Rebuttal to reviewers

Anonymous submission

Paper ID

1. Rebuttal to Reviewer W2qo

1.1. Major Weakness

1.1.1 Answer for question 1

We have acknowledged this issue and have thoroughly proofread and corrected such errors as soon as possible in the current version of the manuscript. We sincerely appreciate your attention and feedback.

1.1.2 Answer for question 2

The generalization capability of the model to unseen objects is grounded in the following two observations: 1) According to the definition and setup of the VQL task, the visual cropping templates during the inference phase differ from those in the training set. Therefore, the understanding of unseen targets can be inferred from the task's feedback, which aligns with the ultimate goal of the few-shot pipeline. 2) We conducted extensive training on the Ego4D dataset, where the VQ2D task data was divided into a training set (13.6k queries, 262 hours of video), a validation set (4.5k queries, 87 hours of video), and a test set (4.4k queries, 84 hours of video). The average target video duration is approximately 140 seconds, while the average target trajectory length is only about 3 seconds. This characteristic of long videos with short target durations necessitates the model's ability to precisely locate targets in time and space within large-scale video data. Consequently, the performance in the VQ2D and VQ3D challenges indirectly demonstrates the model's accurate discrimination and robust localization capabilities for unseen targets. The generalization ability of the hierarchical Transformer benefits from training on the EgoTracks dataset, although the supervised approach does not support overly strong claims, leading us to adopt a more moderate and appropriate phrasing. Detailed descriptions and analyses will be provided in subsequent responses to similar inquiries. We kindly request the reviewers to take note of this clarification.

1.1.3 Answer for question 3

we employ 1×1 and 3×3 convolutional kernels with 64 channels and are followed by ReLU activations to process features of search regions. The 1×1 and 3×3 convolutions employed in DPM are standard operations in deep learning network modules. The 1×1 convolution is primarily used for cross-channel information integration, where it linearly combines feature maps across different channels to reduce channel dimensionality, thereby decreasing computational load and memory usage, achieving dimensionality reduction. It is followed by the ReLU activation function, which introduces additional non-linearity to enhance the network's expressive power. Combining 1×1 and 3×3 convolutions allows for the retention of spatial information while improving the network's non-linear modeling capability, a frequently adopted operational practice in our work.

1.1.4 Answer for question 4

The initial segmentation generated by SAM indeed impacts the performance of DPM. Specifically, we employ a combined center-point and bounding box prompting strategy for segmentation. This leverages the assumption that the target object is centrally located in the template, using its center point along with a bounding box derived from a 2/3-scaled visual crop. We also investigated segmentation pathways: the "everything" prompt, selecting the largest resultant mask; and "positive/negative point" prompts, using the center and edge points, respectively.

1.1.5 Answer for question 5

We experimented with K=[3,5,10] and found that model performance is highly sensitive to the choice of K. The selection of K significantly impacts the results, and the reasons are as follows: 1) drastic motion and blur: egocentric videos often contain significant camera motion, causing targets to become blurred, distorted, or partially occluded. This leads to a decline in similarity rankings. Smaller K values may miss the target, while larger K values introduce noise. 2) Frequent occlusion: Targets are more likely to be occluded by hands, body parts, or environmental objects in first-person views, resulting in unstable tracking. 3) Target scale variation: Changes in camera perspective and distance cause significant scale variations in targets, affecting feature extraction and similarity judgment. 4) Complex and dynamic backgrounds: Complex backgrounds interfere with target recognition and tracking. 5) Interference from similar objects: Similar objects in the scene may mislead the model. On the Ego4D-VQ2D validation set, K=5 performed optimally. The selection of K should be tailored to specific scenarios to determine the best hyperparameters.

1.1.6 Answer for question 6

In VQL-3D, the discrepancies between Matterport3D scan data and real-world environments (especially egocentric videos), referred to as the domain gap, significantly impact angle prediction. These discrepancies primarily manifest in: lighting conditions (controlled lighting in scan data versus dynamic lighting in real-world videos), scene appearance (structured scenes in scan data versus complex and diverse real-world scenes), scan quality (potential missing or low-quality regions in scan data), and motion blur (common in egocentric videos). Crucially, the target locations in scan data and real-world videos are not captured and annotated synchronously, leading to depth estimation errors, feature matching errors, and ultimately, impacting angle calculation accuracy.

1.1.7 Answer for question 7

Hierarchical Transformer are crucial for accurate and temporally consistent segmentation masks. They refine initial coarse masks hierarchically, ensuring fine-grained segmentation that closely aligns with target boundaries. This hierarchical approach leverages temporal context from previous frames, mitigating flickering and jitter. Experiments show that under partial occlusion, this temporal consistency, combined with contextual information, allows inference of occluded regions, leading to more complete masks. Furthermore, the integration of global and local context within the hierarchy enables robust handling of target deformations and improves boundary clarity in low-resolution or blurry images. Ultimately, this hierarchical approach synergistically enhances overall retrieval accuracy.

1.1.8 Answer for question 8

Due to time constraints, we did not explore variations in loss function design.

1.1.9 Answer for question 9

The Success metric is calculated as the percentage of queries within trackable frames where the L2 distance (between the predicted location and the ground truth 3D bounding box) is below a threshold. Our method's lower Success rate compared to EgoLoc-v1 may stem from several factors. First, the error distributions differ. Our method typically exhibits smaller errors, but with occasional large outliers, whereas EgoLoc-v1's errors are more uniformly distributed. Although EgoLoc-v1 has a larger average L2 error, more of its predictions fall within the threshold, resulting in higher Success. Second, a larger Success threshold could also contribute to EgoLoc-v1 achieving higher Success despite a larger L2 error. We adhered to the official evaluation settings for a fair comparison. Furthermore, differences in dataset characteristics and algorithmic strategies also influence the results. Our method may excel on easier samples but exhibit larger L2 errors on challenging samples. EgoLoc-v1 might benefit from its detector settings, demonstrating greater robustness on challenging samples. Finally, EgoLoc-v1's localization strategy may be more conservative, predicting a larger area, which, even with a larger L2 distance, is more likely to encompass the ground truth location, thus increasing Success. A simplified example: with 10 samples and a 1-meter threshold, our method has a 0.5-meter L2 error on 9 samples and a 5-meter error on one, resulting in 90% Success and an average L2 error of 0.95 meters. EgoLoc-v1 has a 0.9-meter L2 error on all samples, achieving 100% Success and an average L2 error of 0.9 meters. This illustrates how even with a smaller average L2 error, a few large errors can negatively impact the Success metric.

1.2. Minor Weakness

1.2.1 Answer for question 1

As detailed in the supplementary material, we utilize a DINO-pre-trained ViT for feature extraction within the search region; please refer to the "Visual Backbone" section of the supplement for further details.

1.2.2 Answer for question 2

We agree on the importance of failure case analysis. We have included a qualitative comparison of visual failure cases on the Ego4D-VQ2D against state-of-the-art algorithms in the updated appendix.

2. Rebuttal to Reviewer 5aMe

Thank you for your review. However, I'm unsure whether you've thoroughly examined my paper and supplementary materials, as the supplement includes pseudo code detailing the core algorithms and key steps, including the DCF components you've inquired about.

2.1. Answer for question 1

We have further elaborated on the relationship between our method and its motivation. The two key insights are not arbitrary but stem from years of dedicated research and exploration by our team. How these insights guide the proposed framework is intrinsically linked to our key components and architectural design. A precise segmentation framework directs our overarching design, while the integration of global context and local information is achieved through a hierarchical transformer, optimized and designed for iterative refinement across levels to reconstruct the final output resolution. I am unclear about the source of your confusion. We have, however, further refined and improved our paper and respectfully request your reconsideration.

2.2. Answer for question 2

The superiority of our framework is evident in two ways: firstly, the unified pipeline for VQL-2D and VQL-3D can further inspire researchers to build upon our algorithm; secondly, our superior performance on both tasks is demonstrably clear. These are not unsubstantiated claims; please refer to the appendix and our updated detailed analysis for further clarification.

2.3. Answer for question 3

We appreciate you bringing the grammatical errors and phrasing issues to our attention. We have made every effort to refine and improve the manuscript.

3. Rebuttal to Reviewer cSGB

We have provided an anonymized link for your review of the updated manuscript and supplementary materials.

3.0.1 Answer for question 1

We have submitted detailed supplementary materials and further refined the logical presentation within the main text. The supplementary materials have also been updated and expanded. We apologize for any confusion caused by the citations and numerical representations. We have undertaken a systematic and focused correction within the limited time available. Please refer to our anonymized link.

3.1. Answer for question 2

We consistently employed two input resolutions: 224×224 and 448×448. In all cases, we observed performance gains with the higher resolution.

3.2. Answer for question 3

We specify the depth estimation model in both the main text and supplementary materials. We utilized a depth-anything-base model, pre-trained on KITTI and NYUv2.