Tem-adapter: Adapting Image-Text Pretraining for Video Question Answer (Supplementary Material)

Guangyi Chen*, Xiao Liu*, Guangrun Wang, Kun Zhang, Philip H.S. Torr,
Xiao-Ping Zhang, Yansong Tang†

S-1. Language Template

In this section, we provide some examples of the language template utilized in the Vision-guided Language Adapter to see how the language template works. The steps of transferring a question and the corresponding answer to a declarative sentence are explained in Section 3.2. Table S-2 shows examples of the 6 challenging traffic-related reasoning tasks including "Basic understanding", "Event forecasting", "Reverse reasoning", "Counterfactual inference", "Introspection", "Attribution". Table S-3 provides examples with various question types including: "Where", "Why", "How", "How many", "What's", "Are there", "Did". It shows that all types of question-answer pairs can be transferred to declarative sentences with our language template.

S-2. Details of Baseline Methods

To investigate different categories of methods that transfer the pre-trained model into downstream tasks, such as finetuning, prompt learning, and adapter. We compare **`Tem-adapter`** with the other 9 baseline methods in Sections 4.2 and 4.3 of the main submission. In this section, we provide a detailed description of these methods as follows:

- Unsupervised CLIP [4]: The most direct manner to use the pre-trained clip model is the unsupervised manner, which uses image and text encoders to obtain the visual and textual features and match them with cosine distance, where the QA pair is connected as one sentence.
- Unsupervised CLIP [4] + Language template: Using a predefined template to transfer the QA pair into a declarative sentence, to reduce the language style gap between pre-train and downstream domains. We use this template for all the following baseline methods on SUTD-TrafficQA. Please note that MSR-VTT-MC doesn't provide QA pairs but a caption, thus we don't use the template for this dataset.
- Totally finetuning: Totally finetuning denotes that we finetune all parameters of the CLIP model.
- LoRA [2]: We add the LoRA module to the text encoder of the CLIP model. Each transformer layer of the encoder is adapted with LoRA.
- Partially finetuning: Partial finetuning indicates that we only finetune a part of model parameters, such as the projection layers.
- CLIP-Adapter [1]: CLIP-Adapter adds a linear layer following the CLIP textual encoder and then freezes the encodes and learns this linear layer with downstream losses (classification).
- Multi-layer CLIP-Adapter [1]: To evaluate the effect of more parameters, we use a multi-layer perceptron as the adapter translation.
- Prompt learning (change words) [5]: Given a sequence of tokens of the QA pair, it changes a part of tokens as the learnable parameters and learns them to better align video and texts. Note that we add the adapter heads in this method.
- Prompt learning (change words) without adapter heads [5]: Similar to the previous Prompt learning (change words), it learns the parameters of tokens but the adapter heads are removed.
- Prompt learning (add words) [3]: Different changing word tokens as learnable parameters, an alternative is to add some learnable word tokens before the QA pair. The adapter heads are also added.

S-3. Parameter Analysis

In Section 4.5, we discussed the importance of each component of the **`Tem-adapter`** , and it is observed that the performance drop if either the VL Adapter or the LV Adapter is removed. In this section, we further investigate hyperparameters' effects in the **`Tem-adapter`** . We implemented experiments on the SUTD-TrafficQA dataset. We tried different hyper-parameters of both the Semantic Aligner and the Temporal Aligner, including the latent dimension and the number of layers in the Semantic Aligner, the number

*Equal contribution.
†Corresponding author.

Table S-1: Parameter analysis on the SUTD-TrafficQA dataset. **D** denotes the latent dimension of our Semantic Aligner; **LN** denotes the transformer layer number in Semantic Aligner; **ELN** and **DLN** respectively denote the layer numbers of the Encoder and Decoder in our Temporal Aligner.

Parameters				Accuracy
Semantic Aligner		Temporal Aligner		
D	LN	ELN	DLN	
64	1	1	1	45.3
64	2	1	1	44.6
256	1	1	1	45.1
256	2	1	1	45.4
128	1	1	2	45.6
128	1	2	1	45.8
128	1	2	2	45.8
128	1	1	1	**46.0**

of encoder layers, and the number of decoder layers in the Temporal Aligner. Results are shown in Table S-1. We set the latent dimension of the Semantic Aligner to 64 and 256. Also, the layer number of the Semantic Aligner is changed to 1 or 2. It is observed that the performance is similar to the best accuracy of our model, which indicates the robustness of the training model. In addition, we adjusted the number of encoder layers and the number of decoder layers in the Temporal Aligner and obtained comparable results. The stable performance illustrates our model is robust enough under different hyper-parameter settings.

S-4. Additional Visualizations

Qualitative results are shown in Section 4.6 in our main submission. To better understand our method, we provide more examples in Figure S-1. Positive examples can support that our **`Tem-adapter`** is able to learn the temporal dependencies of videos with the language information and visual embeddings. An example is shown in the top right of Figure S-1. It can be observed a small car is hit by a van when driving and the rear area is badly damaged. Our model is able to capture the temporal dynamics of the video and align the correct text information with the visual dependencies. In addition, Two more failure cases are included to show that our model can not behave well on samples that need complex reasoning. In the bottom right of Figure S-1, the question and answer are closely related to the causes of the accident. That requires exploring interactions and relationships between the components of the video. We find that our model may fail under the complex causal reasoning and leave it as our further work.

References

[1] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. S-1

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. S-1

[3] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. S-1

[4] 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*, pages 8748–8763. PMLR, 2021. S-1

[5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. S-1

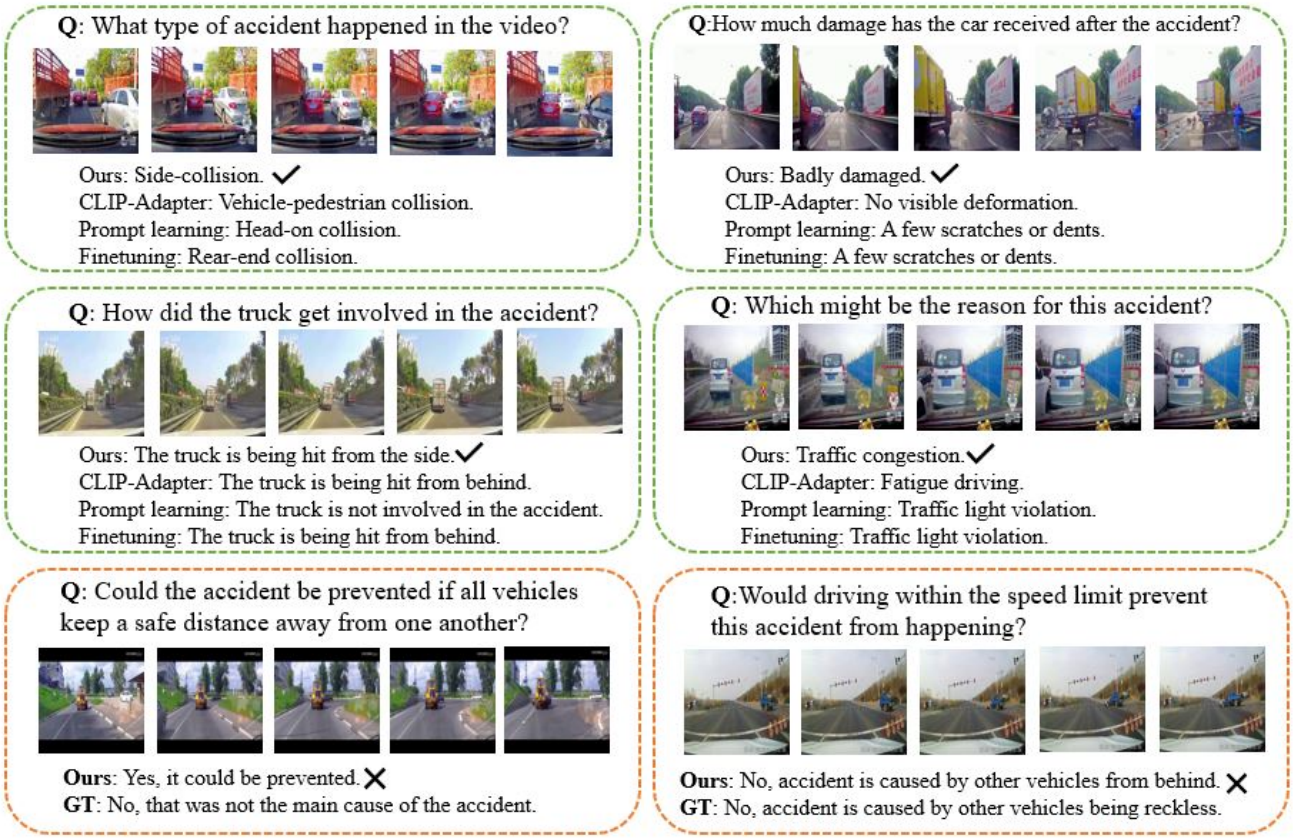

Figure S-1: Visualization of more examples of the VideoQA task from the SUTD-TrafficQA dataset. The top two rows show four positive examples, in which **`Tem-adapter`** learns temporal dependencies of videos to understand the traffic event. Correct candidates are selected by the **`Tem-adapter`** . The bottom row includes two failed cases. Our model can not behave well when encountering complex reasoning scenarios.

Table S-2: Examples of transferring QA pairs to declarative sentences on the SUTD-TrafficQA dataset. Different traffic-related reasoning tasks are included.

Task	Basic Understanding
Question	Which area has been damaged on the vehicle being hit?
Answers	Back Front Side
Declarative sentences	Back has been damaged on the vehicle being hit. Side has been damaged on the vehicle being hit. Front has been damaged on the vehicle being hit.
Task	Attribution
Question	What could possibly cause this accident?
Answers	Obstructed by unexpected objects Sudden braking of a vehicle Violation of traffic rules by pedestrians Sudden or extreme movement by a vehicle
Declarative sentences	Obstructed by unexpected objects could possibly cause this accident. Sudden braking of a vehicle could possibly cause this accident. Violation of traffic rules by pedestrians could possibly cause this accident. Sudden or extreme movement by a vehicle could possibly cause this accident.
Task	Introspection
Question	Can this road infrastructure prevent head-on collision?
Answers	No, the road is unmarked Yes, the divider between two directions is marked clearly
Declarative sentences	This road infrastructure cannot prevent head-on collision, the road is unmarked. This road infrastructure can prevent head-on collision, the divider between two directions is marked clearly.
Task	Counterfactual Inference
Question	Would the accident still occur if the driver slows down in time?
Answers	Yes No
Declarative sentences	The accident still occur if the driver slows down in time. The accident would not occur if the driver slows down in time.
Task	Reverse Reasoning
Question	Which could be the reason for this accident?
Answers	Traffic light violation Retrograde vehicles Improper lane change Obstructed view or limited visibility
Declarative sentences	Traffic light violation could be the reason for this accident. Retrograde vehicles could be the reason for this accident. Improper lane change could be the reason for this accident. Obstructed view or limited visibility could be the reason for this accident.
Task	Event Forecasting
Question	How much damage will the vehicle(s) receive after collision?
Answers	Nearly no damage Significant deformation Some scratches
Declarative sentences	The vehicle (s) will receive significant deformation after collision. The vehicle (s) will receive nearly no damage after collision. The vehicle (s) will receive some scratches after collision.

Table S-3: Examples of transferring QA pairs to declarative sentences. Different type of questions are included.

Question Type	Where
Question	Where was the video taken?
Answers	A crossroad The countryside Road in the city Forest
Declarative sentences	The video was taken in a crossroad. The video was taken in the countryside. The video was taken in the city. The video was taken in Forest.
Question Type	Why
Question	Why did the accident occur when the road is clear?
Answers	Vehicle malfunction. Trying to avoid something on the road. Driver was not paying attention to the road. Uneven road, full of potholes.
Declarative sentences	The accident occurred when the road is clear because of vehicle malfunction. The accident occurred when the road is clear because of trying to avoid something on the road. The accident occurred when the road is clear because driver was not paying attention to the road. The accident occurred when the road is clear because of uneven road, full of potholes.
Question Type	How
Question	How did the truck get involved in the accident?
Answers	The truck is being hit from behind. The truck is being hit from the side.
Declarative sentences	The truck get involved in the accident by being hit from behind. The truck get involved in the accident by being hit from the side.
Question Type	How many
Question	How many lanes does the road have in single direction?
Answers	Two, Only one, Three to five
Declarative sentences	The road has two in single direction. The road has only one in single direction. The road has three to five in single direction.
Question Type	What's
Question	What's the condition of the road surface?
Answers	The road is wet. The road is covered by snow and ice. The road is smooth and clean. The road is dusty or muddy.
Declarative sentences	The condition of the road surface is wet. The condition of the road surface is covered by snow and ice. The condition of the road surface is smooth and clean. The condition of the road surface is dusty or muddy.
Question Type	General Question
Question	Are there any trees along the road?
Answers	Yes, No
Declarative sentences	There are some trees along the road. There are not any trees along the road.
Question Type	General Question
Question	Did a car violate the traffic light?
Answers	Yes, No
Declarative sentences	A car violated the traffic light. A car did not violate the traffic light.