Scanning Only Once: An End-to-end Framework for Fast Temporal Grounding in Long Videos
– *Supplementary Material* –

A. Qualitative Grounding Results

Recall that this work targets the temporal grounding in long videos. It proposes a tailored framework as well as training objectives, alleviating the inefficiency, insufficiency and inflexibility issues caused by the sliding window pipeline. We provide some qualitative results in Fig. 1 and Fig. 2 to illustrate the effectiveness of our method. They suggest, our method can achieve flexible boundary localization for various-length target segments. Besides, thanks to the content-enhanced re-ranking, some inconspicuous objects which usually perturbed by other frames can be detected (*e.g.*, the wooden ferry in the third sample of Fig. 1), hence facilitates accurate temporal localization.

Despite the effectiveness, due to the fact that the sentence feature is pre-extracted considering efficiency, some word-to-object alignment may get lost. Here we provide some failure cases of our method in Fig. 3. We observe from top to bottom, our method misunderstands "*skids*", "*a*", "*kerosene lanterns*", "*smoke*", "*soup*", "*carton*" in turn. They suggest that, the lack of explicit token-level semantic alignment learning leads to inadequate semantic analysis to some extent, which is left as our future work.

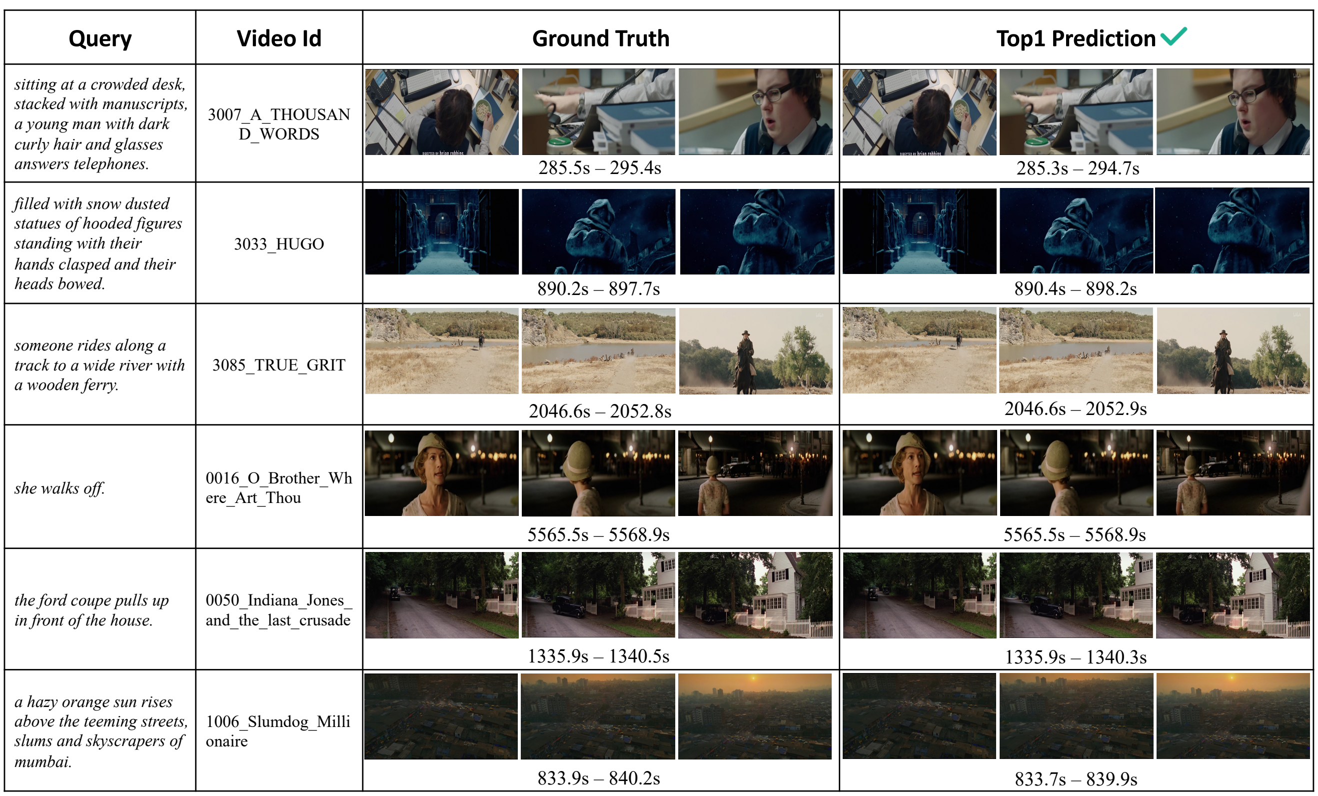


Figure 1. **Visualization of some qualitative results on MAD.**

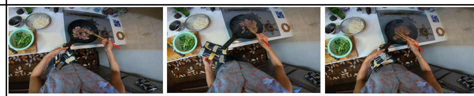
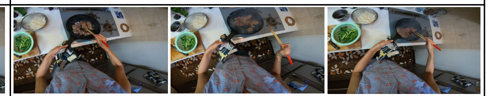
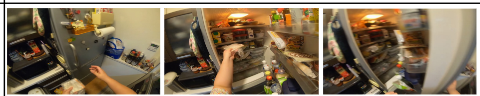
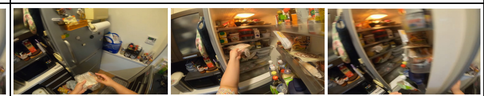
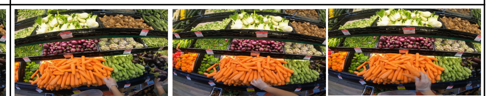
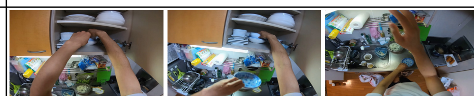
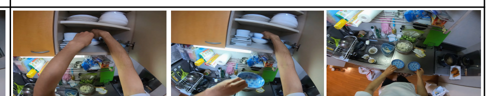
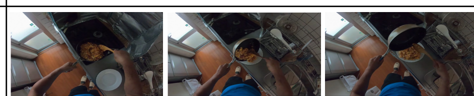
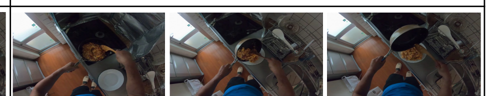
Query	Video Id	Ground Truth	Top1 Prediction ✓
what meat did i fry in the pan?	eb04561c-2ffd-4ea1-aab4-7cadc24db9f9	 90.0s – 144.0s	 98.5s – 149.4s
in what location did i last see the mitten?	ff6d3d52-dda5-46dd-8515-b9b772933030	 194.0s – 198.1s	 192.8s – 197.9s
how many carrots did i pick?	fbc425c4-def6-49a7-8b88-5d5d00b5524e	 463.6s – 495.4s	 466.9s – 493.8s
how many plates did i take from the top shelf?	404cc1c1-f7a0-4e16-9a39-b8e2d5d9ae59	 323.0s – 330.0s	 323.5s – 329.6s
who did i interact with when i was standing?	e4a01f13-4f09-4ee4-ae13-17af72eaca87	 0.0s – 3.0s	 0.0s – 3.1s
what did i put in the plate?	224c3de4-9683-462a-8eb4-224773425a7e	 303.2s – 350.0s	 311.5s – 346.3s

Figure 2. **Visualization of some qualitative results on Ego4d-Video-NLQ.**

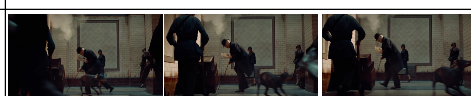
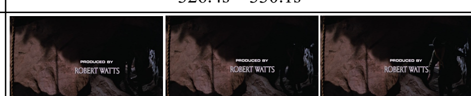
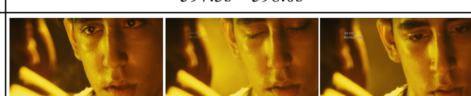
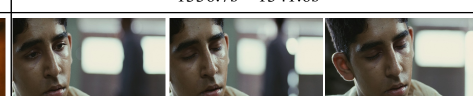
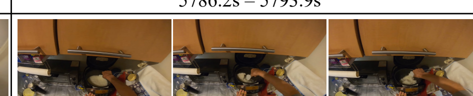
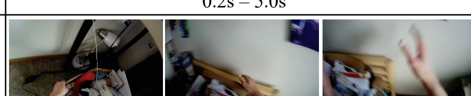
Query	Video Id	Ground Truth	Top1 Prediction ✗
the dog skids across the polished floor as he runs up the hallway opposite the toyshop.	3033_HUGO	 448.9s – 450.9s	 1414.1s – 1422.2s
someone takes a complimentary coffee and leaves.	3007_A_THOUSAND_WORDS	 326.4s – 330.1s	 1767.0s – 1770.8s
the light of kerosene lanterns dances on the tunnel walls ahead.	0050_Indiana_Jones_and_the_last_crusade	 597.3s – 598.0s	 1336.7s – 1341.6s
the prisoner, someone, blinks rapidly as the smoke stings his eyes.	1006_Slumdog_Millionaire	 84.6s – 90.2s	 5786.2s – 5793.9s
what did i use to stir the soup?	413fe086-1745-4573-b75b-e7d26ff72df9	 0.2s – 5.0s	 437.9s – 495.5s
where did i put carton?	38737402-19bd-4689-9e74-3af391b15feb	 808.0s – 814.0s	 1366.9s – 1369.7s

Figure 3. **Visualization of some failure cases on MAD and Ego4d-Video-NLQ.**