

Supplementary data

Ego gesture intention grounding facilitated by video prediction



Yixuan Ma¹, Rui Wang² and Jianwu Fang^{1,*}

¹ Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China

² Shaanxi Dechuang Digital Industrial Intelligence Technology Co., Ltd., Xi'an, China

* Correspondence author; E-mail: fangjianwu@mail.xjtu.edu.cn.

These supplementary figures provide additional qualitative sequences for motion, long occlusion, and out-of-view re-entry. The colors, labels, and frame-level annotations follow the main article.

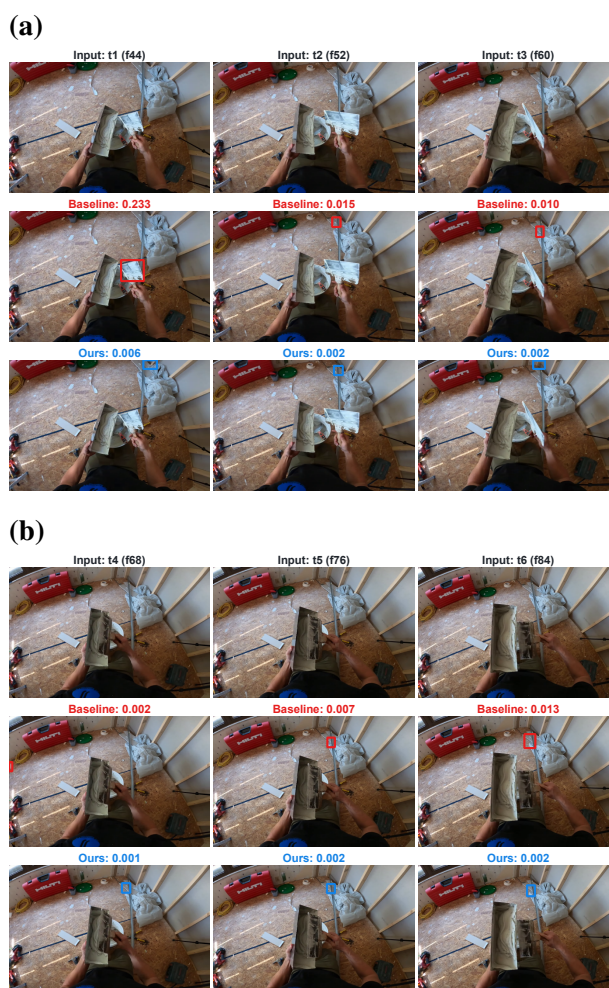


Figure S1. Motion and interaction sequence. Rows show input frames, baseline predictions (red), and proposed predictions (blue); values are confidence scores. (a) First three frames; (b) Last three frames.



Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

(a)



(b)

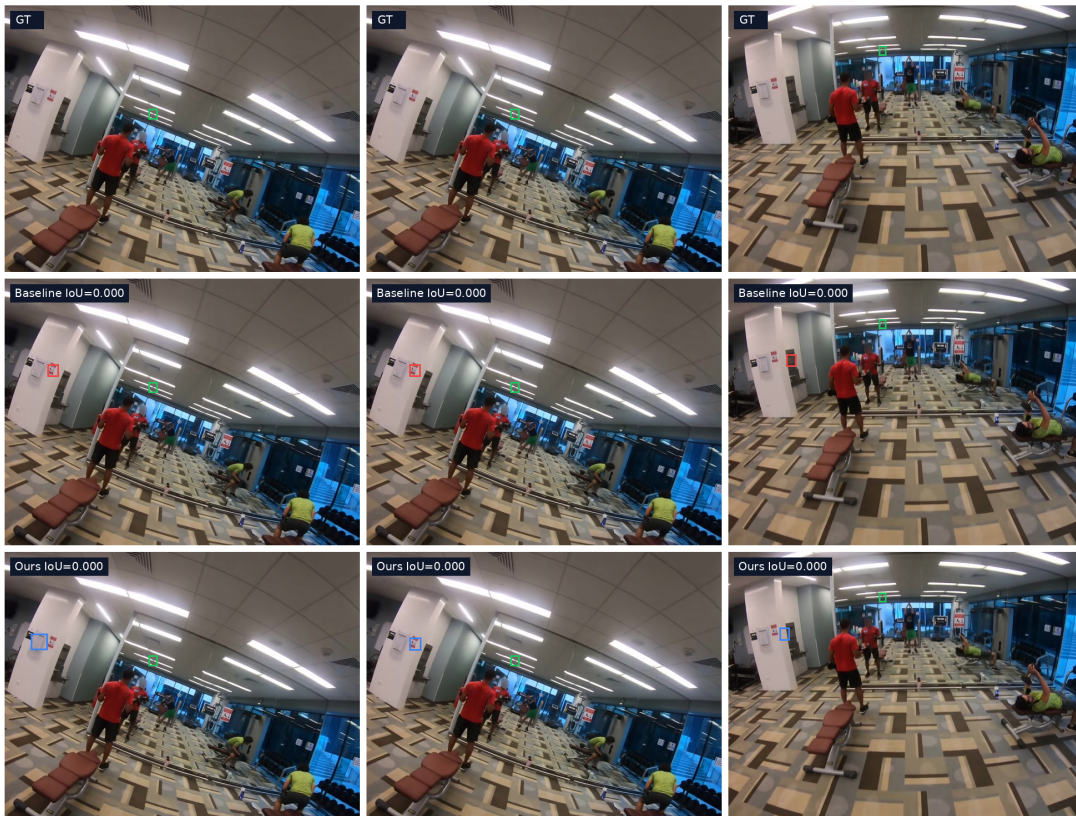
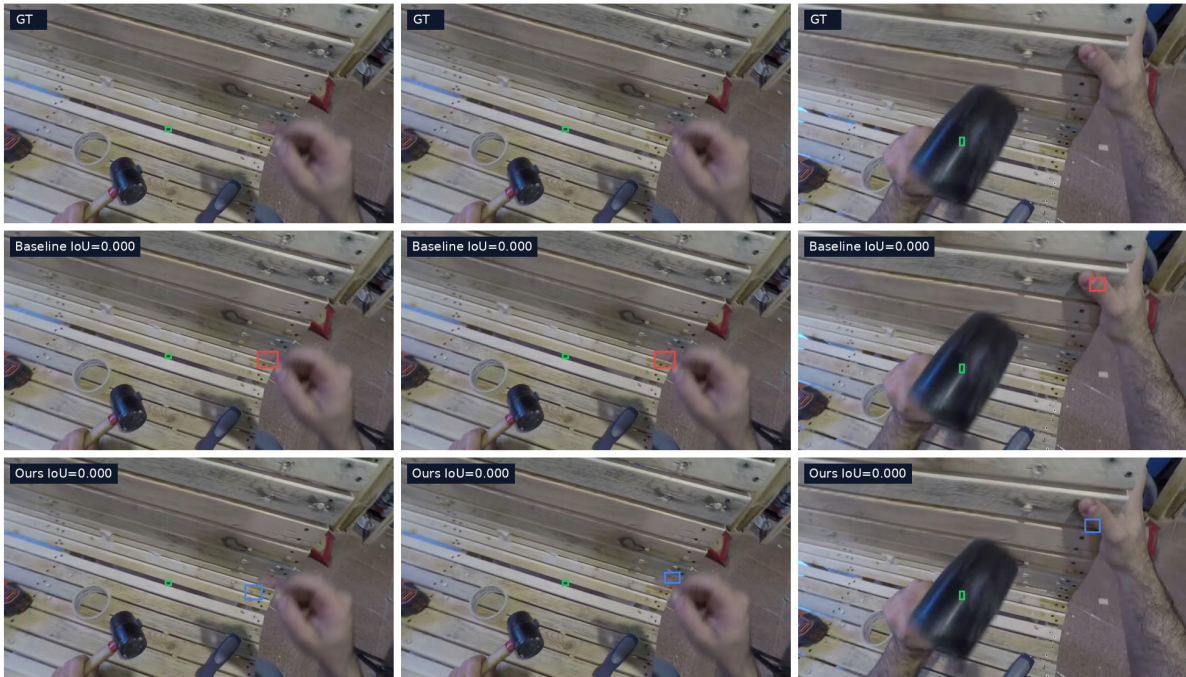


Figure S2. Difficult small-target sequence associated with long occlusion. Rows show ground truth (green), baseline predictions (red), and proposed predictions (blue). Both methods have displayed IoU values of 0.000. (a) First three displayed frames; (b) Last three displayed frames.

(a)



(b)

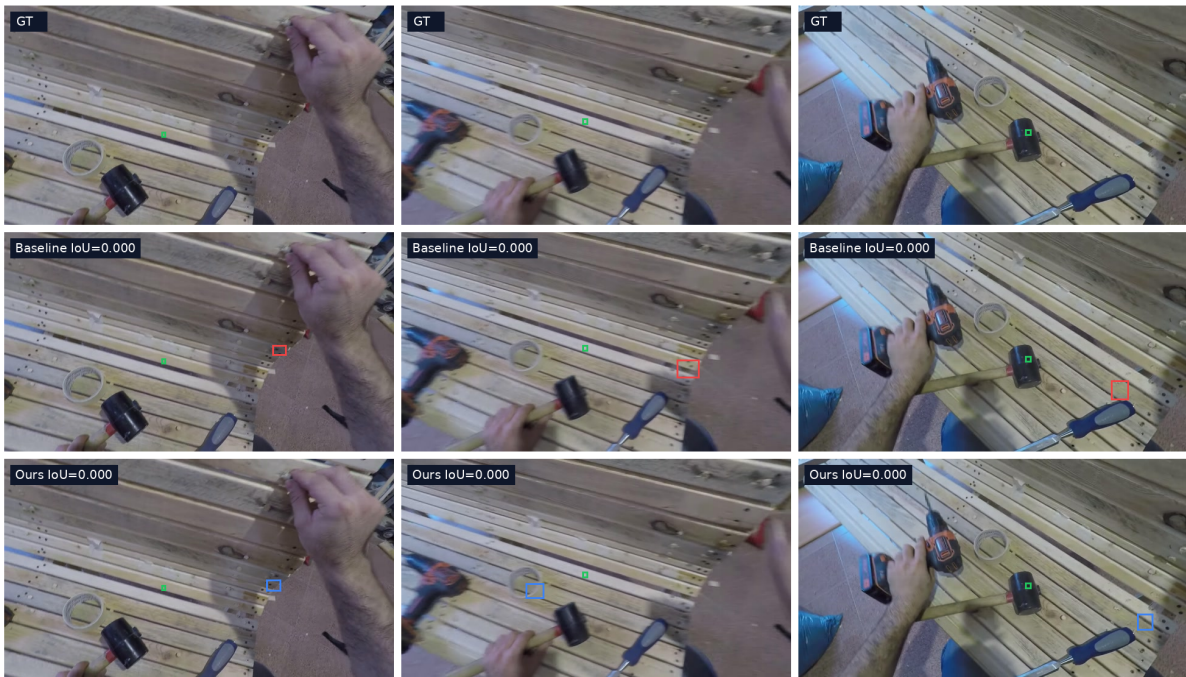


Figure S3. Failure example associated with out-of-view disappearance and re-entry. Rows show ground truth (green), baseline predictions (red), and proposed predictions (blue). Both methods have displayed IoU values of 0.000; smoothing alone does not resolve target recovery. **(a)** First three displayed frames; **(b)** Last three displayed frames.