RELOCATE: A Simple Training-Free Baseline for Visual Query Localization Using Region-Based Representations

Supplementary Material

k	stAP_{25}	tAP_{25}	Success	Recovery
5	0.302	0.371	56.5	49.9
10	0.333	0.409	58.0	50.5
25	0.329	0.404	58.2	50.6
50	0.330	0.409	58.5	50.8

Table 5. **Effect of initially selected candidates on model performance.** Our final evaluations use $k = 10$.

t_{sim}	stAP_{25}	tAP_{25}	Success	Recovery
0.6	0.348	0.446	58.4	47.8
0.7	0.333	0.409	58.0	50.5
0.8	0.258	0.316	52.9	48.0

Table 6. **Effect of candidate selection threshold on model performance.** Our final evaluations use $t_{\text{sim}} = 0.7$.

t_{nms}	stAP_{25}	tAP_{25}	Success	Recovery
0.6	0.308	0.379	57.1	50.9
0.7	0.320	0.393	57.8	51.0
0.8	0.333	0.409	58.0	50.5
0.9	0.324	0.404	58.3	50.8

Table 7. **Effect of inter-frame NMS threshold on model performance.** Our final evaluations use $t_{\text{nms}} = 0.8$.

t_{q}	stAP_{25}	tAP_{25}	Success	Recovery
0.4	0.320	0.402	58.2	50.2
0.5	0.333	0.409	58.0	50.5
0.6	0.320	0.396	58.0	50.4

Table 8. **Effect of query selection threshold on model performance.** Our final evaluations use $t_{\text{q}} = 0.5$.

Method	stAP_{25}	tAP_{25}	Success	Recovery
SAM 2 [29]	0.290	0.329	55.0	42.7
RELOCATE	**0.378**	**0.458**	**63.0**	**49.1**

Table 9. **Evaluating SAM 2 on VQ2D.** Here, we evaluate on 100 randomly sampled examples from the VQ2D validation set.

Category	SAM 2	RELOCATE
Last occurrence localized	54	61
Prior occurrence localized	24	32
Wrong object localized	18	7
No track returned	4	0

Table 10. **Response track prediction analysis of SAM 2 and RELOCATE.** We compare the predictions of SAM 2 and RELOCATE on 100 sampled examples from the VQ2D validation set. Predictions are categorized into four types, and the count for each category is reported.

This supplementary material is structured as follows. In Appendix A we analyze the sensitivity of RELOCATE to its hyperparameters. In Appendix B we study the performance of SAM 2 on the VQL task.

A. Hyperparameter Sensitivity Analysis

We analyze RELOCATE's sensitivity to four key hyperparameters: (1) the maximum number of initially retrieved candidates k, (2) the candidate selection threshold t_{sim}, (3) the inter-frame NMS threshold t_{nms}, and (4) the query selection threshold t_{q}. Tables 5-8 and Figure 7 present model's performance across different hyperparameter configurations.

For the initial retrieval count k, we observe stable performance across values from 10 to 50, with only a slight degradation at $k = 5$. The candidate selection threshold t_{sim} leads to a noticeable decline in performance when set above 0.7. The inter-frame NMS threshold t_{nms} demonstrates consistent performance across the range 0.7-0.9, suggesting robustness to this parameter. Similarly, the query selection threshold t_q shows minimal variation in performance between 0.4 and 0.6.

Overall, these results indicate that our model maintains stable performance across a wide range of hyperparameter values, with selected values of $k = 10$, $t_{\text{sim}} = 0.7$, $t_{\text{nms}} = 0.8$, and $t_{\text{q}} = 0.5$ providing a robust operating point.

B. Evaluating SAM 2 on VQ2D

Jiang et al. [15] demonstrated significant limitations in VQL capabilities among contemporary tracking systems. Specifically, they showed that STARK [41], a state-of-the-art visual tracker at the time, achieves only a 0.04 stAP_{25} score on the VQ2D validation set. Since then, tracking systems have advanced considerably. To evaluate the capabilities of current tracking systems, we test SAM 2 [29] on the VQL task.

To adapt SAM 2 for VQ2D, we prepend the query frame to the target video and use the query bounding box from the annotations as the prompt for mask generation. SAM 2 then propagates the generated mask across all subsequent frames, tracking multiple occurrences of the query object. We select the last contiguous track as the response track prediction.

We evaluate SAM 2 on 100 randomly sampled examples

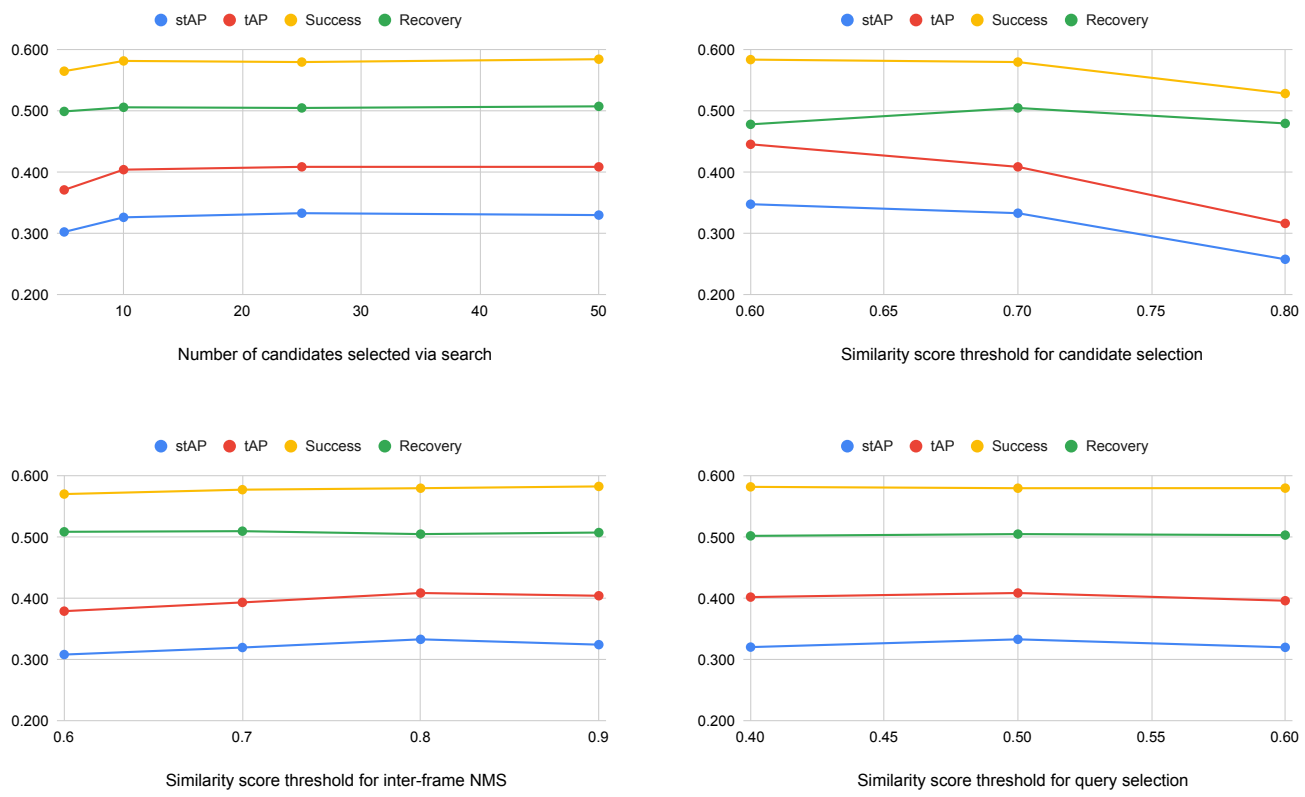

Figure 7. **Hyperparameter sensitivity analysis of** RELOCATE**.** Empirical evaluation demonstrates RELOCATE's robustness across different hyperparameter configurations.

previously used for the manual analysis of RELOCATE reported in Section 4.1, and the results are shown in Tables 9 and 10. While SAM 2 shows competitive performance on VQ2D (Table 9), it underperforms compared to RELOCATE. Our qualitative analysis (Table 10) reveals that SAM 2 has a higher tendency to localize incorrect objects or produce no tracks compared to RELOCATE. On an NVIDIA A40, with our implementation, SAM 2 takes an average of 110.7 seconds to locate a query object in a 1000-frame video. In comparison, RELOCATE incurs a one-time cost of 1422.5 seconds to prepare a 1000-frame video, followed by 73.6 seconds to process each query. However, the processing time of RELOCATE can be significantly reduced by using batch processing and faster SAM variants.