DLCV FALL 2023 FINAL VQ2D

GROUP6

NANI

B08901210 陳祈安　R11945073 林芳瑜

R11942085 許宸睿　R12921040 柯岱佑

(R11942118 李振勳)

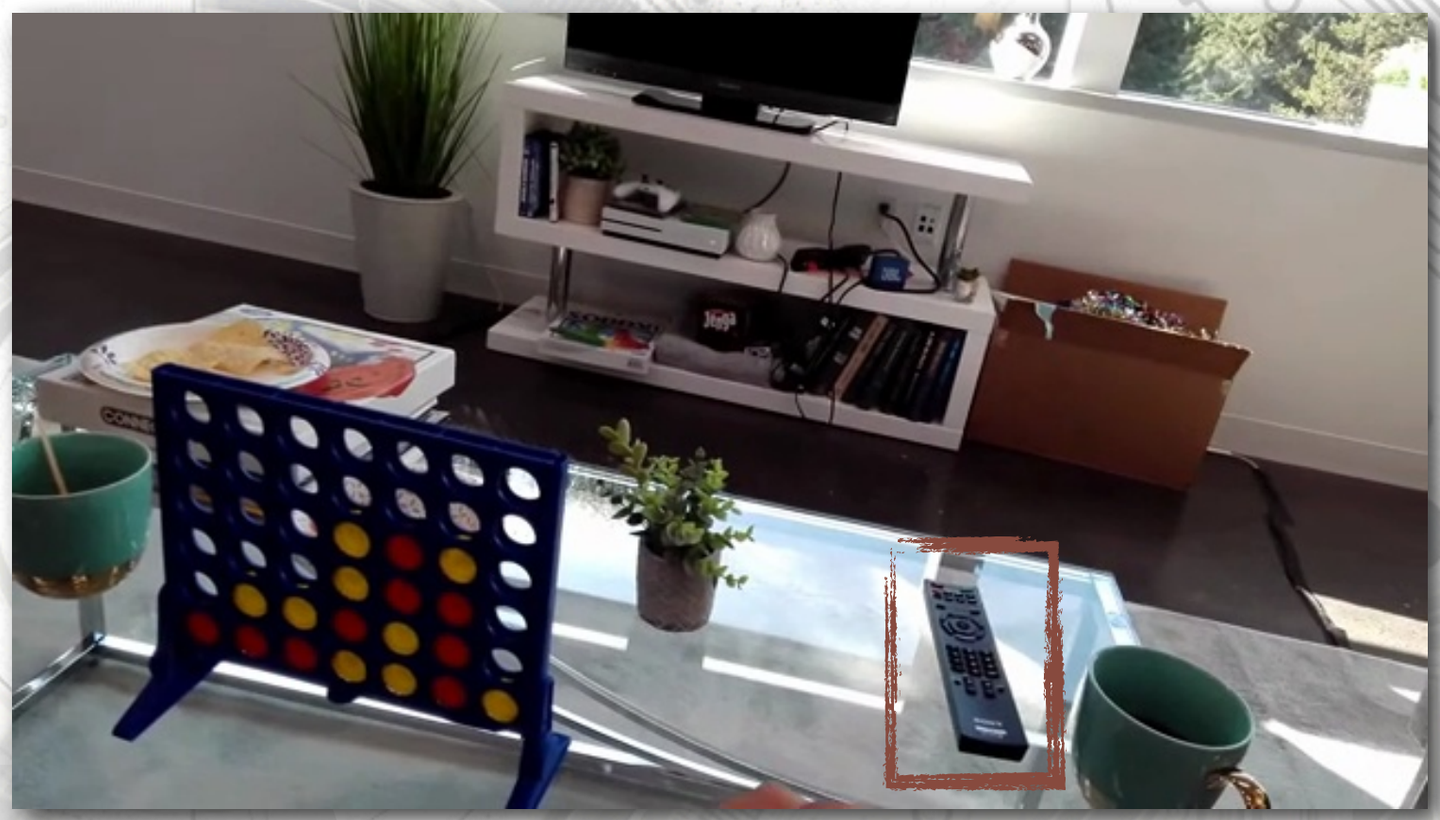
OUTLINE

- Introduction of the Problem
- Baseline: VQLoc
- Study on failure cases
- Our Approach
- Result & Conclusion

INTRODUCTION

Sample from “0e7fba95-22d9-4ab0-9815-4bb7880d8557"


```
"object_title": "remote control",
"visual_crop": {
    "frame_number": 126,
    "x": 411,
    "y": 255,
    "width": 48,
    "height": 83,
    "original_width": 640,
    "original_height": 360
}
```

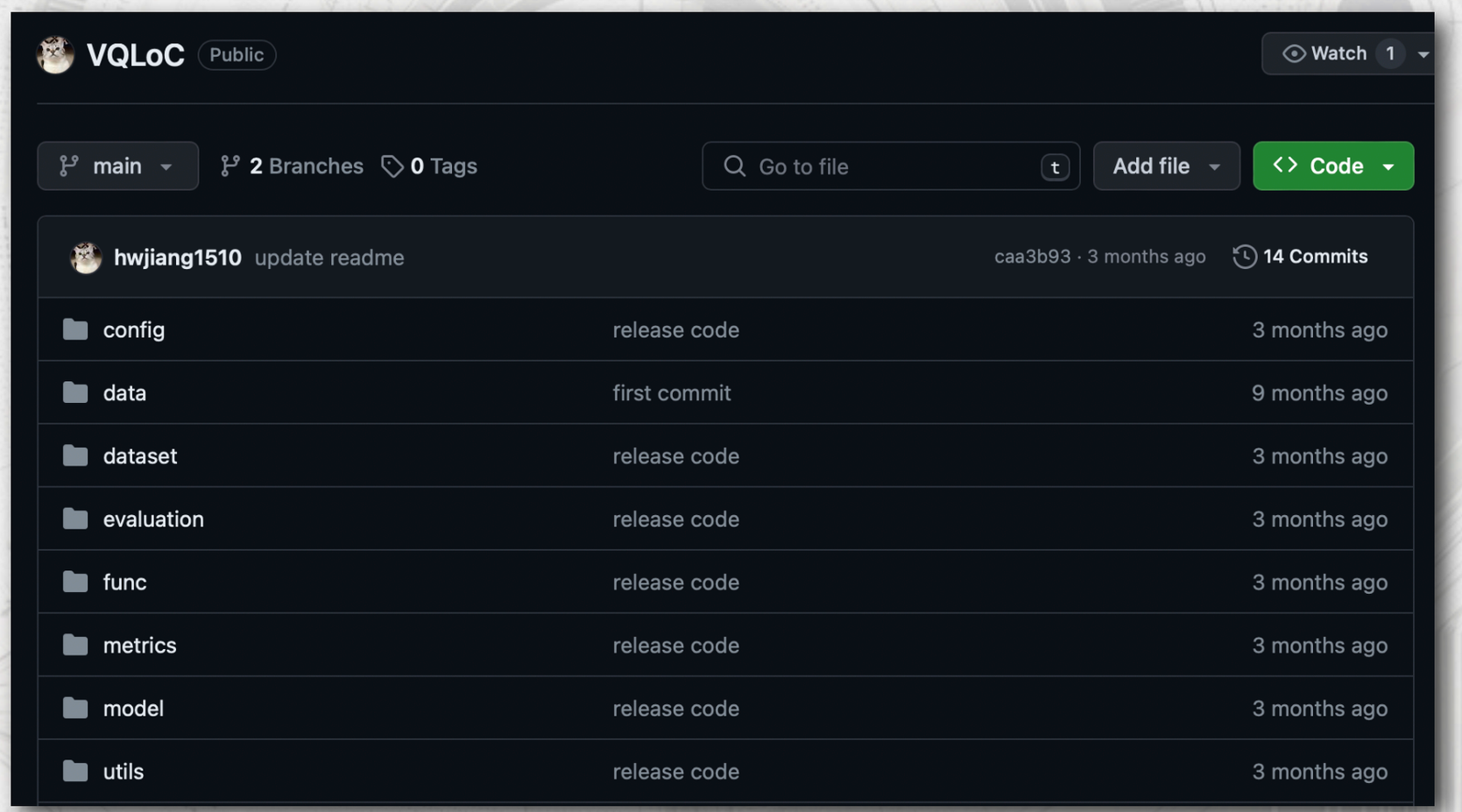



when was the last time that I saw X

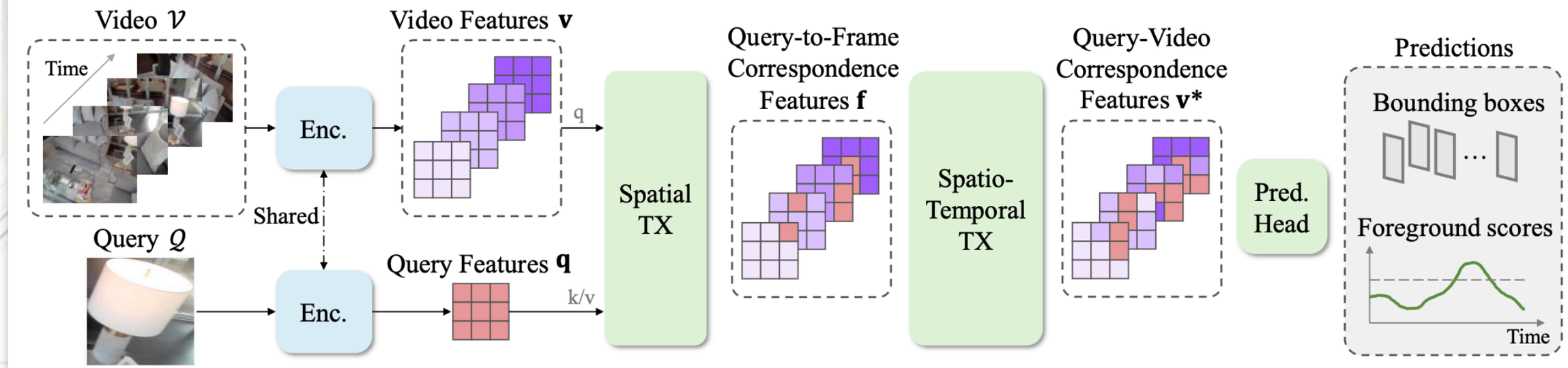
BASELINE

https://github.com/hwjiang1510/VQLoC

- Paper: https://arxiv.org/pdf/2306.09324.pdf
- top entry on the Ego4D VQ2D challenge leaderboard



BASELINE - BRIEF INTRODUCTION



Stage	Configuration	Output
0	Video Frames	$T \times 448 \times 448 \times 3$
0	Visual Query	$448 \times 448 \times 3$
	Encoder	
1	Frame features	$T \times 32 \times 32 \times 256$
1	Query features	$32 \times 32 \times 256$
	Spatial Transformer	
2	Updated frame features	$T \times 32 \times 32 \times 256$
	Spatio-Temporal Transformer	
3	Downsampled frame features	$T \times 8 \times 8 \times 256$
3	Propagated frame features	$T \times 8 \times 8 \times 256$
	Prediction Heads	
4	Anchor refinement	$T \times N \times 4$
4	Anchor scores	$T \times N$
	If inference	
5	Top-1 anchor	$T \times 4$
5	Top-1 score	T

Table 6: Model workflow.

1. Resize and Normalize (query image and each frame)
2. Visual Encoder: DinoV2
3. Cross attention to mix query features to frame features
4. Self attention to capture temporal information
5. Use the features to generate bounding boxes and probability curve
6. Predict the final result

STUDY ON FAILURE CASES

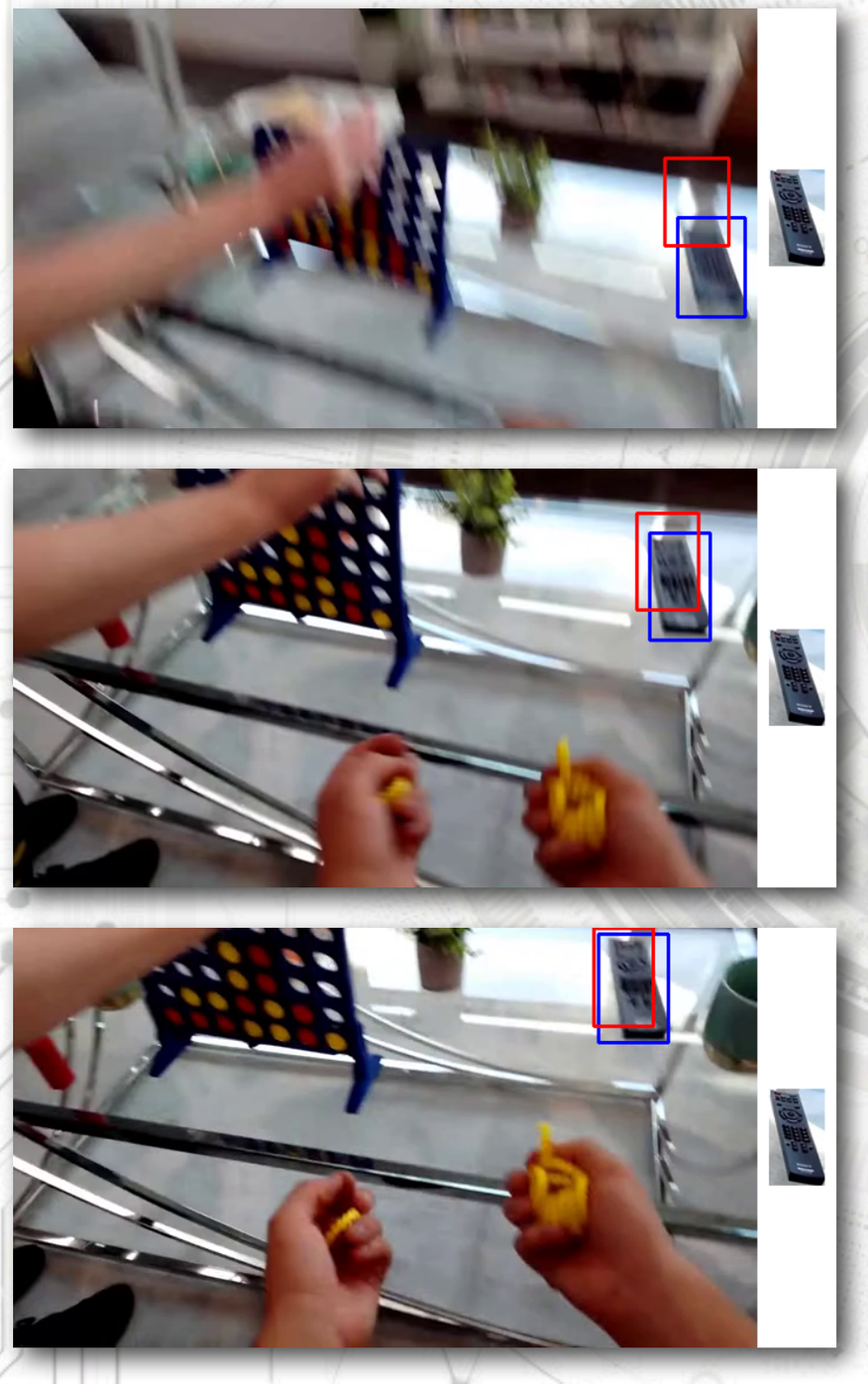
Use the pre-train weights given in the GitHub repo, we modify the code to fit our dataset and inference. We visualize the result, and classify them into four types:

1. Noisy label, misleading query image
2. Similar color and texture
3. Small part of object shown
4. Not cleanly cut

Noisy label, misleading query image

STUDY ON FAILURE CASES

0e7fba95-22d9-4ab0-9815-4bb7880d8557



Blue: Prediction result

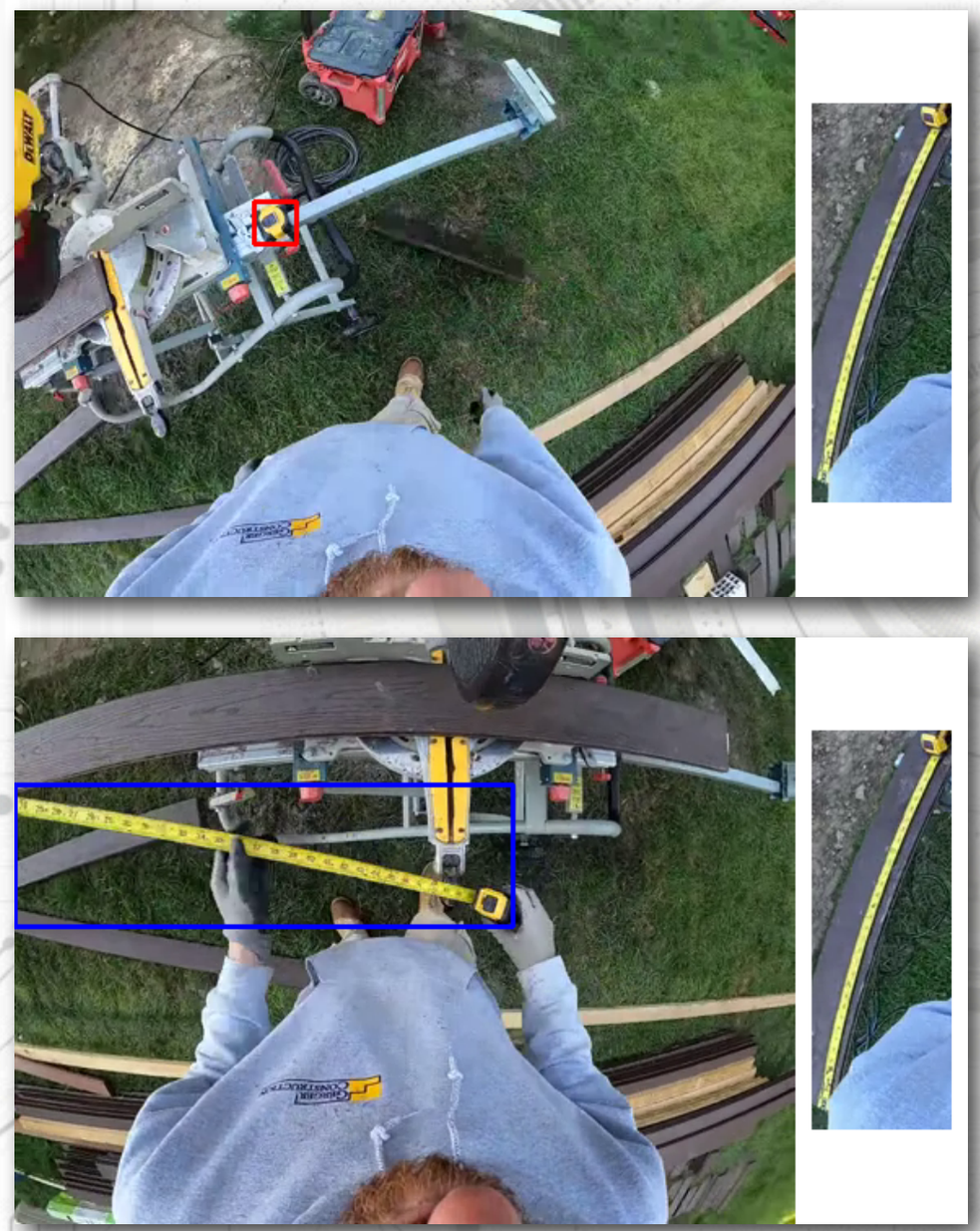
Red: Ground truth

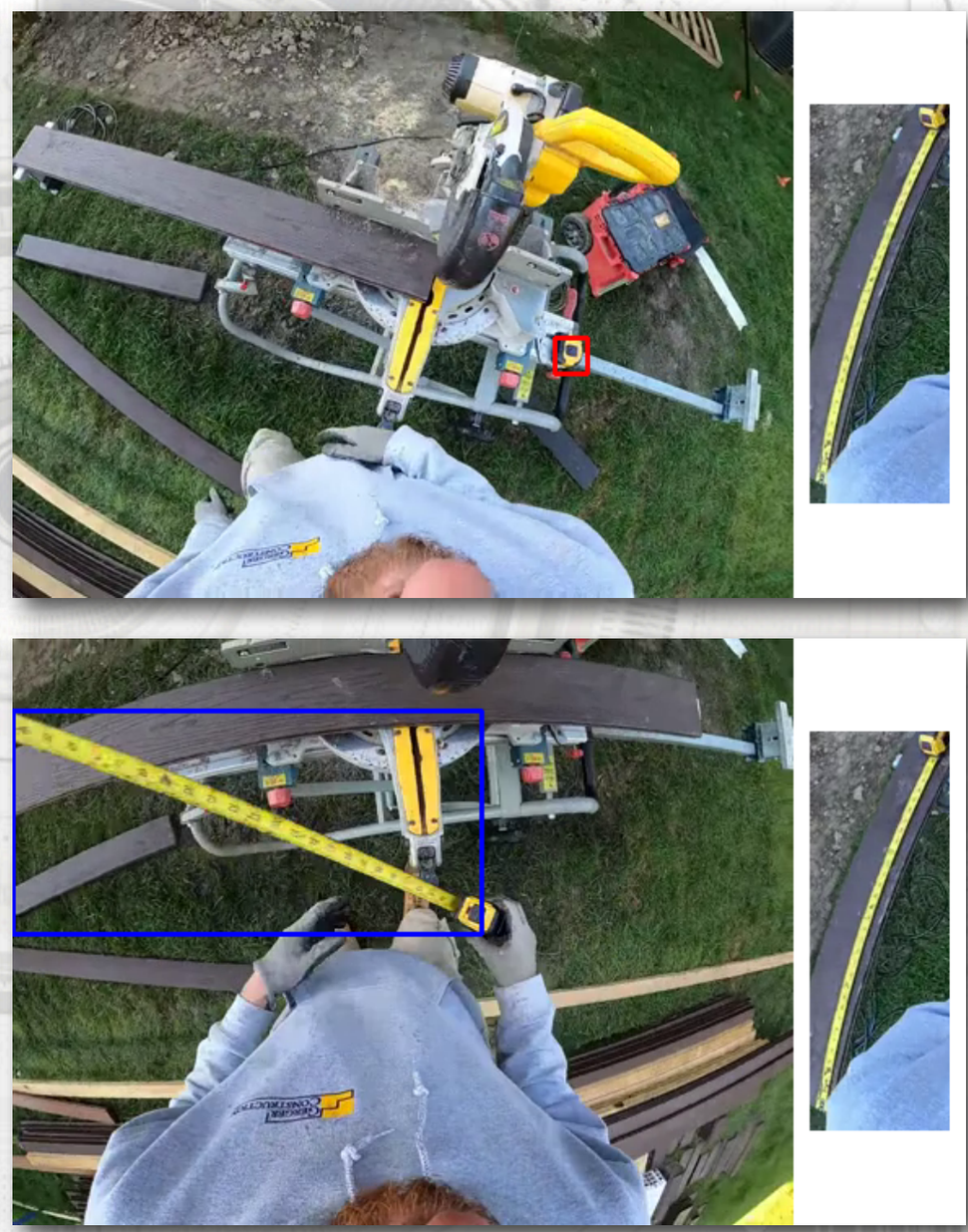
Query image

Noisy label, misleading query image

STUDY ON FAILURE CASES

3d7b86a0-63e7-406a-ab4e-9e31a99898ca



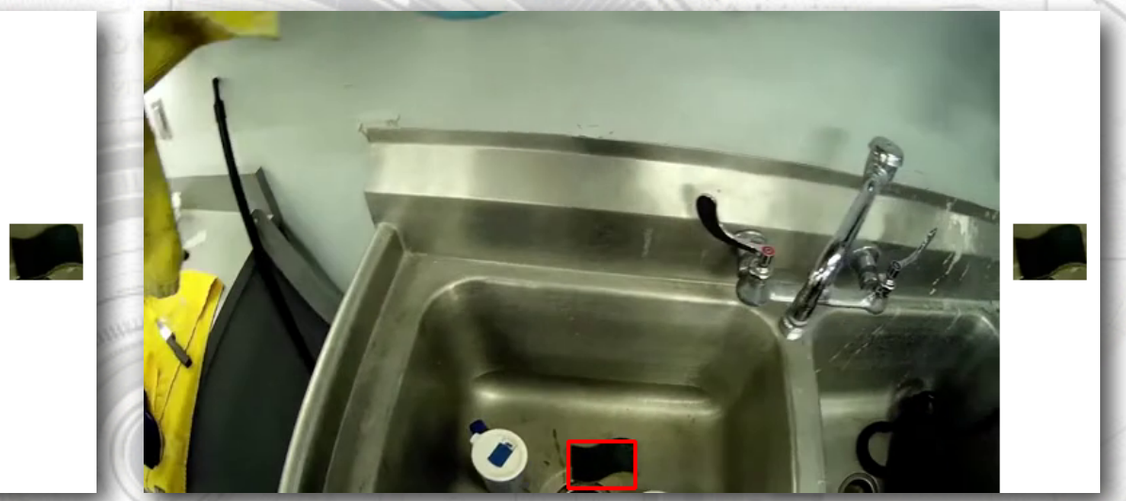


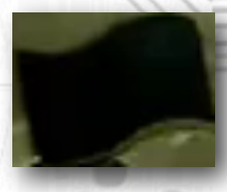
Query image

Similar color and texture

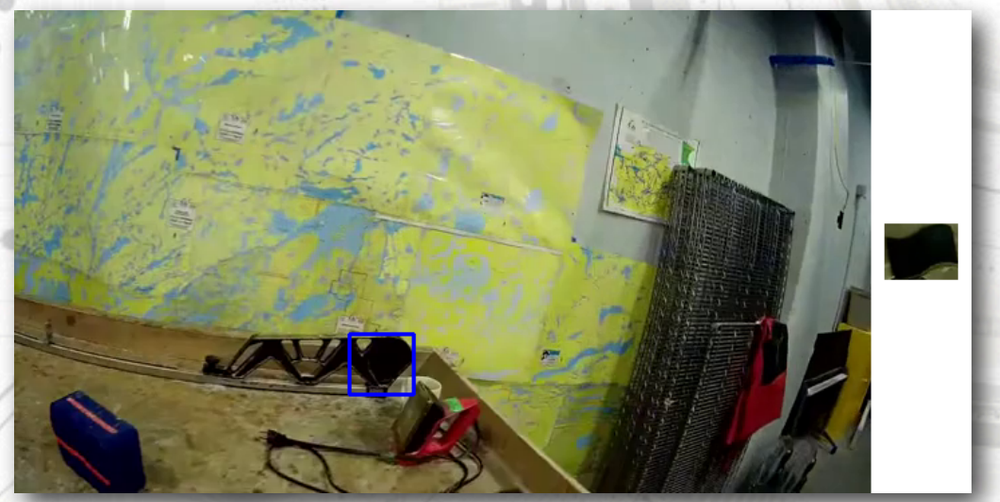
STUDY ON FAILURE CASES

2f790afe-53ca-4fbe-8182-89c0efdccb06





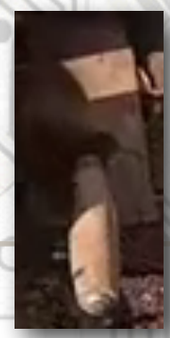
Query image



Similar color and texture

STUDY ON FAILURE CASES

1b7b16dd-4319-44c0-b26c-6bb55de753dd



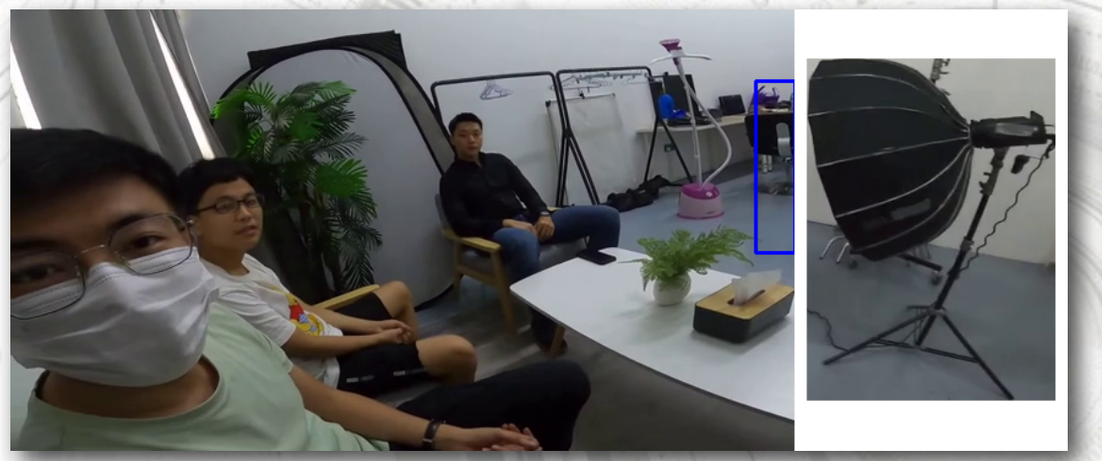
Query image

Small part of object shown

STUDY ON FAILURE CASES

1cdc9561-05fa-4842-a463-9fb6273b4685

Query image

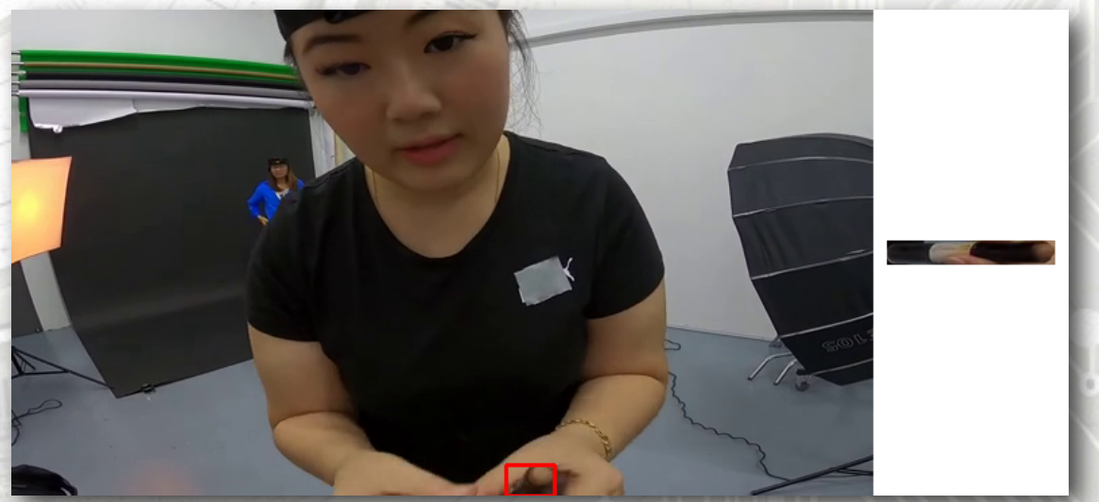


Small part of object shown

STUDY ON FAILURE CASES

1cdc9561-05fa-4842-a463-9fb6273b4685

Query image



Not cleanly cut

STUDY ON FAILURE CASES

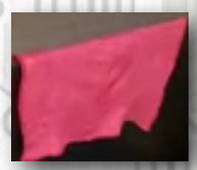
4a3d17d1-e2ec-4b6c-8e14-8908040895bd

Query image

Not cleanly cut

STUDY ON FAILURE CASES

4b5dc015-e442-4a37-a318-65ab2bc9b2e3



Query image

OUR APPROACH

Based on the study on failure cases, we propose 3 methods, one in training stage and the others in post-processing. The overall architecture is unchanged.

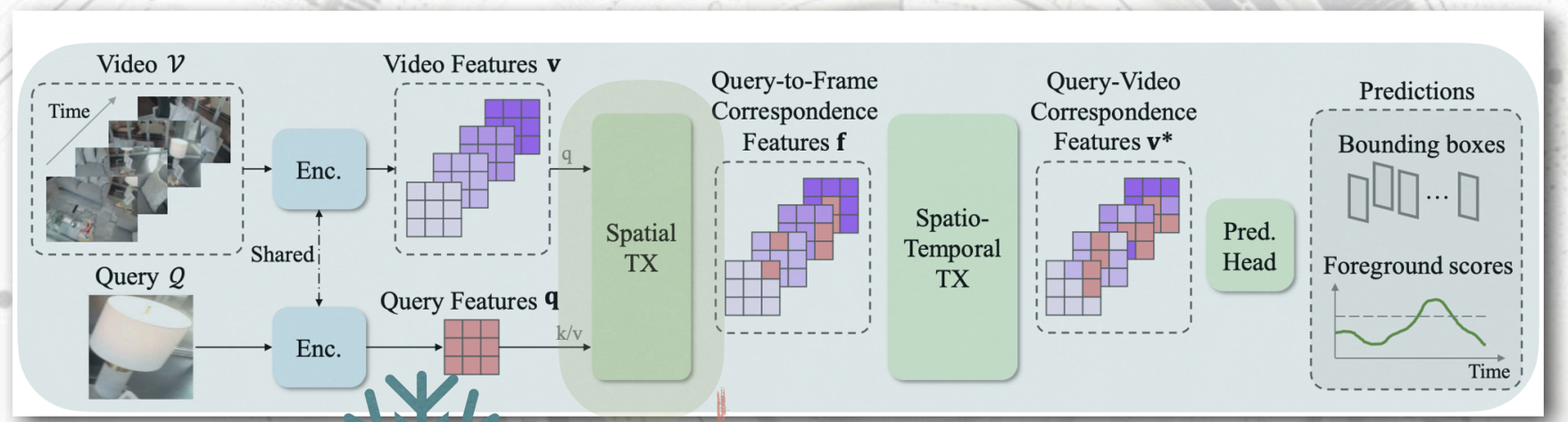
1. Improve the temporal resolution by ensemble pred. Head
2. Post-Processing by CLIP to capture overall features
3. Post-Processing based on continuity

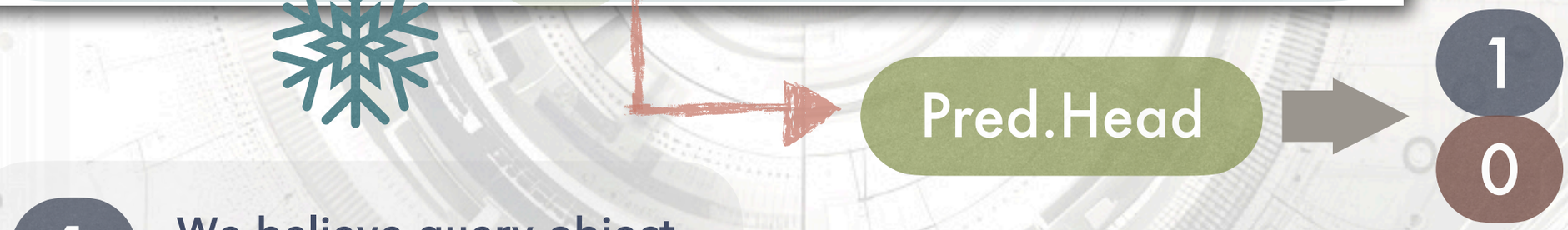
1. Improve the temporal resolution by ensemble pred. Head

OUR APPROACH

stAP	tAP
0.231	0.304

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.



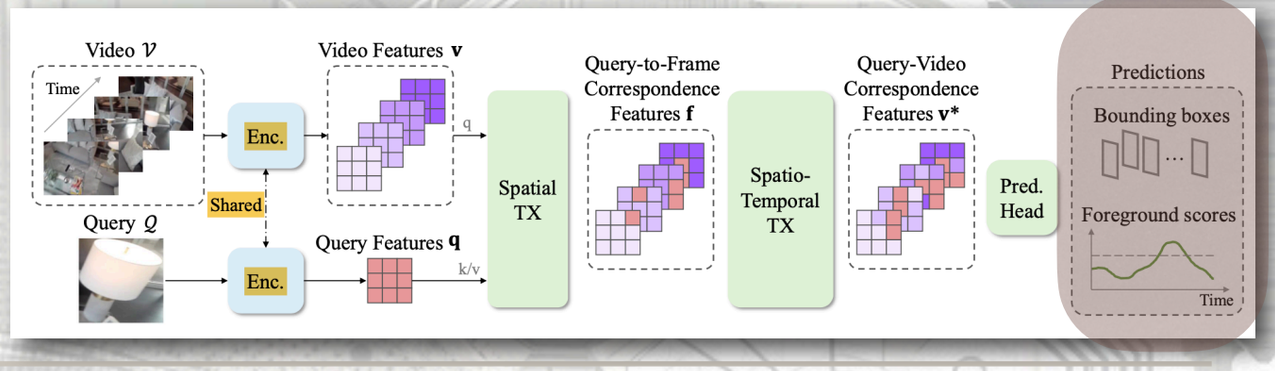


1 We believe query object is in the frame

0 Otherwise

We believe that binary classification is simpler and more efficient (about 2%) compared to training an additional model for ensemble purposes.

OUR APPROACH



2. Post-Processing by CLIP to capture overall features

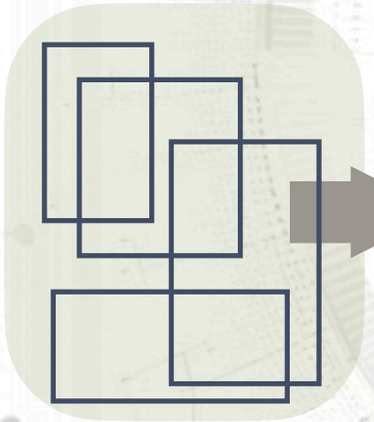
text prompt:
a image of {'title'}

CLIP MODEL

CLIP embedding

SIMILARITY

Find peak



CLIP MODEL

Candidates Bounding Boxes,
Cropped from original video clips.

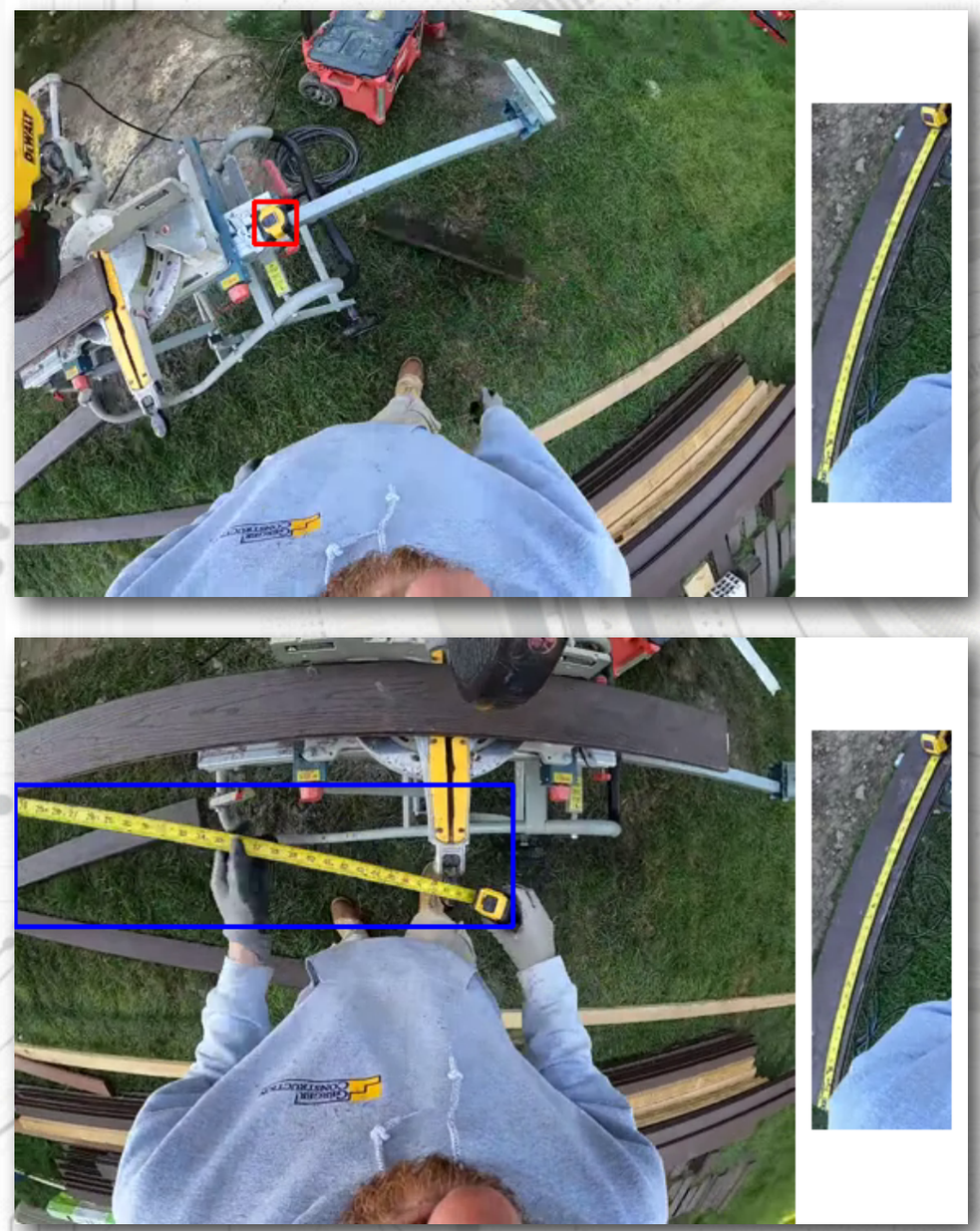
May help

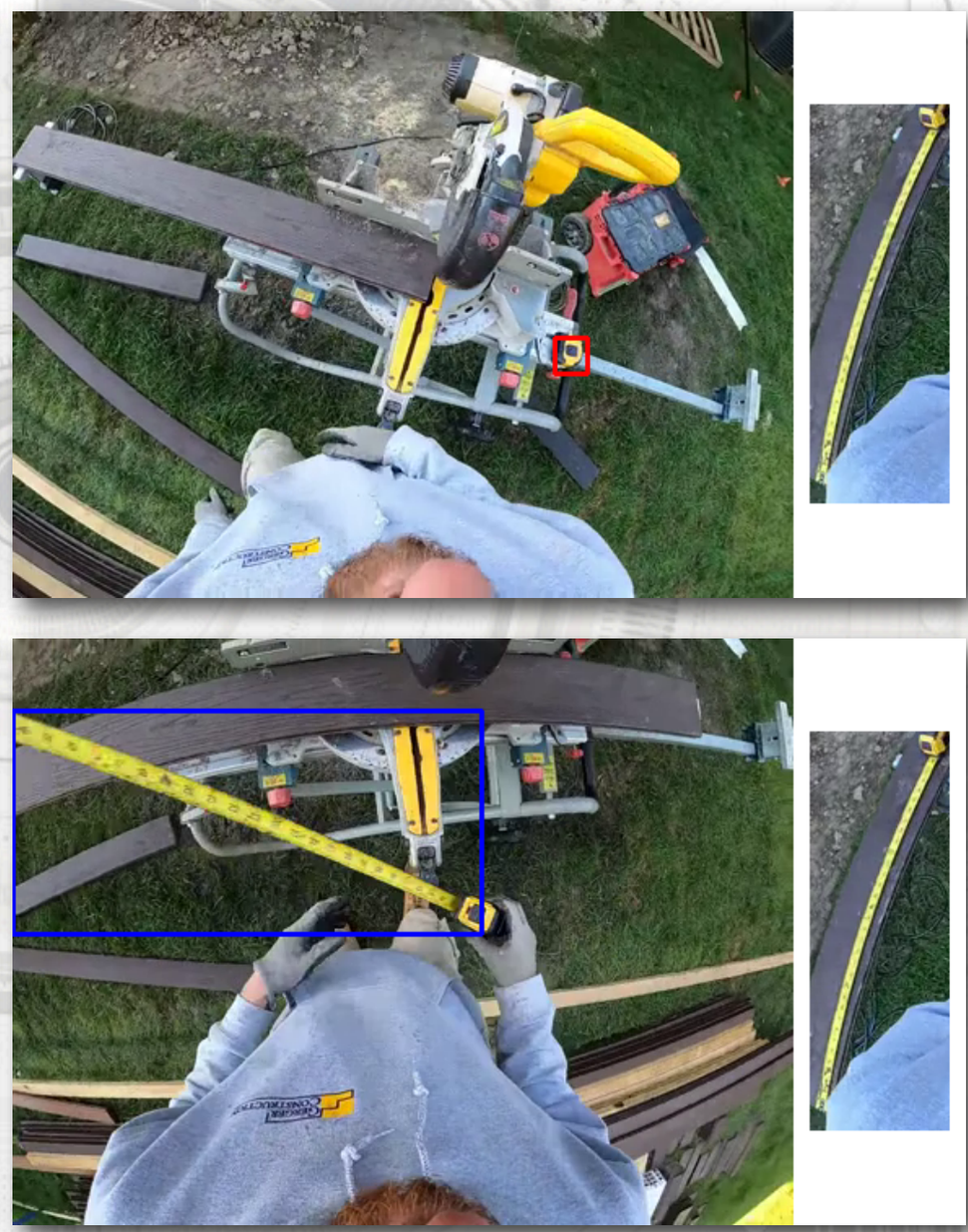
1. Noisy label, misleading query image
2. Similar color and texture
3. Small part of object shown
4. Not cleanly cut

Noisy label, misleading query image

STUDY ON FAILURE CASES

3d7b86a0-63e7-406a-ab4e-9e31a99898ca



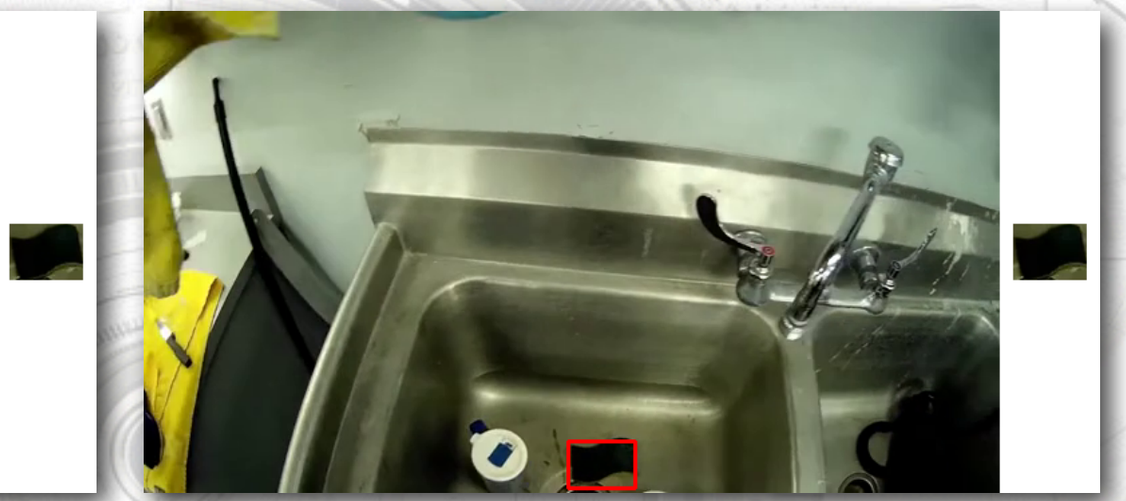


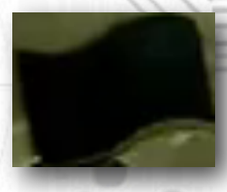
Query image

Similar color and texture

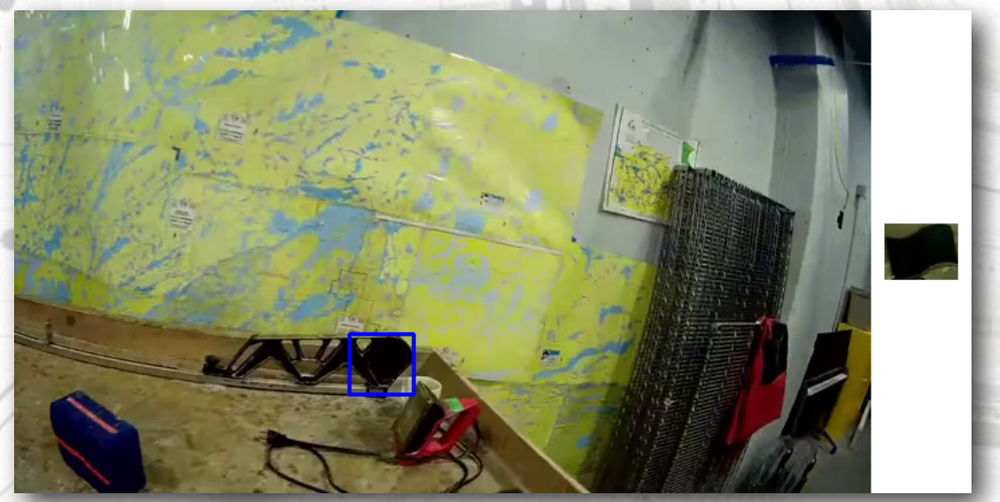
STUDY ON FAILURE CASES

2f790afe-53ca-4fbe-8182-89c0efdccb06



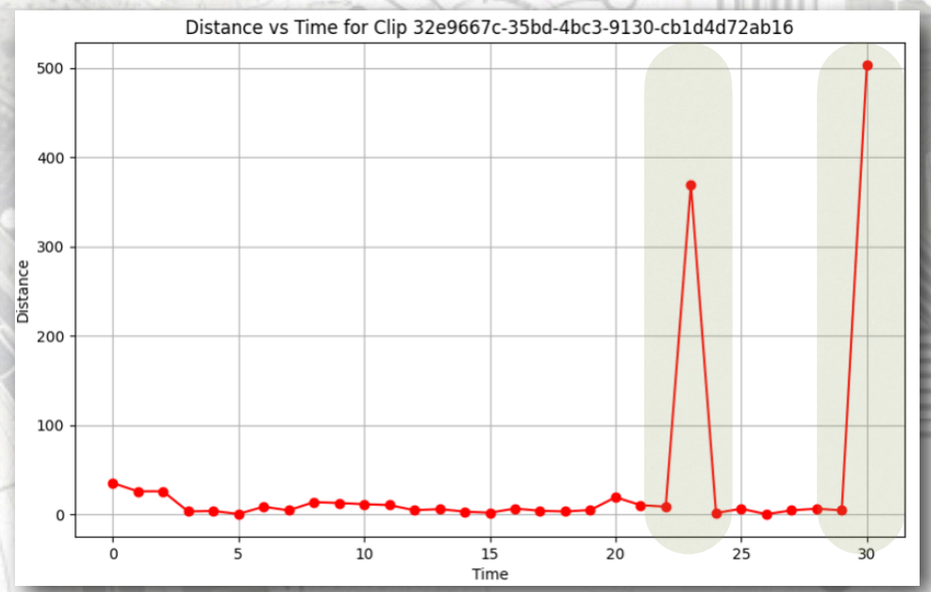


Query image



OUR APPROACH

3. Post-Processing based on continuity



Distance between the center points across frames

May help

1. Noisy label, misleading query image
2. Similar color and texture
3. Small part of object shown
4. Not cleanly cut

Define Discontinuous:

delta_x > image_width * 0.75 or delta_y > image_height * 0.75

RESULT & CONCLUSION

Setting	stAP(val)	tAP(val)	stAP(test)
Baseline	0.2306	0.3039	0.2897
Ours (Improve t-resolution)	0.2742	0.3595	0.2937
Ours (Post-process)	0.2284	0.3049	0.2877

1. Our method achieved an 18% improvement in tAP on the validation set, and stAP saw a comparable increase. However, there was only a slight improvement on the test set.

2. The post-processing did not meet our expectations, as CLIP did not prove to be as effective as anticipated.

Q & A