Supplementary Material: 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

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
Pedestrian	Location	Orientation	Same direction as the vehicle
			Opposite direction to the vehicle
			Diagonally to the left, in the same direction as the vehicle
		Position	Directly in front of the vehicle
			Diagonally to the left in front of the vehicle
			On the right of the vehicle
		Relative distance	5 meter
			10 meter
			15 meter
	Attention	Line of sight	Crossing destination
			Road surface
			To the right of movement direction
		Visual status	Closely watching
			Constantly looking around intently
			Slowly looking around
		Movement direction	In front
			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
Vehicle	Location	Position to pedestrian	In front of the pedestrian
			Diagonally to the left in front of the pedestrian
			Right side of the pedestrian
		Relative distance	5 meter
			10 meter
			15 meter
	Attention	Filed of view	Pedestrian is visible
			Pedestrian is not visible
	Behaviour	General	Going straight
			Stopped
			Parking
			Turning right
			Overtaking
		Light	Hazard lights use
			Direction indicator use
			Brake lights on
		Motion	Deceleration
			Constant speed
			Acceleration
Context	Pedestrian	Gender	Male
			Female
		Age	10s
			20s
			30s
		Height	120cm
			150cm
			170cm
		Attachment	Glasses
			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
		Road type	Intersection w/ singal
			Single road (right curve)
			Single road (left curve)

Table 1: Our pickup checklist examples.

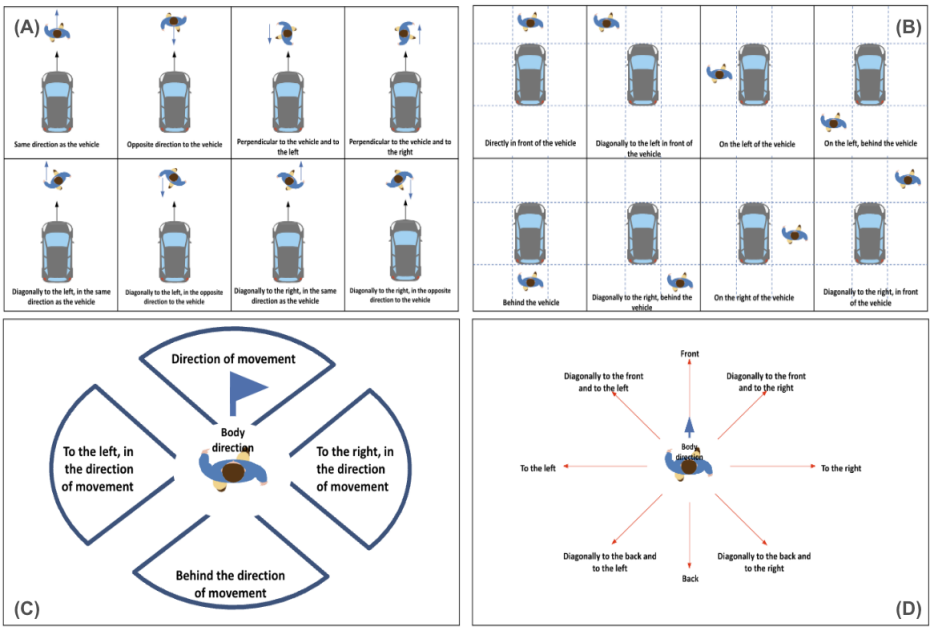

Fig. 1: The illustration for the definition of the pedestrian's direction and position in (A) and (B) respectively. (C) is showing the definition of the Line of Sight direction. (D) is the movement direction definition of the pedestrian.

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.

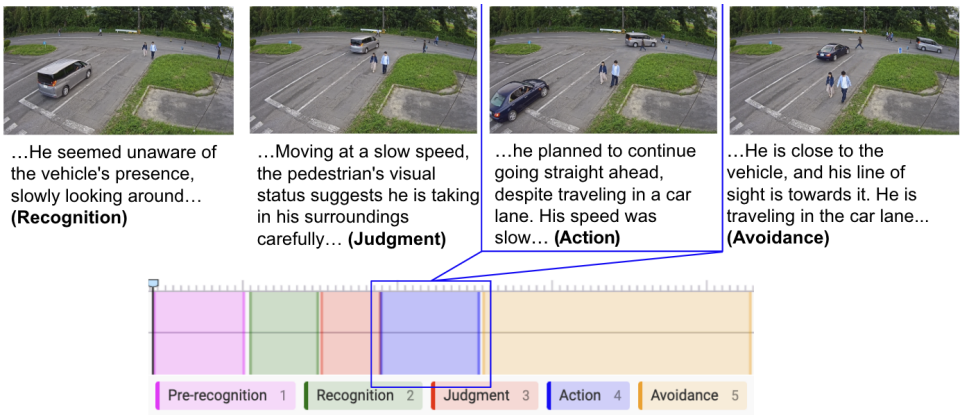

Fig. 2: The segments along the time direction with the captions which have the major difference part. The context (pedestrian, vehicle, and environment attributes) information is almost not changed so frequently along the time direction with a short duration. The major differences regarding the caption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

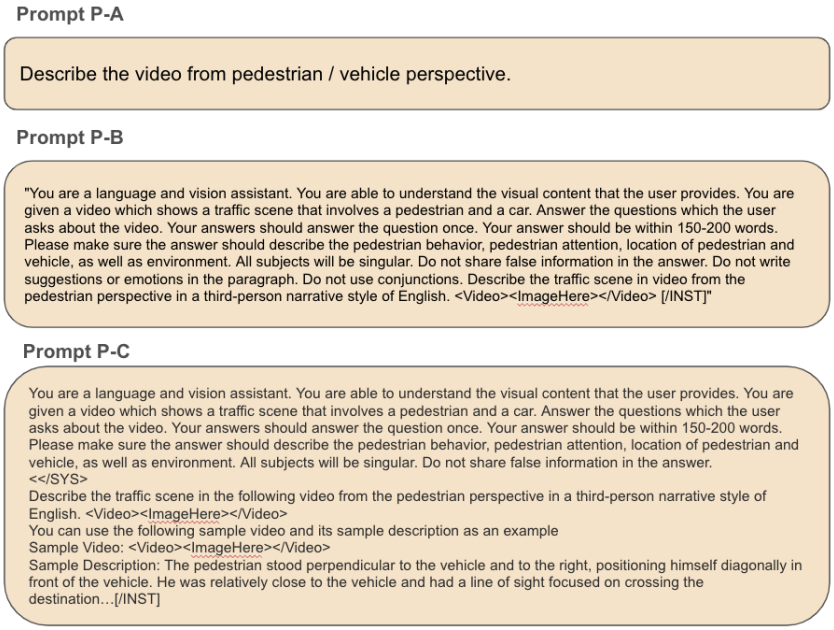

Fig. 3: The segments along the time direction with the captions which have the major difference part. The context (pedestrian, vehicle, and environment attributes) information is almost not changed so frequently along the time direction with a short duration. The major differences regarding the caption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

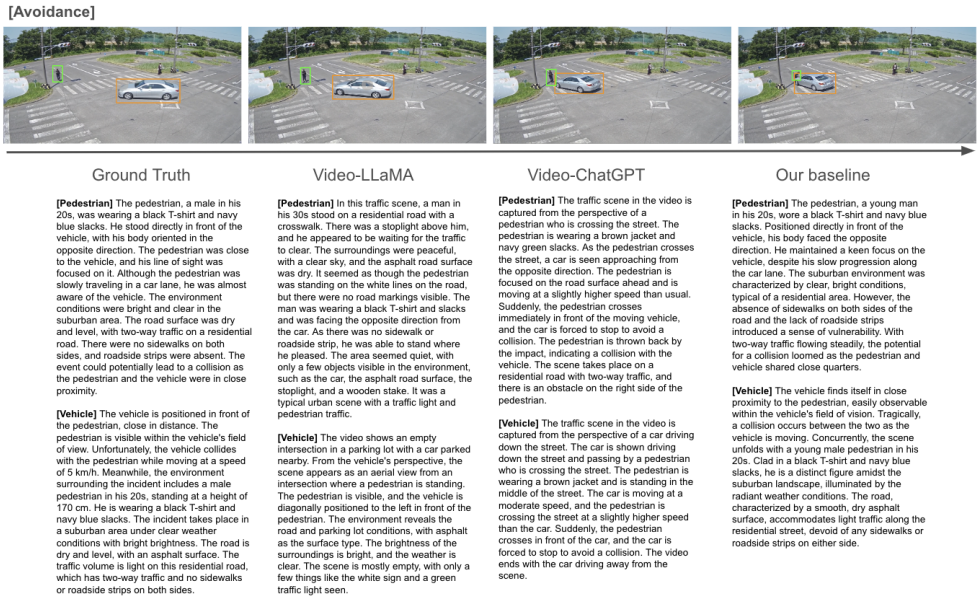

Fig. 4: Sample result for [Avoidance] phase from different baseline methods

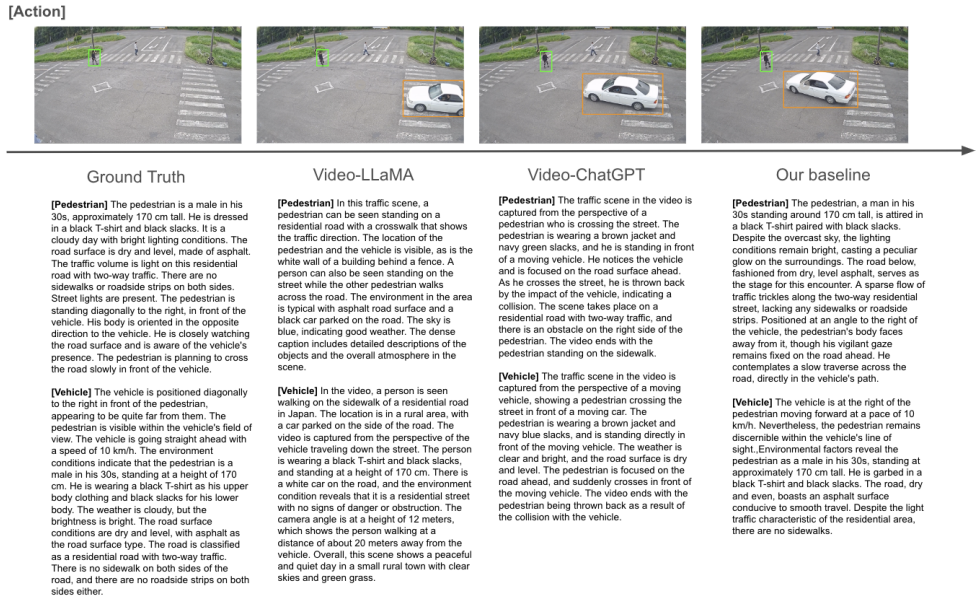

Fig. 5: Sample result for [Action] phase from different baseline methods.

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)
2. 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