ViCaS: A Dataset for Combining Holistic and Pixel-level Video Understanding using Captions with Grounded Segmentation

Ali Athar Xueqing Deng Liang-Chieh Chen
ByteDance Inc.
https://ali2500.github.io/vicas-project/
{ali.athar}@bytedance.com

Abstract

Recent advances in multimodal large language models (MLLMs) have expanded research in video understanding, primarily focusing on high-level tasks such as video captioning and question-answering. Meanwhile, a smaller body of work addresses dense, pixel-precise segmentation tasks, which typically involve category-guided or referral-based object segmentation. Although both directions are essential for developing models with human-level video comprehension, they have largely evolved separately, with distinct benchmarks and architectures. This paper aims to unify these efforts by introducing ViCaS, a new dataset containing thousands of challenging videos, each annotated with detailed, human-written captions and temporally consistent, pixel-accurate masks for multiple objects with phrase grounding. Our benchmark evaluates models on both holistic/high-level understanding and language-guided, pixel-precise segmentation. We also present carefully validated evaluation measures and propose an effective model architecture that can tackle our benchmark.

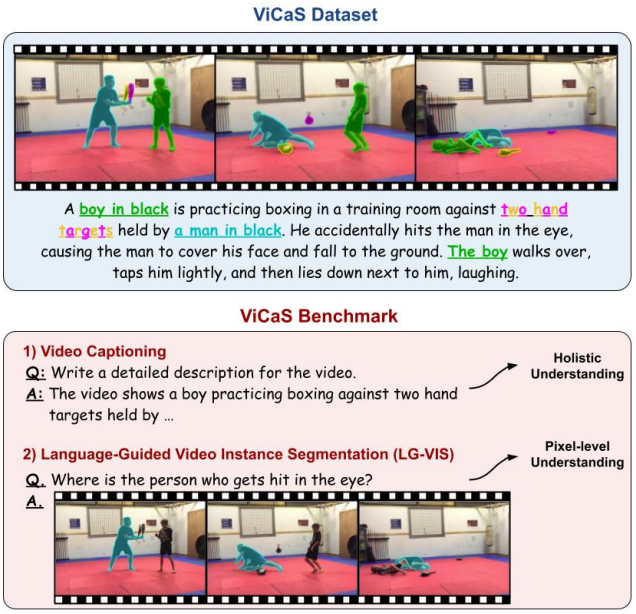


Figure 1. **ViCaS Dataset/Benchmark.** Our dataset contains detailed video captions with phrase-level grounding for accurate object segmentation masks. The benchmark comprises two tasks to evaluate holistic and pixel-level video understanding, respectively.

1. Introduction

The emergence of open-source large lauguage models (LLMs) [9, 32, 96], coupled with progress in aligning vision and language feature spaces [55, 58, 83], has enabled significant research into vision-language models that can jointly reason over visual and language inputs. Earlier advances in image-language models [8, 56, 66] have spurred development of multimodal LLMs capable of reasoning over other modalities such as videos [63] and 3D point-clouds [128].

In the field of video-language models, research has mainly focused on holistic video understanding tasks such as video captioning and question-answering [63, 67, 72, 73, 113, 127], with a particular emphasis on generating temporally dense text output for long video sequences [18, 62, 92, 118, 132]. By contrast, comparatively less attention has been given to predicting spatially dense outputs — such as bounding boxes or segmentation masks — that can localize key objects and actors in videos based on text prompts, even though this has important practical applications, *e.g.*, autonomous robots and video editing. Although recent language-driven benchmarks [29, 47, 89, 117] in the video segmentation community have made progress in this area, these datasets are largely object-centric and lack evaluation of high-level, holistic video understanding.

In this paper, we seek to bridge the gap between benchmarks that evaluate holistic video understanding and those

focused on pixel-level localization. We introduce ViCaS, a new video dataset and benchmark containing thousands of videos with detailed, human-annotated captions. In these captions, words and phrases corresponding to key objects are grounded in human-drawn, pixel-precise segmentation masks that span the entire video duration. To the best of our knowledge, this is the first video dataset to offer human-labeled annotations of this kind.

Our benchmark comprises two tasks: (1) Video Captioning, which requires describing the video events and objects in detail, and (2) our newly proposed Language-Guided Video Instance Segmentation (LG-VIS) task, which requires predicting temporally-consistent segmentation masks for multiple objects based on a text prompt. An example of our dataset annotations and tasks is given in Fig. 1. To effectively evaluate Video Captioning, we conduct a comprehensive user study to validate the evaluation measures used for open-ended text similarity by comparing several existing approaches. Finally, we introduce Video-LLaVA-Seg, an end-to-end architecture designed to effectively tackle our benchmark by integrating insights from recent vision-language models [63, 67] and prompt-based video segmentation approaches [53, 86]. In summary, our contributions are as follows:

- We introduce a large, human-annotated video dataset with detailed text captions with phrase-level grounding for objects, accompanied by pixel-precise segmentation masks. This dataset enables the evaluation of both holistic and pixel-level video understanding.
- We propose accurate, reproducible evaluation measures for open-ended text similarity which are verified by a comprehensive user study.
- We present Video-LLaVA-Seg, an effective, end-to-end trained architecture that can tackle our benchmark.

2. Related Work

Although video understanding is a well-researched topic in computer vision, researchers have traditionally approached holistic and pixel-level video understanding as separate streams, each with its own datasets and benchmarks. Below, we review related work in these areas.

2.1. Holistic Video Understanding

Video Classification. This is one of the earliest tasks in video understanding, popularized by activity recognition datasets [52, 93, 125], and later expanded by larger datasets [37, 38, 91], *e.g.*, Kinetics [46]. Early deep learning approaches employed 3D CNNs [39, 45, 97–99] to capture spatio-temporal information. With the popularization of transformers [100], attention-based architectures [4, 68, 79] emerged, offering improved performance. Although these datasets and architectures have greatly advanced video understanding, they are limited from a language perspective, typically assigning a single, predefined label to each video.

Video Captioning and Question-Answering. Alongside classification, language-oriented tasks such as video captioning and question-answering (Q/A) have gained research attention. Early datasets like MSVD [17], MSR-VTT [111], and TGIF-QA [61] laid the groundwork for video captioning and were later adapted as Q/A benchmarks [110]. Initial approaches [31, 112] combined CNNs for visual reasoning with RNNs for text generation. As transformers gained popularity, architectures like VideoBERT [94] and UniVL [71] were among the first to unify vision and language by learning shared representations for both modalities.

Multimodal Large Language Models (MLLMs). The recent popularization of LLMs [9, 24, 32, 96] has enabled research on multimodal models which extend LLMs to process visual inputs such as images [14, 19, 28, 66, 124] and videos [18, 62, 63, 72, 73, 92]. This research has been supported by large-scale video captioning datasets [11, 20, 116] and multi-task video understanding benchmarks [25, 57].

Overall, the datasets and models discussed above focus primarily on high-level, holistic video understanding and do not address finegrained, pixel-level localization.

2.2. Pixel-level Video Understanding

Object Tracking and Segmentation. Object localization and tracking is a deeply studied problem, even prior to the deep learning era [2, 78]. This research has been propelled by several datasets, with early efforts focused on bounding-box-level object tracking [26, 27, 34, 36, 44, 51, 122]. As model architectures advanced, benchmarks for pixel-precise video segmentation emerged, covering tasks such as object segmentation based on predefined categories [7, 48, 74, 75, 81, 82, 102, 108, 119] or segmenting specific objects given their first-frame ground-truth masks [30, 80, 115] or points [7, 43, 133]. Popular approaches have evolved from tracking-by-detection methods [16, 69, 95, 102] to end-to-end trainable transformer-based [6, 40, 42, 49, 90, 107, 109, 130], and object-level attention-based architectures [5, 21–23, 76]. Despite substantial progress, these works focus solely on tracking and segmentation, without addressing holistic video understanding.

Language-Guided Segmentation. Recent progress in vision-language models has motivated multiple language-guided video segmentation datasets [103]. These include "open-vocabulary" datasets [104, 105] that cover a large set of object classes, and referral-based video segmentation benchmarks [29, 47, 89], which require segmenting objects based on text prompts. Although more language-oriented, these datasets are still object-centric and do not involve predicting a high-level description of the entire video.

Some approaches [29, 41, 123] for these tasks use text and image backbones to encode the text prompt and video

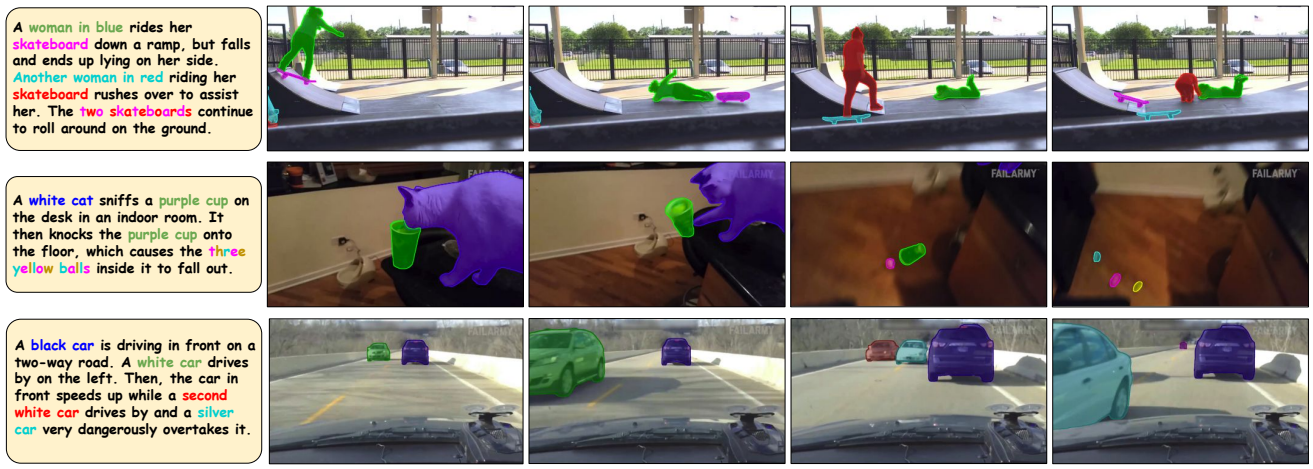


Figure 2. **ViCaS Examples.** Our dataset showcases diverse scenes with a variety of objects and video events, along with detailed captions. Phrases referring to multiple objects are written with multiple colors, *e.g.*, "three yellow balls" in row 2 references three different objects.

frames, followed by a transformer decoder and segmentation head to generate masks. Meanwhile, recent LLM-based approaches [10, 117] use a special `SEG` vocabulary token in conjunction with a segmentation network to predict the target masks. These approaches typically utilize several image-level and video segmentation datasets for training, and utilize other models [13, 62] for dynamic frame selection or temporal mask propagation.

In contrast to these two categories of works, we propose a unified dataset that provides the annotations as well as benchmark tasks to evaluate both holistic/high-level, and pixel-level video understanding. Additionally, we propose a baseline architecture that is end-to-end trainable, and can be effectively trained for segmentation using only our dataset, thereby making it easy to setup and extend.

3. ViCaS Dataset

Our dataset is designed to evaluate both holistic as well as pixel-level video understanding. To this end, we annotate 20,416 videos with detailed captions in which words/phrases referencing salient objects are grounded with temporally consistent segmentation masks. All annotations are done by professional human annotators.

3.1. Video Source

For our objective, it is crucial to annotate videos which contain meaningful events that can only be explained through effective temporal reasoning. At the same time, to effectively evaluate pixel-level understanding, the videos should feature challenging scenes with multiple moving objects.

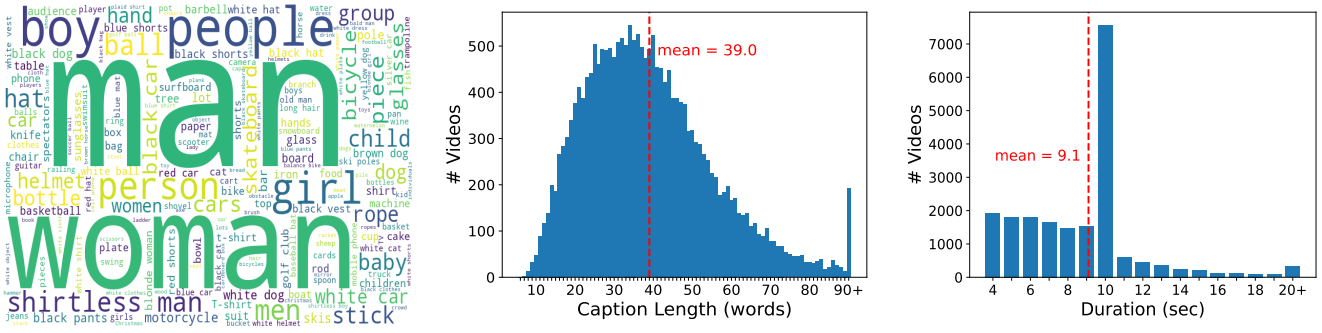
To meet these requirements, we annotate videos from three sources: (1) Oops dataset [33], a collection of 'fail videos' from the internet, (2) Unidentified Video Object (UVO) [105], which is a video segmentation dataset, and (3) Kinetics-700 [46], a popular human action recognition dataset. These sources are suitable for our dataset since they contain in-the-wild videos with diverse objects and backgrounds. Moreover, they contain videos with multiple objects undergoing appearance and shape changes, and fast motion. This combination of attributes makes the videos challenging both in captioning and in segmentation.

We annotate 20,416 videos with durations ranging from 4 to 30 seconds (distribution illustrated in Fig. 3c). Note that we use only the raw videos from these Oops dataset and disregard their annotations. More details on video selection are given in supplementary.

3.2. Annotation Process

The annotation process consists of two main steps: first, a detailed caption is written (Step 1a) in which objects selected for segmentation masks are identified and marked using a specific syntax (Step 1b). In the second step, a different annotator reviews the video and its text, and draws the corresponding segmentation masks (Step 2). The process is illustrated in Fig. 4. Further details are given below.

Step 1a: Text Captions. Professional annotators fluent in English were tasked with writing a detailed caption covering the events in each video. To ensure consistency in the level of detail and writing style, they received comprehensive guidelines with several examples. Annotators were advised to think as if they have ∼30 seconds to convey the video content in as much detail as possible over an audio call. They were instructed to describe the 'fail' event [33], the actions and movements of objects, as well as the background scene and any additional, interesting elements. Since our dataset is video-centric, annotators were explicitly guided to include temporal information, with the instruction: "*Focus on details that require watching the en-*



(a) (b) (c)

Figure 3. **ViCaS Dataset.** (a): Word cloud for nouns in grounding phrases. (b), (c): Histograms for caption length and video duration, respectively. Since these distribution are long-tailed, the right-most bar labeled '$n+$' captures all values $\geq n$ to prevent visual distortion.

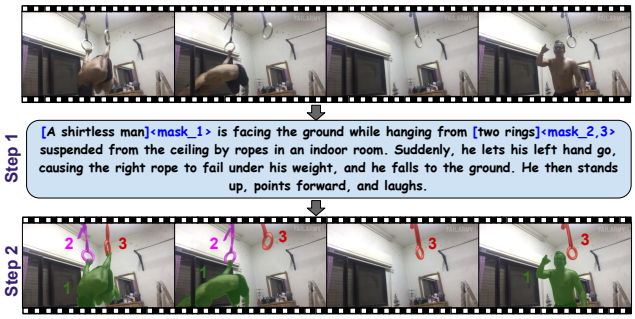


Figure 4. **Annotation Process.** Our annotation process consists of two steps. Step 1: A detailed video caption is written by human annotators. Salient objects are identified and marked using a special syntax (highlighted in blue). Step 2: Segmentation masks are drawn for the salient objects throughout the video.

tire video and cannot be inferred from a few frames."

Step 1b: Phrase Grounding. Alongside writing the text caption, annotators were instructed to identify salient objects/actors in the video, and mark the corresponding phrases or words using a specific syntax, as illustrated in Fig. 4. This syntax is programmatically parsable to extract object IDs that can be removed to produce a standard, human-readable caption. To streamline the mask-drawing process, annotators were advised to mark no more than 10 objects per video.

Step 2: Segmentation Masks. In the second step, a different annotator reviews the video and its caption, and uses a polygon tool to draw segmentation masks for each marked object. The annotators were instructed to read the text carefully and watch the video multiple times to ensure that correct masks were drawn for the each grounding phrase. To balance annotation accuracy and cost, we annotate masks at 1 frame-per-second (fps) and then use an off-the-shelf SAM2 [86] model to propagate these masks to adjacent frames, producing temporally dense mask annotations at 30 fps. A similar approach has been successfully applied in other video segmentation datasets [7, 105].

Annotation Statistics and Quality Control. For Step 1, we employed 24 annotators and 4 quality inspectors who reviewed/corrected the text captions and language grounding. On average, each video required 9 minutes for annotation and quality checking. For Step 2, we hired 40 annotators and 5 quality inspectors, with each video frame requiring an average of 7 minutes to annotate.

3.3. Benchmark Design and Evaluation Tasks

Leveraging our annotations, we create a benchmark to evaluate both high-level video understanding and pixel-precise segmentation. Consequently, it includes two tasks:

1) Video Captioning. In this task, the model is expected to produce an open-ended text summary that explains the events in the video, including descriptions of salient objects and background elements.

2) Language-Guided Video Instance Segmentation (LG-VIS). For this task, the model receives a text prompt describing a specific set of objects, *e.g.*, '*Where is the person who walks from left to right?*', and is expected to predict segmentation masks for the corresponding objects. For cases where multiple objects are referenced, *e.g.*, '*Where is the group of kittens that are playing?*', a separate mask is required for each object. This is in contrast to Referral Video Object Segmentation [29, 47, 89], where only a single, binary mask covering all target objects is required.

From a vision-language perspective, both tasks are forms of Video Question-Answering. For Video Captioning, the question asks for a video description, with the answer provided as open-ended text. For LG-VIS, the question is a '*Where is?*' query regarding specific objects, and the answer is a video-length segmentation mask for each object.

Converting Grounded Captions to LG-VIS Prompts. As outlined in Sec. 3.2, our annotations contain phrase ground-

Type	Dataset Name	Year	Human Ann.	Videos	Duration (sec)	Caption (words)	Object Classes	Object Tracks	Object Masks
Captioning and Q/A	MSVD [17]	2011	✓	2K	9.7	8.7	-	-	-
	MSR-VTT [110]	2016	✓	10K	15.0	9.3	-	-	-
	ActivityNet [125]	2017	✓	20K / 100K	36.0	13.5	-	-	-
	YouCook2 [131]	2018	✓	2K / 14K	19.6	8.8	-	-	-
	VATEX [106]	2019	✓	41K	∼10	15.2	-	-	-
	DiDeMo [3]	2017	✓	10K / 27K	6.9	8.0	-	-	-
	WebVid10M [11]	2021	✗	10M	17.2	14.2	-	-	-
	Panda70M [20]	2024	✗	3.8M / 70.8M	8.5	13.2	-	-	-
Referral VOS	A2D Sentence [35]	2018	✓	3,782	4.9	-	-	4,825	58K
	DAVIS-17 RVOS [47]	2019	✓	90	2.4	-	78	205	13.5K
	Refer-Youtube-VOS [89]	2020	✓	3,978	5.5	-	94	7,451	131K
	MeViS [29]	2023	✓	2,006	13.2	-	-	8,171	443K
	ReVOS [117]	2024	✓	1,042	8.6	-	-	9,084	273K
Class-guided Segmentation	YouTube-VIS [119]	2019	✓	2,883	5.6	-	40	4,883	131K
	OVIS [81]	2021	✓	901	12.8	-	25	5,223	296K
	UVO [105]	2021	✓	11,228	9.6	-	80	104,898	593K
	BURST [7]	2023	✓	2,914	35.7	-	482	16,089	600K
	LV-VIS [104]	2023	✓	4,828	4.6	-	1,196	25,588	544K
	ViCaS (Ours)	2025	✓	20,416	9.1	39.0	11,492	65,588	18.3M

Table 1. **Dataset Comparison.** In the 'Videos' column, entries formatted as 'X / Y' indicate a total of X videos with Y labeled sub-clips, with the average sub-clip duration being reported.

ing for labeled objects. We use GPT-4 [1] to convert these grounded captions into '*Where is?*' style questions, which serve as text prompts for the LG-VIS task. Further details about this process are provided in supplementary.

Dataset Splits. We split the 20,416 videos into train, validation, and test sets with 14,516 videos for training, and 2,950 each for validation and testing. Statistics for each set are given in supplementary.

3.4. Comparison with Related Datasets

Table 1 provides a quantitative comparison between ViCaS and existing datasets, showing that ViCaS uniquely combines text captions and pixel-precise segmentation masks.

Captioning and Q/A. For text descriptions, ViCaS prioritizes quality over quantity, offering detailed, high-quality, human-written captions with phrase grounding for objects. Our captions average 39 words (distribution illustrated in Fig. 3b), significantly longer than those in other datasets. While ViCaS annotates more videos than early datasets like MSVD [17] and MSR-VTT [111], later datasets with temporally dense captions, such DiDeMo [3], and YouCook2 [131], include more labeled video segments. Recently, large-scale datasets such as WebVid10M [11] and Panda70M [20] have emerged, containing millions of *automatically* captioned videos sourced from the internet. While useful for (pre-)training large models, these datasets offer shorter, less detailed descriptions and have neither segmentation masks nor phrase grounding.

Referral VOS. Compared to Referral Video Object Segmentation (VOS) datasets, ViCaS stands out by providing holistic, video-level captions in addition to object-level language expressions. With 20,416 annotated videos, ViCaS is significantly larger than the second-largest Refer-YouTubeVOS [89], which contains only 3,978 videos. Although MeViS [29] features longer videos, averaging 13.2s compared to our 9.1s, ViCaS offers significantly more annotated objects (65,588) compared to MeViS (8,171).

Class-Guided Segmentation. Compared to these datasets, ViCaS provides more nuanced language labels, going beyond single-word category labels typical in Video Instance Segmentation (VIS) [81, 119] datasets. Using NLTK [15], we found that our object-grounding phrases contains 11,492 unique nouns/noun-phrases (illustrated in Fig. 3a), which is significantly higher than the 1,196 object categories in LV-VIS [104]. Furthermore, ViCaS contains more videos than the others, and also contains more labeled object tracks than all other datasets except UVO [105]. Lastly, ViCaS stands out by providing detailed text captions in addition to the segmentation masks.

4. Evaluation Measures

As outlined in Sec. 3.3, our benchmark comprises two tasks: Video Captioning and LG-VIS. Evaluating the video caption quality is challenging since it requires computing

	Evaluation Method	Part 1			Part 2
		r	r_{max}	r_{avg}	Δ_{PN}
LLM	GPT4 [1]	58.9	62.1	**67.3**	**37.6**
	Llama3-70B [32]	**59.3**	**64.3**	65.6	29.2
	Llama3-8B [32]	31.3	-	45.9	2.4
Embedding Sim.	AnglE Lm2-7B [60]	50.5	50.0	50.3	3.3
	STSB Roberta-L [87]	49.5	48.3	49.1	7.5
	AnglE Lm2-13B [60]	45.6	47.4	46.4	4.5
	AnglE UAE-L [60]	44.5	42.7	40.5	0.9
	MPNet-B [87]	41.5	35.4	34.7	0.4
	BERTScore [129]	22.1	22.9	26.2	-2.4
Word mat.	METEOR [12]	22.3	23.9	22.8	-16.4
	BLEU4 [77]	19.9	15.9	15.4	-27.6
	ROUGE-L [64]	18.1	18.2	13.2	-23.7
	CIDEr [101]	2.8	0.1	-0.8	-21.2

Table 2. **Caption Scoring User Study.** Results for Part 1 (Correlation with Human Scores) and Part 2 (Robustness to Hard Positives/Negatives). r, r_{max} and r_{avg} denote Pearson correlation coefficients. Δ_{PN} is the absolute difference in predicted scores.

the similarity between two open-ended text passages—the human-written ground truth and the model's prediction. To address this, we first conduct a comprehensive user study for scoring open-ended text similarity in Sec. 4.1, before introducing the selected evaluation measures in Sec. 4.2.

4.1. User Study for Video Caption Scoring

To decide evaluation measures for video captioning, we carried out a two-part user study comparing several scoring methods which are categorized into three groups: (1) classical text similarity metrics, *e.g.*, METEOR [12] and BLEU4 [77], which rely on word/phrase matching; (2) embedding similarity-based measures, *e.g.*, BERTScore [129] and AnglE [60], which summarize the ground-truth and prediction as an embedding using a language model, followed by computing the cosine similarity between the two; and (3) recent LLM-based scoring methods [73], where an LLM is provided with the ground-truth and predicted captions along with instructions for assessing similarity, and it outputs a similarity score directly as text.

Part 1: Correlation with Human Scores. In this part, 20 participants rated the accuracy of various model-predicted video captions on a scale of 0 to 10, with a minimum interval of 0.5. Similar to the evaluation methods, participants could only read the ground-truth caption and could not watch the video itself. They represented 10 different nationalities and had either native, or professional English proficiency. A total of 131 video captions were evaluated, with each sample scored by two different individuals.

To filter samples with diverging human scores, we calculated the difference Δ between the two assigned scores for each sample, discarding any with $\Delta > \alpha$ (*i.e.*, outlier removal). We used multiple thresholds $\{\alpha_i\}$ ranging from 0.5 to 2.5 in steps of 0.5. For each α_i and each method, we calculated the Pearson correlation coefficient r_i between averaged human scores and the predicted scores. The final correlation score r for each method was computed as the mean of the set $\{r_i\}$, and is reported in Table 2.

Results indicate that LLM-based scoring methods performed best, with Llama3-70B [32] achieving $r = 59.3\%$, closely followed by GPT-4 [1] with 58.9%. Model size significantly impacted performance, as evidenced by Llama3-8B's lower $r = 31.3\%$. Among embedding-based methods, AnglE Llama2-7B [60] performed best at 50.5%, while classical word/phrase matching metrics lagged behind, with METEOR [12] achieving only $r = 22.3\%$.

Pooling Scores Across Multiple Ground Truths. A common strategy to enhance robustness in text similarity evaluation is to compare predictions with multiple, synonymous ground-truths. Using GPT-4 [1], we generated four rephrasings for each human-written caption, resulting in five ground-truth variants per video. For each video, we calculated the similarity between the prediction and all ground-truths, followed by applying either average or maximum pooling to obtain the final score. We report the resulting correlations as r_{avg} and r_{max} in Table 2. As shown in the table, LLM-based scoring methods benefited most from average pooling, with Llama3-70B improving from 59.3% to 65.6% and GPT-4 from 58.9% to 67.3%. Embedding-based methods generally saw minimal improvements for both types of pooling, with some, like MPNet-B and AnglE-UAE-L, experiencing slight declines. Word/phrase matching metrics similarly showed no substantial improvement.

Part 2: Robustness to Hard Positives/Negatives. Here, participants watched videos from our dataset and read the ground-truth captions (GT), and were then asked to write a hard-positive (HP) and a hard-negative (HN) caption for each video. For the HP, they rephrased the caption using different wording and style while retaining all original information. For the HN, they altered critical details while retaining most wording and sentence structure. Each of the 10 participants processed 5 videos, yielding 50 samples.

We computed the similarity between GT-HP and GT-HN pairs for each video using different methods. The scores were scaled to $[0, 100]$[1] and the difference Δ_{PN} between the GT-HP and GT-HN pairs was calculated, and reported in Table 2. Ideally, Δ_{PN} should be high, indicating that methods assign higher similarity to GT-HP pairs and lower similarity to GT-HN pairs. GPT-4 achieved the highest Δ_{PN} (37.6), followed by Llama3-70B (29.2). Most embedding-based models showed low Δ_{PN}, indicating similar scores for GT-HP and GT-HN pairs, while word/phrase matching metrics performed worst, with negative Δ_{PN} values suggesting higher scores for GT-HN pairs than GT-HP pairs.

[1]Except CIDEr, which is not a normalized metric

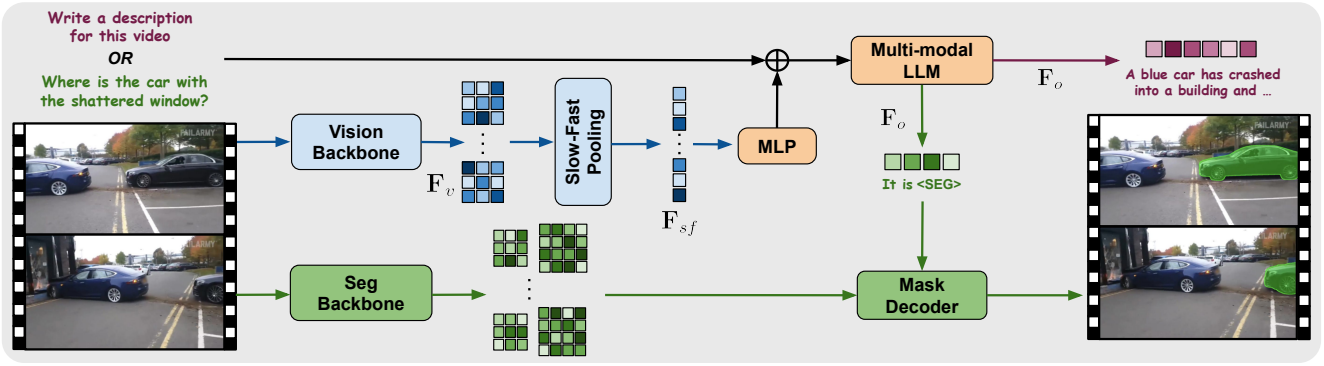


Figure 5. **Video-LLaVA-Seg Architecture.** The vision backbone and projection MLP encode the input video frames into a set of features $\mathcal{F}_{sf}$ which are concatenated with the text embeddings and input to the LLM. For Video Captioning, the output $\mathbf{F}_o$ is decoded into text. For LG-VIS, the <SEG> token in $\mathbf{F}_o$ is applied to the mask decoder along with multi-scale features from the segmentation backbone.

4.2. Selected Evaluation Measures

Video Captioning. Although GPT-4 achieved the highest r_{avg} and Δ_{PN} in our study, it is proprietary and subject to updates, which complicates reproducibility. Moreover, its runtime can be slow due to token limits, and it incurs monetary costs which effectively makes the benchmark 'pay-to-evaluate'. To support open-source, reproducible research, we accept a slight performance trade-off and select Llama3-70B as the evaluation measure for our video captioning task. Since r_{avg} is noticeably higher than r for Llama3-70B (Table 2), we generate 4 reworded variants of each ground-truth caption, and compute the final accuracy by averaging over the prediction-GT scores for each video. This score lies in $[0, 5]$ and is abbreviated as CA (Caption Accuracy).

Language-Guided Video Instance Segmentation (LG-VIS). Unlike text similarity, segmentation mask accuracy is a well-studied problem with several feasible metrics [59, 70, 119]. We select Track mean Average Precision (mAP) as our primary metric since it is widely used by video segmentation benchmarks [81, 105, 119] and accommodates multi-object prediction.

5. Video-LLaVA-Seg Model

We propose an effective baseline called Video-LLaVA-Seg that can tackle both Video Captioning and LG-VIS with a single, end-to-end trained model.

5.1. Architecture

Video-LLaVA-Seg extends the popular LLaVA [65–67] architecture with segmentation ability. It is illustrated in Fig. 5 and comprises 3 main parts: (1) a multi-modal LLM, (2), a vision backbone, and (3) a segmentation network.

Vision Features. We input T uniformly sampled frames from the video to the vision backbone, yielding a set of video features $\mathbf{F}_v \in \mathbb{R}^{T\times H\times W\times C}$. Next, we adopt the workflow proposed by Xu *et al*. [114] and split $\mathbf{F}_v$ into two sets of features, namely 'slow' features $\mathbf{F}_v^s \in \mathbb{R}^{T_s\times H\times W\times C}$ and 'fast features' $\mathbf{F}_v^f \in \mathbb{R}^{T\times H_f\times W_f\times C}$. Fast features are generated by aggressively downsampling the spatial dimensions in $\mathbf{F}_v$ using adaptive average pooling, *i.e.* $H_f << H$ and $W_f << W$. Meanwhile, slow features are obtained by uniformly sampling $T_s < T$ frame features from $\mathbf{F}_v$. These two sets of features are then flattened and concatenated to yield the set of 'slow-fast' video features $\mathbf{F}_{sf} \in \mathbb{R}^{N_v\times C}$ where $N_v = (T \times H_f \times W_f) + (T_s \times H \times W)$. The idea is to retain finegrained information along both temporal and spatial dimensions: the slow features are temporally dense but spatially condensed, whereas the fast features are spatially dense but temporally sparse. This reduces the number of video tokens N_v input to the multimodal LLM.

Multi-modal LLM. The slow-fast video features are then applied to a projection MLP, concatenated with the text prompt embeddings, and then input to the LLM, which outputs a set of embeddings $\mathbf{F}_o$. For the Video Captioning task, $\mathbf{F}_o$ is decoded into text. For LG-VIS, the model is trained to output a special <SEG> token which is used by the segmentation network to regress the mask. Note that when the text prompt references multiple objects, the model generates a separate <SEG> token for each target object.

Segmentation Network. The video frames are encoded at high resolution by a segmentation backbone to yield multi-scale features which, together with <SEG> tokens, are input to a mask decoder to obtain the final segmentation masks. The mask decoder architecture is borrowed from SAM2 [86] and consists of multiple layers of bi-directional cross-attention between the <SEG> tokens and the segmentation backbone features. Finally, the dot-product between the two is computed to obtain the mask logits. We refer to Ravi *et al*. [86] for more details. Our approach for mask prediction using <SEG> tokens generated by an MLLM is inspired from recent image segmentation [53, 85, 120, 126] and video segmentation works [10, 117].

Comparison to Existing Methods. Video-LLaVA-Seg is a compact baseline model that is easy to setup and extend, unlike existing LLM-based video segmentation approaches [10, 117]: VISA [117] involves inference with a pretrained Llama-Vid [62] to select key frames for their own model, and both VISA and VideoLISA [10] rely on pretrained models [13, 21] to perform/improve temporal mask propagation. Moreover, these approaches tackle only segmentation tasks whereas our model is trained to tackle both captioning and segmentation.

5.2. Implementation Details

We employ an AM-RADIO [84] ViT-H [50] model as the vision backbone and Llama3-8B [32] as the LLM. The segmentation backbone is a Hiera-Small [88] network which, together with the mask decoder, is initialized with weights from SAM2 [86]. We sample $T = 32$ frames and use $T_s = 8$ slow frames. We train our model in three stages: (1) a vision-language alignment stage to optimize the projection MLP, (2) a finetuning step to optimize the MLLM, projection MLP and vision backbone for video captioning, and finally (3) the entire model with the segmentation network is fintuned for both tasks. For the first two stages, we utilize a total of 3.5M video caption samples from WebVid10M [11] and Panda70M [20]. In stage 3, we finetune the model on the ViCaS training set for both tasks. Training takes a total of $\sim$ 4 days on 32 A100 GPUs. Further implementation details can be found in supplementary.

5.3. Ablation Studies

We report ablations in Table 3 and discuss them below.

#	Vis Enc.	T_s	SF-Pool.	Cap.	Seg.	CA	mAP
1	AM-RAD.	1	✗	✓	✓	2.1	15.9
2	AM-RAD.	2	✗	✓	✓	2.4	17.0
3	AM-RAD.	4	✗	✓	✓	2.8	18.8
4	AM-RAD.	8	✓	✓	✗	3.0	-
5	AM-RAD.	8	✓	✗	✓	-	18.5
6	CLIP	8	✗	✓	✓	2.9	19.9
	AM-RAD.	8	✓	✓	✓	3.0	20.5

Table 3. Ablation experiments on the ViCaS validation set. CA: Caption Accuracy. SF-Pool: Slow-Fast Pooling.

Input Frames. In rows 1-3, we train our model using $T_s = \{1, 2, 4\}$ frames instead of $Ts = 8$ in the final setting. We see that using fewer frames strongly reduces performance, especially for captioning, showing that our dataset/benchmark is video-centric and requires temporal context to effectively tackle.

Task Synergy. For rows 4 and 5, we train the model for either one of our benchmark tasks. We see that the model trained only for LG-VIS achieves 18.5 mAP which is much worse than the 20.5 achieved by the multi-task model. Meanwhile, the Caption Accuracy remains unchanged at 3.0. This indicates that pixel-level segmentation benefits greatly from holistic video understanding.

LLaVA Baseline. Row 6 shows results for a LLaVA-NeXT baseline trained with our recipe which uses a CLIP vision encoder and no Slow-Fast Pooling. This achieves 19.9 mAP and 2.9 CA, both of which are worse than our final setting.

5.4. Benchmark Results

Model	CA	mAP	AP_{50}	AP_{75}	AP_{90}
LLaVA-OV [54] (ZS)	2.9	-	-	-	-
MiniCPM-o 2.6 (ZS)	3.0	-	-	-	-
LMPM [29]	-	8.4	17.1	7.7	1.8
DsHmp [41]	-	13.3	27.6	11.6	2.9
VideoLISA [10]	-	10.7	23.8	8.5	1.8
Video-LLaVA-Seg	3.0	20.5	38.6	19.0	5.1

Table 4. Benchmark Results on our validation set. Refer to supplementary for test set results. CA: Caption Accuracy

Table 4 compares Video-LLaVA-Seg to other methods on the validation set. For video captioning, we evaluate off-the-shelf LLaVA-OneVision [54] and MiniCPM-o[121] models which achieve 2.9 and 3.0 CA scores, respectively. This is the same as Video-LLaVA-Seg, but these models are trained on significantly more data and cannot tackle LG-VIS. For LG-VIS, we evaluate multiple existing Referral-VOS approaches. LMPM is an earlier transformer-based approach which only achieves 8.4 mAP. DsHmp [41] performs better with 13.3 mAP since the architecture decouples static and motion cues to improve video segmentation performance. Finally, VideoLISA [10] is a recent LLM-based approach which only achieves 10.7 mAP when finetuned on ViCaS. By contrast, Video-LLaVA-Seg achieves 20.5 mAP while tackling both captioning and LG-VIS. Furthermore, unlike existing Referral-VOS approaches, it can predict multiple segmentation masks for a single prompt.

6. Conclusion

We introduce ViCaS, a first-of-its-kind, human-annotated dataset that provides detailed captions for videos along with phrase-grounded segmentation masks for salient objects. Our associated benchmark comprises two tasks: (1) Video Captioning, which evaluates high-level, holistic understanding, and (2) Language-Guided Video Instance Segmentation (LG-VIS), which evaluates pixel-level understanding. We propose evaluation measures for open-ended caption accuracy which are based on open-source models, and are experimentally verified through a comprehensive user study. Furthermore, we propose Video-LLaVA-Seg, a compact baseline which effectively tackles both tasks.

Acknowledgements. We thank the participants of the user study and all the annotation personnel for their effort.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *Arxiv*, 2023. 5, 6

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008. 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 5

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2

[5] Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Hodor: High-level object descriptors for object re-segmentation in video learned from static images. In *CVPR*, 2022. 2

[6] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified architecture for target-based video segmentation. In *CVPR*, 2023. 2

[7] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 2, 4, 5

[8] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *Arxiv*, 2023. 1

[9] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *Arxiv*, 2023. 1, 2

[10] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 2024. 3, 7, 8

[11] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2, 5, 8

[12] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005. 6

[13] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. *Arxiv*, 2023. 3, 8

[14] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *Arxiv*, 2024. 2

[15] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009. 5

[16] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *CVPR*, 2020. 2

[17] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 2, 5

[18] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024. 1, 2

[19] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *CVPR*, 2024. 2

[20] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 2, 5, 8

[21] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 2, 8

[22] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 2021.

[23] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *CVPR*, 2024. 2

[24] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[25] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *Arxiv*, 2024. 2

[26] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 2

[27] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 2021. 2

[28] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *CVPR*, 2024. 2

[29] 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*, 2023. 1, 2, 4, 5, 8

[30] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2

[31] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[32] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *Arxiv*, 2024. 1, 2, 6, 8

[33] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, 2020. 3

[34] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 2

[35] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 5

[36] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[37] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2

[38] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[39] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 2

[40] Ju He, Qihang Yu, Inkyu Shin, Xueqing Deng, Alan Yuille, Xiaohui Shen, and Liang-Chieh Chen. Maxtron: Mask transformer with trajectory attention for video panoptic segmentation. *Arxiv*, 2023. 2

[41] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024. 2, 8

[42] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *NeurIPS*, 2022. 2

[43] Namdar Homayounfar, Justin Liang, Wei-Chiu Ma, and Raquel Urtasun. Videoclick: Video object segmentation with a single click. *Arxiv*, 2021. 2

[44] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 2021. 2

[45] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[46] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *Arxiv*, 2017. 2, 3

[47] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ECCV*, 2019. 1, 2, 4, 5

[48] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 2

[49] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *CVPR*, 2022. 2

[50] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 8

[51] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE TPAMI*, 2016. 2

[52] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2

[53] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 2, 7

[54] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Arxiv*, 2024. 8

[55] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[56] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[57] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 2

[58] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *Arxiv*, 2019. 1

[59] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*, 2022. 7

[60] Xianming Li and Jing Li. Angle-optimized text embeddings. *Arxiv*, 2023. 6

[61] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, 2016. 2

[62] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024. 1, 2, 3, 8

[63] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024. 1, 2

[64] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004. 6

[65] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *Arxiv*, 2023. 7

[66] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 1, 2

[67] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2, 7

[68] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 2

[69] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 2

[70] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 2020. 7

[71] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *Arxiv*, 2020. 2

[72] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *Arxiv*, 2023. 1, 2

[73] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 1, 2, 6

[74] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022. 2

[75] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 2

[76] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2

[77] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[78] Nikos Paragios and Rachid Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE TPAMI*, 2000. 2

[79] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. In *CVPR*, 2023. 2

[80] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *Arxiv*, 2017. 2

[81] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 2022. 2, 5, 7

[82] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, 2021. 2

[83] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[84] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024. 8

[85] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 7

[86] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *Arxiv*, 2024. 2, 4, 7, 8

[87] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019. 6

[88] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023. 8

[89] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 1, 2, 4, 5

[90] Inkyu Shin, Dahun Kim, Qihang Yu, Jun Xie, Hong-Seok Kim, Bradley Green, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen. Video-kmax: A simple unified approach for online and near-online video panoptic segmentation. In *WACV*, 2024. 2

[91] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[92] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. 1, 2

[93] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Arxiv*, 2012. 2

[94] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[95] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017. 2

[96] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *Arxiv*, 2024. 1, 2

[97] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[98] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[99] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 2

[100] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[101] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[102] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019. 2

[103] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *CVPR*, 2023. 2

[104] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *ICCV*, 2023. 2, 5

[105] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 2, 3, 4, 5, 7

[106] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 5

[107] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[108] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljoša Ošep, Laura Leal-Taixé, and Liang-Chieh Chen. Step: Segmenting and tracking every pixel. *NeurIPS*, 2021. 2

[109] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 2

[110] 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 *ACM MM*, 2017. 2, 5

[111] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 5

[112] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[113] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *Arxiv*, 2024. 1

[114] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *Arxiv*, 2024. 7

[115] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *Arxiv*, 2018. 2

[116] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 2

[117] 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*, 2024. 1, 3, 5, 7, 8

[118] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1

[119] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 5, 7

[120] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *Arxiv*, 2023. 7

[121] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *Arxiv*, 2024. 8

[122] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2

[123] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 2

[124] Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Towards open-ended visual recognition with large language models. In *ECCV*, 2024. 2

[125] 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 *AAAI*, 2019. 2, 5

[126] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024. 7

[127] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Arxiv*, 2023. 1

[128] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 1

[129] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. 6

[130] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *Arxiv*, 2023. 2

[131] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 5

[132] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *CVPR*, 2024. 1

[133] Idil Esen Zulfikar, Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Point-vos: Pointing up video object segmentation. In *CVPR*, 2024. 2