

Analysis of supplementary Cases

To further evaluate the performance of our model, we conducted qualitative experiments using examples from other datasets, such as Charades-STA, and even included a clip from *Tom and Jerry*. The comparisons involved existing VTG models (e.g., UniVTG), our baseline model, and the proposed DDM-VTG model. These experiments covered three aspects described in Figure 2 (a), (b), and (d) with a total of four comparative cases. Below, we provide a detailed analysis of each case.

Case 1 (Figure 2a):

We selected a clip from *Tom and Jerry* as a representative out-of-distribution (OOD) sample. This animation video differs significantly from the QVHighlight training dataset. As shown in the results, none of the baseline models successfully localized the segment corresponding to the query "a mouse is sitting on the cat's paw." In contrast, our method not only estimates uncertainty to reflect the model's lack of knowledge but also demonstrates improved uncertainty reasoning with DDM-VTG compared to the baseline.

Case 2 (Figure 2a):

This case uses a video from Charades-STA with a 90-degree rotated frame, serving as another OOD sample. Results indicate that while all models struggled to localize the target segment, our approach significantly improved uncertainty estimation. DDM-VTG, compared to the baseline, showcased superior sensitivity to visual anomalies in the video.

Case 3 (Figure 2d):

This case originates from Charades-STA and highlights a video segment where a man transitions from a dimly lit living room to a brightly lit kitchen, causing significant lighting changes. This abrupt change adversely affects the accuracy of the baseline models, which fail to identify and respond to this uncertainty. In contrast, DDM-VTG assigns a high uncertainty score in response to the drastic low-level feature variation, illustrating its robustness against such environmental changes.

Case 4 (Figure 2b):

We evaluated a video from Charades-STA depicting a man using a computer. When the query was changed from "a man is watching computer" to "a man is watching television," our model significantly increased uncertainty estimates. This reflects its sensitivity to the rarity of televisions in the video. By comparison, the baseline failed to adequately adjust its uncertainty. Furthermore, the inability to localize the original query "a man is watching computer" likely stems from the target segment encompassing the entire video, a labeling scenario that is rare in the training data and presents additional challenges as an OOD sample.