Object Reprojection Error (ORE): Camera pose benchmarks from lightweight tracking annotations - Supplementary Materials

Xingyu Chen[1,*], **Weiyao Wang**[1,*], **Hao Tang**[1], **Matt Feiszli**[1]
FAIR, Meta[1], Equal Technical Contribution[*]
{xingyuchen, weiyaowang, haotang, mdf}@meta.com

A Supplementary List

We include the following contents in the supplementary materials,

1. **supplemental.pdf**: This PDF contains more experiments/ablations, visualizations as well as information about EgoStatic dataset.
2. **visualization_1.mp4**, **visualization_2.mp4** and **visualization_3.mp4**: visualization videos of EgoStatic benchmark where colored bounding boxes are groundtruth tracklet annotations, colored dots on the left are reprojected points tracked by estimated camera trajectory and green dots on the right are visualization of estimated camera trajectory (from COLMAP[3]).
3. **egostatic.zip**: This repo contains (a) code snippets for ORE evaluation on EgoStatic and (b) sample pose files from 7 baseline methods. Detailed instruction is included in **README.md**.

B Experiments on ScanNet (Continued)

ScanNet [2] is an RGB-D video dataset containing more than 1500 3D scans, each with 3D camera pose annotations and point-clouds. We pick high confidence instance mask from Mask-RCNN [6], and construct object tracklets by back-projecting masks to the point cloud. We benchmark methods on 20 test sequences using ATE_{trans}, ATE_{rot} and ORE, and summarized the per-scene breakdown in Table 1, 2 and 3.

C Extended results on ranking statistics

In Section 4.2, we used two ranking statistics to capture the relationship between ORE and standard metrics: Spearman's rank correlation coefficient (Spearman Coef.) [8] and Kendall's τ coefficient (Kendall Coef.) [1].

Spearman's rank correlation coefficient Given n raw scores X_i and Y_i, ranking function $R(\cdot)$,

$$\rho_{R(x),R(Y)} = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{1}$$

where $\text{cov}(R(X), R(Y))$ is the covariance, $\sigma_{R(X}$ and $\sigma_{R(Y)}$ are standard deviations.

We compute Spearman's ranking statistics on each scene in Table 4, then average across all scenes.

Kendall's τ rank correlation coefficient measures for ordinal association between two sets of data. Given $\{R(X_i)\}_{i=1}^n$ and $\{R(Y_i)\}_{i=1}^n$,

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

Supp. Table 1: **ORE** performance of 7 methods on ScanNet dataset

	DROID	COLMAP	ORB2	ORB3	Particle	TartanVO	Monodepth2
scene0707_00	0.000	0.000	0.002	0.001	0.057	0.149	0.276
scene0708_00	0.169	0.066	0.357	0.060	0.169	0.176	0.206
scene0709_00	0.000	0.035	0.034	0.017	0.129	0.155	0.230
scene0710_00	0.013	0.012	0.069	0.033	0.336	0.533	0.366
scene0711_00	0.004	0.004	0.232	-	0.203	0.374	0.460
scene0712_00	0.186	0.062	0.143	0.141	0.292	0.409	0.569
scene0713_00	0.000	0.000	0.211	0.402	0.264	0.362	0.411
scene0714_00	0.001	0.001	-	0.001	0.004	0.123	0.237
scene0715_00	0.000	0.000	0.130	0.000	0.024	0.076	0.321
scene0716_00	0.000	0.006	0.000	0.000	0.000	0.083	0.083
scene0717_00	0.000	0.000	0.002	-	0.000	0.070	0.103
scene0718_00	0.000	0.000	-	0.163	0.002	0.079	0.069
scene0719_00	0.083	0.082	0.085	0.078	0.113	0.090	0.224
scene0720_00	0.000	0.170	0.013	0.185	0.261	0.532	0.436
scene0721_00	0.077	0.204	0.128	0.186	0.114	0.369	0.425
scene0722_00	0.002	0.002	0.074	0.085	0.079	0.194	0.311
scene0723_00	0.040	0.040	0.077	0.381	0.098	0.542	0.641
scene0724_00	0.001	0.001	0.115	0.123	0.314	0.237	0.201
scene0725_00	0.007	0.006	0.161	-	0.130	0.209	0.112
scene0726_00	0.000	0.000	0.012	0.002	0.003	0.258	0.202
Average	0.029	0.035	0.102	0.109	0.130	0.251	0.294
STD	0.055	0.057	0.093	0.122	0.111	0.159	0.155

Supp. Table 2: **ATE**$_{trans}$ performance of 7 methods on ScanNet dataset

	DROID	COLMAP	ORB2	ORB3	Particle	TartanVO	Monodepth2
scene0707_00	0.052	0.111	0.123	0.111	0.184	0.589	0.851
scene0708_00	0.638	0.380	1.497	0.380	0.233	0.389	1.196
scene0709_00	0.060	0.455	0.110	0.455	0.237	0.385	0.737
scene0710_00	0.070	0.167	0.102	0.167	0.271	0.395	0.549
scene0711_00	0.049	0.393	0.632	-	0.708	0.888	0.928
scene0712_00	0.636	0.656	0.194	0.393	0.570	0.787	0.698
scene0713_00	0.199	0.156	0.612	0.656	0.580	0.538	0.630
scene0714_00	0.040	0.108	-	0.156	0.497	0.443	0.782
scene0715_00	0.057	0.669	0.500	0.108	0.243	0.122	0.483
scene0716_00	0.529	0.455	0.647	0.669	0.527	0.222	0.867
scene0717_00	0.072	0.038	0.243	-	0.253	0.374	0.529
scene0718_00	0.170	0.995	-	0.455	0.256	0.092	0.230
scene0719_00	0.026	2.262	0.067	0.038	0.132	0.184	0.628
scene0720_00	0.053	0.537	0.100	0.995	0.797	0.635	0.943
scene0721_00	0.098	0.878	1.380	2.262	1.399	0.813	2.200
scene0722_00	0.034	0.744	0.501	0.537	0.529	0.246	0.492
scene0723_00	0.055	0.058	0.121	0.878	0.544	0.561	0.830
scene0724_00	0.031	0.001	0.527	0.744	0.713	0.478	0.743
scene0725_00	0.039	0.006	0.857	-	0.799	0.613	0.908
scene0726_00	0.039	0.000	0.111	0.058	0.216	0.270	0.568
Average	0.147	0.331	0.462	0.533	0.484	0.451	0.790
STD	0.196	0.468	0.420	0.518	0.298	0.222	0.385

a pair of random variables (X, Y), where $X = \{x_1, x_2, ...x_n\}$ and $Y = \{y_1, y_2, ...y_n\}$. If either both $x_i > x_j$ and $y_i > y_j$ or both $x_i < x_j$ and $y_i < y_j$ holds, the pair (x_i, y_i) and (x_j, y_j) are said to

Supp. Table 3: **ATE**$_{rot}$ performance of 7 methods on ScanNet dataset

	DROID	COLMAP	ORB2	ORB3	Particle	TartanVO	Monodepth2
scene0707_00	0.035	0.066	0.094	0.097	0.176	2.403	1.251
scene0708_00	0.783	0.079	1.564	0.164	0.158	2.271	1.926
scene0709_00	0.029	1.134	0.086	0.178	0.220	2.265	1.586
scene0710_00	0.058	0.057	0.203	0.231	0.449	1.704	2.382
scene0711_00	0.033	0.204	1.068	-	0.976	1.962	2.251
scene0712_00	0.450	0.829	0.233	0.308	1.836	1.918	2.096
scene0713_00	0.211	0.077	1.573	2.527	0.581	2.299	2.646
scene0714_00	0.041	0.055	-	0.147	0.154	2.118	1.044
scene0715_00	0.056	0.100	1.638	0.105	0.295	2.105	0.814
scene0716_00	0.296	0.784	1.557	0.238	0.458	2.156	1.165
scene0717_00	0.033	0.071	0.116	-	0.266	2.018	1.087
scene0718_00	0.226	0.225	-	1.899	1.089	2.049	0.606
scene0719_00	0.051	0.058	0.118	0.063	0.104	1.998	1.937
scene0720_00	0.049	1.438	0.197	1.995	0.633	2.155	2.582
scene0721_00	0.065	2.824	1.966	2.747	0.842	2.038	1.982
scene0722_00	0.034	0.037	0.746	2.407	0.673	1.945	1.717
scene0723_00	0.051	0.056	0.146	2.551	0.328	2.069	2.746
scene0724_00	0.050	0.135	2.349	2.359	1.806	2.020	1.906
scene0725_00	0.086	0.591	2.352	-	1.266	2.343	2.605
scene0726_00	0.060	0.134	0.114	0.084	0.266	2.081	2.899
Average	0.135	0.448	0.896	1.065	0.629	2.096	1.861
STD	0.184	0.676	0.830	1.096	0.513	0.161	0.669

Supp. Table 4: **Spearman Coef. and Kendall Coef. for ScanNet**. Here we show Spearman and Kendall Coef. between (a) ORE and ATE$_{trans}$, (b) ORE and ATE$_{rot}$, as well as (c) ATE$_{trans}$ and ATE$_{rot}$ for each scene in ScanNet test set. The table is summarized as Table 3 of main paper.

	Spearman Coef.			Kendall Coef.		
ORE	vs. ATE$_{trans}$	vs. ATE$_{rot}$	ATE$_{trans}$ vs ATE$_{rot}$	vs. ATE$_{trans}$	vs. ATE$_{rot}$	ATE$_{trans}$ vs ATE$_{rot}$
scene0707_00	0.964	0.893	0.929	0.905	0.714	0.810
scene0708_00	0.857	0.750	0.750	0.714	0.524	0.619
scene0709_00	0.607	0.893	0.714	0.524	0.714	0.619
scene0710_00	0.893	0.929	0.964	0.714	0.810	0.905
scene0711_00	0.964	1.000	0.964	0.905	1.000	0.905
scene0712_00	0.643	0.750	0.821	0.429	0.619	0.619
scene0713_00	0.786	0.929	0.857	0.524	0.810	0.714
scene0714_00	0.857	0.857	0.893	0.714	0.714	0.810
scene0715_00	0.889	0.815	0.786	0.720	0.617	0.714
scene0716_00	0.039	0.670	-0.214	0.056	0.620	-0.238
scene0717_00	0.889	0.852	0.929	0.823	0.720	0.810
scene0718_00	0.721	0.919	0.643	0.586	0.781	0.619
scene0719_00	0.821	0.679	0.929	0.619	0.429	0.810
scene0720_00	0.679	0.857	0.750	0.524	0.714	0.619
scene0721_00	0.464	0.679	0.607	0.333	0.524	0.429
scene0722_00	0.464	0.821	0.643	0.429	0.619	0.429
scene0723_00	0.857	0.929	0.964	0.714	0.810	0.905
scene0724_00	0.536	0.321	0.714	0.333	0.143	0.619
scene0725_00	0.643	0.714	0.964	0.429	0.524	0.905
scene0726_00	0.739	0.739	1.000	0.586	0.586	1.000
Average (**STD**)	0.716 ($\pm$ 0.216)	0.800 ($\pm$ 0.145)	0.780 ($\pm$ 0.216)	0.579 ($\pm$ 0.205)	0.650 ($\pm$ 0.173)	0.681 ($\pm$ 0.261)

be concordant; otherwise they are said to be discordant. Assume for random variable (X, Y), there exists A concordant pairs and B discordant pairs, the Kendall τ coefficient is defined as $\tau = \frac{A-B}{\binom{n}{2}}$.

We consider (X, Y) as the performance of two different methods on the same scene, resulting in 420 such pairs across 20 scenes in ScanNet.

C.1 Relationship with RPE

We summarize the ranking statistics between ORE and RPE in Supp. Tab 5. We observe despite ORE has strong correlation with RPE, it is weaker than the correlation with ATE. This is expected,

Supp. Table 5: ORE's ranking correlation with RPE. Spearman Coef. (upper triangle) and Kendall Coef. (lower triangle) between ORE and standard metrics.

Kendall \ Spearman	ORE	RPE_{trans}	RPE_{rot}
ORE	-	0.425	0.750
RPE_{trans}	0.334	-	0.473
RPE_{rot}	0.615	0.386	-

Supp. Table 6: ORE acheives very strong correlation with ATE and RPE with inconsistent pairs removal.

	ATE	RPE
Kendall	0.895	0.571

since both ORE and ATE measure the quality of the entire trajectory, whereas RPE measures a small local window. In addition, similar to ATE, we observe ORE to have stronger correlation with rotation compared to translation.

C.2 Inconsistent pairs removal in Kendall's τ

As discussed in the section 4.2 in the main paper, standard metrics ATE_{trans} and ATE_{rot} may often disagree with each other. This indicates for two methods, one may perform better on translation while the other performs better on rotation. This limits ORE's ranking statistic values to have an upper bound: since ORE factors in both translation and rotation, it can agree with either ATE_{trans} or ATE_{rot} when ATE_{trans} and ATE_{rot} mismatch. We capture this problem by removing such pairs in Kendall's τ. In total, we remove 16% of such pairs.

As summarized in Supp. Tab. 6, ORE and ATE has extremely high Kendall ranking coefficients. This implies only 5% of the pairs may disagree between ORE and ATE. ORE also acheives strong correlation with RPE, but is weaker than the global ATE measurement.

D Visualization

In this section, we provided more visualization of the ORE metrics and its associated EgoStatic benchmark.

Visualization for ORE metrics For each of the 7 baseline methods, we plot the estimated trajectory along with the ground truth trajectory, as well as its associated ORE, ATE_{trans} and ATE_{rot} metric. From Figure 1, we can see both qualitatively and quantitatively that ORE correlated highly with the quality of estimated camera trajectory.

Visualization for EgoStatic Benchmark In Figure 2, we visualized 24 sampled screenshot from 12 arbitrary scenes in EgoStatic benchmark (each with 2 images). The scenarios span across various categories including cooking, crafting, yardwork, gardening, lab, board game etc, which demonstrate the diverse scenes and complex actions EgoStatic is able to cover.

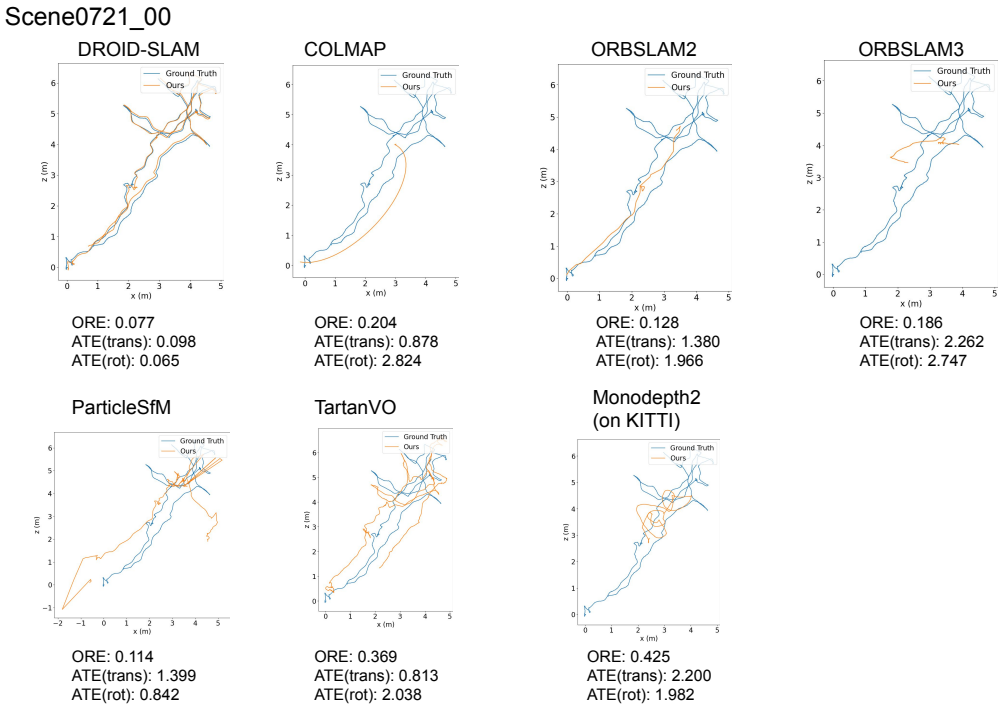
E Annotation Limitations

The annotation for whether an object is static was done by human labelers. Although quality check has been placed for all jobs, it is prune to human bias and error. Also, the videos are sourced from Ego4D VQ benchmark, so it may not represent the entire data disribution of Ego4D.

F Potential negative societal impacts

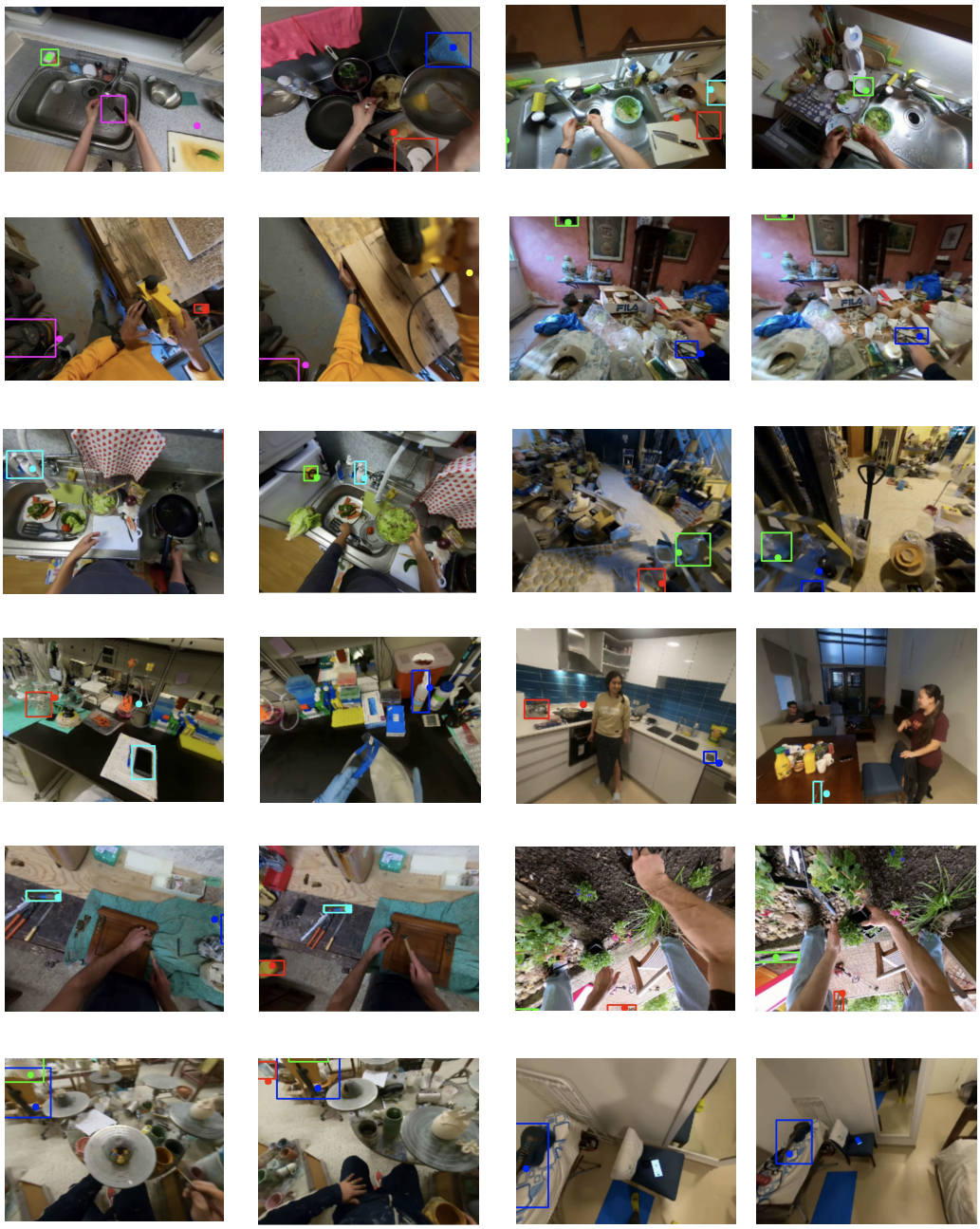
Since our videos are sourced from Ego4D and EgoTracks, we inherit many of the potential negative impacts. Details are discussed in the original paper [5] Appendix K. We summarize several here:

- There may be risks surrounding privacy, such as personnel being recorded during the video. Consent was obtained in the original dataset, and a user agreement is enforced for Ego4D. Our dataset follow the same protocol as Ego4D.



Supp. Figure 1: **Visualization of the camera trajectory** The ORE, ATE$_{trans}$ and ATE$_{rot}$ metric of 7 methods discussed in Sec 4.1 are listed below the trajectory to show that ORE highly correlates with ATE$_{trans}$ and ATE$_{rot}$ in its capability to describe the trajectory quality.

- The existing efforts may inspire future data collection with less attention to privacy and ethics. Best practices were detailed in the original paper [5]. We did the same with our paper to include the instructions to help mitigate this risk.



Supp. Figure 2: **Visualization of EgoStatic benchmark** We include 24 screenshots from 12 videos in the proposed EgoStatic benchmark. Colored bounding boxes are ground truth tracklet annotations; dots are reprojected points tracked by estimated camera trajectory. Ideally, the points should fall in the bounding boxes.

- There may be data imbalances, such as geographical distribution. This risk can be mitigated with future work that grows the collaboration in underrepresented areas.

G Data sheet

We follow [4] for writing the data sheet of EgoStatic.

1. Motivation
   (a) *For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)*
   3D spatial understanding is highly valuable in the context of semantic modeling of environments, agents, and their relationships. Semantic modeling approaches employed on monocular video often ingest outputs from off-the-shelf SLAM/SfM pipelines, which are anecdotally observed to perform poorly or fail completely on some fraction of the videos of interest. These target videos may vary widely in complexity of scenes, activities, camera trajectory, etc. Unfortunately, such semantically-rich video data often comes with no ground-truth 3D information, and in practice it is prohibitively costly or impossible to obtain ground truth reconstructions or camera pose post-hoc.
   This paper proposes a novel evaluation protocol, Object Reprojection Error (ORE) to benchmark camera trajectories; ORE computes reprojection error for static objects within the video and requires only lightweight object tracklet annotations. These annotations are easy to gather on new or existing video, enabling ORE to be calculated on essentially arbitrary datasets. We show that ORE maintains high rank correlation with standard metrics based on ground truth. Leveraging ORE, we source videos and annotations from Ego4D-EgoTracks, resulting in EgoStatic, a large-scale diverse dataset for evaluating camera trajectories in-the-wild.
   (b) *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?* We prefer to stay anonymous in this submission. We refer to Ego4D [5] and EgoTracks [7] for information on dataset collection and annotations of bounding boxes. We will release this information upon paper acceptance.
   (c) *Who funded the creation of the dataset? (If there is an associated grant, please provide the name of the grantor and the grant name and number.)* We prefer to stay anonymous in this submission. We will release this information upon paper acceptance.
   (d) *Any other comments?*
   None.
2. Composition
   (a) *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)*
   Each instance is a video; our annotations are labeling object tracklets from EgoTracks with a static vs. non-static label.
   (b) *How many instances are there in total (of each type, if appropriate)?*
   There are 5708 instances in total.
   (c) *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).)*
   The dataset does not contain all videos in Ego4D. Instead, we only have videos that are present in the Ego4D VQ benchmark [5]. Since these videos are the same, we share the same human and geographic bias and limitations in video selection.
   (d) *What data does each instance consist of? ("Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.)*
   Each object tracklet in a video is labeled as "static" or not.
   (e) *Is there a label or target associated with each instance? If so, please provide a description.*
   Yes, see answer above.

(f) *Is any information missing from individual instances? (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)*

No, we do not remove any information for individual instances already present in Ego4D [5].

(g) *Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? (If so, please describe how these relationships are made explicit.)*

No. Since videos in Ego4D can be hours long, we follow the same procedure to split a video into multiple shorter video clips [5]. Some videos clips may may have been captured by the same individual, but we do not expose this information in the dataset.

(h) *Are there recommended data splits (e.g., training, development/validation, testing)? (If so, please provide a description of these splits, explaining the rationale behind them.)*

No. The benchmark is mainly designed for evaluation. However, given the large-scale of this benchmark, one may leverage this to train models. We leave this as promising future works.

(i) *Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.)*

Since the labels are annotated by human raters, they are prone to human bias and errors. We applied quality assurance procedures to minimize such errors where we can.

(j) *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)*

The underlying videos are sourced from [5]. The [5] dataset is maintained by Ego4D consortium which can be regarded as guaranteed to exist and remain constant. The license can be found https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf.

(k) *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.)*

No. Ego4D [5] was collected with careful consideration of ethics and consent. We largely inherit these characteristics. The additional attribute ("static") annotations of our dataset do not reveal any confidential information.

(l) *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.)*

We share the same underlying videos as [5] and objects as [7], so we inherit its privacy and ethics standards, as well as any potential risks. The additional attribute annotations of our dataset do not add any additional objectionable content.

(m) *Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)*

Yes. The annotations we provide do not contain any information relating to people; however, they are labeled from Ego4D [5] videos, which do contain people, and indirectly are related to people due to its egocentric nature.

(n) *Does the dataset identify any subpopulations (e.g., by age, gender)? (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)*

EgoStatic does not identify subpopulations, but the underlying dataset Ego4D [5] does. See Appendix C of [5] for more details.

(o) *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? (If so, please describe how.)*
EgoStatic does not contain tracks of people, focusing instead on objects. Ego4D [5] does contain some visually identifiable individuals, who provided their consent to appear in the dataset. Other individuals have their faces blurred.

(p) *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? (If so, please provide a description.)*
EgoStatic does not add any labels for people, focusing instead on static objects. Ego4D [5] does contain demographics information. See (n) above.

(q) *Any other comments?*
N/A.

3. Collection Process

(a) *How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)*
The attribute annotation for each object track was acquired by human annotators. They were instructed to label a track as "static" or not.

(b) *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)*
We use the proprietary annotation software to collect "static" attribute annotations. The software shows a video frame by frame, and the annotator is able to select from a dropdown menu whether the object is "static". The software will then record the response for each object track.

(c) *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*
We used the entirety of Ego4D's Visual Queries benchmark [5] and object tracks from EgoTracks [7] as the basis of EgoStatic.

(d) *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*
The participants were contractors employed by a third-party vendor and are compensated based on the agreement with their employer.

(e) *Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)*
The dataset was created in winter and spring of 2023, which is not the time the videos were collected.

(f) *Were any ethical review processes conducted (e.g., by an institutional review board)? (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)*
Yes. For proprietary reasons, we are not able to provide supporting documentation. Our internal review was conducted thoroughly vetted potential privacy and ethical related concerns.

(g) *Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)*
Yes. The annotations we provide in EgoStatic do not contain any information relating to people; however, they are labeled from Ego4D [5] videos, which do contain people, and indirectly are related to people due to its egocentric nature.

(h) *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*
Ego4D [5] collected the data from the individual directly. We do not collect any additional videos beyond those in Ego4D.

(i) *Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)*
Ego4D [5] was collected by willing and consenting individuals. See Section 3.4 in [5].

(j) *Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)*
Ego4D [5] was collected by willing and consenting individuals. See Section 3.4 in [5].

(k) *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)*
Yes, see Section 3.4 in [5].

(l) *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)*
No.

(m) *Any other comments?*
None.

4. Preprocessing/cleaning/labeling

(a) *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)*
No.

(b) *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.)*
No.

(c) *Is the software used to preprocess/clean/label the instances available? (If so, please provide a link or other access point.)*
No.

(d) *Any other comments?*
None.

5. Uses

(a) *Has the dataset been used for any tasks already? (If so, please provide a description.)*
The dataset has been used for understanding the performance of SLAM methods.

(b) *Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)*
N/A.

(c) *What (other) tasks could the dataset be used for?*
The dataset could possibly be used for training SLAM system that focus on the semantic relationships between objects.

(d) *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide*

a description. Is there anything a future user could do to mitigate these undesirable harms?)

The selection of objects is inherited from the Ego4D visual-query benchmark and EgoTracks, which is biased towards objects with appearances of at least modest duration. Consequently, our benchmark may contain less instances that only appear very briefly in the video.

(e) *Are there tasks for which the dataset should not be used? (If so, please provide a description.)*

We currently do not foresee anything such tasks.

(f) *Any other comments?*

None.

6. Distribution

(a) *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.)*

Yes, we will share the same license and distribution as [5].

(b) *How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)*

The dataset will be distributed via GitHub.

(c) *When will the dataset be distributed?*

Upon acceptance of this work.

(d) *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)*

The dataset will share the same license as Ego4D. Please see https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf.

(e) *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)*

Not to our knowledge.

(f) *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)*

Not to our knowledge.

(g) *Any other comments?*

None.

7. Maintenance

(a) *Who is supporting/hosting/maintaining the dataset?*

All authors will maintain the dataset.

(b) *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?*

Authors can be contacted via emails.

(c) *Is there an erratum? (If so, please provide a link or other access point.)*

Not currently. Future versions of the dataset may be released if we find errors, which will be provided within the same GitHub.

(d) *Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)*

See previous question.

(e) *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)*

N/A. The dataset is not related to people.

(f) *Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)*
Yes, all data will be versioned.

(g) *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)*
Errors/features can be submitted as issues/pull requests on GitHub.

(h) *Any other comments?*
None.

References

[1] Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510, 2007. 1

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[3] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L Palmer. Colmap: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021. 1

[4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 7

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 4, 5, 7, 8, 9, 10, 11

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1

[7] Hao Tang, Kevin Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *arXiv preprint arXiv:2301.03213*, 2023. 7, 8, 9

[8] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005. 1