

Fine-grained Spatiotemporal Grounding on Egocentric Videos

Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, Liwei Wang[†]
 The Chinese University of Hong Kong

Abstract

Spatiotemporal video grounding aims to localize target entities in videos based on textual queries. While existing research has made significant progress in exocentric videos, the egocentric setting remains relatively underexplored, despite its growing importance in applications such as augmented reality and robotics. In this work, we conduct a systematic analysis of the discrepancies between egocentric and exocentric videos, revealing key challenges such as shorter object durations, sparser trajectories, smaller object sizes, and larger positional shifts. To address these challenges, we introduce EgoMask, the first pixel-level benchmark for fine-grained spatiotemporal grounding in egocentric videos. It is constructed by our proposed automatic annotation pipeline, which annotates referring expressions and object masks across short-, medium-, and long-term videos. Additionally, we create EgoMask-Train, a large-scale training dataset to facilitate model development. Experiments demonstrate that the state-of-the-art spatiotemporal grounding models perform poorly on our benchmark EgoMask, but fine-tuning on EgoMask-Train yields significant improvements, while preserving performance on exocentric datasets. Our work thus provides essential resources and insights for advancing egocentric video understanding. Our code is available at <https://github.com/LaVi-Lab/EgoMask>.

1. Introduction

Video grounding has been extensively studied in recent years [11, 12, 16, 22, 44, 47, 62]. It requires models to generate a temporal tube for target entities based on a given language query. As illustrated in Figure 1, this tube can be represented as a sequence of consecutive frames [11, 22] (temporal grounding), or a set of bounding boxes and masks that localize entities within these frames [12, 44] (spatiotemporal grounding). The mask output further requires a pixel-level localization on the fine-grained object geometry. It is worth noting that existing works [1, 59] predominantly focus on spatiotemporal grounding in exocentric videos [9, 44],

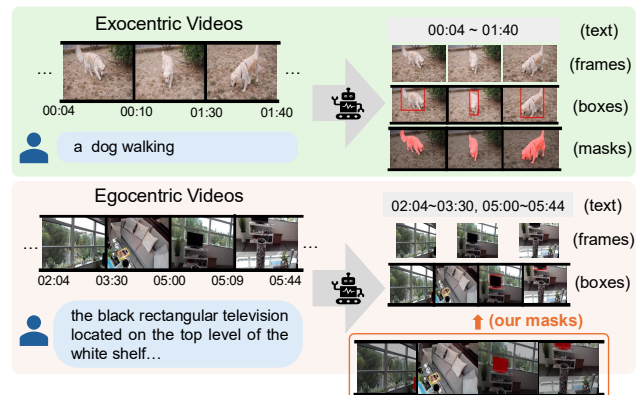


Figure 1. Comparison of video grounding tasks. We propose the first pixel-level benchmark for fine-grained spatiotemporal grounding in egocentric videos.

while egocentric video grounding remains relatively underexplored. This is despite the growing interest in egocentric video understanding [4, 7, 56], driven by its potential applications in augmented reality (AR) glasses, household robotics, and other real-world scenarios.

Meanwhile, existing studies have observed a gap between egocentric and exocentric videos [11, 20, 46]. For instance, egocentric videos are characterized by rapid camera movements, causing objects to frequently enter and exit the field of view, while also undergoing rapid appearance changes. However, a systematic and statistical analysis of these differences remains lacking. Therefore, we conducted a study that quantitatively analyzes the discrepancies between egocentric and exocentric videos. Our findings reveal several key differences: entities in egocentric videos exhibit **shorter** total durations, **sparser** continuous trajectories, **smaller** object sizes, and **larger** positional shifts compared to their counterparts in exocentric videos. This study raises a key question: *Can existing models perform well on fine-grained spatiotemporal grounding in egocentric videos?*

To answer the question, however, there lacks a benchmark specifically designed for fine-grained spatiotemporal grounding in egocentric videos. The most relevant datasets are EgoTracks [46] and RefEgo [20]. EgoTracks focuses on long-term object tracking in egocentric videos but only provides object category labels and bounding box annota-

[†] Corresponding author

tions. RefEgo, on the other hand, evaluates referring expression comprehension in egocentric videos and includes text queries, yet it also provides only bounding box annotations. Importantly, RefEgo consists solely of short video segments (under 1 minute), which do not align with the typical visual inputs in egocentric AI applications, where models process long egocentric video streams.

In this work, we propose an automatic annotation pipeline to construct a benchmark that provides pixel-level annotations and enables the evaluation of spatiotemporal grounding in egocentric videos across various durations. To generate mask annotations, we leverage the pre-trained segmentation model SAM2 [42], using the bounding boxes provided by the EgoTracks dataset. To generate object expressions as language queries, we employ the vision-language model GPT-4o [33] through two strategies to ensure data diversity: (1) prompting GPT-4o to directly generate both short and long expressions, and (2) first prompting GPT-4o to produce metadata about the target object, such as visual attributes and world knowledge, which is then combined to construct referring expressions. Finally, both mask and expression annotations are refined and verified by human annotators. As a result, we introduce the first egocentric, fine-grained spatiotemporal grounding benchmark, EgoMask, built upon a subset of EgoTracks and RefEgo. The benchmark comprises 700 queries across 315 videos, spanning **short-term** (under 1 minute), **medium-term** (1 to 3 minutes), and **long-term** (over 3 minutes) durations to enable a comprehensive evaluation. Moreover, to support model training, we create *EgoMask-Train*, a large-scale training set created by our automatic pipeline. It includes 2,624 videos with mask annotations for 9,592 objects and a total of 47,968 referring expressions.

With our newly proposed benchmark, we systematically evaluate the state-of-the-art (SOTA) spatiotemporal grounding models on egocentric videos. Experiment results reveal that existing SOTA models [1, 59] perform significantly worse on our benchmark compared to their performance on existing exocentric benchmarks. Going further, we fine-tune pre-trained SOTA models on our created training dataset, *EgoMask-Train*, leading to large performance improvements (*e.g.*, an average relative increase of 41.30%). These results highlight the effectiveness of our dataset as a valuable resource for advancing egocentric spatiotemporal grounding. Notably, we observe that models fine-tuned on our egocentric dataset still retain their performance on existing exocentric benchmarks. This suggests that our dataset is complementary to existing exocentric datasets and can serve as a unique and orthogonal data source for training future video foundation models. In addition to end-to-end trained SOTA models, we also evaluate a baseline method, Grounded-SAM2 [43], a pipeline framework that first runs GroundingDino [27] to localize objects based on text queries, followed by SAM2 [42]

for object tracking. Even though it uses the same tracking model as our annotation pipeline, this method also fails to perform well on egocentric videos, which further verifies the difficulty of our proposed benchmark and the challenges of fine-grained spatiotemporal grounding on egocentric videos.

Overall, our contributions are as follows:

- We explore spatiotemporal video grounding for egocentric videos and develop an automatic data annotation pipeline, resulting in the first pixel-level benchmark *EgoMask* and a large-scale training dataset *EgoMask-Train*.
- We conduct in-depth analysis to quantitatively and systematically measure the gap between exocentric and egocentric videos, providing insights for future modeling.
- Extensive experiments reveal that existing spatiotemporal grounding models fail to perform effectively on egocentric videos, and our collected training data can remarkably improve existing models.

2. Related work

Referring Video Segmentation Dataset. Referring video object segmentation (RVOS) [2, 10, 23, 31, 49, 50], a subtask of spatiotemporal video grounding. It aims to segment the target objects in a video based on a given natural language expression. Compared to standard video object segmentation [55], RVOS is more challenging as it requires models to effectively integrate both visual and linguistic information. The RVOS task was first introduced by [10] along with the A2D-Sentences and J-HMDB Sentences datasets, which primarily focus on actor segmentation. Subsequent works [1, 9, 18, 44] have expanded the scope of RVOS by scaling up datasets, increasing task complexity, or enhancing task diversity. Specifically, Ref-DAVIS [18] extends the DAVIS datasets [35, 37] by replacing segmentation masks with language descriptions, introducing a multi-object segmentation setting. Given its relatively small scale, Refer-YouTube-VOS [44] was built on YouTube-VOS [53], featuring diverse object categories and longer videos. MeViS [9] incorporates more complex scenes with a higher density of objects, with expressions focusing on object motion. ReasonVOS [1] emphasizes complex reasoning over language queries and temporal object tracking with explicit motion understanding. However, these datasets predominantly feature exocentric videos which are generally shorter and exhibit less rapid camera movement. In contrast, we focus on egocentric videos which present additional challenges due to their longer durations, frequent and large camera movements, and increased object density.

Egocentric Dataset. Egocentric video understanding [11, 24, 34, 38] has recently emerged as a pivotal research area in Embodied AI [32, 36], presenting challenges distinct from those in exocentric video analysis. Unlike conventional exocentric datasets [14, 51, 52, 58], egocentric videos are

captured from a first-person perspective using wearable devices, resulting in long-form, dynamic footage characterized by frequent camera motion. To advance egocentric video understanding, numerous datasets have been introduced [6–8, 11, 20, 29, 45, 46]. Ego4D [11] laid the foundation by providing a large-scale, comprehensive egocentric dataset. Building upon Ego4D, several datasets have been developed to extend the scope of egocentric video understanding. For instance, EgoSchema [29] focuses on long-form egocentric video question-answering (VideoQA), while [7, 8] addresses grounding in VideoQA. EgoTracks [46] is designed for long-term object tracking, and RefEgo [20] evaluates referring expression comprehension. However, these datasets do not support pixel-level spatiotemporal grounding due to the lack of mask annotations. EPIC-Visor [6] provides pixel-level masks. However, it has only 158 videos that are in kitchens, and its clips with consistent mask annotations are 12 seconds on average, failing to support long video grounding. In contrast, our EgoMask provides diverse textual queries alongside mask annotations, covering video lengths ranging from seconds to minutes. This facilitates more fine-grained and comprehensive spatiotemporal grounding.

Multimodal Large Language Model. Multimodal large language models (MLLMs) [25, 33, 48] have made significant progress in visual-language tasks [13, 57, 60]. Recently, many studies explore their capabilities in pixel-level understanding [1, 21, 59, 61]. A common practice is using special tokens to enable MLLMs with grounding ability, such as the Segment Anything Model (SAM) [19, 42]. LISA [21] and VideoLISA [1] introduce a special “[SEG]” token to connect the MLLM with SAM and perform independent image segmentation, while Sa2VA [59] integrate SAM2 [42] into MLLMs, allowing the utilization of other frames to perform video segmentation. However, these models primarily focus on exocentric videos. It is still unclear whether they can perform well on egocentric videos. In this work, we create a benchmark to evaluate them on egocentric spatiotemporal grounding and a training dataset to enhance their performance. The fine-tuned models achieve large improvements on egocentric videos, without performance loss on exocentric videos.

3. Method

Our work fills a critical gap in pixel-level spatiotemporal grounding for egocentric videos, better supporting egocentric applications in augmented reality and robotics. Specifically, we design an automatic pipeline that utilizes existing segmentation models to generate pixel-level object masks and leverages visual-language models to produce object expressions as text queries. This pipeline enables the construction of both a training dataset and an evaluation benchmark, laying the foundation for advancing egocentric video understanding.

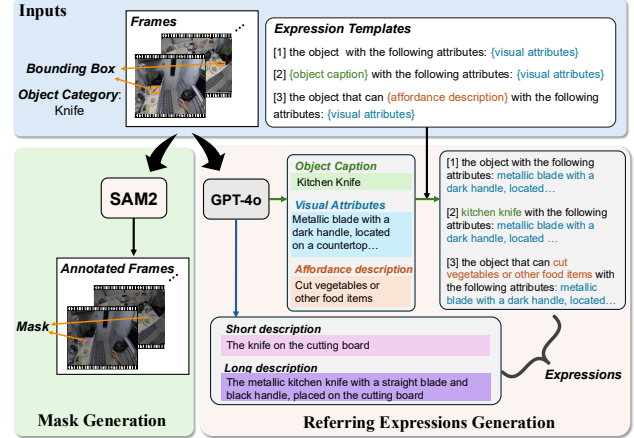


Figure 2. Automatic Annotation Pipeline. The inputs include video frames, the annotated bounding boxes, and the object category from the EgoTracks dataset. The annotation process contains two parts: (1) **Mask Generation** (Bottom Left): utilizing SAM2 as annotator to generate mask throughout the input frames; (2) **Referring Expression Generation** (Bottom Right): prompting GPT-4o to directly generate referring expressions (blue arrow) or first generate metadata about the labeled object and then adopt pre-defined templates to generate expressions (green arrow).

3.1. Automatic Annotation Pipeline

Existing egocentric datasets are not directly suitable for pixel-level spatiotemporal grounding tasks. For instance, EgoTracks [46] provides long-term object tracking with only bounding box annotations and lacks referring expressions for the annotated objects. Given the high cost of manually annotating object masks, we build upon the densely annotated bounding boxes in EgoTracks and introduce additional annotations to address the absence of pixel-level masks and language queries. Specifically, we propose an automatic annotation pipeline that: (1) generates object masks using existing segmentation tools guided by the provided bounding boxes, and (2) produces referring expressions (*i.e.*, language queries) for the target objects using vision-language models, as illustrated in Figure 2.

Pixel-level mask generation. In order to reduce the annotation time and avoid segmentation errors caused by object absence, we only select the clip segments that contain the target object to annotate. For each clip segment, the bounding box in the first frame is selected as the box prompt input of SAM2 [42], a powerful unified model that can segment objects across images and generate their masks throughout the video clip. We then post-process the generated masks by only keeping the areas that overlap with box annotations. In this way, we can minimize hallucination errors and ensure that the generated masks are within or near the area of the target objects.

Referring expression generation. In grounding tasks, the

referring expression should correspond to the unique target object without any ambiguity, which requires the annotation tool to fully understand the distinction of the target object from the surroundings. Furthermore, for better supporting egocentric applications in complex situations, the referring expressions should cover diverse aspects of the objects [63], including simple object captions, various visual attributes, and world knowledge about the objects. Thus, to guarantee the *diversity* of referring expressions, we propose two different strategies to generate referring expressions.

- Prompting GPT-4o to directly generate a short description and a long description. We first select three frames that have the most clear objects (*e.g.*, large bounding boxes) from the video and make sure they come from different trajectories. Then, We draw boxes of the target object on video frames and prompt GPT-4o to generate descriptions.
- Prompting GPT-4o to generate metadata first and then produce expressions based on designed templates. Inspired by [63], the metadata includes object caption, visual attributes (*e.g.*, physical locations, color, shape, dynamics), and world knowledge (*e.g.*, object affordance). Once these metadata are obtained, we utilize the templates shown at the top of Figure 2 to form expressions.

The detailed prompts can be found in the Appendix.

3.2. Dataset Curation

Our automatic pipeline is used to construct both an evaluation set and a training set. The evaluation set undergoes additional verification by human annotators to ensure annotation quality, while the training set is used to train spatiotemporal grounding models.

Training dataset curation. Our training set *EgoMask-Train* comes from the filtered subset of Egotracks. It contains 2,624 videos and mask annotations of 9,592 objects with 47,968 expressions. The detailed statistics are shown in Table 1, alongside the comparison with the existing exocentric training dataset. Note that we annotate the video at 1 FPS, while the sampling rate of the original Egotracks dataset is 5 FPS. From Table 1, we can observe that our EgoMask-Train is characterized by:

- **Shorter** total duration: On average, the referred objects show in the video for only **21.56%** time, compared to over 75% time in previous datasets.
- **Sparser** continuous trajectories: On average, the length of continuous trajectory (*i.e.*, consecutive appearance) is only **1.33%** of the whole video, while in the exocentric video, the object trajectory length is over **65%**. Also, the disappearance of target objects is about six times longer than their appearance on average (**655.82%** Disappear Ratio), while in exocentric videos, the disappearance duration is much shorter (under 21%).
- **Smaller** object size: The average mask area is only **1.20%** of the whole frame, while objects in the previous bench-

marks are five times larger (over 5%).

- **Larger** positional shift: The average IoU of masks on the adjacent frames is only **14.96%**, while in the exocentric video, the IoU value is over 50%. It is caused by the large and frequent camera movement in egocentric videos.

Evaluation benchmark curation. Our benchmark *EgoMask* includes 315 videos ranging from 5 seconds to 16 minutes, categorized into short-term (under 1 minute), medium-term (between 1 minute and 3 minutes), and long-term (over 3 minutes). It contains 700 expressions in total with an average of 15 words. All samples are manually refined and verified by using the semi-automatic labeling tool [15] that supports SAM2 tracking.

- EgoMask-Short: The short videos and the corresponding expressions are sampled from RefEgo [20]. The annotators manually label the object masks and refine the expressions to make them more precise.
- EgoMask-Long: The long videos are sampled from the validation set of Egotracks [46]. We use the above pipeline to generate the masks and expressions for human annotators to refine, which can largely reduce the annotation time.
- EgoMask-Medium: The medium videos are extracted from the above annotated long videos. For each object in the long video, we randomly extract two different clip segments that contain the target object.

The statistics of EgoMask are shown in Table 2, alongside the comparison with existing exocentric benchmarks. Same as our training dataset, our egocentric benchmark is characterized by **shorter** total duration, **sparser** continuous trajectories, **smaller** object size, and **larger** positional shift.

4. Experiments

In this section, we first introduce the evaluation dataset, evaluation metrics, the baseline models, and implementation details. Then we present our results, including the results of our egocentric benchmark, exocentric benchmarks, detailed analysis, and visualizations.

4.1. Setup

Dataset. We evaluate pixel-level spatiotemporal video grounding on one egocentric benchmark (our proposed EgoMask) and three exocentric benchmarks (Ref-Davis [18], Mevis [9], and ReasonVOS [1]). For model fine-tuning, apart from our proposed dataset EgoMask-Train, we also use a mixture of image-based and video-based segmentation datasets [1, 3, 5, 9, 17, 18, 30, 39, 44, 53, 54, 64], following the previous works [1, 59]. These segmentation datasets focus on exocentric videos, while our training dataset is tailored for egocentric videos and has distinct characteristics.

Metrics. For fine-grained spatiotemporal grounding on egocentric videos, we adopt four metrics. We refer “target frame” as the video frame where the target (ground-truth)

Training dataset	Video Length (s)	Total Duration (%)	Mask Area (%)	#Traj.	Avg. Traj. Length (%)	Disappear Ratio (%)	Adj. Mask IoU (%)	#Video	#Object	# Expr.	Avg. Expr. Length
<i>Existing Exocentric Training Dataset</i>											
Ref-Davis	71.18	94.33	5.84	1.22	87.45	5.85	64.96	60	142	572	6.27
Mevis	69.67	77.51	5.34	1.43	68.52	20.29	54.83	1,662	6,719	23,051	7.05
Ref-YT-VOS	26.53	93.57	10.23	1.11	89.65	5.97	61.62	3,471	6,459	12,913	9.63
<i>Our Egocentric Training Dataset</i>											
EgoMask-Train	369.94	21.56	1.20	17.00	1.33	655.82	14.96	2,624	9,592	47,968	21.67

Table 1. The statistics of our proposed training dataset and the comparison with existing egocentric training datasets. The “**Total Duration (%)**” means the percent of the total appearance of the objects. The “**Mask Area (%)**” means the average area of the annotated mask over the frame size, which can reveal the **object size**. The “**# Traj.**” means the number of object’s continuous trajectories throughout the video, where the trajectory is defined as one consecutive appearance in 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 disappearance duration over trajectory duration. These two values can reveal the **sparsity of the continuous trajectory**. the “**Adj. Mask IoU(%)**” shows the **positional shifts** over the adjacent frames by calculating the IoU value of the masks of the target object. The notion “**Expr.**” refers to the expressions.

Benchmark	Video Length (s)	Total Duration (%)	Mask Area (%)	#Traj.	Avg. Traj. Length (%)	Disappear Ratio (%)	Adj. Mask IoU (%)	#Video	#Object	#Expr.	Avg. Expr. Length	
Existing Exocentric Benchmark												
Ref-Davis	65.31	98.77	5.92	1.10	95.88	1.09	71.13	30	61	244	6.11	
Mevis	76.55	88.14	4.92	1.36	74.77	11.44	60.82	50	198	793	7.70	
ReasonVOS	98.12	89.34	1247.75	1.34	72.03	15.73	66.61	91	105	458	14.09	
Our Egocentric Benchmark												
EgoMask	Short	12.15	80.31	1.83	1.66	57.13	21.92	8.51	200	200	400	13.27
	Medium	116.30	36.69	1.87	4.62	11.10	187.84	21.15	100	100	200	17.34
	Long	361.32	27.48	1.86	14.60	1.81	450.29	19.53	15	50	100	17.34

Table 2. The statistics of our proposed benchmark and comparison with existing exocentric benchmarks.

entity presents, “background frame” as the video frames that do not have the target entity, and “predicted frame” as the video frame with predicted masks from the models.

- T_{recall} : It is defined as the proportion of predicted frames among target frames. It evaluates the temporal grounding performance of models.
- IoU_{all} : It is calculated by the average of Intersection-over-Unions (IoUs) value among all the video frames, which is also called region similarity (\mathcal{J}) and commonly adopted in the previous works [1, 37].
- IoU_{gold} : It is defined by the average of IoU among target frames. This metric only focuses on the prediction results over target masks and *ignores* the predictions on the background frames.
- $IoU_{gold-pred}$: It is defined by the average of IoUs among all target frames and predicted frames. It excludes the frames where the ground-truth mask and predicted mask do not exist. This metric is more challenging as it *penalizes* the hallucinated predictions on the background frames.

Baselines. We choose one pipeline tracker and two end-to-end VideoLLMs as our baselines.

- Grounded-SAM2 [43], a pipeline tracker. It first uses GroundingDino [26] to perform image-based object detection and output a bounding box. Then, the box is viewed as the box prompt for the SAM2 model to get the segmen-

tation masks. For the first stage, we apply GroundingDino on all the frames and select the frame whose generated bounding box has the highest detection confidence score as the initial frame for SAM2 tracking.

- Sa2VA [59], an open-sourced VideoLLM which links LLaVA and SAM2 to perform video grounding. It first generates the mask for a few key frames, and then uses SAM2 to perform video propagation based on these frames. During inference, it uses the first five frames as the key frames. We test its 4B and 26B versions in our evaluation.
- VideoLISA [1], an open-source VideoLLM which uses “[SEG]” token to link LLaVA-Phi-3-V [40] and SAM. It performs independent image segmentation on the frames using the same [SEG] token. Its model size is 3.8B.

Fine-tuning Implementation. We build two fine-tuned VideoLLMs, which are shown as follows,

- Sa2VA-4B (+FT): We fine-tune Sa2VA-4B [59] on our proposed egocentric dataset EgoMask-Train, alongside three video segmentation datasets (Mevis [9], ReasonVOS [1], Ref-YouTube-VOS [44]) that the original model is trained on. The fine-tuning is deployed on 8 NVIDIA 80G A100 GPUs for about 10 hours, with AdamW [28] optimizer and learning rate as 4e-6. The batch size is set to 16.
- VideoLISA-3.8B (+FT): We fine-tune VideoLISA [1] using 4 NVIDIA 80G A100 GPUs based on DeepSpeed [41]

for a total of 20 epochs, with AdamW [28] optimizer and learning rate as $3e-5$. The fine-tuning data contains two parts, with 80% sampled from our proposed egocentric dataset EgoMask-Train and 20% sampled from a suite of image and video segmentation datasets that the original model is trained on. The batch size is set to 16, and the number of steps for each epoch is set to 500. The whole process takes about 12 hours.

4.2. Main Results

We conduct experiments with various grounding methods on our proposed egocentric benchmark EgoMask (Table 3) and three existing exocentric benchmarks (Table 4).

Existing models perform poorly on egocentric videos.

As Table 3 shows, all grounding methods fail to achieve satisfactory performance on our benchmark. For the least challenging subset of our benchmark (EgoMask-Short), the best IoU_{gold_pred} score is under 50%. And for EgoMask-Medium and EgoMask-Long, the challenging subsets with longer video lengths, the performance drops largely for all models (*i.e.*, less than 30% IoU_{gold_pred} score). Our Benchmark annotations have been manually verified to minimize the bias from annotation tools. Even though some baselines adopt the SAM2 model, which is also used in our annotation pipeline, they still fail to perform well on our benchmark (*i.e.*, Grounded-SAM2 only achieves 49.95%, 25.73%, and 24.80% on Short, Medium, and Long, respectively). These results reveal that our pixel-level egocentric grounding benchmark is challenging, and there is a significant gap between egocentric and exocentric videos.

Our training dataset is beneficial. The models fine-tuned on our training set achieve large improvements on our benchmark, while maintaining their grounding ability in exocentric benchmarks. As in Table 3, VideoLISA-3.8B (+FT) achieves a noticeable performance improvement on all subsets, with an average relative increase of 41.30% and absolute increase of 3.67%. Similarly, Sa2VA-4B (+FT) achieves an average relative increase of 5.74% and absolute increase of 1.20%. On the other hand, when evaluated on exocentric benchmarks, the fine-tuned models achieve comparable or even better results than the original pre-trained models. VideoLISA-3.8B (+FT) achieves a 1.77% performance improvement in the ReasonVOS benchmark and drops slightly in Ref-Davis (-0.22%). Sa2VA-4B (+FT) surpasses its baseline on all benchmarks. These results indicate that our training dataset is complementary to previous exocentric datasets. Overall, these experiments verify the effectiveness of our proposed training dataset EgoMask-Train for improving fine-grained egocentric spatiotemporal grounding, providing a unique resource for future video foundation models.

Our defined metrics can better reflect the grounding capabilities on both egocentric and exocentric videos.

The commonly-used metric IoU_{all} (also known as \mathcal{J}) is not

proper for our egocentric benchmark, especially in EgoMask-Medium and EgoMask-Long. This arises from the large portion of background frames. The IoU_{all} metric considers all frames and thus is dominated by the background frames, failing to reflect the model performance at target frames. In contrast, our proposed metrics $IoU_{gold}/IoU_{gold_pred}$ will ignore/penalize the predictions on background frames, thereby better reflecting model performance. Take Sa2VA in Table 3 as an example. We can observe that from short to long, the model retrieves fewer target frames (T_{recall} is small), while the IoU_{all} score gets higher. In comparison, our defined metrics are more stable and align with the subset difficulty (*i.e.*, $IoU_{gold}/IoU_{gold_pred}$ drops as T_{recall} decreases). Meanwhile, exocentric benchmarks include fewer background frames, and thus, our proposed metrics degenerate to the old metric. As shown in Table 4, our proposed metrics are consistent with IoU_{all} . Therefore, our defined metrics well reflect grounding abilities on both ego and exo videos.

Model design tailored for egocentric videos is lacking.

The comparison between VideoLISA and Sa2VA suggests that transferring video segmentation ability from the SAM2 model is better than performing independent image segmentation on individual frames. We also compare Sa2VA and Grounded-SAM2, both of which are SAM2-based methods. Their major difference lies in their architectures: Sa2VA is an end-to-end model while Grounded-SAM2 follows a pipeline design. We find that Sa2VA shows better text query understanding than the pipeline model, as reflected by the results on ReasonVOS, and performs better on exocentric benchmarks (Table 4). The main reason is that Grounded-SAM2 relies on the boxes predicted from GroundingDino, which fails to fully understand the expressions. It thus achieves low grounding scores due to the wrong boxes and the error propagation. In contrast, Sa2VA can take advantage of the reasoning capabilities of LLMs to better understand the queries. However, on EgoMask, even the largest Sa2VA model (26B) achieves inferior performance to Grounded-SAM2. These observations indicate that the existing modeling fails to explore the full potential of the pre-trained grounding models on egocentric videos.

4.3. Analysis

We conduct further analysis to evaluate the modeling design and reveal corresponding challenges.

Effects of Initialization State of SAM2. From Table 3, the SAM2-based models (Grounded-SAM2 and Sa2VA) perform better than the SAM-based model (VideoLISA) as they can utilize information from the context frames. However, for these SAM2-based models, the initialization of the inference state is of vital importance and influences the performance a lot due to the information propagation. We conduct experiments with the SAM2-based methods (shown in Table 5 and Table 6). (1) For Grounded-SAM2, its infer-

Methods	Short				Medium				Long			
	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$
Pipeline Baseline												
Grounded-SAM2	91.31	54.75	51.00	49.95	65.85	54.35	28.23	25.73	61.44	61.54	27.36	24.80
End2End Open-Source VideoLLMs												
Sa2VA-26B	70.08	48.23	39.20	37.30	44.15	73.06	28.83	25.83	28.53	72.01	15.45	12.96
Sa2VA-4B	69.33	40.41	31.01	29.00	36.83	66.10	18.68	17.02	21.55	68.67	8.68	8.11
Sa2VA-4B (+FT)	72.62	41.70	32.92	30.97 (+1.97)	39.27	67.42	20.12	18.52 (+1.50)	21.60	69.50	9.14	8.24 (+0.13)
VideoLISA-3.8B	98.37	18.14	20.94	17.85	96.99	9.28	11.87	6.48	96.13	8.48	12.11	5.15
VideoLISA-3.8B (+FT)	97.95	23.98	27.33	23.36 (+5.51)	95.52	14.05	17.28	9.98 (+3.50)	95.33	12.03	14.82	7.16 (+2.01)

Table 3. Experimental results on our egocentric benchmark EgoMask. FT means the fine-tuning on our training dataset EgoMask-Train.

Methods	Ref-Davis				Mevis				ReasonVOS			
	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$	T_{recall}	IoU_{all}	IoU_{gold}	$IoU_{gold,pred}$
Pipeline Baseline												
Grounded-SAM2	98.48	62.74	62.41	62.39	97.27	41.07	40.69	40.65	0.00	9.91	0.00	0.00
End2End Open-Source VideoLLMs												
Sa2VA-26B	96.90	73.77	74.12	73.58	95.98	56.38	55.24	54.80	79.84	57.34	53.21	52.47
Sa2VA-4B	96.21	69.87	70.22	69.75	96.54	51.24	50.41	50.01	77.69	47.50	43.35	42.35
Sa2VA-4B (+FT)	96.22	70.09	70.45	69.97 (+0.22)	96.21	57.45	56.05	55.55 (+5.54)	79.10	50.11	46.34	45.54 (+3.19)
VideoLISA-3.8B	99.99	65.84	66.26	65.82	99.75	49.36	51.18	49.20	99.66	42.43	45.48	42.41
VideoLISA-3.8B (+FT)	99.96	65.60	66.01	65.60 (-0.22)	99.54	49.46	51.31	49.20 (-0.00)	99.09	44.27	47.13	44.18 (+1.77)

Table 4. Experimental results on existing exocentric benchmarks. FT means the fine-tuning on our training dataset EgoMask-Train.

EgoMask	Ground with highest detection confidence	Detection		ST-Grounding	
		Accuracy	IoU	T_{recall}	$IoU_{gold,pred}$
Short	✓	87.00	52.97	91.31	49.95
	✗	82.75	42.49 (-10.48)	87.64	40.42 (-9.53)
Medium	✓	51.50	31.56	65.85	25.73
	✗	67.00	18.24 (-13.32)	65.38	15.11 (-10.62)
Long	✓	47.00	32.74	61.44	24.80
	✗	34.00	14.20 (-18.54)	54.36	11.65 (-13.15)

Table 5. Experimental results for different initialization states of SAM2 in Grounded-SAM2, where ST means spatiotemporal, and the initialization state refers to the box prompt detected by GroundingDino.

Type	Valid Key Frames	#Test sample	Sa2VA-26B		Sa2VA-4B	
			T_{recall}	$IoU_{gold,pred}$	T_{recall}	$IoU_{gold,pred}$
Short	✓	394	70.93	37.87	70.38	29.44
	✗	6	14.58	0.00	0.00	0.00
Medium	✓	150	54.64	34.15	42.96	22.44
	✗	50	12.67	0.86	18.42	0.78
Long	✓	48	46.24	25.59	32.73	16.42
	✗	52	12.18	1.29	11.23	0.44

Table 6. Experimental results for different initialization states of SAM2 in Sa2VA, where the initialization state means the mask generated for key frames. The “Valid Key Frames” means the first five frames contain targeted frames.

ence state is initialized by the detected bounding box from GroundingDino. Thus, we implement a naive variant that, instead of using the detected bounding box with the highest confidence for grounding, we simply use the first detected bounding box as the prompt, regardless of its confidence

Methods	VideoLLM	SAM-2	Variation	Speed (FPS)
Grounded-SAM2	✗	✓	w. highest conf.	3.17
			w.o highest conf.	7.14
Sa2VA-4B	✓	✓	-	6.47
VideoLISA-3.8B	✓	✗	-	0.42

Table 7. Comparison of grounding speed. For Grounded-SAM2, the “w. highest conf.” means it performs grounding with the highest detection confidence score and “w.o highest conf.” means the naive variant that uses the first detected object as the box prompt.

score. The results are shown in Table 5. As the detection performance drops in the naive variant, the grounding performance also drops, essentially at the average of 11.1%. We also witness the same situation on exocentric video grounding tasks, with the average of 4.65% score reduced. (2) For Sa2VA, the inference state of SAM2 is initialized by the generated mask on the input key frames. As we mentioned before, the Sa2VA model usually takes key frames as the first five frames of the input video. When the referred entity does not show in these frames, the key frames are viewed as invalid. The performance in Table 6 shows that when the model fails to initialize the SAM2 properly, the overall grounding performance will suffer from rapid deduction and drop to nearly 0.00%. These results reveal that *Utilizing SAM2 is beneficial for grounding, yet a proper initialization state is critical to fully unleash its capability.*

The initialization state poses a challenge for modeling design. On one hand, the limited input tokens of the current grounding models restrict them from taking more frames to



Figure 3. Visualization of one example from EgoMask-Short with sampled frames. The language query is “black container bottle on the left side of a wooden table behind computer tablet”. The fine-tuned models perform better than their zero-shot counterparts.

provide initial information for SAM2. On the other hand, the ego videos are usually long, and the target objects have shorter total durations and sparser continuous trajectories. So the sampled frames are more likely to be the background frames. Possible solutions could be: (1) enhancing long video understanding capabilities, and (2) optimizing frame selection to capture the target entities.

Inference Speed. We measure the inference speed of each method in Table 7. The results demonstrate that performing image-level segmentation is inefficient (*e.g.*, only **0.42 FPS**), while SAM2-based methods with video-level segmentation generally achieve higher speed (*e.g.*, at least **3.17 FPS**).

The above analysis suggests that:

Utilizing SAM2 for spatiotemporal grounding offers advantages in both effectiveness and efficiency; however, careful initialization is crucial to fully unleash its pre-trained capability.

4.4. Visualization examples

We visualize a few examples from EgoMask and compare the results from different models (Figure 3 and Figure 4).

Small object size poses a challenge to existing methods.

As Figure 4 shows, the object is very small and usually near the edge of the view, which challenges existing methods. Both Sa2VA-4B and VideoLISA fail to ground the target.

Our training dataset is beneficial. The fine-tuned VideoLISA model can correct grounding errors compared with its zero-shot counterpart. In Figure 3, where the pre-trained model ignores the less salient target near the edge and locates



Figure 4. Visualization of one example from EgoMask-Long with sampled frames. The language query is “small blue cylindrical container near the floor”. The small target poses a challenge to existing methods.

the wrong object (#2 frame), the fine-tuned model identifies the object correctly. Also in Figure 4, the fine-tuned model successfully locates the small object (#3 - #5 frame), which the pre-trained model fails to ground. Similarly, the fine-tuned Sa2VA-4B performs better than its original version and outputs more precise masks (*i.e.*, #2 frame in Figure 3). These results indicate that our training data can facilitate the models in addressing the challenges in egocentric videos.

Utilizing SAM2 provides benefits. From Figure 3, we notice inconsistent segmentation from two VideoLISA models (#5 frame), while other SAM2-based models can correctly segment the target across video frames without hallucinations. This verifies the effectiveness of the SAM2 model as it can utilize information from the context frames through its memory module.

5. Conclusion

To address the absence of benchmarks for egocentric spatiotemporal grounding, we introduce EgoMask, a benchmark with pixel-level annotations, alongside a large-scale training dataset EgoMask-Train. Through extensive experiments, we demonstrate that existing state-of-the-art models struggle with egocentric videos but significantly improve when fine-tuned on our training data, while retaining performance on exocentric benchmarks. These findings, along with our in-depth analysis on ego-exo video gaps, highlight the importance of egocentric training data and suggest that our benchmark can serve as a valuable resource for future research in video grounding and embodied AI.

Acknowledgements

This work was supported by National Key R&D Program of China (Project No. 2022ZD0161200, 2022ZD0161201). This work has also been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 24200223) as well as partially supported by Hong Kong Innovation and Technology Commission Project No. ITS/228/22FP.

References

- [1] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Jia Chen, liulei, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *NeurIPS*, 2024. 1, 2, 3, 4, 5
- [2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multi-modal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 2
- [3] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218. Computer Vision Foundation / IEEE Computer Society, 2018. 4
- [4] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. In *NeurIPS*, 2024. 1
- [5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1979–1986. IEEE Computer Society, 2014. 4
- [6] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 3
- [7] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, pages 12934–12943. IEEE, 2024. 1, 3
- [8] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, pages 12934–12943. IEEE, 2024. 3
- [9] 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. 1, 2, 4, 5
- [10] Kirill Gavriluk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5958–5966, 2018. 2
- [11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrahm Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Kartikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanov, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18973–18990. IEEE, 2022. 1, 2, 3
- [12] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *CVPR*, pages 18330–18339. IEEE, 2024. 1
- [13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3
- [14] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2
- [15] Shuwei Ji and Hongyuan Zhang. ISAT with Segment Anything: An Interactive Semi-Automatic Annotation Tool, 2024. Updated on 2025-02-07. 4
- [16] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, 2022. 1
- [17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798. ACL, 2014. 4
- [18] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV (4)*, pages 123–141. Springer, 2018. 2, 4
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, pages 3992–4003. IEEE, 2023. 3
- [20] 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. 1, 3, 4
- [21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via

- large language model. In *CVPR*, pages 9579–9589. IEEE, 2024. 3
- [22] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *CoRR*, abs/2107.09609, 2021. 1
- [23] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 2
- [24] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296. IEEE, 2024. 3
- [26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5
- [27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019. 5, 6
- [29] Kartikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023. 3
- [30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20. IEEE Computer Society, 2016. 4
- [31] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 2
- [32] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, 2023. 2
- [33] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-07-29. 2, 3
- [34] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *CoRR*, abs/2406.18070, 2024. 2
- [35] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732. IEEE Computer Society, 2016. 2
- [36] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *Int. J. Comput. Vis.*, 132(11):4880–4936, 2024. 2
- [37] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *CoRR*, abs/1704.00675, 2017. 2, 5
- [38] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, pages 5262–5274. IEEE, 2023. 2
- [39] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: parts and attributes of common objects. In *CVPR*, pages 7141–7151. IEEE, 2023. 4
- [40] Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S Khan. Llava++: extending visual capabilities with llama-3 and phi-3 (2024). URL <https://github.com/mbzuai-oryx/LLaVA-pp>, 2024. 5
- [41] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, pages 3505–3506. ACM, 2020. 5
- [42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *CoRR*, abs/2408.00714, 2024. 2, 3
- [43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 5
- [44] 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. 1, 2, 4, 5
- [45] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *NeurIPS*, 2023. 3
- [46] Hao Tang, Kevin J. Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. In *NeurIPS*, 2023. 1, 3, 4
- [47] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8238–8249, 2022. 1

- [48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024. 3
- [49] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023. 2
- [50] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2
- [51] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2
- [52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2
- [53] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 2, 4
- [54] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: reasoning video object segmentation via large language models. In *ECCV (15)*, pages 98–115. Springer, 2024. 4
- [55] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(4):36:1–36:47, 2020. 2
- [56] Hanrong Ye, Haotian Zhang, Erik A. Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. Mm-ego: Towards building egocentric multimodal llms. *CoRR*, abs/2410.07177, 2024. 1
- [57] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3
- [58] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2
- [59] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying SAM2 with llava for dense grounded understanding of images and videos. *CoRR*, abs/2501.04001, 2025. 1, 2, 3, 4, 5
- [60] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 3
- [61] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3
- [62] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, pages 10665–10674. Computer Vision Foundation / IEEE, 2020. 1
- [63] Yiwu Zhong, Zi-Yuan Hu, Michael Lyu, and Liwei Wang. Beyond embeddings: The promise of visual table in visual reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6876–6911. Association for Computational Linguistics, 2024. 4
- [64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 5122–5130. IEEE Computer Society, 2017. 4