Supplementary Material:

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

Quan Kong^{1*}, Yuki Kawana¹, Rajat Saini¹, Ashutosh Kumar¹, Jingjing Pan¹, Ta Gu^{2**}, Yohei Ozao¹, Balazs Opra¹, Yoichi Sato², and Norimasa Kobori¹

Woven by Toyota
 The University of Tokyo

1 Caption Generation and Checklist Examples

The checklist is prepared for pedestrian and vehicle respectively. There are four supercategories named "Location", "Attention", "Behavior" and "Context". Each supercategory includes several sub-categories such as "Orientation(pedestrian)", "Position", and "Relative distance (to the vehicle)" for "Location". Furthermore, each sub-category consists of several check items for the annotator to check by watching the video segments. The detail of our checklist example used for caption generation is listed in Table 1 for reference.

2 Instruction for Annotation

As "left" or "right" etc., as the position items are relatively defined by the reference object. To remove the bias from the annotator to judge the location and attention-related check items. We prepared an instruction guideline for the annotator to follow up. The location and attention are based on the anchor position of the vehicle to define the pedestrian positions. Figure 1 shows the instructions about our guideline for the annotator to judge this information.

3 Difference for Phase Segments

Our caption includes the "Location", "Attention", "Behavior" and "Context" information in a long paragraph. The major difference across each segment is mainly regarding the "Location", "Attention", and "Behavior" parts. "Context" is the static information to show the attributes of the environment, pedestrian, and vehicle during a short duration. It is not frequently changed along a short time direction. From Figure 2 we could see the fine-grained level difference changed along the time direction across each segment in our captions. These fine-grained level changes will be the key used for the traffic safety reasoning / causal analysis for the downstream task e.g., accident prediction as well.

^{*} Corresponding author: quan.kong@woven.toyota

^{**} Work done while Ta Gu was an intern at Woven by Toyota.

Target	Category	Sub-Category	Check Item
			Same direction as the vehicle
Pedestrian	Location	Orientation	Opposite direction to the vehicle
			Diagonally to the left, in the same direction as the vehicl
			Directly in front of the vehicle
		Position	Diagonally to the left in front of the vehicle
			On the right of the vehicle
			5 meter
		Relative distance	10 meter
			15 meter
	Attention	Line of sight	Crossing destination
			Road surface
			To the right of movement direction
			Closely watching
		Visual status	Constantly looking around intently
			Slowly looking around
			In front
		Movement direction	To the left
			To the right
	Behaviour	General	Crossing
			Squatting
			Going straight
		Special	Crossing the street ignoring the signal
			Crossing immediately in front of or behind a moving vehicle
			Rushing out
		Abnormal	Drunk
			Loitering
			Lying
		Position to pedestrian	In front of the pedestrian
	Location		Diagonally to the left in front of the pedestrian
			Right side of the pedestrian
			5 meter
		Relative distance	
			10 meter
			15 meter
	A + + + :	E:1- 4 -f:	Pedestrian is visible
	Attention	Filed of view	Pedestrian is not visible
			Going straight
	Behaviour	General	Stopped
			Parking
			Turning right
Vehicle			Overtaking
		Light	Hazard lights use
			Direction indicator use
			Brake lights on
			-
		Motion	Deceleration
			Constant speed
			Acceleration
		C1	Male
Context	Pedestrian	Gender	Female
		Age	10s
			20s
			30s
		Height	120cm
			150cm
			170cm
			Glasses
		Attachment	Hat
		Cloth (upper body)	Jacket
			Coat
			T-shirt
	Environment	Road surface type	Asphalt
			Gravel
			Dirt
		Number of lanes	Two-way
			One way, one lane
			One way, two lane
			Intersection w/ singal
		Road type	
			Single road (right curve)
	I .	I	Single road (left curve)

Table 1: Our pickup checklist examples.

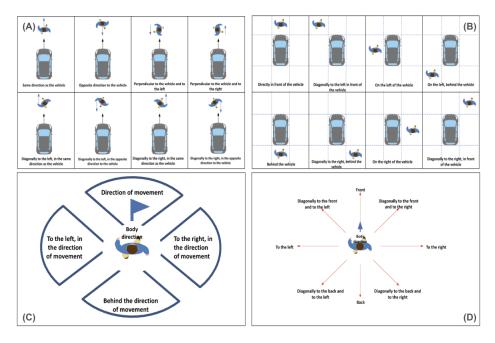


Fig. 1: The illustration for the definition of the pedestrian's direction and position in (A) and (B) respectively. (C) is showing the definition of the Line of Sight direction. (D) is the movement direction definition of the pedestrian.

4 Hard Prompt Example for Baseline Methods

The methods of Video-LLaMA [2], Video-ChatGPT [1], and our baseline need the video and user query prompt as input. User prompt is the pure text information treated as the hard prompt. The prompt setup will heavily affect the performance of the model. Thus we need to make sure to use a promise prompt setup for the experiment. Figure 3 shows the detailed prompt setup for P-A, P-B, and P-C respectively.

5 Comparison about Baseline Generated Captions

Figure 4 and 5 show more example results from different baseline methods with Ground Truth for both pedestrian and vehicle captions.

4 Q.Kong et al.

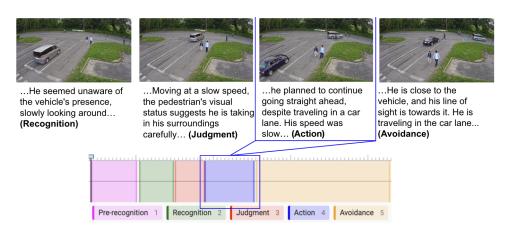


Fig. 2: The segments along the time direction with the captions which have the major difference part. The context (pedestrian, vehicle, and environment attributes) information is almost not changed so frequently along the time direction with a short duration. The major differences regarding the caption for each segment are the location, attention, and behavior parts.

Prompt P-A

Describe the video from pedestrian / vehicle perspective.

Prompt P-B

"You are a language and vision assistant. You are able to understand the visual content that the user provides. You are given a video which shows a traffic scene that involves a pedestrian and a car. Answer the questions which the user asks about the video. Your answers should answer the question once. Your answer should be within 150-200 words. Please make sure the answer should describe the pedestrian behavior, pedestrian attention, location of pedestrian and vehicle, as well as environment. All subjects will be singular. Do not share false information in the answer. Do not write suggestions or emotions in the paragraph. Do not use conjunctions. Describe the traffic scene in video from the pedestrian perspective in a third-person narrative style of English. <Video>kmageHere></kmageHere></kmage/kmarrative-reson-resonance-resonan

Prompt P-C

You are a language and vision assistant. You are able to understand the visual content that the user provides. You are given a video which shows a traffic scene that involves a pedestrian and a car. Answer the questions which the user asks about the video. Your answers should answer the question once. Your answer should be within 150-200 words Please make sure the answer should describe the pedestrian behavior, pedestrian attention, location of pedestrian and vehicle, as well as environment. All subjects will be singular. Do not share false information in the answer Describe the traffic scene in the following video from the pedestrian perspective in a third-person narrative style of

English. <Video><ImageHere></Video>

You can use the following sample video and its sample description as an example Sample Video: <Video><ImageHere></Video>

Sample Description: The pedestrian stood perpendicular to the vehicle and to the right, positioning himself diagonally in front of the vehicle. He destination...[/INST] . He was relatively close to the vehicle and had a line of sight focused on crossing the

Fig. 3: The segments along the time direction with the captions which have the major difference part. The context (pedestrian, vehicle, and environment attributes) information is almost not changed so frequently along the time direction with a short duration. The major differences regarding the caption for each segment are the location, attention, and behavior parts.

[Avoidance]









Ground Truth

Predestrian] The pedestrian, a male in his 20s, was wearing a black T-shirt and navy blue slacks. He stood directly in front of the vehicle, with his body oriented in 18 stood to the vehicle, with his body oriented in 18 stood to the vehicle, and his line of sight was focused on it. Although the pedestrian was slowly traveling in a car lane, he was almost aware of the vehicle. The environment conditions were bright and clear in the suburban area. The road surface was dry and level, with two-way traffic on a residential conditions are no sidewalks on both The event could potentially lead to a collision as the pedestrian and the vehicle were in close proximity.

Nehicle] The vehicle is positioned in front of the pedestrian, close in distance. The pedestrian is visible within the vehicle's field of view. Unfortunately, the vehicle collides with the pedestrian while moving at a speed of 5 km/h. Meanwhile, the environment surrounding the incident includes a male pedestrian in his 20s, standing at a height of 170 cm. He is wearing a black 7-birth and navy blue stacks. The incident bishesp place in conditions with order to the properties of the conditions with order to the properties of the conditions with order to the conditions with the conditions with the conditions with the conditions with order to the conditions with the condit a suburbal rate under tear weather conditions with bright brightness. The road is dry and level, with an asphalt surface. The traffic volume is light on this residential road, which has two-way traffic and no sidewalks or roadside strips on both sides.

Video-LLaMA

VICEO-LLAMA

[Pedestrian] In this traffic scene, a man in his 30s stood on a residential road with a crosswalk. There was a stoplight above him, and he appeared to be waiting for the traffic and he appeared to be waiting for the traffic with a clear sky, and the appeared to see which was stopling to the white lines on the road, was dry, it seemed as though the pedestrian was standing on the white lines on the road was dry, it seemed as though the visible. The man was wearing a black T-shirt and slacks and was facing the opposel decision from roadside strip, he was able to stand where he pleased. The area seemed quiet, with only a few objects visible in the environment, such as the car, the asphalt road surface, the stoplight, and a wooden stake. It was a typical uthan soene with a traffic light and pedestrian traffic.

Vehicle] The video shows an empty intersection in a parking lot with a car parked nearby. From the vehicle's perspective, the intersection where a pedestrian is standing. The pedestrian is visible, and the vehicle is diagonally positioned to the left in front of the pedestrian. The environment reveals the road and parking lot conditions, with asphalt as the surface type. The brightness of the surroundings is bright, and the weather is clear. The scene is mostly empty, with only a few things like the white sign and a green traffic light seen.

Video-ChatGPT

VIGEO-CNATSP1

[Pedestrian] The traffic scene in the video is captured from the perspective of a captured from the perspective of a captured from the perspective of a captured from the pedestrian is wearing a brown jacket and navy green slacks. As the pedestrian crosses the street, a car is seen approaching from the opposite direction. The pedestrian is focused on the road surface ahead and is moving at a slightly higher speed than usual. Suddenly, the pedestrian crosses immediately in front of the moving vehicle, collision. The pedestrian is thrown back by the impact, indicating a collision with the vehicle. The scene takes place on a residential road with how-way traffic, and there is an obstacle on the right side of the pedestrian.

Previous The traffic scene in the video is captured from the perspective of a car driving down the street. The car is shown driving down the street. The car is shown driving down the street and passing by a pedestrian who is crossing the street. The pedestrian is moderate speed, and the pedestrian is crossing the street at a slightly higher speed and the pedestrian crosses in front of the car, and the car is front of the car, and the car is frorted to step to avoid a collision. The video ends with the car driving away from the second.

Our baseline

[Pedestrian] The pedestrian, a young man in his 20s, wore a black T-shirt and rany blue sakes. Positioned directly in front of the vehicle, his body faced the opposite direction. He maintained a keen focus on the vehicle, his body faced the opposite of the carl ane. The suburban environment was characterized by clear, bright conditions, typical of a readerial area. However, the absence of sidewalks on both sides of the absence of sidewalks on both sides of the introduced a sense of vulnerability. With two-way traffic flowing steadily, the potential for a collision loomed as the pedestrian and vehicle shared close quarters.

[Vehicle] The vehicle finds itself in close proximity to the pedestrian, easily observable within the vehicle's field of vision. Tragically, a collision occurs between the two as the vehicle is moving. Concurrently, the scene unfolds with a young male pedestrian in his 20s. Clad in a black T-shirt and navy blue 20s. Clad in a black T-shirt and navy blue slacks, he is a distinct figure amidst the suburban landscape, illuminated by the radiant weather conditions. The road, characterized by a smooth, dry asphalt surface, accommodates light traffic along the residential street, devoid of any sidewalks or

Fig. 4: Sample result for [Avoidance] phase from different baseline methods

[Action]









Ground Truth

[Pedestrian] The pedestrian is a male in his 30s, approximately 170 cm tall. He is dressed in a black 1-Saint and black slacks. It is a cloudy day with bright lighting conditions. The cloud viday with bright lighting conditions. The traffic volume is light on this residential road with two-way traffic. There are no sidewalks or roadside strips on both sides to street lights are present. The pedestrian is street lights are present. The pedestrian is which. His body is oriented in the opposite direction to the whiche. He is closely watching the road surface and is aware of the vehicle's presence. The pedestrian is planning to cross the road slowly in front of the vehicle.

speed of 1b Inh. The environment is a conditional inhibit. The environment is a conditional inhibit. The environment is a conditional inhibit. The environment is a make in his 30s, standing at a height of 170 cm. He is wearing a black. T-shirt has his upper body clothing and black slacks for his lower body. The weariner is cloudy, but he brightness is bright. The road surface conditions are different in conditions are discovered in the brightness is bright. The road surface conditions are discovered in the brightness is bright. The road surface conditions are discovered in the brightness is bright. The road surface is conditionally and level, with support the brightness in brightness in the brightness in

Video-LLaMA

[Pedestrian] In this traffic scene, a pedestrian can be seen standing on a residential road with a crosswalk that shows the traffic direction. The location of the pedestrian and the vehicle is visible, as is the white wall of a building behind a fence. A person can also be seen standing on the white wall of a building behind a fence. A person can also be seen standing on the across the road. The environment in the area is typical with asphalt road surface and a black can parked on the road. The sky is blue, indicating good weather. The dense caption includes detailed descriptions of the objects and the overall atmosphere in the scene. [Pedestrian] In this traffic scene, a

scene.

[Vehicle] In the video, a person is seen wailing on the sidewalk of a residential road in Japan. The location is in a rural area, with video is captured from the perspective of the vehicle traveling down the street. The person is wearing a black-Tehit and black slacks, and standing at a height of 170 cm. There is a white car on the road, and the environment condition reveals that it is a residential street with no signs of danger or obstruction. The camera angle is at a height of 12 meters, with the camera angle is at a height of 12 meters, distance of about 20 meters away from the vehicle. Overall, this scene shows a peaceful and quiet day in a small rural town with clear skies and green grass.

Video-ChatGPT

Predestrian] The traffic scene in the video is captured from the perspective of a pedestrian who is crossing the street. The pedestrian is weeking a brown jacket and predestrian is weeking a brown jacket and respective of a moving whiche. He notices the vehicle of a moving whiche. He notices the vehicle and is focused on the road surface ahead. As he crosses the street, he is thrown back by the impact of the vehicle, inclassing a collision. The scene takes place on a residential road with two-way traffic, and there is an obstacle on the right side of the pedestrian. The video ends with the pedestrian standing on the sidewalk.

Peoclesian saturing or in the sideous. Pehicle] The raffic scene in the video is captured from the perspective of a moving whicle, showing a pedestrian crossing the street in front of a moving car. The pedestrian is wearing a brown jacket and navy blue slacks, and is standing directly in front of the moving vehicle. The vesther is clear and bright, and the road surface is dry and level. The pedestrian is focused on the the moving vehicle. The video ends with the moving vehicle that we will be the moving vehicle the video ends with the moving vehicle that we will be moving vehicle the video ends with the pedestrian being thrown back as a result of the collision with the vehicle.

Our baseline

[Pedestrian] The pedestrian, a man in his 30s standing around 170 cm tall, is attired in a black. T-shift paired with black stacks. Despite the overcast sky, the lighting conditions remain bright, casting a peculiar glow on the surroundings. The road below, the shift period of the surroundings is series as the stage for this encountier. A sparse for the surroundings of the surroundings are the stage for this encountier. A sparse for the stage of the stage the vehicle, the pedestrian's body face away from it, though his vigilant gaze remains fixed on the road ahead. He contemplates a slow traverse across t road, directly in the vehicle's path.

Vehicle) The whicle is at the right of the pedestrian moving forward at a pace of 10 km. Nevertheless, the pedestrian remains discernible within the vehicle's line of sight. Environmental factors reventing at pedestrian as at 10 km. In the set and the pedestrian as a 10 km. In the set and the pedestrian as a 10 km. In the set and the pedestrian as a 10 km. In the set and the

Fig. 5: Sample result for [Action] phase from different baseline methods.

6 Q.Kong et al.

References

- 1. 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)
- 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