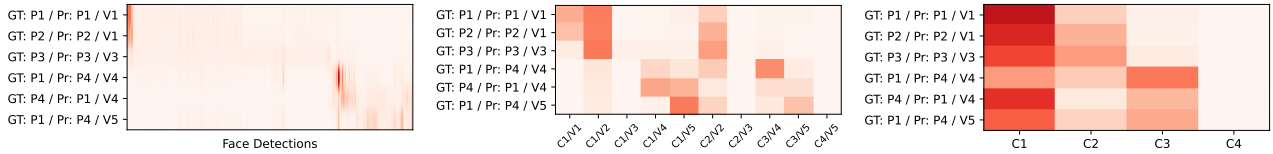
Id-aware captioning examples. Fig. 6 shows 2 examples where our model does relatively well, while Fig. 7 shows 2 difficult examples where our model makes mistakes.

In the left column of Fig. 6 we see that the model rightly

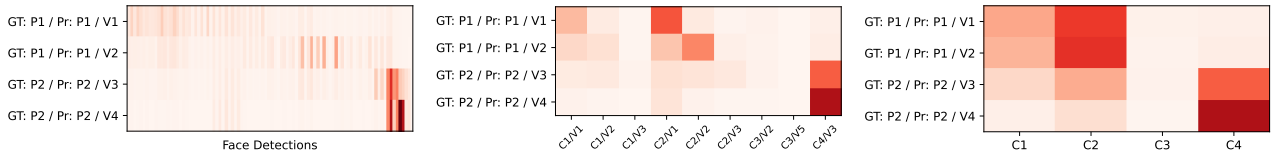
Captionset 1: Someone watches the aliens draw closer. _____ sits back in the doorway clutching a radio. _____ watches from his position several yards away. _____ squeezes the detonator the bus blows apart.



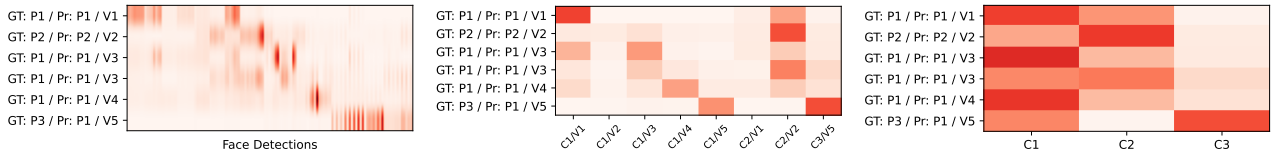
Captionset 2: _____ and _____ killed their first witch. They advance cautiously. Suddenly _____ is thrown to the ground with a jolt. _____ whips around a weapon poised to find _____ holding her wand to neck. _____ begins to put the gun on the ground.



Captionset 3: _____ pulls her phone from her bag and answers. _____ frowns uncertainly. _____ leans on a wall and slips. _____ lowers his phone and folds it shut. The next morning two women stroll across the street in front of apartment building.



Captionset 4: _____ scrutinizes his earnest face. His eyes gleaming in the dim light. _____ abruptly gets to his feet and heads for the door now. _____ talks on his cell as _____ steps into the daylight silhouetted against the sunny day. _____ faces the door frame and leans his head against it now. In a hotel suite a woman applies makeup to _____.



Captionset 5: _____ turns and spots the brown chevy 4x4 parked on a short driveway. _____ approaches the vehicle cautiously across a lawn leaning over to get a view of its occupant. The passenger side window is lowered. _____ puts both hands on the sill and leans in with an inquisitive frown. _____, the asian man who in town sits with one hand clamped to the steering wheel rocking nervously and staring numbly ahead.

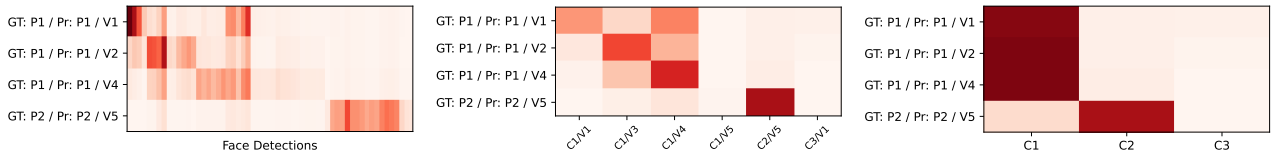


Figure 2. We show 5 examples of our model's attention scores on the FITB task. For each example (row), we show the captionset (with blanks) and the attention scores grouped in various ways. The **left** column shows the attention score for each blank across all face detections in the video. The **middle** column shows attention scores for face detections grouped by clusters in each video. C1/V1 indicates faces appearing in cluster 1 and video 1, while C1/V2 indicates faces of the same cluster 1 appearing in video 2. The **right** column shows attention scores of each blank for face clusters (across videos). For each row in the attention scores, we indicate the ground-truth (GT) and predicted (Pr) person id label and the video index (V1 .. V5) in which this blank appeared. See Appendix A.2 for a discussion.

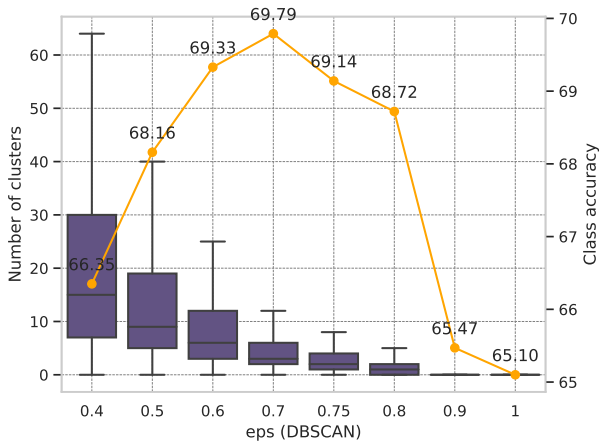


Figure 3. Class-accuracy for the FITB task by varying the DBSCAN eps distance threshold. We also show a box-plot for the number of clusters created at each threshold across samples of the validation set.

identifies P1 as the male character and P2 as the female. The last caption is quite interesting – while the GT points to P1 giving P2 a bowl, our model predicts that P2 gives a sad smile, which is not wrong. This also illustrates some of the challenges of evaluating captioning. In the right column of Fig. 6, the predicted caption uses P2 to refer to the man, and is consistent across videos 3, 4, and 5 in the videoset.

In the complex visual example of Fig. 7 (left), our model assigns P1 to all blanks. Similarly, in the multi-character example of Fig. 7 (right), we observe some confusion between characters. Nevertheless, P2, identified as the man on the left in video 3, is correctly identified for the first 3 videos. The model is also able to predict that they are on a plane (caption for video 2). Nevertheless, these examples illustrate the challenges of id-aware captioning. As future work, they also highlight the need to evaluate visual grounding of the identities beyond captioning performance.

C. Limitation and Future Work

One limitation of our work, inherited from the task definition in LSMDC, is restricting videosets to local groups of 5 videos. In the future, we would like to extend this to larger videosets, perhaps spanning the entire movie. However, the approach will need to be modified to operate on full movies as: (i) providing features of all movie frames as decoder memory creates a huge number of embeddings; (ii) face clustering across the entire movie could be error-prone; and (iii) auto-regressively generating one caption at a time for hundreds of clips seems challenging, as the model needs to be cognizant of all previously generated captions. We believe that a hierarchical model that builds from shots to scenes to the full movie may be more appropriate here.

Second, the tasks for FITB and full captioning do not learn at the same pace, and choosing a single best checkpoint for both may be difficult. We posit that the user may choose two checkpoints, one for each task. Furthermore, we observe that by weighing the FITB and full captioning losses appropriately, additional performance improvements can be achieved for one task at the cost of the other task.

We have also not considered using external knowledge or pre-trained large language models (LLMs) or vision-language models (VLMs) built for captioning. We believe that it is interesting to learn what can be achieved by training on LSMDC alone. As seen in multiple examples throughout Appendix B, MICap does perform quite well given the challenging scenarios.

References

[1] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. Identity-aware multi-sentence video description. In *European Conference on Computer Vision (ECCV)*, 2020. 2

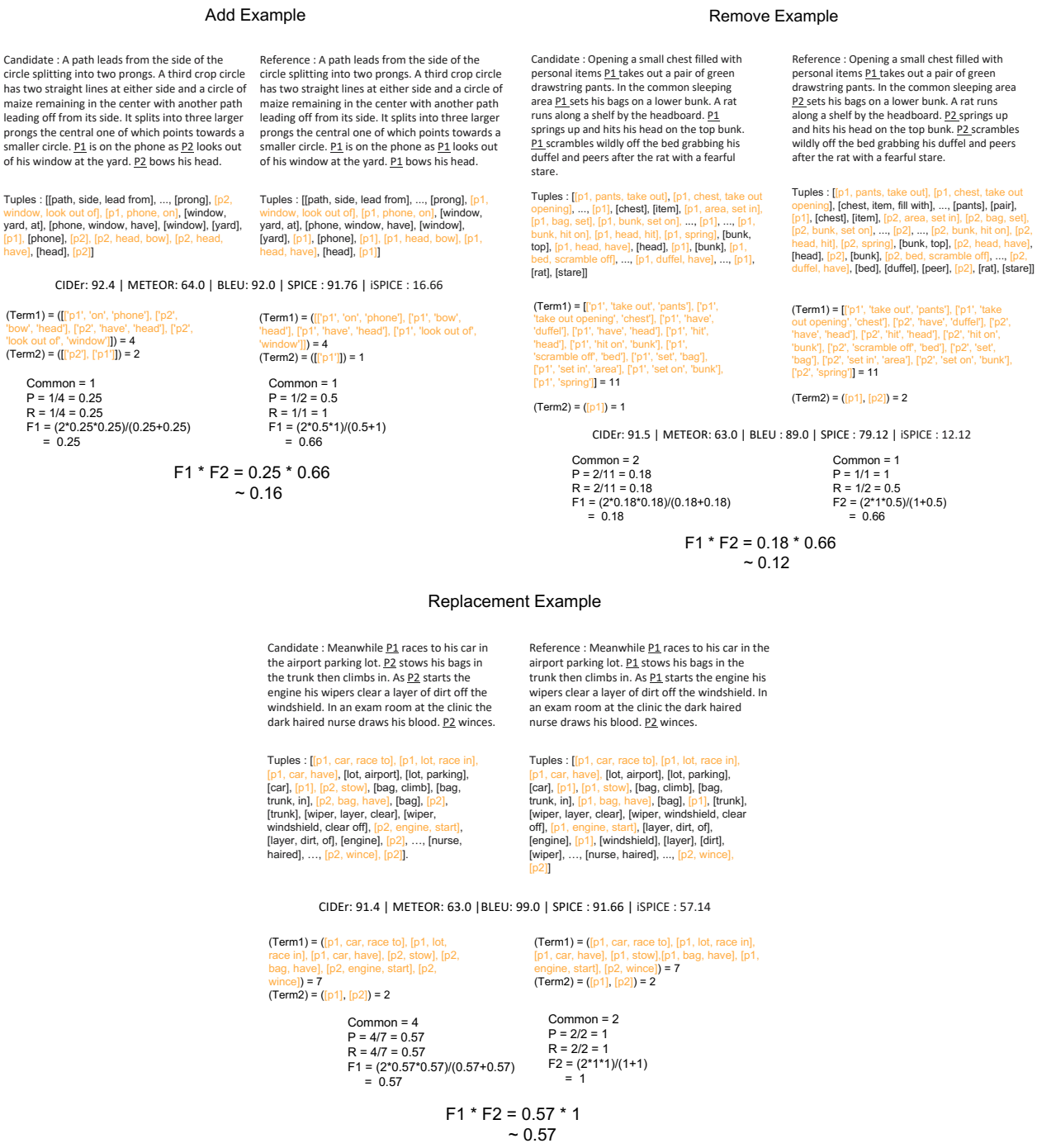


Figure 4. We show the effect of identity on captioning metrics using add, remove, and replacement examples. This corresponds to the validation experiment conducted in Tab. 1 of the main paper. For each example, the identity labels are underlined in the candidate and reference captionsets. We also show how iSPICE works by illustrating the tuples, highlighting tuples with identities, and showing the computation of term 1 (left) and term 2 (right) corresponding to tuples with size ≥ 1 and $= 1$ respectively. iSPICE takes into the account the identity whereas the other metrics show a high score due to high number of n-gram matches.

Figure 5. Examples from the Fill-in-the-blanks (FITB) task. On the left, we show one frame from each video of the videoset and the corresponding caption with blanks. In the middle, we show the ground-truth and predicted person id labels. On the right, we show the cross-attention maps (face detections, clusters, and clusters by video ids), presented in Fig. 2. We pick the examples corresponding to captionset 3 and 4 of Fig. 2 for better understanding. In general, we observe that person predictions depend strongly on the cluster features and their attention. In some cases, the identity may be difficult to predict as seen in the last row of the second example, where our model predicts P1 instead of P3, even though the attention masks are correctly focusing on C3/V5.

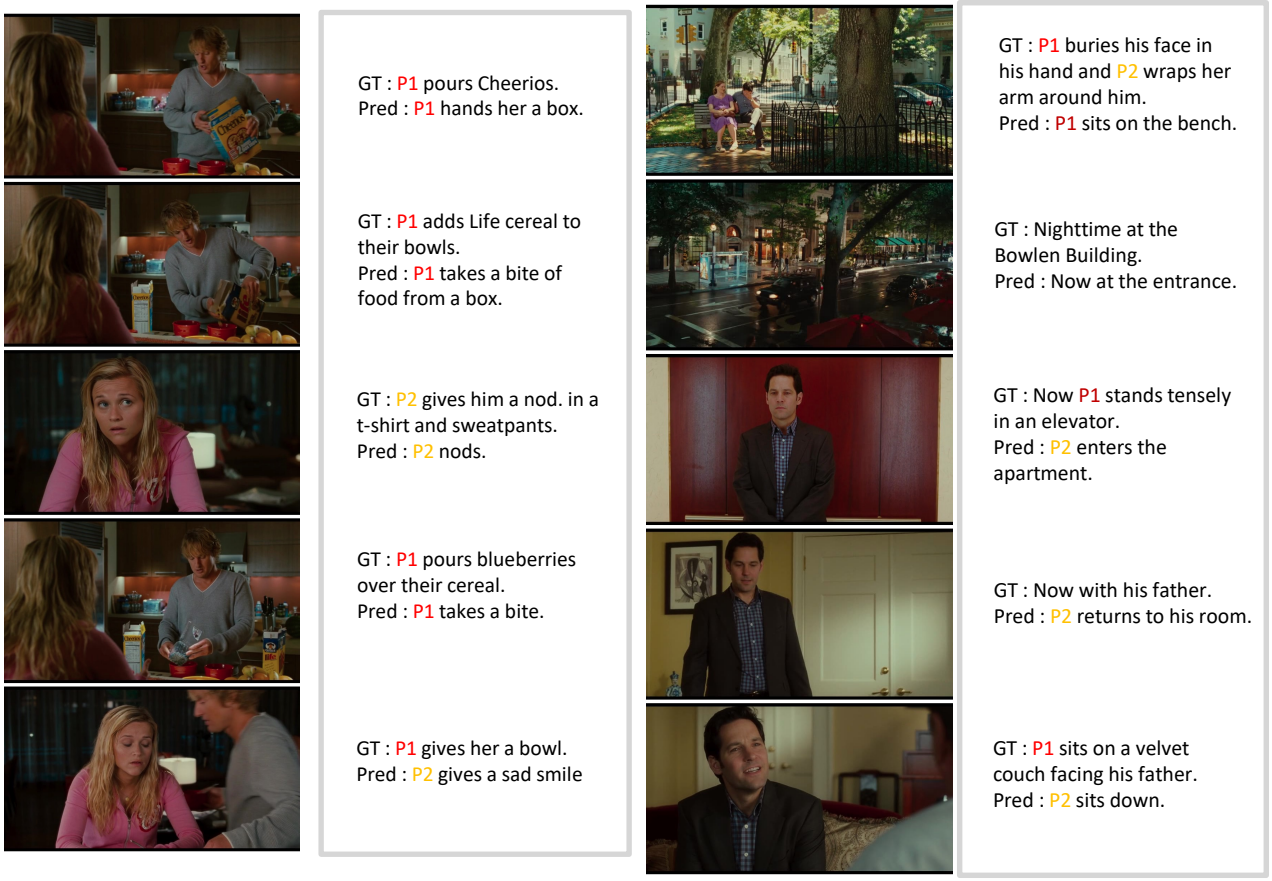


Figure 6. The above examples showcase MICap's ability to perform id-aware captioning. We see that the predicted captions are quite good, although terse. While the GT captions tend to be more descriptive in nature, we believe that such behavior may be introduced in the future by incorporating Large Language Models for captioning.

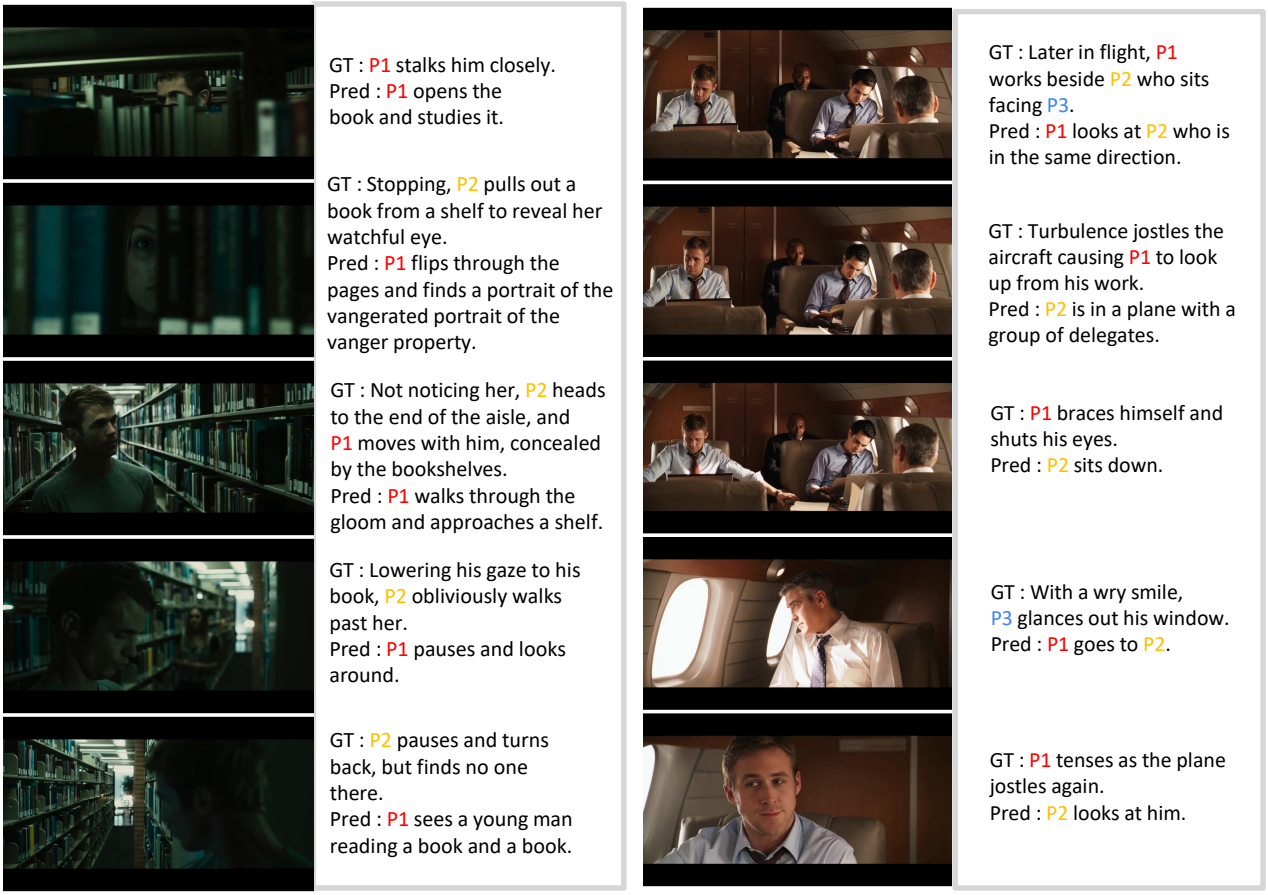


Figure 7. The above examples are relatively difficult cases where there are multiple characters involved with lot of drama or action happening in quick succession. The characters faces are also occluded or partly visible (left example) making it harder to predict identity. We observe that the predicted captions do not capture the tension (*e.g.* plane turbulence) and the identities.