Localizing Moments in Video with Temporal Language

Lisa Anne Hendricks[1]*, **Oliver Wang**[2], **Eli Shechtman**[2],
Josef Sivic[2,3*] , **Trevor Darrell**[1], **Bryan Russell**[2]
[1] UC Berkeley, [2] Adobe Research, [3] INRIA

In this supplemental, we include an example illustrating how we create TEMPO-TL sentences and include additional qualitative examples.

1 TEMPO-Template Language

Figure 1 shows examples from our TEMPO-TL dataset. At the top, we show two sentences from the original DiDeMo dataset ("the cross is seen for the first time" and "window is first seen in room"). A "then" moment (left, green) is created by concatenating the two adjacent moments from the DiDeMo dataset and combining the sentences using the word "then" ("The cross is seen for the first time then window is first seen in room."). An "after" moment (right, pink) is constructed by referencing the moment which occurs first ("After the cross is seen for the first time") and then adding the base moment "window is first seen in room." Finally, a "before" moment is constructed by concatenating the two adjacent sentences with the word before to form the sentence "The cross is seen for the first time before window is first seen in room."

2 Qualitative Examples on TEMPO-Human Language

In this section we show qualitative results for when multiple sentences for the same video are localized properly (Figure 2), examples in which our model properly localizes sentence queries (Figures 3-7), and failure cases (Figure 8).

Comparing Different Temporal Words for the Same Video. Figure 2 shows an example where the original DiDeMo moment is localized correctly ("a cat jumps up and spazzes out") as well as temporal sentences "the cat sniffs the floor *before* it jumps ip [sic] and spazzes out", "a cat jumps up and spazzes out *then* it goes under the counter", and "the cat put it's head under the shelf *after* it jumps up and spazzes out." Our MLLC model is able to properly localize each sentence type. Notice that when the annotators construct temporal sentences, they include pronouns like "it" as opposed to repeating the word "cat" multiple times.

Additional qualitative examples. In Figures 3, 4, 5, 6, 7, and 8, we show additional qualitative examples of the moments localized by MLLC as well as the corresponding context moments. In general, we observe the context moment is accurately localized and is sensible for each temporal sentence.

Figure 3 shows correctly localized moments which include the temporal word "before". The top example shows an example in which the context considered by the MLLC model ("cars move forward when traffic lights are green") consists of multiple GIFs. Figure 4 shows correctly localized moments which include the temporal word "after". The bottom example shows an example ("after the camera zoom out from the dancers, the camera zooms back in") where the localized moment and localized context are not contiguous. Figure 5 shows correctly localized moments which include the temporal word "then". In contrast to the temporal words "before" and "after", to correctly localize "then" the chosen context occurs *within* individual moments. Finally, Figure 6 shows examples in which "while" is used and the context moment corresponds to the global context moment. Figure 7 shows an example from the original DiDeMo dataset in which global context is not chosen by the MLLC model. Rather, for the sentence "last view of the ocean", the context corresponds to a moment earlier in the video in which the ocean appears. Because the MLLC model learns to choose context most appropriate for each query, it is not restricted to always using global context. For sentences like "last view of

*Work done at Adobe during LAH's summer internship.

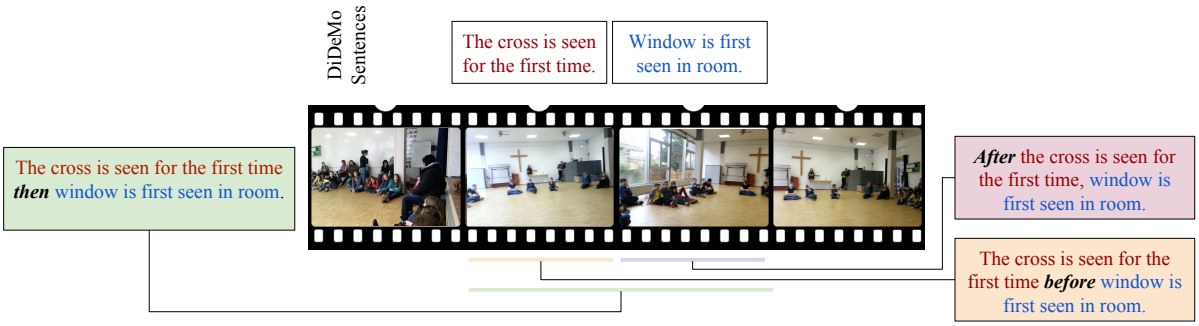

Figure 1: TEMPO - Template Language (TL). We use sentence templates for "before", "after", and "then" to transform human provided sentences from the DiDeMo dataset (top) to temporal language queries (left and right).

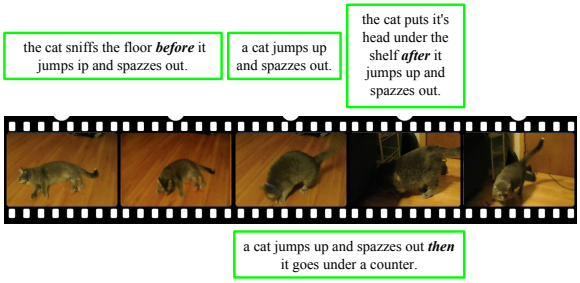

Figure 2: Our MLLC model is able to localize original sentences from DiDeMo (a cat jumps up and spazzes out) as well as newly collected sentence from TEMPO-HL which include temporal words.

the ocean", choosing context which corresponds to when the ocean is seen earlier in the video is sensible. This observation may partially explain why the MLLC model does better on the original DiDeMo dataset.

Finally, Figure 8 shows interesting failure cases. Figure 8 (top) shows an example where the context was localized properly, but the moment was not correctly localized. This could be in part because the query is particularly complex and uses two temporal words ("then" and "before"). Finally, Figure 8 (bottom) shows a failure case for the word "after" where the context moment is temporally related to the localized moment in a sensible way; for an "after" sentence we expect the context to occur *before* the localized moment.

Figure 3: TEMPO - Human Language (HL). Localized moments using the temporal word "before". The blue line shows the context considered when localizing the moment, and the correctly predicted moment is highlighted in green.

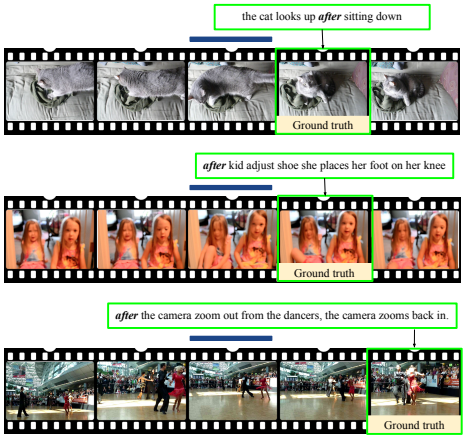

Figure 4: TEMPO - Human Language (HL). Localized moments using the temporal word "after". The blue line shows the context considered when localizing the moment, and the correctly predicted moment is highlighted in green.

Figure 5: TEMPO - Human Language (HL). Localized moments using the temporal word "then". The blue line shows the context considered when localizing the moment, and the correctly predicted moment is highlighted in green.

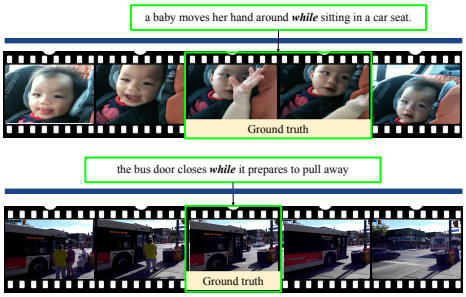

Figure 6: TEMPO - Human Language (HL). Localized moments using the temporal word "while'. The blue line shows the context considered when localizing the moment, and the correctly predicted moment is highlighted in green.

Figure 7: TEMPO - Human Language (HL). Localized moment from the DiDeMo dataset. The blue line shows the context considered when localizing the moment, and the correctly predicted moment is highlighted in green.

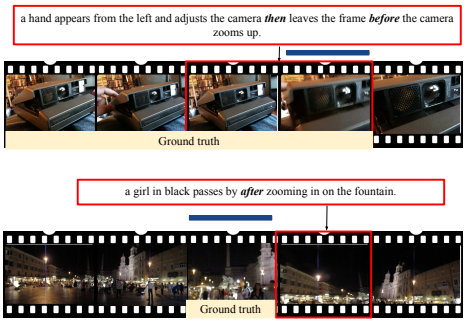

Figure 8: TEMPO - Human Language (HL). Example failure cases. The blue line shows the context considered when localizing the moment, and the (incorrectly) predicted moment is highlighted in red. In the top example, the context is localized correctly, but the moment is not. In the bottom example, the temporal relationship between the context and localized moment is sensible, but the localized moment is incorrect.