DLCV Fall 2023 Final Challenge 2
Visual Queries 2D Localization Task

Group6 **Nani**

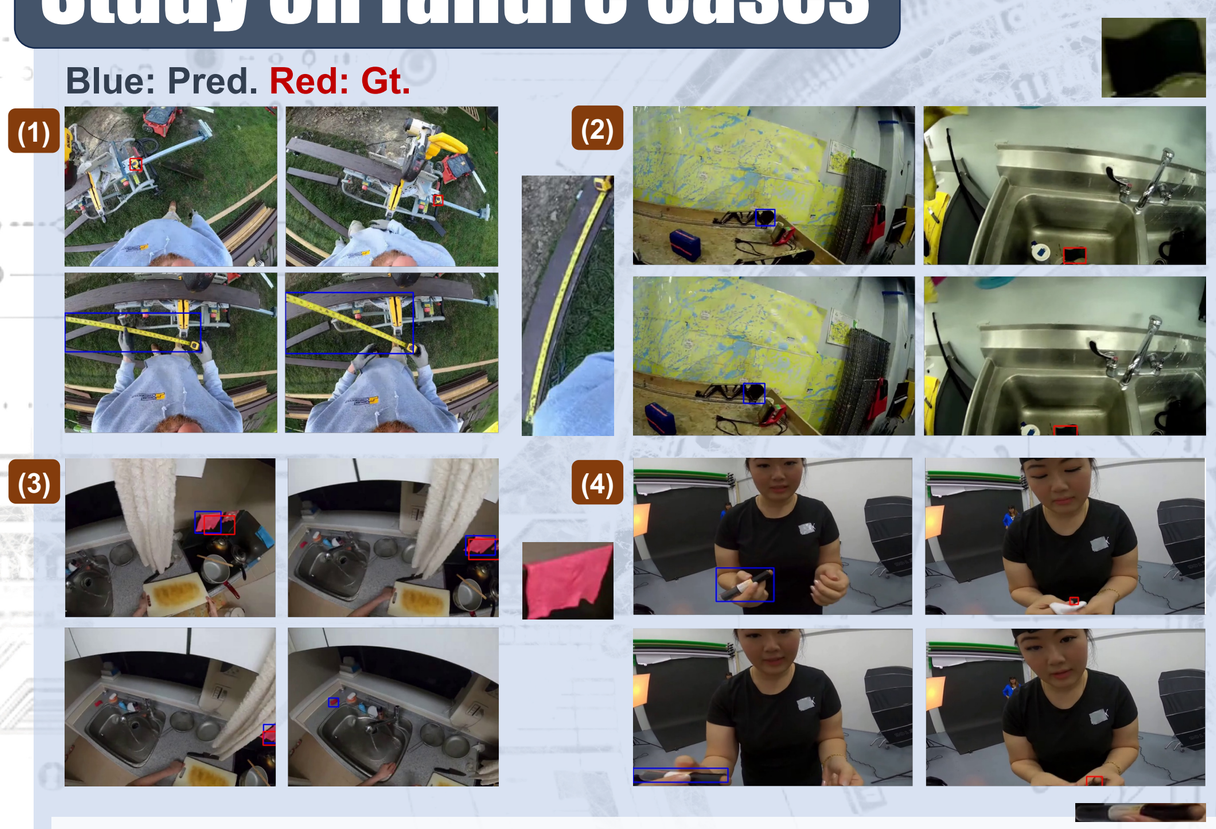
B08901210 陳祈安 R11945073 林芳瑜
R11942085 許宸睿 R12921040 柯岱佑

Abstract

We started with the state-of-the-art architecture, VQLoc, and used the GitHub code/pre-trained weights to reproduce the results presented in the paper. Following this, we conducted error analysis, categorizing mistakes into four types and proposing potential improvements for each. Our initial findings showed that the values of tAP and stAP were closely matched; hence, our enhancements focused primarily on increasing temporal precision. We extracted features from the original model to train a simple binary classifier to determine the presence of the target object within frames, integrating these findings with initial results for prediction. On the validation set, we achieved an 18% increase in both tAP and stAP. Additionally, we proposed and implemented two potential post-processing methods.

Study on failure cases

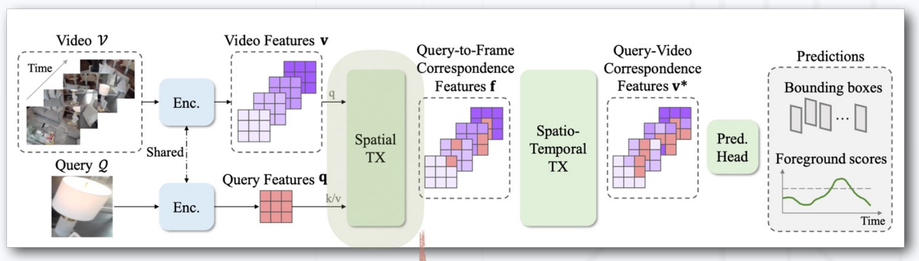
Blue: Pred. Red: Gt.



(1) Misleading Query Image: the query shows an extended tape measure, but the label indicates a retracted tape measure, leading to a misjudgment. (2) Similar Color and Texture: we incorrectly identified a black shelf with a similar shape and color. (3) Not Cleanly Cut: after the query disappears, Pred immediately replaces it with another object in the space.(4) Small Part of Object Shown: only a very small part of the query is visible in the Gt. label, making it difficult to recognize.

Temporal resolution

For the baseline model, when the temporal Average Precision (tAP) is accurate, the spatial accuracy exceeds 70%. Therefore, we are concentrating solely on improving temporal resolution. Our approach is outlined as follows:

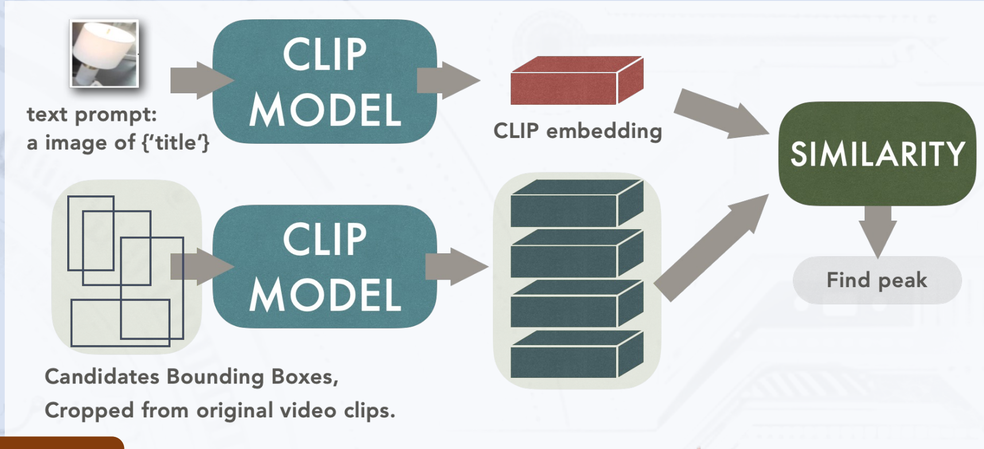


1 We believe query object is in the frame
0 Otherwise

We believe that binary classification is simpler and more efficient compared to training an additional model for ensemble purposes.

Post-Processing

1. CLIP



2. Continuity

1. Post-Processing by CLIP

We aim to resolve failure cases (1), (2), and (4) by utilizing the titles in the annotation files of images. Since CLIP broadly captures the association between text and images, we introduce Similarity as the basis for determining the peak algorithm.

2. Post-Processing based on continuity

Theoretically, the predicted location of the Query object in the video should have a certain continuity. Therefore, we plan to make improvements and corrections to failure case (4) by a Rule-based approach.

Results & Conclusion

Setting	stAP(val)	tAP(val)	stAP(test)
Baseline	0.2306	0.3039	0.2897
Ours	0.2742	0.3595	0.2937
Ours (PP)	0.2284	0.3049	0.2877

Our method has significantly improved tAP on the validation set, with stAP progressing at a similar rate. This aligns with our initial observation that the primary bottleneck in the VQ2D task stems from inadequate temporal resolution. Regrettably, the post-processing techniques we attempted don't seem to address the cases we organized during our error analysis effectively.

Corrected Example