

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.

Highlights:

- Preference tuning and box fusion improve final egocentric target localization.
- Final-output mean IoU rises from 0.177 to 0.232 in the reported RefEgo comparison.
- Confidence-weighted fusion and EMA combine hypotheses into one output sequence.

Abstract: Egocentric interaction grounding requires a system to identify one target when several alternatives are plausible. We propose a framework that combines video-conditioned interaction hypotheses, localization-guided direct preference optimization (DPO), confidence-weighted box fusion, and exponential moving average (EMA) smoothing. We construct preference pairs from training-set localization quality to adapt the hypothesis generator, then combine grounded hypotheses using detector confidence at inference. On RefEgo, our method improves final-output mean intersection over union (mIoU) from 0.177 to 0.232 compared with an adapted reasoning-and-grounding baseline. The detector-proposal oracle reference increases from 0.422 to 0.434, while the reference-to-output gap decreases from 0.245 to 0.202. The fractions of target-present evaluation records meeting intersection over union thresholds of 0.5 and 0.7 are 0.257 and 0.193, respectively. Candidate-generation analysis, a separate pre-/post-DPO comparison, and component ablations complement the main comparison. These results support combining localization-guided preference learning with confidence-weighted spatial aggregation for more accurate final target localization in egocentric video.

Keywords: ego gesture intention grounding; video prediction; egocentric video; target selection; preference learning; box fusion; temporal consistency

1. Introduction

Egocentric interaction grounding asks a system to localize the object relevant to a wearer's intended action. This capability matters for first-person assistants and embodied agents, which must connect an action-centered query to a visible interaction target [1–3]. We study a query-conditioned setting: given first-person video and a textual query, the system returns a sequence of target boxes. We use RefEgo's referring expressions and object annotations to evaluate query-conditioned target localization [4].



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.

Intent-oriented queries can be harder to ground than explicit object descriptions [5–7]. Several tools or hand-adjacent objects may support the same goal, so useful evidence includes the query, visible interaction context, and motion across frames. Reasoning-based grounding and recent surveys of egocentric scene understanding motivate using interaction context to resolve ambiguity [8,9].

Egocentric videos add frequent out-of-frame disappearance, visually similar objects, small target scales, and abrupt positional changes [4,10]. A plausible box in one frame can therefore correspond to the wrong object or drift away from the intended target. The distinction between having an overlapping detector proposal and producing an accurate final output motivates our analysis.

Figure 1 illustrates both an improvement case and a persistent limitation. In the visible-target example, the proposed output is more accurately localized. After a long disappearance, however, both methods remain unreliable when the target returns at very small scale. These observations motivate examining final localization quality alongside the quality of the detector’s available proposals.

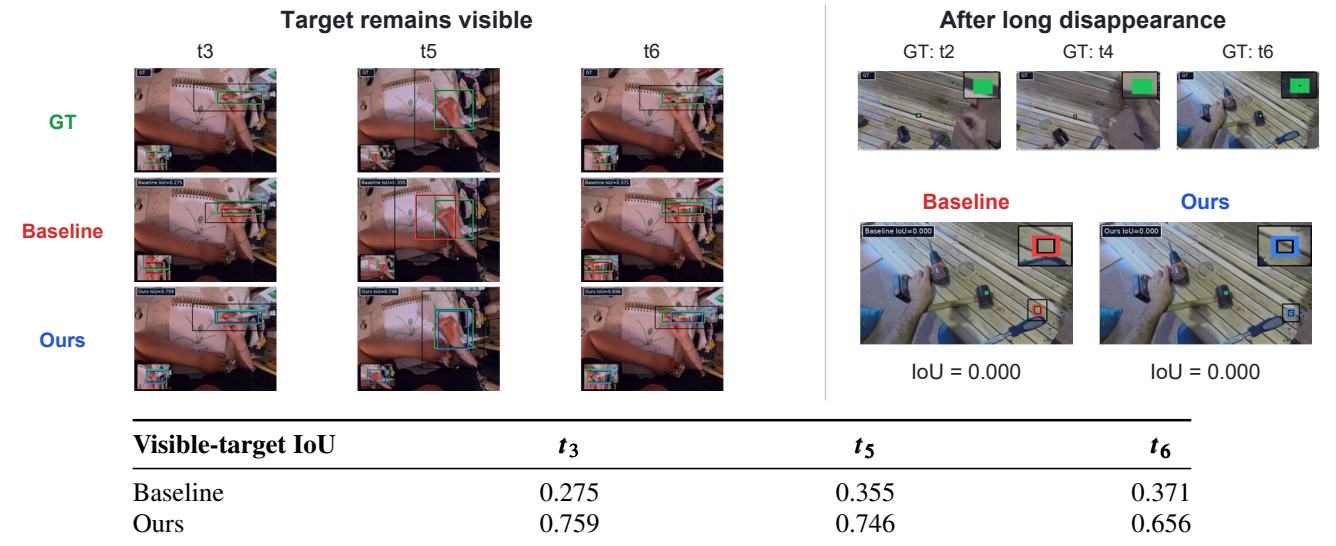


Figure 1. Selected qualitative examples. On the left, the proposed predictions have higher intersection over union (IoU) for a visible target; the corresponding values are listed in the table. On the right, both methods have IoU values of 0.000 for a small target after a long disappearance.

Our pipeline uses video-conditioned generation of prospective interaction descriptions, localization-derived preference tuning of the generator, confidence-weighted fusion of grounded boxes, and exponential moving average (EMA) smoothing. We use prediction to formulate prospective interaction descriptions from visual context for subsequent grounding. Direct preference optimization (DPO) [11] changes the distribution of generated hypotheses, while fusion and smoothing determine how their spatial outputs are combined over time.

Relative to our RefEgo adaptation of a Reason-to-Ground (RoG)-style reasoning baseline [8], the reported mIoU(single) increases from 0.177 to 0.232, and the detector-proposal reference gap decreases from 0.245 to 0.202. IoU@0.5 increases from 0.191 to 0.257, and IoU@0.7 increases from 0.137 to 0.193. These results are consistent with improved use of available localization evidence. We analyze hypothesis quality and spatial aggregation jointly through the main comparison and component configurations.

We make three contributions:

- (1) We distinguish final-output quality from detector-proposal oracle quality and report their gap and ratio as descriptive diagnostics on RefEgo.
- (2) We combine interaction-hypothesis generation, IoU-derived DPO supervision, box fusion weighted by confidence, and EMA smoothing for egocentric grounding.
- (3) We report complementary analyses with explicit evaluation record counts, candidate diagnostics, component ablations, and qualitative analysis, alongside published RefEgo reference results.

2. Related work

Open-vocabulary image grounding links linguistic descriptions to spatial regions. Modulated Detection for End-to-End Multi-Modal Understanding (MDETR) [5], Grounded Language-Image Pre-training (GLIP) [6], and Grounding DINO [7] use multimodal pretraining to improve language-conditioned detection. One For All (OFA) [12] formulates grounding within a unified sequence-to-sequence framework. These approaches provide strong spatial localization components; our framework complements them by generating and combining interaction hypotheses before returning a single box.

Video grounding introduces temporal information into the localization process. TubeDETR [13], VideoGrounding-DINO [14], and Context-Guided Spatio-Temporal Video Grounding (CG-STVG) [15] predict spatio-temporal object regions. Co-grounding networks and DCNet exploit cross-frame semantic or correspondence information for video referring expression comprehension [16,17]. Unified Video-Language Temporal Grounding (UniVTG) and Scalable and Accurate Video Grounding (SnAG) address video-language temporal grounding, where the output is a relevant time interval [18,19]. We distinguish these output definitions from the frame-wise target boxes evaluated here. Our temporal stage applies EMA to the spatial outputs of grounded interaction hypotheses.

Egocentric benchmarks emphasize camera motion, hand-object interactions, and targets that leave the field of view. Ego4D [1], Ego-Exo4D [2], and EgoExoLearn [3] support first-person and cross-view understanding, while RefEgo [4] provides referring expressions and frame-wise object annotations, including target-absence intervals. Fine-grained spatiotemporal grounding and OSGNet further study object localization in egocentric video [10,20]. Two published 2026 surveys provide complementary perspectives: Li *et al.* organize egocentric understanding around subjects, objects, environments, and hybrid tasks [9], while He *et al.* examine cross-view collaboration between egocentric and exocentric vision [21]. These studies position interaction-target grounding within object understanding and embodied perception.

Reasoning and preference learning offer additional supervision for ambiguous decisions. RoG introduces intention reasoning before visual grounding [8], and Intention-Driven Visual Perception via Reinforced Reasoning (VisRL) combines intention-driven perception with reinforced reasoning [22]. Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) enable efficient adaptation of large pretrained models [23,24], while DPO learns from pairwise preferences without training a separate reward model [11]. We use localization quality to construct preferences for the interaction-hypothesis generator. At inference, detector confidences determine the fusion weights, and EMA aggregates the fused coordinates over time.

This organization distinguishes four related directions: image grounding, video grounding, egocentric scene understanding, and reasoning or preference optimization. Our contribution lies in combining localization-guided hypothesis adaptation with confidence-weighted spatial aggregation and in evaluating final outputs alongside an all-proposal oracle reference. We report the original RefEgo results for OFA, MDETR variants, CG-SL-Att, and DCNet as published reference values [4], and evaluate our component configurations under a common internal protocol.

3. Method

3.1. Problem formulation

Given first-person video $V = \{I_t\}_{t=1}^T$ and a textual query x , we predict one target box $\bar{b}_t$ per evaluated frame. The resulting sequence is $\bar{B} = \{\bar{b}_t\}_{t=1}^T$. Let $\mathcal{C}_t$ denote all detector proposals retained by the evaluator for the record at frame t . We report both final-output overlap and the best overlap available in $\mathcal{C}_t$, with the descriptive gap

$$G = \text{mIoU}(\text{upper}) - \text{mIoU}(\text{single}). \quad (1)$$

We use $\text{mIoU}(\text{upper})$ to denote a ground-truth-selected detector-proposal oracle reference. It is not a mathematical upper bound on a box created by fusion or smoothing, which can differ from every individual proposal. We interpret G as a descriptive measure of the difference between proposal-oracle quality and final-output quality.

3.2. Inference pipeline

The implemented pipeline contains a hypothesis generator, a grounding model, confidence-weighted box fusion, and EMA smoothing. DPO is applied to the generator during training; the deployed decision stage uses grounding confidences. Figure 2 separates the training-only preference loop from inference.

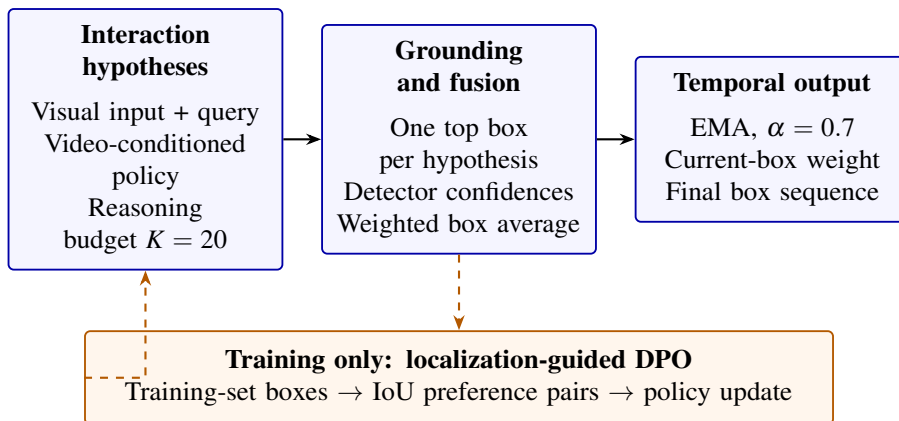


Figure 2. Conceptual workflow of the implemented method. Solid arrows show inference: interaction hypotheses are grounded, confidence-weighted boxes are fused, and EMA produces the output sequence. The dashed branch constructs IoU-derived preference pairs on the training set and updates the generator through DPO. Ground-truth boxes do not enter inference.

Ground-truth boxes and ground-truth IoU are used to construct DPO preference pairs during training and to compute evaluation metrics; they are not inputs to inference. In particular, the oracle reference uses ground truth only after predictions have been produced.

- (1) Decode and align the input video using the fixed preprocessing protocol, and supply its textual query.
- (2) Generate video- or image-conditioned interaction hypotheses with the trained policy, using a reasoning candidate budget of $K = 20$.
- (3) Ground each hypothesis in the current frame and retain its highest-confidence box and confidence score.
- (4) Form a confidence-weighted average of these boxes using model outputs only.
- (5) Apply EMA with $\alpha = 0.7$ and return the final box sequence.

3.3. Interaction hypotheses and box fusion

The generator produces prospective interaction descriptions r_k conditioned on visual input and the query. These descriptions express plausible interpretations of the intended interaction and are grounded frame by frame. The reasoning budget $K = 20$ counts language hypotheses; it is distinct from the number of detector proposals used to compute mIoU(upper).

For hypothesis k at frame t , the detector outputs boxes $b_{t,k,j}$ with confidences $s_{t,k,j}$. We retain the highest-confidence box for each grounded hypothesis:

$$j_{t,k}^* = \arg \max_j s_{t,k,j}, \quad b_{t,k}^* = b_{t,k,j_{t,k}^*}, \quad s_{t,k}^* = s_{t,k,j_{t,k}^*}. \quad (2)$$

For the K_t usable hypotheses, the fused box is

$$\omega_{t,k} = \frac{\exp(s_{t,k}^* - s_t^{\max})}{\sum_{\ell=1}^{K_t} \exp(s_{t,\ell}^* - s_t^{\max})}, \quad b_t^{\text{fuse}} = \sum_{k=1}^{K_t} \omega_{t,k} b_{t,k}^*, \quad (3)$$

where $s_t^{\max} = \max_k s_{t,k}^*$. Fusion is performed on box coordinates in a common image coordinate system. It combines candidate evidence into one output; it does not necessarily select one of the original boxes. The deployed weights are obtained directly from detector confidences.

3.4. Localization-guided preference tuning

For a training record with ground-truth box g_t , we ground the generated hypotheses and compare their localization quality. We denote a better-localized hypothesis by r^+ and a worse-localized hypothesis by r^- . We retain their preference pair when the localization-quality difference satisfies the pair-construction criterion. The preference dataset is $\mathcal{D}_p$.

Using a frozen reference policy π_0 , we optimize the generator π_θ with DPO [11]:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(z,r^+,r^-) \sim \mathcal{D}_p} \left[\log \sigma \left(\beta \left[\log \frac{\pi_\theta(r^+ | z)}{\pi_0(r^+ | z)} - \log \frac{\pi_\theta(r^- | z)}{\pi_0(r^- | z)} \right] \right) \right], \quad (4)$$

where z denotes the input context and $\beta = 0.1$ scales the preference objective. DPO adapts the hypothesis-generation policy using relative localization quality as supervision. The resulting hypotheses are then processed by the grounding and fusion stages.

The final reported training uses the training split only: validation and test records do not participate in parameter updates. The pre-/post-DPO comparisons use the same evaluation records within each reported split. They belong to a separate experiment block from the principal comparison, as detailed in Section 4.5.

3.5. Temporal smoothing

The implementation applies an exponential moving average to fused boxes:

$$\bar{b}_1 = b_1^{\text{fuse}}, \quad \bar{b}_t = \alpha b_t^{\text{fuse}} + (1 - \alpha)\bar{b}_{t-1}, \quad \alpha = 0.7. \quad (5)$$

Thus, 0.7 weights the current fused box and 0.3 weights the preceding smoothed state. EMA provides a lightweight temporal aggregation of fused coordinates. Its fixed weighting balances the current localization estimate with the preceding output state; long-term identity recovery remains the responsibility of the visual grounding stage.

3.6. Training procedure

Training comprises supervised adaptation of the hypothesis generator and grounding branch, followed by localization-guided DPO of the generator. The autoregressive objective trains hypothesis generation; the grounding branch uses box localization losses. Preference tuning follows Equation (4). Box fusion and EMA are inference operations. Model and optimization settings are provided in Appendix A.

4. Experiments

4.1. Dataset and metrics

We evaluate on RefEgo [4], using 2-FPS decoding, annotation alignment, and target-validity filtering. The full validation evaluation contains 76,940 query-frame records, including 58,644 target-present records. A record is one textual query paired with one annotated frame. The validation set contains 38,470 distinct annotated frames, including 29,322 target-present frames; each is evaluated with two referring expressions. The pre-/post-DPO test evaluation uses a test-derived subset of 5120 records, including 3696 target-present records. The candidate diagnostic uses 410 validation records, including 300 target-present records.

Let $\mathcal{E}$ be the target-present evaluation records and $N = |\mathcal{E}|$. For record i , g_i is its ground-truth box, $\bar{b}_i$ is the final single output, and $\mathcal{C}_i$ contains all detector proposals evaluated for that record.

$$\text{mIoU}(\text{single}) = \frac{1}{N} \sum_{i \in \mathcal{E}} \text{IoU}(\bar{b}_i, g_i), \quad (6)$$

$$\text{mIoU}(\text{upper}) = \frac{1}{N} \sum_{i \in \mathcal{E}} \max_{b \in \mathcal{C}_i} \text{IoU}(b, g_i), \quad (7)$$

$$\text{IoU}@ \tau = \frac{1}{N} \sum_{i \in \mathcal{E}} \mathbf{1}\{\text{IoU}(\bar{b}_i, g_i) \geq \tau\}, \quad \tau \in \{0.5, 0.7\}. \quad (8)$$

$\text{IoU}@ \tau$ denotes the final-output threshold hit rate. The oracle reference covers all detector proposals, rather than only the $K = 20$ language hypotheses. Ground truth is used to compute this reference after inference. We use this oracle reference to describe the proposal quality available to each configuration.

We also report G from Equation (1) and the descriptive ratio

$$R = \frac{\text{mIoU}(\text{single})}{\text{mIoU}(\text{upper})}. \quad (9)$$

Because fusion and smoothing operate on box coordinates, their output can differ from each individual detector proposal. We therefore interpret G and R as descriptive reference-to-output diagnostics.

Records with $\text{is_obj_in}=0$ remain in total-record accounting and are excluded from conditional box-overlap averages. All reported overlap metrics are therefore conditional on target presence. For a trajectory, center jitter is defined as $\mathcal{J}^c = |\mathcal{P}|^{-1} \sum_{(t-1,t) \in \mathcal{P}} \|c(\bar{b}_t) - c(\bar{b}_{t-1})\|_2$, in pixels, where $c(\cdot)$ is the box center and $\mathcal{P}$ contains consecutive target-present frame pairs for the same query and target within one video clip. This definition separates trajectory coordinate variation from localization accuracy. Our quantitative tables report the localization metrics defined above; temporal aggregation is specified by Equation (5).

Table 1 identifies the split and denominator of each result block. Internally compared configurations use common preprocessing and overlap definitions within each block. The main comparison, candidate diagnostic, and DPO analysis are distinct experiment blocks; their absolute scores should not be combined as one sequential ablation.

Table 1. Split assignment and evaluation-record accounting for the reported analyses.

Result block	Split	Accounting and protocol
Main comparison, component ablation, and reviewed failures	Validation	76,940 total records; 58,644 target-present records
Candidate diagnostic	Validation subset	410 total records; 300 target-present records
DPO pre-/post check	Validation	76,940 total records; 58,644 target-present records
DPO pre-/post check	Test subset	5120 total records; 3696 target-present records

4.2. Implementation details

The reasoning branch uses Qwen2.5-VL-7B-Instruct with QLoRA adaptation, and the grounding branch uses MDETR with an EfficientNet-B3 visual backbone and a RoBERTa text encoder. The reasoning candidate budget is $K = 20$, the EMA coefficient is $\alpha = 0.7$, and the random seed is 42. Appendix A summarizes the model and optimization settings. The main comparison and component ablations each use one run with seed 42. This single-run setting does not quantify run-to-run variance.

The baseline is a RefEgo adaptation of the RoG-style reasoning-and-grounding approach [8], using our backbones and evaluation pipeline. We use common backbones, preprocessing, and metric definitions for the internal comparison and component configurations. Each configuration generates its own hypotheses and detector proposals.

4.3. Main comparison

We organize the comparative analysis into two complementary parts. Table 2 evaluates our method against the adapted reasoning-and-grounding baseline under the common internal validation protocol. Table 3 presents published target-present RefEgo results for image and video referring expression comprehension

methods. These reference values retain the source training and evaluation protocols. Together, the tables place the internal improvement in the broader RefEgo literature while identifying the provenance and evaluation setting of each result.

Table 2. Internal comparison on 58,644 target-present validation records. Boldface indicates the better value for each metric in this two-method comparison; arrows show the preferred direction.

Method	mIoU(single) $\uparrow$	mIoU(upper) $\uparrow$	Gap $\downarrow$	Ratio $\uparrow$	IoU@0.5 $\uparrow$	IoU@0.7 $\uparrow$
RoG-style baseline	0.177	0.422	0.245	0.419	0.191	0.137
Ours	0.232	0.434	0.202	0.535	0.257	0.193

Table 3. Published RefEgo target-present referring expression comprehension (REC) results, transcribed from the RefEgo paper and its video-REC baseline table and converted from percentages to fractions. mAP@50 retains the source notation. Values follow the original training and evaluation protocols and are provided as published reference results.

Method	Val mIoU	Val mAP@50	Test mIoU	Test mAP@50
OFA (RefEgo fine-tuned) [4,12]	0.532	0.584	0.519	0.568
MDETR+BH (all) [4,5]	0.452	0.526	0.454	0.527
MDETR+BH (all) + ByteTrack [4,5,25]	0.454	0.529	0.457	0.530
CG-SL-Att [4,16]	0.332	0.380	0.329	0.380
DCNet [4,17]	0.342	0.370	0.332	0.365

The reported mIoU(single) gain is 0.055, the oracle reference gain is 0.012, and the gap reduction is 0.043 after rounding. Figure 3 shows this relationship. The ratio R increases from 0.419 to 0.535. The substantially larger gain in final-output mIoU supports the combined hypothesis-generation and aggregation framework. We interpret the oracle gap alongside the component analysis, since hypothesis adaptation, fusion, and smoothing all contribute to the final prediction.

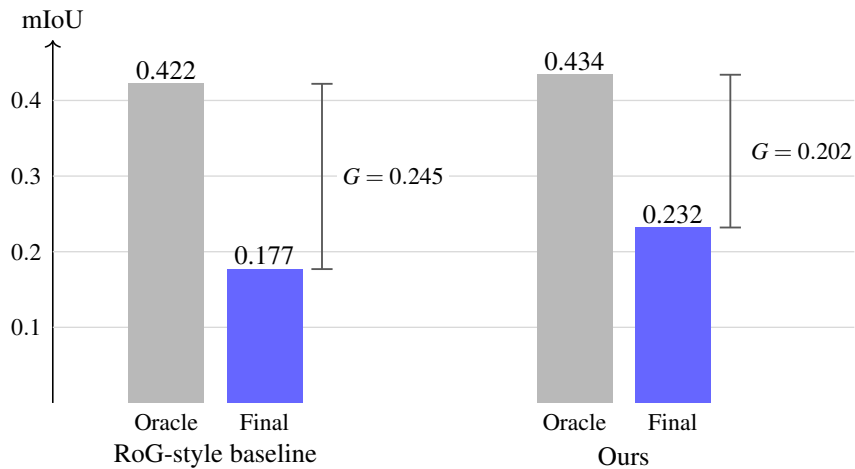


Figure 3. Final-output and detector-proposal oracle mIoU for the internal validation comparison. The displayed reference gaps are 0.245 and 0.202. The oracle is computed using ground truth for evaluation and is not an inference output.

4.4. Candidate generation analysis

Table 4 reports all three available metrics for each configuration on the same 300 target-present validation records. This subset is distinct from the full validation evaluation and is used as a descriptive diagnostic.

Table 4. Complete candidate diagnostic on a shared validation subset: 410 total query-frame records and 300 target-present records. All configurations report both final-output and detector-proposal oracle mIoU.

Configuration	mIoU(single) $\uparrow$	mIoU(upper) $\uparrow$	Gap $\downarrow$
Direct grounding	0.023	0.209	0.186
Single-path reasoning	0.030	0.224	0.194
RoG-style multi-path reasoning	0.019	0.240	0.221

Single-path reasoning improves final-output mIoU from 0.023 to 0.030 on this subset. Multi-path reasoning raises the oracle reference to 0.240, while its final-output mIoU is 0.019 and the gap is 0.221. Thus, a higher proposal oracle does not necessarily coincide with a better returned box. This diagnostic motivates evaluating candidate availability and final-output accuracy together.

4.5. DPO analysis

Table 5 reports the absolute pre- and post-DPO mIoU values as well as their differences. This complementary experiment compares the generator before and after preference tuning on identical evaluation records within each split. The final DPO training used training records only; the reported validation and test-subset records were reserved for evaluation.

Table 5. Separate pre-/post-DPO comparison on target-present records. Absolute values and gains are rounded to three decimals; differences are computed before rounding.

Evaluation set	Records	Pre-DPO mIoU	Post-DPO mIoU	Δ mIoU $\uparrow$
Validation	58,644	0.157	0.171	+0.014
Test subset	3,696	0.075	0.081	+0.006

Preference tuning improves final-output mIoU on both evaluation sets, from 0.157 to 0.171 on validation and from 0.075 to 0.081 on the test subset. These complementary results support using localization quality to supervise the interaction-hypothesis generator.

4.6. Ablation studies

Table 6 compares the contributions of preference tuning and temporal aggregation on the common validation evaluation. We report final-output mIoU, the oracle-reference gap and ratio, and the IoU@0.5 hit rate to assess localization across all configurations.

Adding DPO raises mIoU(single) from 0.177 to 0.199 and IoU@0.5 from 0.191 to 0.218; the gap decreases from 0.245 to 0.218 and the ratio increases from 0.419 to 0.477. EMA alone gives mIoU(single) of 0.185, IoU@0.5 of 0.200, and a ratio of 0.436. The full configuration reaches mIoU(single) of 0.232 and IoU@0.5 of 0.257. These observations support the combined configuration for final localization

accuracy. DPO and EMA produce complementary improvements, with their combination giving the strongest final-output result among these configurations.

Table 6. Component configurations on 58,644 target-present validation records. The full model gives the highest final-output mIoU and IoU@0.5 among the evaluated configurations. Each configuration uses its own generated hypotheses and proposals.

Setting	mIoU(single) $\uparrow$	Gap $\downarrow$	Ratio $\uparrow$	IoU@0.5 $\uparrow$
Multi-candidate generation only	0.177	0.245	0.419	0.191
+ DPO tuning	0.199	0.218	0.477	0.218
+ EMA smoothing	0.185	0.239	0.436	0.200
Full model	0.232	0.202	0.535	0.257

4.7. Qualitative results

Figure 1 illustrates a visible-target improvement, and Figure 4 examines clutter and semantic ambiguity. The complete motion and interaction sequence is provided in Supplementary Figure S1, whereas Supplementary Figures S2 and S3 present the long-occlusion and out-of-view re-entry sequences, respectively. Each supplementary sequence is displayed in two chronological three-frame panels, with predicted boxes and frame-level annotations for direct inspection.

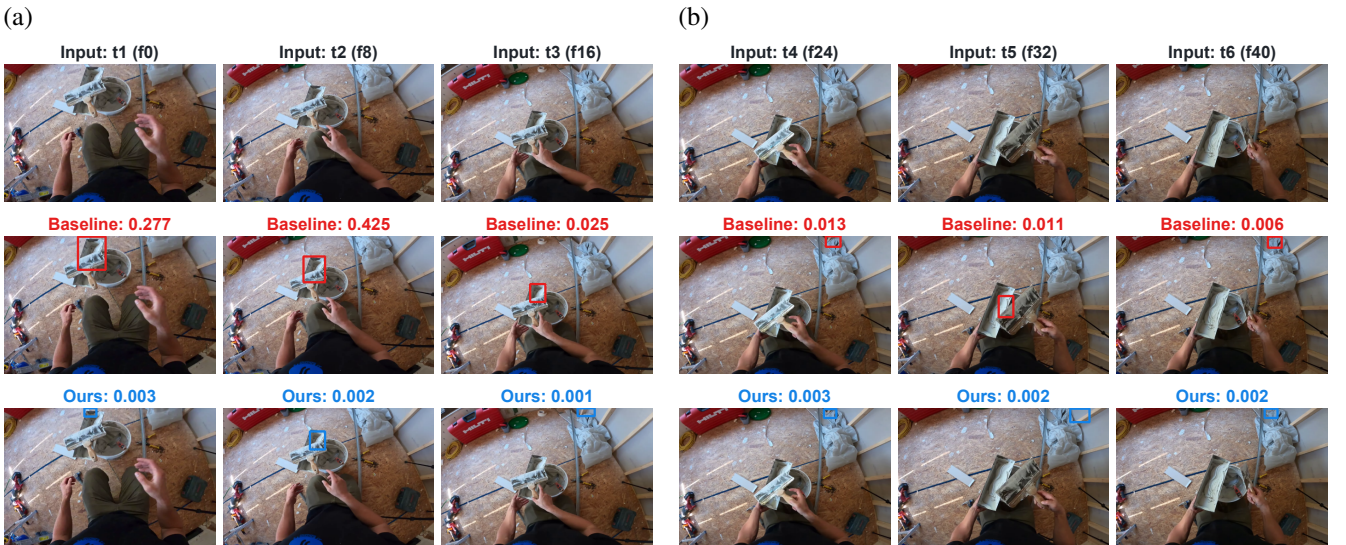


Figure 4. Clutter and semantic ambiguity. Rows show input frames, baseline predictions (red), and proposed predictions (blue); values are confidence scores. **(a)** First three frames; **(b)** Last three frames.

The clutter example and the motion sequence in Supplementary Figure S1 highlight how ambiguous grounded hypotheses affect the fused output. In Supplementary Figures S2 and S3, both methods have displayed IoUs of 0.000. These recovery challenges motivate longer-term visual memory and explicit re-detection as extensions to EMA-based temporal aggregation.

4.8. Failure analysis

Figure 5 summarizes a manually reviewed set of 156 validation failures. The reported proportions are rounded descriptive shares within this reviewed set: 45% long-term occlusion, 30% out-of-view

re-entry, and 25% semantic distractors. These shares characterize the reviewed cases and complement the quantitative localization results.

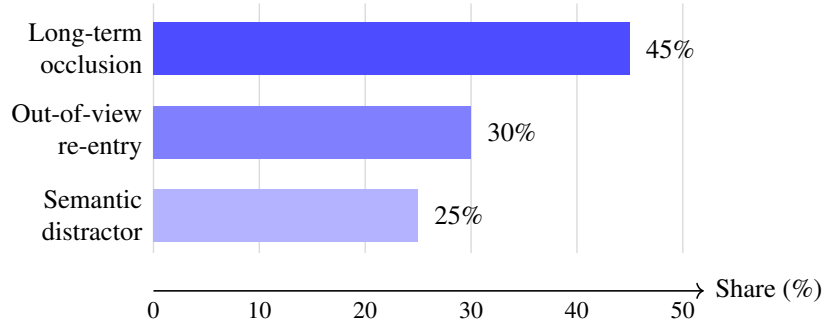


Figure 5. Rounded descriptive proportions in 156 manually reviewed validation failures. The reviewed cases are grouped into long-term occlusion, out-of-view re-entry, and semantic distractors.

Long occlusion and out-of-view intervals weaken the visual evidence needed for target recovery. Similar neighboring objects create ambiguity even when the target remains visible. These conditions motivate future work on memory, re-detection, and interaction-aware grounding, as extensions to the present framework.

5. Conclusion

We presented an egocentric interaction-target grounding framework that combines video-conditioned hypotheses, localization-guided DPO, confidence-weighted box fusion, and EMA. In the internal RefEgo validation comparison, final-output mIoU improves from 0.177 to 0.232, while the detector-proposal oracle reference increases from 0.422 to 0.434 and the reference-to-output gap decreases from 0.245 to 0.202. Candidate diagnostics and component comparisons complement this result by examining how hypothesis quality and aggregation affect the returned box. The separate pre-/post-DPO analysis also shows positive localization gains on validation and the reported test subset.

The evaluation focuses on target-present localization in RefEgo. The main comparison uses one run with seed 42, so run-to-run variability remains a limitation. The test-derived subset provides a complementary preference-tuning check, and the published reference table supplies results under the original source protocols. These settings define the scope of the reported evidence. Future work will examine broader benchmarks and evaluation coverage, multiple random seeds, and longer-term memory for small-target recovery, occlusion, and out-of-view re-entry.

Data availability statement

RefEgo annotations and implementation resources are available from the official project repository, <https://github.com/shuheikurita/RefEgo>. The underlying videos originate from Ego4D and are subject to its access and license conditions, described at <https://ego4d-data.org/>. The Supplementary Materials accompanying this article contain Supplementary Figures S1–S3, which present the motion and interaction,

long-occlusion, and out-of-view re-entry sequences, respectively, together with aggregate result tables and evaluation-accounting details that support the reported analyses.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, we used OpenAI Codex to assist with language editing and figure layout. We reviewed and revised the resulting content and take full responsibility for the accuracy and integrity of the manuscript and its figures.

Acknowledgments

Yixuan Ma thanks Prof. Jianwu Fang for guidance on problem formulation, experimental design, and manuscript revision. The authors thank laboratory members for helpful discussions and assistance with experiment verification, result checking, and figure preparation, and acknowledge the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, for its academic support. This work was supported by the National Natural Science Foundation of China (Grant No. 62273057), the Outstanding Youth Foundation of Shaanxi Province (Grant No. 2025JC-JCQN-092), and the Key Research and Development Program of Shaanxi Province (Grant No. 2025PT-ZCK-66).

Authors' contribution

Conceptualization, Y.M. and J.F.; methodology, Y.M. and J.F.; software, Y.M.; validation, Y.M. and R.W.; formal analysis, Y.M. and R.W.; investigation, Y.M.; resources, J.F.; data curation, Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, Y.M., R.W., and J.F.; visualization, Y.M.; supervision, J.F.; project administration, J.F.; funding acquisition, J.F. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Rui Wang is an employee of Shaanxi Dechuang Digital Industrial Intelligence Technology Co., Ltd. The author declares that this affiliation does not influence the design, analysis, interpretation, or presentation of the results reported in this manuscript. The authors declare no other competing interests related to this work.

Ethical statement

This study uses existing RefEgo annotations and Ego4D video data. It does not report new participant recruitment, interventions, or collection of human-subject data. The original datasets' access, consent, and license conditions apply to their reuse; no identifying participant information is added by this study.

Appendix A. Implementation details

Table A1 summarizes the model, optimization, and evaluation settings. Sections 4.3–4.6 specify the evaluation records used by each analysis.

All internal comparisons use the preprocessing and metric definitions in Section 4.1. We apply preference training to the training split and use the validation and test-subset records exclusively for evaluation.

Table A1. Model, optimization, and evaluation settings used in the experiments.

Category	Setting
Reasoning model	Qwen2.5-VL-7B-Instruct adapted with QLoRA
Quantization and LoRA	4-bit NF4 quantization, LoRA rank 64, alpha 128, dropout 0.05
Candidate generation training	3 epochs, per-device batch size 1, gradient accumulation 8, learning rate 10^{-4} , maximum sequence length 1024, image longer side 448
Box-head training	8 epochs, batch size 2, learning rate 2×10^{-4} , normalized xywh prediction with L1 and generalized IoU losses
Preference tuning	1 DPO epoch, batch size 1, gradient accumulation 8, learning rate 5×10^{-6} , query-context limit 256, maximum sequence length 512, preference scale $\beta = 0.1$
Grounding model	MDETR with EfficientNet-B3 visual backbone and RoBERTa text encoder
Grounding training	RefCOCO-pretrained initialization, 5 epochs on RefEgo, batch size 4, learning rates $10^{-4}/10^{-5}/5 \times 10^{-5}$ for main/backbone/text modules, weight decay 10^{-4} , lr drop at epoch 3, EMA decay 0.9998
Hypothesis budget and smoothing	$K = 20$ reasoning hypotheses; confidence-weighted box fusion; EMA with current-box weight $\alpha = 0.7$
RefEgo task options	Binary head enabled; no-target setting enabled
Randomness control	Fixed random seed 42
Evaluation definitions	Common preprocessing and overlap definitions within each internal comparison; all detector proposals for the oracle; published values retain their source protocol
Validation accounting	76,940 validation query-frame records, 58,644 target-present records
Test accounting	5120 test-subset query-frame records, 3696 target-present records for the pre-/post-DPO comparison

References

- [1] Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, *et al.* Ego4D: around the world in 3000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, June 21–24, 2022, pp. 18995–19012.
- [2] Grauman K, Westbury A, Torresani L, Kitani K, Malik J, *et al.* Ego-Exo4D: understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 19–21, 2024, pp. 19383–19400.
- [3] Huang Y, Chen G, Xu J, Zhang M, Yang L, *et al.* EgoExoLearn: a dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 19–21, 2024, pp. 22072–22086.
- [4] Kurita S, Katsura N, Onami E. RefEgo: referring expression comprehension dataset from first-person perception of Ego4D. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, October 2–6, 2023, pp. 15214–15224.

- [5] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, *et al.* MDETR—modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Online, October 12–15, 2021, pp. 1780–1790.
- [6] Li LH, Zhang P, Zhang H, Yang J, Li C, *et al.* Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, June 19–24, 2022, pp. 10965–10975.
- [7] Liu S, Zeng Z, Ren T, Li F, Zhang H, *et al.* Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, Milan, Italy, September 29–October 4, 2024, pp. 38–55.
- [8] Sun P, Xiao J, Tse THE, Li Y, Akula A, *et al.* Visual intention grounding for egocentric assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, USA, October 19–23, 2025, pp. 2512–2522.
- [9] Li X, Qiu H, Wang L, Zhang H, Qi C, *et al.* Challenges and trends in egocentric vision: a survey. *Mach. Intell. Res.* 2026, 23(1):1–33.
- [10] Liang S, Zhong Y, Hu Z, Tao Y, Wang L. Fine-grained spatiotemporal grounding on egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, USA, October 19–23, 2025, pp. 9385–9395.
- [11] Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, *et al.* Direct preference optimization: your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, New Orleans, USA, December 10–16, 2023, pp. 53728–53741.
- [12] Wang P, Yang A, Men R, Lin J, Bai S, *et al.* OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, July 17–23, 2022, pp. 23318–23340.
- [13] Yang A, Miech A, Sivic J, Laptev I, Schmid C. TubeDETR: spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, June 19–24, 2022, pp. 16442–16453.
- [14] Wasim ST, Naseer M, Khan S, Yang MH, Khan FS. VideoGrounding-DINO: towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 17–21, 2024, pp. 18909–18918.
- [15] Gu X, Fan H, Huang Y, Luo T, Zhang L. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 17–21, 2024, pp. 18330–18339.
- [16] Song S, Lin X, Liu J, Guo Z, Chang SF. Co-grounding networks with semantic attention for referring expression comprehension in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Online, June 23–25, 2021, pp. 1346–1355.
- [17] Cao M, Jiang J, Chen L, Zou Y. Correspondence matters for video referring expression comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa, Portugal, October 10–14, 2022, pp. 4967–4977.

- [18] Lin KQ, Zhang P, Chen J, Pramanick S, Gao D, *et al.* UniVTG: towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, October 2–6, 2023, pp. 2794–2804.
- [19] Mu F, Mo S, Li Y. SnAG: scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 17–21, 2024, pp. 18930–18940.
- [20] Feng Y, Zhang H, Liu M, Guan W, Nie L. Object-shot enhanced grounding network for egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, June 11–15, 2025, pp. 24190–24200.
- [21] He Y, Huang Y, Chen G, Lu L, Pei B, *et al.* Bridging perspectives: a survey on cross-view collaborative intelligence with egocentric-exocentric vision. *Int. J. Comput. Vision* 2026, 134(2):62.
- [22] Chen Z, Luo X, Li D. VisRL: intention-driven visual perception via reinforced reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, USA, October 19–23, 2025, pp. 2545–2555.
- [23] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, *et al.* LoRA: low-rank adaptation of large language models. In *Proceedings of International Conference on Learning Representations (ICLR 2022)*, Online, April 25–29, 2022.
- [24] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. In *Proceedings of Advances in Neural Information Processing Systems 36*, New Orleans, USA, December 10–16, 2023, pp. 10088–10115.
- [25] Zhang Y, Sun P, Jiang Y, Yu D, Weng F, *et al.* ByteTrack: multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision*, Tel Aviv, Israel, October 23–27, 2022, pp. 1–21.