Fine-grained Spatiotemporal Grounding on Egocentric Videos

Supplementary Material

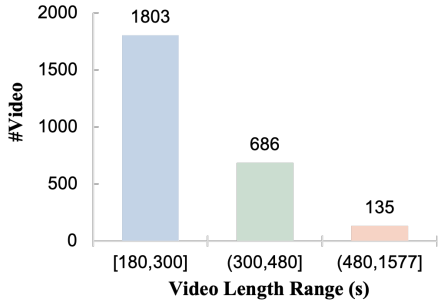

Figure A. Video Length Distribution of **EgoMask-Train**.

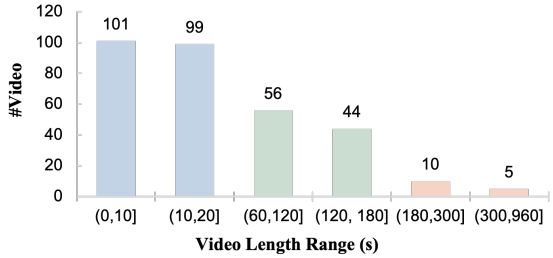

Figure B. Video Length Distribution of **EgoMask**.

A. Comparison of Existing Datasets

We present the detailed comparison of some existing datasets related to spatiotemporal grounding tasks in Table A to show the distinguished difference between egocentric videos and exocentric videos.

B. Annotation verification of EgoMask

We verified 20% test annotations with three experts to score 1 to 5, where 5 is the best. The average scores of expressions/masks are 4.65/4.92. If we set 3 as the threshold score, the error rate is 2.5%/0%, suggesting high quality of our annotations.

C. Video Length Distributions of our datasets

The video length distribution of our training set *EgoMask-Train* and test set *EgoMask* is shown in Figure A and B.

D. Prompts for Expression Generation

To guarantee diversity, we use two different strategies to generate the referring expressions as the language queries for our dataset. We use the prompt shown in Figure C to directly instruct the GPT-4o to generate a short expression and a longer expression. We use the prompt shown in Figure D to first instruct the GPT-4o to generate the metadata of the target objects and then use templates to form expressions. The length statistics are shown in Figure B.

E. Evaluation results of closed-source models

We conduct experiments with GPT-4o and Gemini-Pro on our benchmark (10% subset). Due to their inability to support dense segmentation, we prompt them to generate the corners of boxes, which are then evaluated using IoU_{gold_pred}. The result is 3.47%/1.47% for GPT-4o/Gemini, demonstrating their limitations on our task.

F. Effects of Characteristics of Egocentric Entities

We further provide an in-depth analysis of four key factors. Specifically, we investigate the relations between model performance and the key factors in each benchmark subset.

Total duration. Table C shows the effects of total duration. It is defined as the ratio of total appearance time over the whole video. For all types of benchmarks, the model performs better when the objects have larger total durations (see Below Avg. ✗).

Object size. Table D shows the effects of object size. Generally, the model achieves higher performance when the referred objects have larger sizes (see Below Avg. ✗).

Continuous trajectories. Table E and Table F show the effects of continuous trajectories. The trajectory is defined as one consecutive appearance, and the trajectory length is calculated as the average time of each appearance over the whole video. We define the non-trajectory length as the average time of each disappearance over the whole video. And then, the ratio of disappearance over appearance is calculated as the ratio of non-trajectory length over trajectory length. When the average trajectory length is longer (see Below Avg. ✗ in Table E) and with less disappearance (see Below Avg. ✓ in Table F), the model performs better.

Positional shift. Table G shows the effects of positional shifts. When the objects have fewer shifts in the video (see Below Avg. ✗ in Table G), the performance improves a lot.

Dataset	Egocentric	Video Length (s)	Total Duration (%)	BBox Area (%)	# Traj.	Avg. Traj Length. (%)	Disappear. Ratio (%)	Adj. Bbox IoU (%)	Anno. type
Egotracks [5]	✓	369.00	25.23	2.42	22.96	1.35	496.31	45.07	Bbox
RefEgo [3]	✓	12.27	76.82	2.84	2.01	50.52	24.20	22.16	Bbox
Mevis [1]	✗	69.83	77.78	10.72	1.42	68.88	19.93	65.69	Mask
Ref-Davis [2]	✗	69.41	95.66	13.00	1.18	89.84	4.41	83.30	Mask
Ref-YT-VOS [4]	✗	26.53	93.57	18.49	1.11	89.65	5.97	72.81	Mask

Table A. Comparison of existing datasets related to spatiotemporal grounding task. The "**Total Duration** (%)" means the percent of the total appearance of the referred objects. The "BBox Area (%)" means the average area of the annotated bounding box over the frame size, which can reveal the **object size**. The "# Traj." means the number of object's continuous trajectories throughout the video. The "Avg. Traj. Length (%)" means the average of each trajectory duration over the whole video and the "Disappear. Ratio(%)" is formulated as the mean of each disappearance duration over each trajectory duration. These two values can reveal **the sparsity of the continuous trajectory**. the "Adj. Bbox IoU(%)" shows the **positional shifts** over the adjacent frames by calculating the IoU value of the bounding boxes of the target object.

Short and long expression generation

{frame_0} {frame_1} ...
Please help me generate referring expressions for object segmentation.
There are {total_frames} frames from a video. Each frame contains a red bounding box that corresponds to the same object. Based on the object in the red bounding box and its object tag, please generate the descriptions that uniquely identify the object throughout the video.
Object tag: {object_category}
- Output should consist of two lines, separated by a newline:
1. A short expression with no more than 10 words, starting with "Short expressions: ".
2. A longer expression with more detailed illustrations, starting with "Long expressions: ".
Restriction Policies:
- The referring expressions should be concise and informative. They can be spatial location in the physical world, OCR characters on the object, spatial relations to surrounding objects, action relations to surrounding objects, relative size compared to surrounding objects, color, geometry shape, material, texture pattern, motion or dynamics of objects, and so on.
- The generated referring expressions should clearly identify the object to avoid any ambiguity without referencing bounding boxes in the video.
- Do not use "red bounding box", "image", or "frame" in the answer.

Figure C. Expression generation prompts.

Type	Average Length
Expression	
Short expression	7.75
Long expression	26.31
Metadata	
Caption	2.98
Visual attributes	16.19
Affordance	4.72

Table B. Statistics of the generated expressions and metadata.

Based on the above analysis, we can safely deduce that spatiotemporal grounding on egocentric videos is much harder than that in exocentric videos. We also notice that in most cases, our fine-tuned models, Sa2VA-4B(+FT) and VideoLISA-3.8B(+FT), surpass their pre-trained models. It can verify the effectiveness of our proposed training dataset EgoMask-Train.

G. More Visual Examples

We present more data examples from our proposed benchmark, along with the predictions from different grounding methods in Figure E, F, G, H.

Our fine-tuned models perform better than the pre-trained models. After fine-tuning our proposed training dataset, the VideoLISA-3.8 (+FT) model can **avoid some grounding hallucinations** (#1 frames in Figure E), **perform more precise grounding** (#2- #5 frames in Figure F, #1- #4 frames in Figure G, and #1/ #3 frames in Figure H). Such performance improvements verify the effectiveness of our proposed training dataset EgoMask-Train.

Query understanding ability matters. The SAM2-based model has strong object-tracking ability. However, the capabilities of understanding the queries and knowing the correct object to ground are also important for spatiotemporal grounding tasks. Grounded-SAM2 has an inferior query un-

Object metadata generation

{frame_0} {frame_1} ...
Please help me generate object descriptions. These are {total_frames} frames from a video. Each frame contains a red bounding box that corresponds to the same object. Based on the object in the red bounding box and its object tag, please generate its caption, visual attributes and affordance description (if applicable).
Object tag: {object_category}
- Output should consist of three lines, separated by a newline:
1. A clear object caption with no more than 10 words, starting with "Object Caption: ".
2. The visual attributes of the object, starting with "Visual Attributes: ".
3. A concrete affordance description of the object, starting with "Object: Affordance: ".
Restriction Policies:
- Use the provided object tag selectively, as it may contain noise.
- The object caption should be a noun phrase.
- The object caption should clearly identify the object with minimal words to avoid any ambiguity without referencing bounding boxes.
- Visual attributes characterize the objects in images. They can be spatial location in the physical world, OCR characters on the object, spatial relations to surrounding objects, action relations to surrounding objects, relative size compared to surrounding objects, color, geometry shape, material, texture pattern, motion or dynamics of objects, and so on.
- The affordance description should focus on the object's potential actions, interactions, or functions, describing how the object can be utilized or manipulated in a given context. Avoid generic statements and provide specific and practical insights into the object's affordances.
- The affordance description should be a verb phrase, e.g., cut vegetables, clean the tables, etc. If there is no affordance about the object, output "None".
- Do not use "red bounding box", "image", or "frame" in the answer.

Figure D. Prompts for generating metadata of the labeled object.

Type	Avg. Total Duration(%)	Below Avg.	#Test Sample	Grounded-SAM2	Sa2VA-26B	Sa2VA-4B	Sa2VA-4B (+FT)	VideoLISA	VideoLISA-3.8 (+FT)
Short	80.31	✓	190	40.14	29.15	21.47	22.10 (+**0.63**)	9.71	14.13 (+**4.42**)
		✗	210	58.84	44.67	35.81	39.00 (+**3.19**)	25.21	31.71 (+**6.49**)
Medium	36.69	✓	118	14.71	14.18	10.51	12.95 (+**2.44**)	0.85	1.29 (+**0.44**)
		✗	82	41.59	42.58	26.40	26.54 (+**0.15**)	14.57	22.48 (+**7.91**)
Long	27.48	✓	62	13.13	5.25	1.28	2.60 (+**1.32**)	0.54	0.48 (-**0.06**)
		✗	38	43.86	25.53	19.25	17.43 (-**1.83**)	12.68	18.07 (+**5.39**)

Table C. Performance Comparison over different subsets of total durations. The Avg. Total Duration means the average of total duration (%).

Type	Avg. Mask Area (%)	Below Avg.	#Test Sample	Grounded-SAM2	Sa2VA-26B	Sa2VA-4B	Sa2VA-4B (+FT)	VideoLISA	VideoLISA-3.8 (+FT)
Short	1.83	✓	308	48.63	31.40	22.29	24.31 (+**2.02**)	12.91	18.45 (+**5.54**)
		✗	92	54.38	57.06	51.44	53.28 (+**1.83**)	34.40	39.80 (+**5.39**)
Medium	1.87	✓	142	17.08	20.01	12.18	16.04 (+**3.86**)	5.24	8.46 (+**3.22**)
		✗	58	46.91	40.07	28.88	24.59 (-**4.29**)	9.50	13.69 (+**4.19**)
Long	1.86	✓	76	17.18	11.10	5.27	6.97 (+**1.71**)	3.03	4.67 (+**1.64**)
		✗	24	48.95	18.83	17.11	12.23 (-**4.88**)	11.87	15.05 (+**3.18**)

Table D. Performance Comparison over different subsets of object size. The Avg. Mask Area refers to the average mask area (%) of the queried objects.

Type	Avg. Traj. Length (%)	Below Avg.	#Test Sample	Grounded-SAM2	Sa2VA-26B	Sa2VA-4B	Sa2VA-4B (+FT)	VideoLISA	VideoLISA-3.8 (+FT)
Short	57.13	✓	196	44.17	29.94	21.72	23.65 (+**1.93**)	10.56	15.54 (+**4.99**)
		✗	204	55.52	44.37	35.99	38.01 (+**2.02**)	24.86	30.86 (+**6.01**)
Medium	11.52	✓	156	20.83	21.81	14.69	16.20 (+**1.51**)	4.03	5.60 (+**1.57**)
		✗	44	43.08	40.08	25.30	26.76(+**1.46**)	15.14	25.50 (+**10.36**)
Long	1.81	✓	64	18.01	4.74	0.82	2.16 (+**1.34**)	1.16	1.49 (+**0.33**)
		✗	36	36.89	27.56	21.07	19.03 (-**2.04**)	12.26	17.25 (+**5.00**)

Table E. Performance Comparison over different subsets of our test data over object continuous trajectories. The Avg. Traj. Length refers to the average of each consecutive appearance duration (%).

Type	Avg. Disappear. Ratio (%)	Below Avg.	#Test Sample	Grounded-SAM2	Sa2VA-26B	Sa2VA-4B	Sa2VA-4B (+FT)	VideoLISA	VideoLISA-3.8(+FT)
Short	21.92	✓	190	58.26	45.03	37.37	40.51 (+**3.14**)	25.98	32.00 (+**6.02**)
		✗	210	42.44	30.30	21.42	22.34 (+**0.92**)	10.49	15.53 (+**5.04**)
Medium	179.47	✓	88	38.76	39.97	24.69	24.83 (+**0.15**)	13.69	21.11(+**7.42**)
		✗	112	15.49	14.72	11.01	13.56 (+**2.56**)	0.81	1.23 (+**0.42**)
Long	450.29	✓	46	38.43	21.09	15.91	14.40 (-**1.51**)	10.66	15.05 (+**4.39**)
		✗	54	13.19	6.03	1.47	2.99 (+**1.52**)	0.46	0.45 (-**0.01**)

Table F. Performance Comparison over different subsets over the ratio of disappearance over appearance. The Avg. Disappear. Ratio refers to the mean value of the ratio of average disappearance duration over the average trajectory length.

Type	Avg. Adj. Mask IoU (%)	Below Avg.	#Test Sample	Grounded-SAM2	Sa2VA-26B	Sa2VA-4B	Sa2VA-4B (+FT)	VideoLISA	VideoLISA-3.8(+FT)
Short	8.51	✓	268	45.78	28.91	19.83	21.88 (+**2.05**)	13.29	18.75 (+**5.46**)
		✗	132	58.43	54.34	47.62	49.43 (+**1.81**)	27.12	32.71 (+**5.59**)
Medium	20.98	✓	116	16.70	19.00	12.99	14.93 (+**1.94**)	1.54	3.09 (+**1.54**)
		✗	84	38.20	35.26	22.59	23.48 (+**0.89**)	13.29	19.50 (+**6.21**)
Long	19.53	✓	58	18.31	7.24	1.65	3.31 (+**1.66**)	2.23	2.75 (+**0.52**)
		✗	42	33.77	20.85	17.03	15.04 (-**1.99**)	9.19	13.26 (+**4.07**)

Table G. Performance Comparison over different subsets over object position shifts. The Avg. Adj. Mask IoU refers to the mean IoU value of spatial position over the adjacent frames.

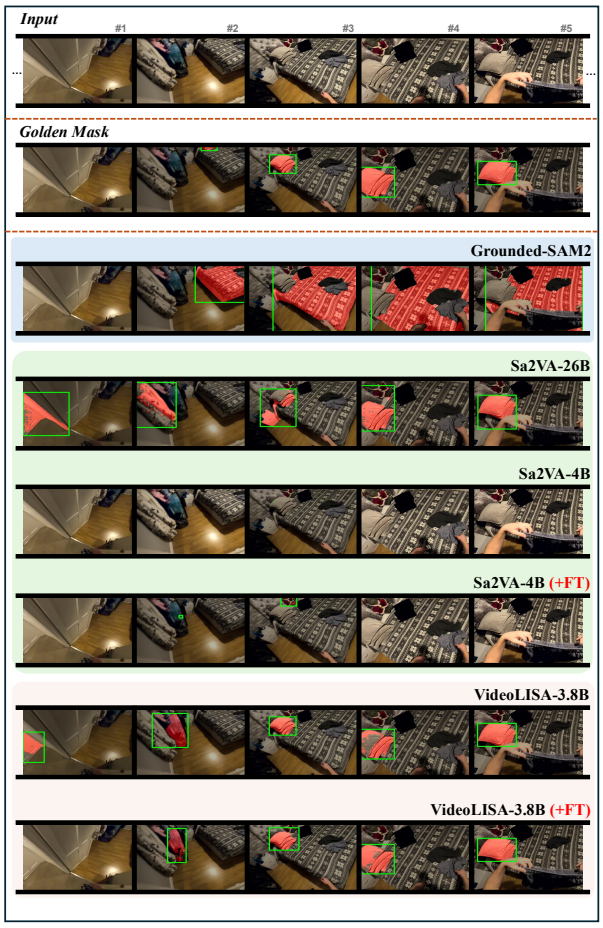

Figure E. Visualization of example from EgoMask-Short. The language query is "the pillows stacked on top of bed".

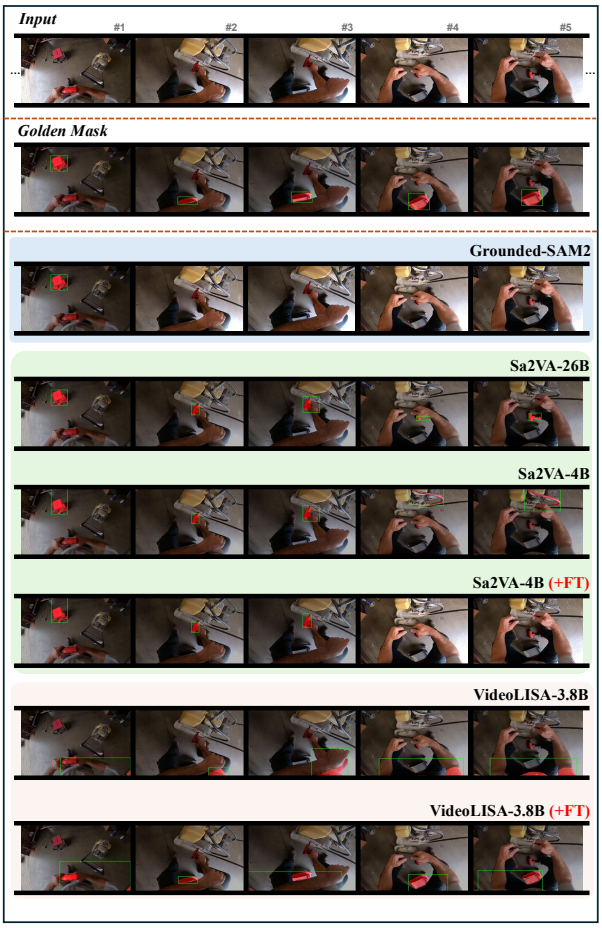

Figure F. Visualization of one example from EgoMask-Medium with sampled frames. The language query is "the snap-on stool with a red cushion".

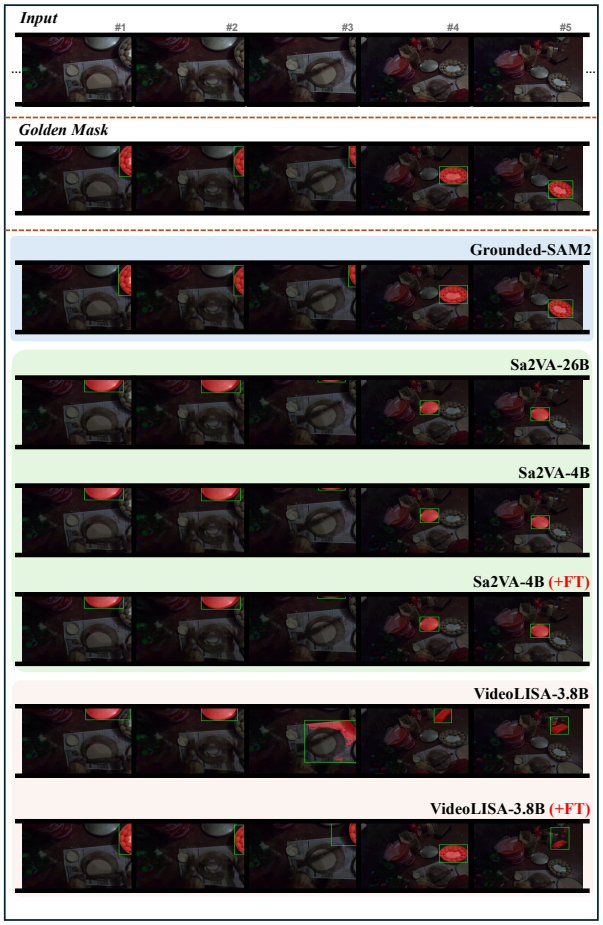

Figure G. Visualization of example from EgoMask-Medium. The language query is "the circular silver-colored metal platter containing evenly arranged small, oval-shaped dough balls, placed on a table near a red container labeled "deepak" and surrounded by other kitchen items ".

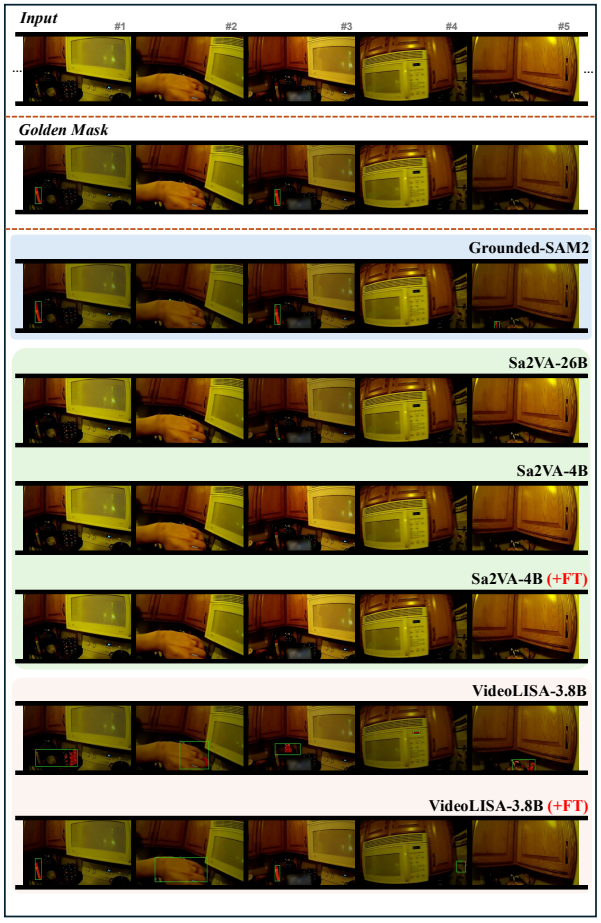

Figure H. Visualization of one example from EgoMask-Long. The language query is "the tall, cylindrical white bottle with a red cap, located on the kitchen counter near the sink and surrounded by dishes and other kitchen items ".

derstanding compared to VideoLLMs. As shown in Figure E, it tracks the wrong object bed instead of the referred object pillow.

All the above visual examples can show the difficulty of fine-grained spatiotemporal grounding on egocentric videos.

References

[1] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703. IEEE, 2023. 2

[2] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV (4)*, pages 123–141. Springer, 2018. 2

[3] Shuhei Kurita, Naoki Katsura, and Eri Onami. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *ICCV*, pages 15168–15178. IEEE, 2023. 2

[4] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: unified referring video object segmentation network with a large-scale benchmark. In *ECCV (15)*, pages 208–223. Springer, 2020. 2

[5] Hao Tang, Kevin J. Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. In *NeurIPS*, 2023. 2