

## A SUPPLEMENTARY MATERIAL

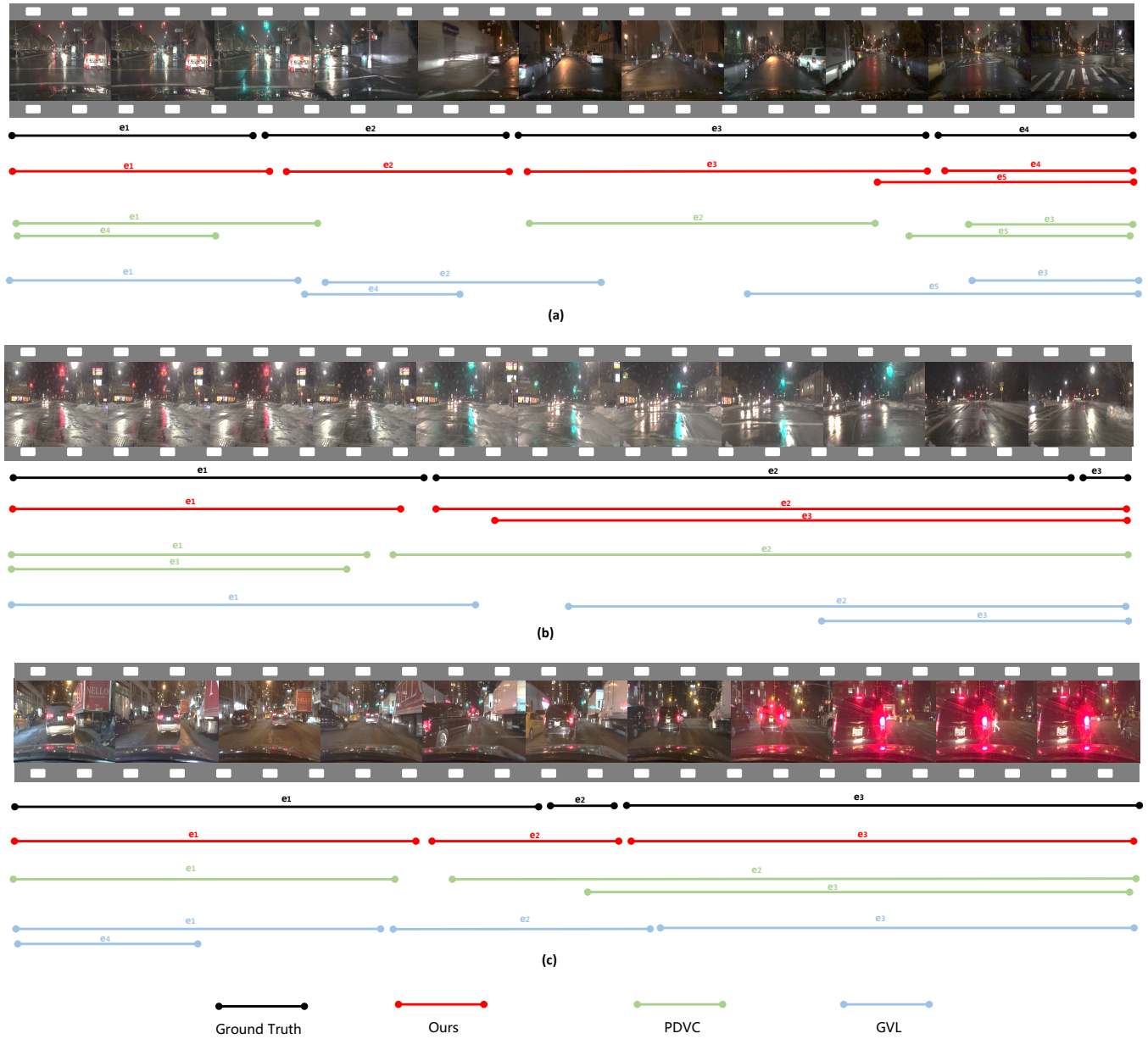
### A.1 Visualization of Examples

In Figure 4, we present several cases of caption generation performance comparing our approach with others. In e5 of the ground

truth in Figure 4 (a), our method detects that the reason for "The car is slowing down" is noticing a stop sign, while GVL mistakes for a red light. In e1 of the ground truth in Figure 4 (b), our method detects the reason for "The car is driving forward" is because it meets a green light while PDVC and GVL mistakenly believe it is



Figure 4: Visualization of several captioning performances.



**Figure 5: Visualization of several location performances. (Note that due to the size of the frame, the visualized locations do not accurately correspond with the frame sequence in this visualization).**

due to a clear road. These cases verify that our method has extracted the driving decision information.

We further show several cases of location performance in Figure 5. Take the ground truth  $e1$  in Figure 5 (a) for instance, its start and end times are more precise compared to other methods of

corresponding events, which proves the validity of our proposed novel adjacent contrastive learning strategy. Additionally, as demonstrated in our cases, our method is more effective at night, a time when the boundary between events is more ambiguous.