WTS: A Pedestrian-Centric Traffic Video Dataset for Fine-grained Spatial-Temporal Understanding

Quan Kong[1*], Yuki Kawana[1], Rajat Saini[1], Ashutosh Kumar[1], Jingjing Pan[1], Ta Gu[2†], Yohei Ozao[1], Balazs Opra[1], Yoichi Sato[2], and Norimasa Kobori[1]

[1] Woven by Toyota
[2] The University of Tokyo

Abstract. In this paper, we address the challenge of fine-grained video event understanding in traffic scenarios, vital for autonomous driving and safety. Traditional datasets focus on driver or vehicle behavior, often neglecting pedestrian perspectives. To fill this gap, we introduce the WTS dataset, highlighting detailed behaviors of both vehicles and pedestrians across over $1.2k$ video events in over hundreds traffic scenarios. WTS integrates diverse perspectives from vehicle ego and fixed overhead cameras in a vehicle-infrastructure cooperative environment, enriched with comprehensive textual descriptions and unique 3D Gaze data for a synchronized 2D/3D view, focusing on pedestrian analysis. We also provide annotations for $5k$ publicly sourced pedestrian-related traffic videos. Additionally, we introduce LLMScorer, an LLM-based evaluation metric to align inference captions with ground truth. Using WTS, we establish a benchmark for dense video-to-text tasks, exploring state-of-the-art Vision-Language Models with an instance-aware VideoLLM method as a baseline. WTS aims to advance fine-grained video event understanding, enhancing traffic safety and autonomous driving development. Dataset page: `https://woven-visionai.github.io/wts-dataset-homepage/`.

1 Introduction

Understanding fine-grained information from videos has been a paramount challenge in computer vision, especially in mission-critical applications like autonomous driving and traffic safety scenario analysis [23, 36, 42]. This challenge hinges on interpreting complex spatial-temporal data swiftly and accurately, encompassing environmental context and individual behaviors for robust decision-making and causal understanding of user intentions. Despite significant advancements in this domain, several gaps persist, which we aim to address in our work.

Existing research extensively focuses on vehicle and driver behavior, but pedestrian behavior—a critical aspect of traffic safety—remains underexplored, despite statistics showing over 20% [26] of traffic accidents involve pedestrians. Current traffic event models lack granularity in behavior definition, limiting nuanced

*Corresponding author: quan.kong@woven.toyota
†Work done while Ta Gu was an intern at Woven by Toyota.

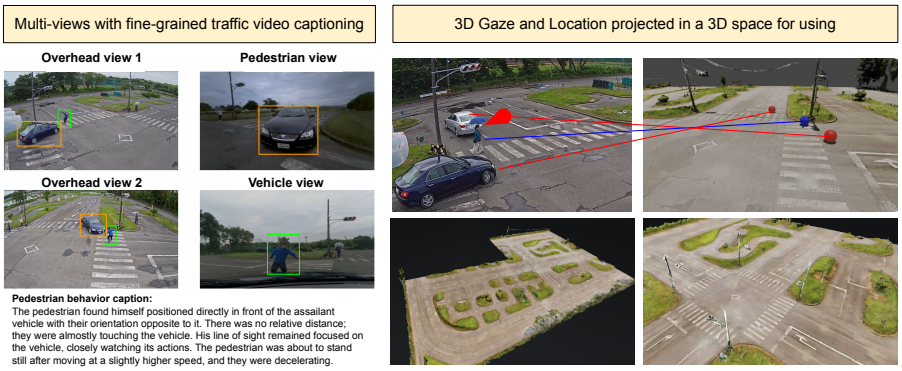

Fig. 1: The overview of WTS dataset features. We provide multi-view videos with fine-grained video captions focusing on pedestrian behavior and the 3D gaze and location information for a further detailed understanding of the traffic-related videos.

decision-making. The rise of Multi-modal Large Language Models (MLLMs), which integrate large language models with multi-modality, advances the generation of detailed textual descriptions from images or videos. However, applying MLLMs to interpret fine-grained, domain-specific details within traffic scenarios is still challenging and underdeveloped for real-world applications.

To address these gaps, we propose the WTS dataset, a pedestrian-centric traffic video dataset with detailed textual descriptions of both pedestrian and vehicle behaviors. We recorded traffic scenarios from multiple views, using overhead and vehicle drive recorder cameras, with five video segment labels for traffic accident analysis. The dataset includes $1.2K$ well-annotated dense descriptions across 255 traffic scenarios. Multiple-view videos are synchronized using an audio signal from a radio on the same channel attached to each camera. Additionally, we provide a 2D/3D synchronized space for our recording environment, offering accurate 3D gaze annotation data of pedestrians using Tobii Pro Glasses 3. The dataset includes 132 traffic accident patterns described using the ISO34502 standard, with videos in high $1080p$ resolution at 24 fps. Figure 1 provides an overview of our dataset features. For broader experimental purposes, we also offer detailed textual descriptions for approximately $5K$ publicly sourced pedestrian-related traffic videos.

As a benchmark for fine-grained video-to-text tasks using the WTS dataset, captions cover four high-level categories: *Location, Attention, Behavior, and Context*, each with detailed textual information. The average caption length for one video segment is about 58.7 words. Figure 2 shows a full caption example. Traditional metrics for video/image caption evaluation, which use text embedding similarity [21, 29, 33], struggle with long descriptions as they measure word-level rather than semantic similarity between inference and ground truth sentences. To address this, we propose an LLM-based video caption scorer focusing on semantic similarity. Additionally, we introduce an instance-aware approach based

Fig. 2: A full caption example with its structure design for the [action] phase

on the Video Large Language Model, serving as a baseline for the fine-grained video-to-text challenge in WTS.

As a summarization of our contribution to the field in several significant ways:

A novel pedestrian-centric traffic video dataset: we introduce a unique dataset focusing on pedestrian-related traffic scenarios. Each traffic event in this dataset is accompanied by detailed textual descriptions of both vehicle and pedestrian behavior, annotated with structured knowledge from traffic safety analysts. 3D Gaze of the pedestrian as a meta-analysis factor in traffic safety is also provided.

An LLM-based video caption evaluation scorer: we introduce new metrics cards composited LLM-based scorer for better alignment with evaluating the semantic correctness than only word-level similarity.

Empirical Evaluation with Vision-Language Models: to demonstrate the efficacy of our dataset, we conduct extensive experiments using cutting-edge vision-language models, including a proposed instance-aware VideoLLM.

2 Related Works

In the evolution of video captioning and behavior-understanding datasets, a significant focus has been placed on varying domains and the granularity of annotations. Our dataset, WTS, stands out in its comprehensive coverage of traffic scenarios with a pedestrian-centric focus. We now draw comparisons with other datasets to highlight WTS's unique contributions to the field.

2.1 Related Datasets

Video Captioning: TACoS [31] offers fine-grained cooking activities, while the MSVD [6], MPII-MD [32], and M-VAD [30] datasets present a broad open domain with a substantial volume of clips. Although the MSR-VTT [41] dataset is rich in movie scene captions and provides a fundamental scene-based approach, it lacks the specificity required for fine-grained descriptions. The Charades [35] and Charades-Ego [34] datasets contribute valuable insights into daily indoor activities with lengthy captions. The ActivityNet Captions [9] dataset broadens the domain

Datasets	Videos (total)	Type	Domain	Captions num.	Avg. caption len.	Year
MSVD [6]	1,970	scene	open	80,380	7.14	2011
TACoS [31]	7,206	scene	cooking	18,227	8.27	2013
MPII-MD [32]	68,327	scene	movie	68,375	11.05	2015
M-VAD [30]	46,589	scene	movie	46,589	12.44	2015
MSR-VTT [41]	507,502	scene	open	200,000	9.27	2016
Charades [35]	9,848	scene	daily indoor	25,032	23.91	2015
Charades-Ego [34]	7,860	scene	daily indoor	14,039	26.30	2016
TGIF [20]	125,782	scene	open	125,781	11.28	2016
ActivityNet Caps. [9]	19,994	instance	human activity	72,976	14.72	2017
VATEX [39]	34,991	scene	open	349,910	15.25	2019
HowTo100M [24]	139,668,840	scene	instruction	139,668,840	4.16	2019
TRECVID-VTT'20 [2]	9,185	scene	open	28,183	18.90	2020
BDD-X [16]	6,984	scene	traffic + outdoor	26,228	14.5	2018
WTS	**6,061 (1,200+4,861)**	**instance**	**traffic + outdoor**	**49,860**	**58.7**	**2023**

Table 1: Comparison between different video caption and 3D gaze-related datasets.

of instance-based activities as a dense captioning task with a significant number of clips, but it does not match the level of detail in pedestrian behavior that WTS offers. The large-scale instructional dataset HowTo100M [24] encompasses a vast array of activities, but it provides limited length in caption information. TRECVID-VTT'20 [2] offers a noteworthy volume of open domain clips, yet it does not approach the intricacy of pedestrian-vehicle interactions as WTS does. In the context of traffic-specific datasets, BDD-X [16] marks a significant step with its focus on traffic scenes and considerable annotation detail for driver action explanation. However, WTS surpasses it with higher granularity in pedestrian behavior analysis and a larger volume of clips and annotations focusing on pedestrians. Notably, WTS is pioneering in its inclusion of 3D gaze data, providing unparalleled insights into pedestrian attention and behavior in traffic scenarios.

2.2 Video Captioning Methods

Video and image captioning are fundamental tasks in video understanding. Vid2Seq [43] introduces a model that integrates special time tokens in a language model to predict event boundaries and textual descriptions in the same sequence. T. Wang et al. [38] present PDVC, a framework for dense video captioning that uses parallel decoding and treats dense caption generation as a set prediction task. MPLUG-2 [40] leverages large-scale pre-training for a deep understanding of complex visual-language interactions. VALOR [7] is a framework for object retrieval tasks involving video and language input, excelling in processing complex queries and locating items based on descriptions. Recent foundation models have significantly improved performance in video-to-text tasks due to prior knowledge alignment. DriveGPT-4 [42], based on the GPT-4 architecture, integrates visual data and contextual understanding for autonomous driving scenarios, showing strong performance on the BDD-X [16] dataset. Caption Anything [37] generates accurate, context-aware captions for a wide range of video content, leveraging the

segment anything model for descriptive and relevant captions. Video-LLaMA [47] combines linguistic, visual, and audio data for comprehensive video content understanding. We benchmarked recent Video LLM-based methods' performance on the WTS dataset to evaluate their potential for fine-grained video-to-text tasks.

2.3 Evaluation Metrics

Video/image caption evaluation metrics including, reference-based ones such as BLUE [29], ROUGE [21], CIDER [33], METEOR [5], SPICE [1] and Rankgen [18]. However, due to the above methods focused on the word level similarity and its order, for the long caption, especially its semantic meaning evaluation is relatively difficult to judge that two paragraphs represent the same thing but different words. Recently, for video-language understanding benchmark, GPT-based metrics [3, 11, 19] have been developed for use, which are better for aligning the semantic meaning for evaluation. The main difference between our proposed LLMScorer and the above ones is LLMScore is designed with a customizable specified aspect considering the semantic meaning as well as syntactic structure similarity for a holistic caption correctness evaluation.

3 WTS Dataset

WTS is a novel pedestrian-centric traffic video dataset featuring 255 traffic scenarios, including staged pedestrian-related accidents across $1.2k$ video segments. Each scenario spans 1 to 3 minutes, with segments ranging from 1 to 15 seconds. It covers 5 phases of pedestrian behavior (Pre-recognition, Recognition, Judgement, Action, Avoidance). Detailed textual descriptions of pedestrian and vehicle behaviors are provided for each segment, along with bounding box annotations. We also curated approximately $5k$ pedestrian-related videos from BDD100K [45] using the same annotation approach as WTS. Additionally, we include synchronized 3D gaze and location annotations for each scenario video, totaling $52,823$ frames across 6 subjects in outdoor environments.

3.1 Data Construction

Camera views: Our recording setup includes three types of cameras: overhead, driver recorder, and ego-centric cameras. The multi-view setup is designed for vehicle-infrastructure cooperation, such as in smart cities, enhancing the accuracy of fine-grained descriptions and improving AD system safety features. It also helps avoid false negatives from vehicle blind spots and offers promising avenues for future research. We selected 18 out of 24 overhead views after removing occluded viewpoints. Each view records at $1080p$ resolution and 24 fps, with calibrated camera parameters. The driver recorder uses a GoPro Hero10 with linear model settings at $1080p$ and 24 fps. The ego-centric camera, Tobii Pro Glasses 3, captures $720p$ resolution videos at 24 fps, providing accurate 2D gaze ground truth. Sample views of these cameras are illustrated in Figure 1.

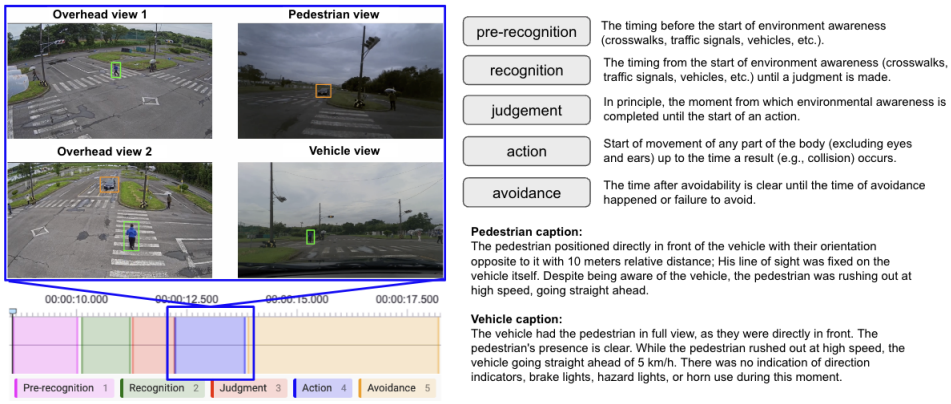

Fig. 3: The overview of WTS video caption data structure: 1) the left figure shows multiple views from overhead to ego vehicle view with 5 phases. 2) the right figure shows the definition of our phase segment and the GT captions corresponding with action segment about the target pedestrian and vehicle respectively as an example.

Subjects: There are total 14 subjects who joined the recording with 7 females and 7 males, whose ages are ranged from 16∼50 years old.

Scenarios: We follow the ISO34502 standard, a scenario-based safety evaluation framework used for automated driving systems as well as our scenario guideline. We created at least one scenario for each of the 138 pedestrian-vehicle relative position patterns defined in it to construct the recording scenario. The scenario patterns and its recorded video sample frames are shown in Figure 4(a).

Captions: We provide fine-grained pedestrian-related behavior captions for each video segment in the traffic scenarios. It starts from the phase segmentation to find each behavior phase temporal localization part, then moves to describe the event in the segment into text along the temporal directions. The features of this caption can be drawn as 1. long paragraph; 2. fine-grained observation regarding the position, action, attributes, and attention of pedestrians and vehicles with the surrounding context; 3. focus on the target objective.

3D Gaze: 3D Gaze is provided for the target pedestrians as extra data for further use, such as using it as a prior for traffic accident analysis. We provide both 2D and 3D gaze annotations for the corresponding frames with the target pedestrian. Figure 1 shows the example of 2D/3D gaze annotations in the frame. Except for the above-ground truth, we also provide the 3D head position and raw 2D gaze ground truth acquired from the Tobii glasses. To check and visualize the 3D information appropriately, a 2D paired 3D scanned map is given.

3.2 Data Collection

Generally, collecting natural traffic events in the real world is challenge. WTS ensures a controlled environment for collecting traffic events, adhering to ISO34502 scenarios and using pre-defined context settings based on actual accident videos.

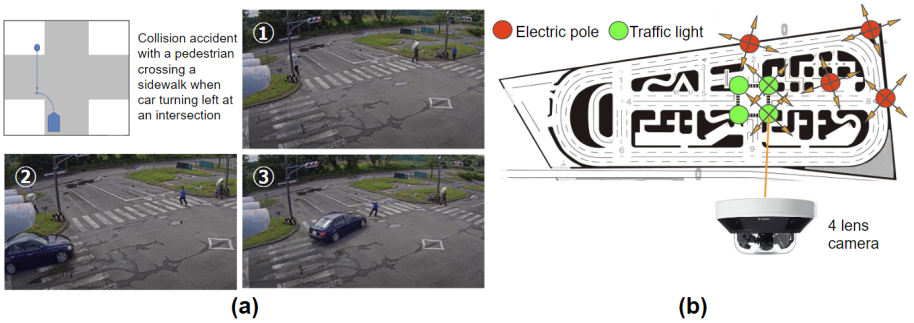

Fig. 4: (a).Sample of scenario pattern. 3 frames are sampled from the video along the temporal direction with the order 1 to 3 at the upper left of the frame. (b).Our recording environment map and camera position.

All accident-related scenarios are staged by professional stuntmen to replicate real-world conditions accurately.

Recording environment: We use a driving school with several intersections and single roads as our recording environment. It is a 72×84 meters outdoor space in total, and 15×15 meters for intersection and $11m$ as width and $81m$ as length for the single road. We have installed 6 multi-lens (4 lenses) overhead view cameras *BOSCH FLEXIDOME multi 7000i* in our recording place, and each overhead view camera is attached to the electric pole around the road. Figure 4(b) shows the map of our recording place with the camera placement positions.

Recording process: A series of scenarios was listed on a worksheet. Random walk or standby action was performed by subjects before and after each event. Each event includes each phase from Pre-recognition to Avoidance. The traffic light in the intersection is operated normally without any pre-defined behavior. The events occurred at various positions to ensure the diversity of the dataset. In each event, target pedestrians will be involved in the scenario, and non-target pedestrians who only performed the random walk in the video to make the task setting close to the real-world setup.

Synchronization: Synchronizing multiple, heterogeneous videos is challenging, especially under unconstrained outdoor environments, where we have multiple dynamic cameras. We utilize the audio input that most commodity video cameras are equipped to synchronize the videos. We use analog radio signals as the audio syncing signals. The average sync delay is 0.015 seconds, well below our frame interval of 24 FPS. For quality assurance, human annotators checked for audio delays (echoing when two videos are out of sync by more than 0.03 seconds) and labeled such videos for human modification.

3.3 Ground Truth Generation

Annotation of the detailed description of traffic video is not easy to ensure accuracy and bias from the human. To resolve this kind of challenge and provide

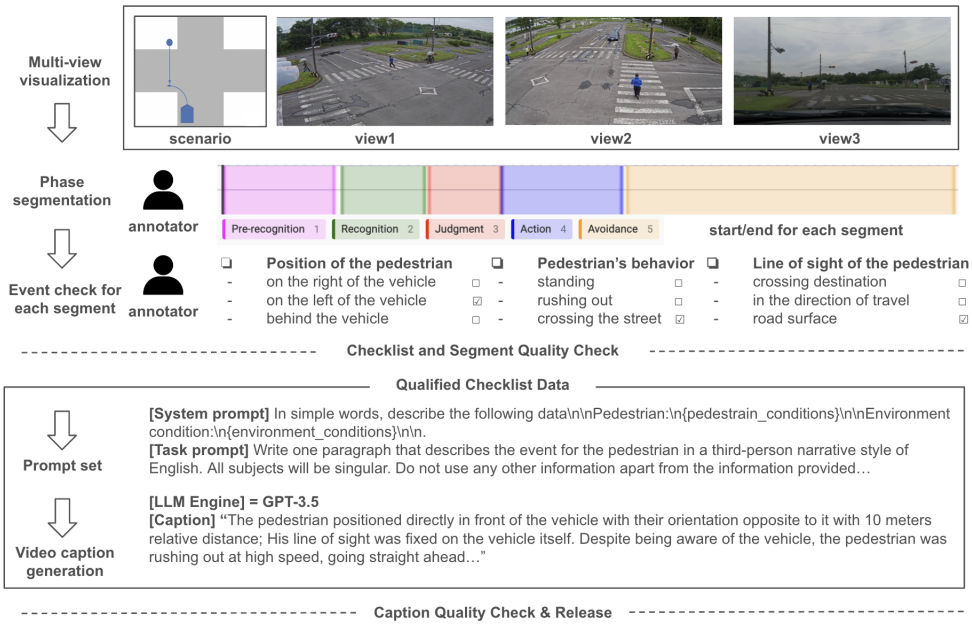

Fig. 5: Annotation pipeline for generating the traffic domain-related captions.

high-quality annotations for each data, we introduce several novel manners for the annotation.

Captions: For the traffic event-related caption, two challenge points are: 1) the caption information requires professional knowledge and viewpoint that is hard to create by a normal annotator correctly. 2) to give the temporal localization information to each segment and create the related temporal-spatial behavior caption is still based on the intuition of the analyst's experience, which is biased. 3) writing a long description of the video from humans requires a high concentration which is error-prone. To remove the bias and make the annotator perform the annotation without deep knowledge of the traffic safety analysis area, we introduced a semi-auto caption generation manner with the following flow as shown in Figure 5: 1. We cooperated with the professional analyst in the insurance company to regularize and unify the guidelines for the segment temporal localization manner. Annotators are asked to do the phase segmentation according to the guidelines. 2. We make structured knowledge from the existing traffic event textual description data. The structured knowledge will be a breakdown of over 180 factual items as a checklist related to the environment context, attributes, position, action, and attention from target pedestrians and vehicles. For the position (left, right, front, etc.) items regarding the pedestrian and vehicle, we use the vehicle-centric as the relative anchor to define whether the pedestrian is "left" or right" to the vehicle to remove the bias. Samples of our checklist are shown in Figure 5. Details can be referred from the supplementary. 3. Annotators are asked to check the items from the structured checklist for each annotated segment

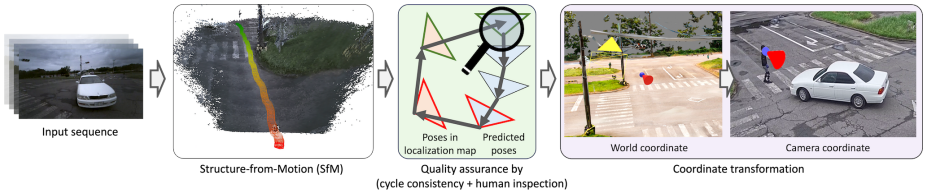

Fig. 6: Illustration of the 3D gaze GT pipeline. Note that the radius of the red cone, illustrating the 3D gaze direction, is solely for visualization purposes. It represents an approximate eccentric gaze FOV of 15°, attainable without head movement.

Fig. 7: Qualitative sample from our tracking results for bbox generation. Using human annotated bbox as visual prompt input to the track-anything.

according to facts that occurred in the video. 4. Once the event check process is done, a double-check process happens to verify whether the checked items match the video or not as the first quality verification. 5. Then, the checked items are fed to a Large Language Model with an appropriate prompt setting (detailed can be referred from the supplementary) for generating the natural sentence, including all the checked items as the caption ground truth. We use GPT-3.5 [28] as the LLM engine for the caption generation. 6. Finally, a double-check for the generated caption is done manually as the second quality verification.

Notice that GPT is used solely to summarize human-annotated scenario-describing checklist results into captions, ensuring that the diversity of scenario descriptions is not limited by GPT's diversity.

3D Gaze: Previous studies on 3D gaze ground truth have typically been conducted in controlled environments or with people under controlled conditions. These studies often rely on numerous AR markers to localize ego-view cameras [25] or restrict the point of gaze, providing instructions to subjects for gaze estimation [13, 15]. However, these methods can make the environment appear unnatural in third-person view videos, or hinder subjects from acting naturally.

To overcome these limitations, our approach involves localizing in the environment through structure-from-motion (SfM) from ego-view video, inspired by [12]. The pipeline is outlined in Figure 6. Input videos are sub-sampled at 5 fps, and each frame is localized in world coordinates using SfM and a pre-built localization map. We utilize the GTSfM open source SfM library [4] and a Matterport camera with LiDAR scanning [‡] to create the pre-built localization map. Finally, the 3D gaze direction in ego-view is transformed into each surveillance camera's

[‡]https://matterport.com/

third-person view. This is done using the pre-calibrated surveillance camera pose, the ego-view frame pose, and the 3D gaze direction from the Tobii glasses in local coordinates. The transformation process for the i-th surveillance camera is defined as:$\mathbf{d}_i = \mathbf{R}_i^{-1}\mathbf{R}_{\text{ego}}\mathbf{d}_{\text{ego}}$, where $\mathbf{d}_{\text{ego}} \in \mathbb{S}^2$ represents the 3D gaze direction in ego-view, $\mathbf{R}_{\text{ego}} \in \text{SO}(3)$ the ego-view pose in world coordinates, $\mathbf{R}_i \in \text{SO}(3)$ the i-th surveillance camera pose in world coordinates, and $\mathbf{d}_i \in \mathbb{S}^2$ the 3D gaze direction in the i-th surveillance camera.

We evaluated the 3D gaze annotation quality in four aspects: (1) ego-view pose estimation using SfM, (2) transformation to world coordinates, (3) overhead camera pose estimation, and (4) Tobii's accuracy with moving subjects. For (1), sampled videos with rapid motion were used, resulting in an average error of 4.19 degrees. For (2) and (3), we estimated relative poses between real and rendered images from scanned 3D scenes, achieving an error of 0.18 degrees. We applied a sanity check to remove visually perceptible errors and aggregated ego-view poses using RANSAC and Procrustes. For (4), Tobii's gaze accuracy for walking subjects was 1.74 degrees [27]. Finally, the combined error was 6.11 degrees, within the human eye's 15 degree eccentric gaze FOV.

Bounding Box Generation We also provide bounding boxes of target vehicles and pedestrians. We choose a semi-supervised approach based on the human prompt in the first frame and tracking the target in the rest of the frames. We leverage Track Anything [44], an interactive tool for segmentation and video object tracking based on Segment Anything [17] and XMem [8] respectively, which only takes several clicks on the target in the first frame as input.

4 Baseline Approach

Based on our dataset, we prepared three baselines for testing the fine-grained video captioning task. **Video-LLaMA** [48] is a multi-modal LLM framework with the capability of understanding both visual and auditory content in the video. The video-language branch is composed of a frozen pre-trained image encoder from EVA-CLIP ViT-G/14 to extract features from video frames In our experiment, we do not use the audio branch but only the pre-trained video-language branch for the video caption benchmark without fine-tuning on the WTS dataset under several different prompt settings. **Video-ChatGPT** [22] use CLIP ViT-L/14 as the visual encoder. To acquire the video-level feature, it uses frame-level embeddings are average-pooled along the temporal dimension to obtain a video-level temporal representation. Similarly, the frame-level embeddings are average-pooled along the spatial dimension to yield the video-level spatial representation. The temporal and spatial features are concatenated to obtain the video-level features with a linear layer to project the video embedding Q_v into the language decoder's embedding space. The text queries are tokenized to the same dimensions as Q_t concatenated with the Q_v input to the language decoders.

Instance-VideoLLM is our proposed baseline with fine-tuning on our training dataset. The main framework follows a similar architecture to Video-LLaMA, thus we use the same visual encoder, positional embedding, and Video Q-former as Video-LLLaMA. As the caption is targeted at the specific pedestrian

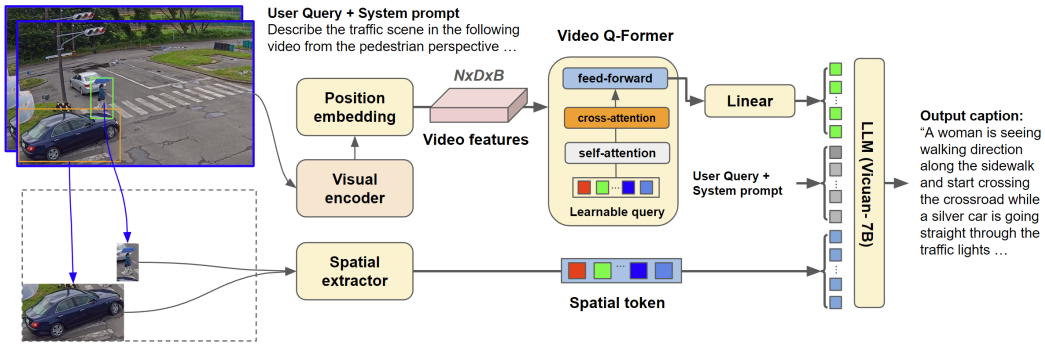

Fig. 8: Instance-VideoLLM approach overview.

or vehicle, we introduce a spatial token that represents the target instance region to the language encoder. The overview of the approach is shown in Figure 8. For a given video V, the segments of the video M is defined as M_{seg}. The frame $F_t, t \in N$ from M_{seg} is fed to the visual encoder and then applied with a positional embedding to the representations of different frames. The position-encoded frame representations with dimension D and frame number N in a batch size B are denoted with tensor $N \times D \times B$ are fed to the Video Q-Former to obtain the video embeddings. A linear layer to transform the video embeddings into the video query vectors V. The video query vectors are of the same dimension as the text embeddings T of LLM. Inspired by Pixel LLM [46] using the binary mask to introduce the instance spatial information. Each video segment's first frame target bounding box region will be used for generating a bbox mask binary map, which will be resized into 224×224 fulfilled with a binary mask as 1 if the pixel in the bbox. Then flatten and project it to generate the spatial token S. In the forward pass, video query embedding V and spatial token S will be concatenated to text embeddings T as a video-instance soft prompt to guide the frozen LLMs to generate the text conditioned on video content

5 Experiment

5.1 Experiment Setup

Dataset for benchmarking: There are around 120 scenarios from staged data and 2000 scenarios from selected BDD data as training sets, 60 scenarios from staged, and 800 scenarios from BDD as validation sets. Each scenario will have ~ 5 segments with ~ 5 captions. For staged data, as there are multiple overhead views, we picked up a main view that covers the whole scenario in one view with a clear visible angle for the pedestrian and vehicle from the dataset for training and validation purposes. We think multiple-view caption consistency is also a new aspect for pushing forward more accurate video-to-text performance.

LLMScore Evaluation protocol: Semantic and syntactic information is crucial to measure the relatedness of two sentences. Various studies [14] [10]

Method	Prompt	LLM	Fine-tune	BLUE-4	METEOR	ROUGE-L	CIDER
Video-LLaMA [48]	P-A	LLaVA-7B	No	0.022	0.201	0.195	0.119
Video-ChatGPT [22]	P-A	Vicuna-7B	No	0.096	0.117	0.171	0.009
Video-LLaMA [48]	P-B	LLaVA-7B	No	0.027	0.210	0.211	0.143
Video-ChatGPT [22]	P-B	Vicuna-7B	No	0.024	0.178	0.208	0.053
Video-LLaMA [48]	P-C	LLaVA-7B	No	0.045	0.247	0.226	0.210
Video-ChatGPT [22]	P-C	Vicuna-7B	No	0.072	0.267	0.266	0.282
Ours(VideoLLM)	P-C	Vicuna-7B	YES	0.101	0.389	0.407	0.363
Ours(Instance-VideoLLM)	P-C	Vicuna-7B	YES	0.121	0.409	0.417	0.389

Table 2: Average performance comparison on the WTS dataset. Both WTS staged data and BDD are used for evaluation.

Method	Fine-tune	Semantic	Syntactic	LLMScore
Video-LLaMA [48]	No	0.008	0.373	0.117
Video-ChatGPT [22]	No	0.004	0.468	0.143
Ours(Instance-VideoLLM)	YES	0.285	0.508	0.351

Table 3: Comparison methods on LLMScorer metric.

learn to disentangle the semantic and syntactic representations. To achieve this, inspired by the evaluation protocol in GPTScore [11], our LLMScore has a prompt template includes task description (comparing two captions), ground truth caption (G), inferred caption (C), and consideration aspects (location, attention, behavior of pedestrian/vehicle, and environment). **Semantic score** quantifies the degree of semantic similarity between two sentences. In our approach, we instruct the LLM (*GPT-3.5-turbo*) to evaluate the semantic accuracy of the caption for each aspect *Location, Attention, Behavior, Context* with respect to the ground truth. We assign a score of 1 to the aspect that is semantically correct, otherwise 0. We take an equal-weighted average of these scores and call this Semantic Score ($Score_{sem}$). **Syntactic score** quantifies the syntactic similarity between the answers. First, for each aspect, we prompt LLM (*GPT-3.5-turbo*) to give the answers to subjective questions for candidate caption as well as ground truth. Subsequently, we compute the cosine similarity score between the embeddings of these answers for the subject questions associated with candidate caption and ground truth. We use OpenAI's *text-embedding-3-small* for generating embeddings. We then calculate the syntactic score ($Score_{syn}$) by taking the equally weighted average of the cosine similarities. Semantic score and Syntactic score are defined as $Score_{sem}$ and $Score_{syn}$ respectively as,

$$Score_{sem} = \frac{1}{n} \cdot \sum_{i=1}^{n} P_{sem}(T, A, G, C, Q_{sem}, O_{sem}, S_{sem}) \tag{1}$$

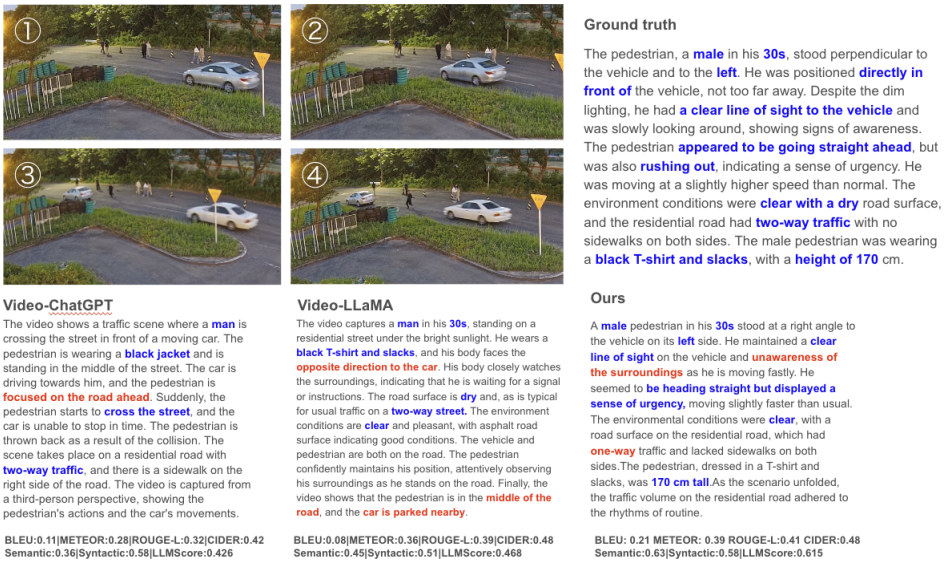

Fig. 9: Comparison of the caption from each method evaluated with each metric. Blue font means the evaluation target aspect in GT and correct representation, and Red font means the error representation.

$$Score_{syn} = \frac{1}{n} \cdot \sum_{i=1}^{n} P_{syn}(T, A, G, C, Q_{syn}, O_{syn}, S_{syn}) \tag{2}$$

where P_{sem} and P_{syn} are the prompt template with the input, T is task definition, A is definition for aspects, G is ground truth, C is inferred caption. Q_{sem} defines the queries, S_{sem} is scoring criteria, O_{sem} is output format for semantic scoring and Q_{syn} defines the queries, S_{syn} is scoring criteria, O_{syn} is output format for syntactic scoring. Therefore, LLMScore is defined as $LLMScore = w_1 * Score_{sem} + w_2 * Score_{syn}$, where w_1 and w_2 are the weights for the scores.

5.2 Implementation details

For the Video-LLaMA and Video-ChatGPT, we did not fine-tune the model but used the input video and the user query prompt fed to the LLM for generating the captions. There are three kinds of prompts we used, *P-A* is a simple task prompt like "Describe the video from a pedestrian perspective". *P-B* is a prompt with system settings and more constraints for the traffic domain for the task. *P-C* is a system prompt with a task description following a demonstration sample. More detail can be referred from the supplementary. For our proposed baseline, during the fine-tuning, LLM is frozen only to fine-tune the Q-Former, linear translation part. We use the AdamW optimizer and a weight decay of 0.05. We

use a cosine learning rate decay with a peak learning rate of $2e-5$ and a linear warmup of $2k$ steps. We use images of size 22×224. For LLMScore, the w_1 and w_2 are set to 0.7 and 0.3 respectively.

5.3 Evaluations

Table 2 shows the comparison result on the WTS dataset for each method with different prompt settings on 4 kinds of traditional popular metrics. It is obvious that the prompt $P-C$ is the best setup and thus all the methods could achieve the best results under this setting. However, based on this best prompt setting, Video-LLaMA and Video-ChatGPT still worked not well for the WTS dataset traffic event domain showing that the fine-grained description in WTS is not generalizable from the common sense knowledge trained Video LLM model without fine-tuning.

To evaluate the impact of fine-tuning on the WTS dataset, we compared *Ours(VideoLLM)* to Video-LLaMA in Table 2 of our paper. Both models share similar architectures, isolating the effect of fine-tuning. The comparison showed that fine-tuning improves performance but is still a challenge. In *Ours(Instance-VideoLLM)*, we added region-specific information. Despite this, results indicate significant room for improvement, suggesting our approach as an initial idea for developing more advanced methods for fine-grained instance-level video understanding with LLMs.

Table 3 shows the result using our LLMScore for each method. For Video-LLaMA and Video-ChatGPT, the semantic score is relatively low and the syntactic score is high, is because almost all the critical semantic meaning regarding the location, attention, behavior, and context are not correct according to the GT even though the whole paragraph looks similar to each other. More samples can be referred from the supplementary. Figure 9 shows a success case sample of how the caption looks like from each method. It is hard to tell the difference from the traditional metrics but highly be separated by using the LLMScore for fine-tuned results, as more semantic meanings are correct for this case.

To compare the LLMScore with the human evaluation result, We use 50% of the validation set for human evaluations. Human evaluators score the correctness of information in C and G using pre-set questions, extracting sub-texts that best describe the aspects. The average human score is 0.243 (variance is 0.002), close to the LLMScore of 0.242 (variance is 0.0007) for the same samples.

6 Conclusion

We introduced the WTS dataset a large-scale pedestrian-centric traffic dataset accompanied by detailed textual descriptions of both vehicle and pedestrian behavior and 3D gaze meta-information for pedestrian use. A new LLM-based video caption evaluation scorer and an Instace-VideoLLM baseline are proposed as well. WTS is a challenging dataset with long descriptions for the traffic video domain, experiment shows that there is a large space for pushing forward the spatial-temporal language understanding into the next stage.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation (2016)
2. Awad, G., Butt, A.A., Curtis, K., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Zhang, J., Godard, E., Chocot, B., et al.: Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. arXiv preprint arXiv:2104.13473 (2021)
3. Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., Zhou, J.: Touchstone: Evaluating vision-language models by language models (2023)
4. Baid, A., Driver, T., Jiang, F., Krishnan, A., Lambert, J., Liu, R., Singh, A., Upadhyay, N., Venkataramanan, A., Warrier, S., Womack, J., Wu, J., Wu, X., Dellaert, F.: GTSFM: Georgia tech structure from motion. `https://github.com/borglab/gtsfm` (2021)
5. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), `https://aclanthology.org/W05-0909`
6. Chen, D., Dolan, W.: Collecting highly parallel data for paraphrase evaluation. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 190–200. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), `https://aclanthology.org/P11-1020`
7. Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345 (2023)
8. Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022)
9. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
10. Fei, H., Ren, Y., Ji, D.: Improving text understanding via deep syntax-semantics communication. In: Findings (2020), `https://api.semanticscholar.org/CorpusID:226283615`
11. Fu, J., Ng, S.K., Jiang, Z., Liu, P.: Gptscore: Evaluate as you desire (2023)
12. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021)
13. Hu, Z., Yang, Y., Zhai, X., Yang, D., Zhou, B., Liu, J.: Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8907–8916 (2023)
14. Huang, J.Y., Huang, K.H., Chang, K.W.: Disentangling semantics and syntax in sentence embeddings with pre-trained language models (2021)
15. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: IEEE International Conference on Computer Vision (ICCV) (October 2019)

16. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. Proceedings of the European Conference on Computer Vision (ECCV) (2018)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
18. Krishna, K., Chang, Y., Wieting, J., Iyyer, M.: RankGen: Improving text generation with large ranking models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 199–232. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.15, https://aclanthology.org/2022.emnlp-main.15
19. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark (2024)
20. Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., Luo, J.: Tgif: A new dataset and benchmark on animated gif description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4641–4650 (2016)
21. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
22. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv:2306.05424 (2023)
23. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: DRAMA: joint risk localization and captioning in driving. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023. pp. 1043–1052. IEEE (2023). https://doi.org/10.1109/WACV56688.2023.00110, https://doi.org/10.1109/WACV56688.2023.00110
24. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
25. Nonaka, S., Nobuhara, S., Nishino, K.: Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2192–2201 (2022)
26. Onishi, Hirofumi, H.T.K.R.I.H., Murase, T.: Analysis of pedestrian-fatality statistics in japan and the us and vehicle-pedestrian communication for vehicle-pedestrian crash-warnings. International Journal of Automotive Engineering (2018)
27. Onkhar, V., Dodou, D., de Winter, J.: Evaluating the tobii pro glasses 2 and 3 in static and dynamic conditions. Behavior research methods (August 2023). https://doi.org/10.3758/s13428-023-02173-7
28. OpenAI: GPT-3.5 (2023), https://platform.openai.com/docs/models/gpt-3-5-turbo
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040

30. Pini, S., Cornia, M., Bolelli, F., Baraldi, L., Cucchiara, R.: M-VAD Names: a Dataset for Video Captioning with Naming. Multimedia Tools and Applications **78**(10), 14007–14027 (2019)
31. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013). https://doi.org/10.1162/tacl_a_00207, https://aclanthology.org/Q13-1003
32. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
33. Oliveira dos Santos, G., Colombini, E.L., Avila, S.: CIDEr-R: Robust consensus-based image description evaluation. In: Xu, W., Ritter, A., Baldwin, T., Rahimi, A. (eds.) Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). pp. 351–360. Association for Computational Linguistics, Online (Nov 2021). https://doi.org/10.18653/v1/2021.wnut-1.39, https://aclanthology.org/2021.wnut-1.39
34. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7396–7404 (2018)
35. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 510–526. Springer (2016)
36. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering (2023)
37. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., Zheng, F.: Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677 (2023)
38. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6847–6857 (2021)
39. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4581–4591 (2019)
40. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., Zhou, J.: mplug-2: A modularized multi-modal foundation model across text, image and video. ArXiv **abs/2302.00402** (2023)
41. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
42. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)
43. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR (2023)
44. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos (2023)

45. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
46. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning (2023)
47. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023), https://arxiv.org/abs/2306.02858
48. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding (2023)