STOP: Integrated Spatial-Temporal Dynamic Prompting for Video Understanding

Supplementary Material

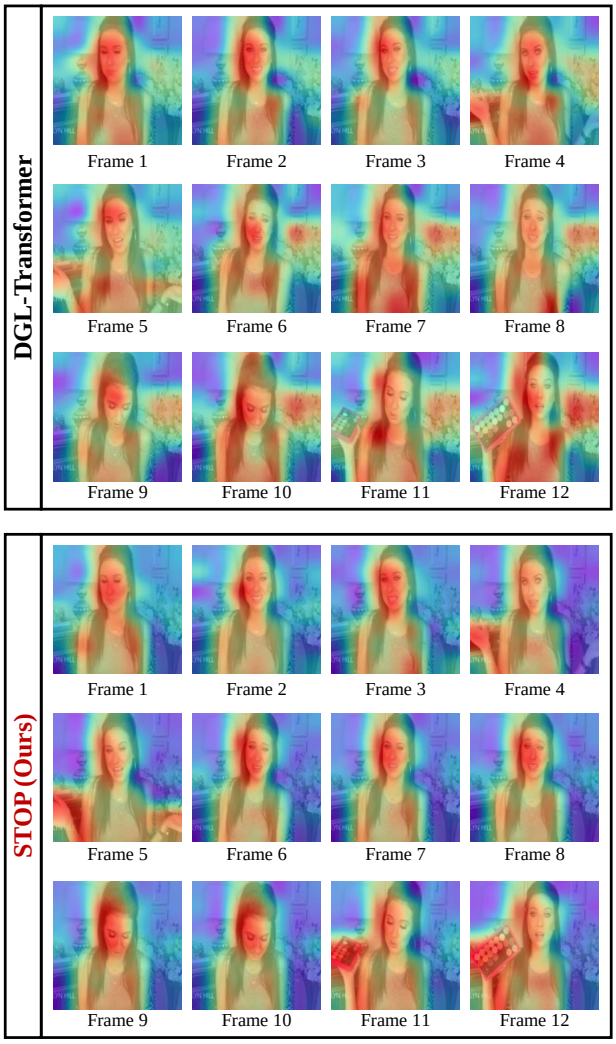

Figure 1. More Attention map visualization results of existing method DGL-Transformer [1] and STOP (Ours).

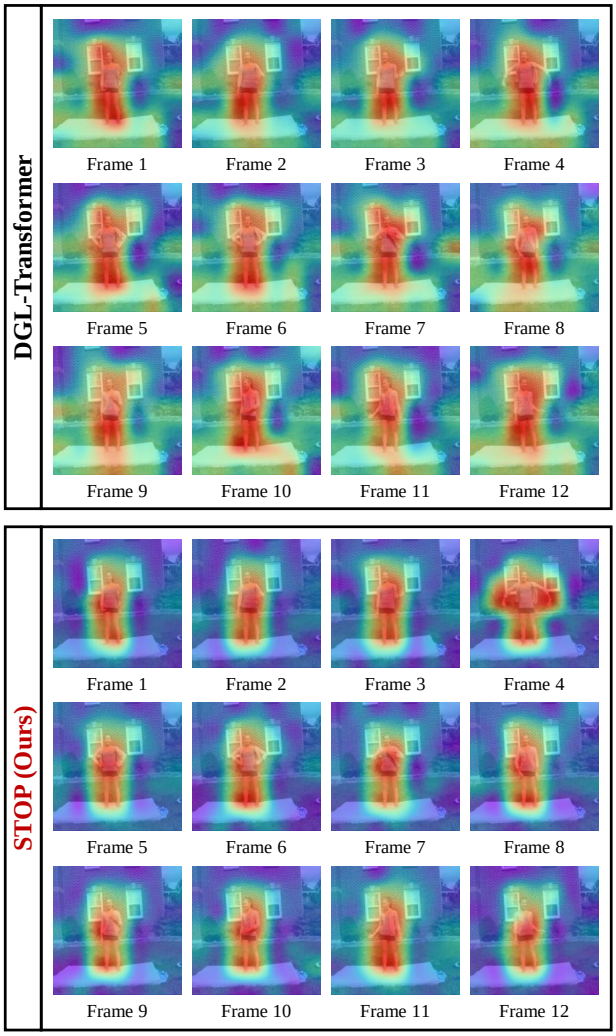

Figure 2. More Attention map visualization results of existing method DGL-Transformer [1] and STOP (Ours).

1. Attention Map Visualization of More Cases

To further explore the impact of our intra-frame spatial prompting and inter-frame temporal prompting, we visualized the attention maps for more video cases. As shown in Figure 2 and Figure 1, existing video prompting methods (e.g., DGL-Transformer) use the same static prompt for all videos. This causes the pre-trained CLIP model to focus on static objects and backgrounds in the video, making it challenging to accurately understand the actions of people in the video.

In contrast, our intra-frame spatial prompting and inter-frame temporal prompting highlight the key regions with dynamic changes in the video, enabling the pre-trained model to focus on the people and their actions. This leads to a more accurate understanding and shows a similar trend to the visualized results presented in the main text.

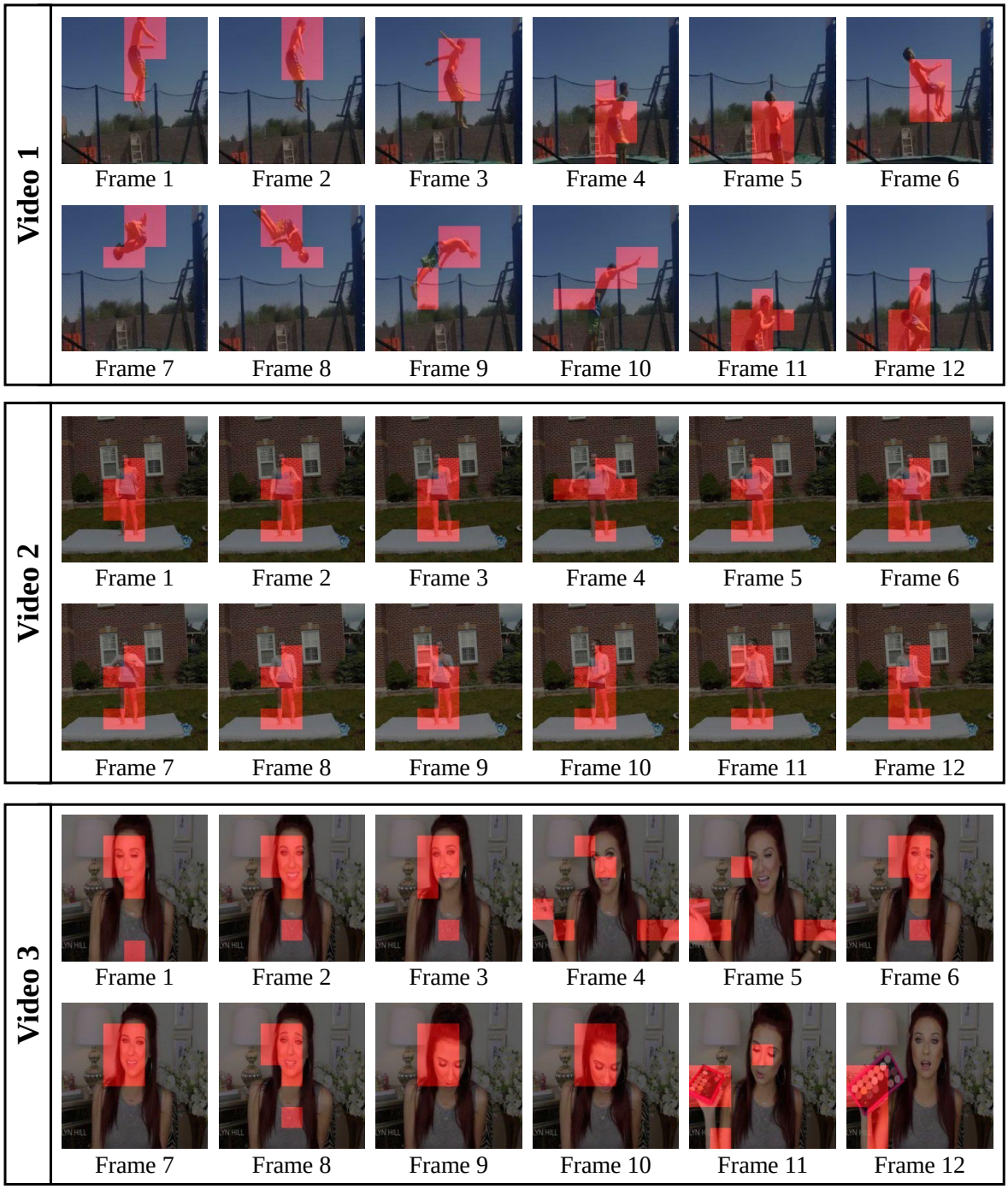

Figure 3. The visualization results of the positions where the intra-frame spatial prompt is added.

2. Visualization of Intra-Frame Spatial Prompt

To verify the effectiveness of our intra-frame spatial prompting, we visualized the positions where it is added. As shown in Figure 3, the red patches indicate the locations where the intra-frame spatial prompt is applied. It can be observed that our method comprehensively considers intra-frame attention weights and temporal variations, enabling accurate localization of discriminative regions in the video.

This facilitates the pre-trained vision-language model to focus accurately on these discriminative regions, enhancing the model's ability to extract temporal information.

References

[1] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *AAAI*, pages 6540–6548, 2024. 1