Thanks very much for your overall recognition of the methodology and experimental results of our work, as well as your positive feedback on our writing and expression. Below, we provide a detailed explanation of your feedback and questions. Should any aspects remain unclear or if you have further questions, please do not hesitate to inform us. We are committed to addressing them promptly and comprehensively.

> **[W1]** The paper lacks an in-depth discussion on addressing language ambiguity within the proposed method, which is a critical aspect in vision-language models.
> **[Q1]** Could the authors clarify how language ambiguity is managed in ST-CLIP, especially given the challenges it presents in vision-language models?

Thanks for raising this worthwhile question. In fact, addressing the problem of language ambiguity in prompts of large models is just a motivation of our work. As you mentioned, natural language-based prompts often contain inherent ambiguity, which may cause performance degradation of large models. To eliminate ambiguity, laborious prompt engineering is required for to design highly precise prompts manually. Besides, even minor changes in the prompt could have a significant impact on task performance [1].

Considering these factors, we employ learnable prompts instead of explicitly designed textual prompts in our model. The prompts adopted by our model, *i.e.*, the ST context aware multi-aspect prompts, are constructed from spatio-temporal data and are automatically learned from traffic scene data through minimizing the cross-entropy loss function according to Eq. (23). Our model **does not** use natural language data as inputs and avoids the laborious hand-crafted prompt engineering, so it can address the ambiguity problem in natural language-based prompts.

> **[W2]** Given that the main framework heavily relies on the pretrained CLIP encoder for understanding, the novelty of the approach could be considered somewhat limited.

Thank you for giving us the opportunity to clarify this point. As you mentioned, our model is indeed based on the classic large vision-language CLIP. As with most large models, the performance of the CILP model depends heavily on the quality of the prompts. However, as you mentioned in [W1] and [Q1], most prompts are based on natural language, which have language ambiguity problem. Moreover, the CLIP model is also widely used in many traffic related multimodal applications. And many studies proved that spatio-temporal information in helpful for improving the performance of these traffic related applications [2-3]. However, integrating spatio-temporal information into the prompt of pre-trained multimodal models for traffic-related applications is an area that has received limited attention in current research. The primary contribution of our work is not simply the application of a pre-trained CLIP encoder for traffic scene understanding, but rather the introduction of an automatic prompts learning method that incorporate crucial spatio-temporal information into pre-trained multimodal models.

Here, we proposed the SCAMP module that construct learnable representations of prompt from a spatio-temporal context representation vector, and adopt a multi-aspect prompt attention to model relations between different aspects of a traffic scene. To the best of our knowledge, this is the first attempt to integrate spatio-temporal information into pre-trained multimodal models to facilitate the task of TSU. We believe this innovative design has the potential for application in other prompt-based large models.

> **[W3]** The rationale behind using only a single initial image of a traffic scene in conjunction with trajectories is not convincingly explained. Integrating video clips with GPS trajectories might offer a more coherent and informative approach to traffic scene understanding (but it should be out of the scope of this paper).
>
> **[Q2]** In the context of traffic scene understanding, what are the specific advantages of using only a single image with GPS trajectories over a more dynamic approach like video clips?

I appreciate your insightful comments. The main reason for using single images rather than video clips to understand traffic scenes is the practical constraints of real-world scenarios. In many online ride-hailing apps, the terminal only uploads a screenshot of videos, i.e., single image, captured by vehicle-mounted cameras to servers. Therefore, understanding traffic scene from single image is a more prevalent approach that aligns closely with real-world applications. As an interesting thought, it seems more reliable to use video clips to understand traffic scenes, considering that image data is sparser than videos. We believe it will be a valuable research direction. We may consider it in future studies. Thank you very much for providing this insight.

> **[Q3]** Are there potential scalability issues with the SCAMP module when dealing with large-scale datasets or diverse traffic scenarios? How does it maintain performance efficiency in such cases?

Thank you very much for your insightful question.

- **Diverse traffic scenarios**

  Due to the excellent scalability, ST-CLIP can easily adapt to diverse traffic scenarios, whether by expanding the number of aspects or increasing the number of class-specific words for a particular aspect.

  We extend three additional class-specific words for the "Scene" aspect, refer to **link**:

  - **Expressway**: It represents road segments with a large number of lanes and high average speed.
  - **Construction Road**: It represents road segments under construction, typically featuring barriers or traffic cones.
  - **Avenue**: It represents road segments significantly shaded by trees.

  We integrate these data and conduct few-shot experiments using the same settings and report the results for these three scenes.

Scene Class	Total Number	Correct Number	Accuracy
Expressway	761	724	**0.951**
Construction Road	797	784	**0.984**
Avenue	121	87	**0.719**

  The experimental results indicate that our model can generalize to diverse traffic scenes, particularly those with distinct features such as expressways and construction roads. The recognition accuracy exceeds 95% in these scenarios. For avenues, there are fewer images available, and due to factors such as lighting, the image clarity is lower, resulting in a lower classification accuracy.The experimental

results indicate that our model can generalize to diverse traffic scenes, particularly those with distinct features such as expressways and construction roads. The recognition accuracy exceeds 95% in these scenarios. For avenues, there are fewer images available, and due to factors such as lighting, the image clarity is lower, resulting in a lower classification accuracy.

- **Large-scale datasets**

  In the experiments of our paper, we mainly focus testing our model over a few-shot training set. In fact, for large-scale dataset, our model can also achieve excellence performance. In the following table, we expand the Beijing training set to 5,000 images and test performance of our model over test set with 7464 images (detailed information about the test set can be found in Appendix A). As shown in the table, our ST-CLIP model achieves an improved performance compared with the ST-CLIP with few-short training data (ST-CLIP_FS). The performance improvement rates for the four aspects are 3.4%, 1.7%, 2.5%, and 3.0% respectively. But the marginal benefit of performance improvements is very low. Our performance is still better than Tip-Adapter-F, which is the best baseline.

Model	Beijing Scene	Beijing Surface	Beijing Width	Beijing Accessibility
Tip-Adapter-F	0.687	0.860	0.589	0.745
ST-CLIP_FS	0.757	0.859	0.596	0.799
ST-CLIP	**0.783**	**0.874**	**0.611**	**0.823**
Improved	**3.4%**	**1.7%**	**2.5%**	**3.0%**

References

[1] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[2] Jiang J, Pan D, Ren H, et al. Self-supervised trajectory representation learning with temporal regularities and travel semantics[C]//2023 IEEE 39th international conference on data engineering (ICDE). IEEE, 2023: 843-855.

[3] Fu T Y, Lee W C. Trembr: Exploring road networks for trajectory representation learning[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2020, 11(1): 1-25.