Thanks very much for your feedback and your recognition of the significance of our research and the novelty of our SCAMP module. We elaborate on each of the comments and questions below. If any of our responses fail to adequately address your concerns, please do not hesitate to inform us, and we will promptly follow up with further clarification or assistance.

> **[2. Originality]** However, there are several prior works, such as ... I am not sure if applying a temporal CLIP model to traffic scenes is a novel contribution

Thank you for the comment and sharing the related work. There is a ambiguity caused by the different understanding of the concept of “spatio-temporal information” in different research communities.

In the field of **multimodal** research, as evidenced by the papers you have referenced, the term "spatial information" within the context of "spatio-temporal information" generally pertains to the positional relationships among image patches. Meanwhile, "temporal information" typically denotes the sequential relationships between frames in a video [1-5].

In the field of **spatio-temporal data mining**, however, the spatial information usually refers to geographical locations, while the temporal information represents the sequential relationships when entities appearing at different geographical locations. [6-10].

The "spatio-temporal information" in our paper indicates **the latter**. We incorporate trajectory data that contain both geographical locations and sequential relationships between locations into the prompt of pre-trained multimodal models. Therefore, the contribution of our work should be “the first attempt to integrate **geographical spatio-temporal information** into pre-trained multimodal models to facilitate the task of TSU.” Thank you for the reminder. If this paper is accepted, we will make sure to clarify this point in the final version.

> **[4. Need clarification]** I wonder what the Tip-Adapter-F's clustering results are, considering it also has relatively good performance among the rest of the baselines.

Thank you for this valuable feedback. In clustering experiment in Figure 6, we visualize the clustering result of feature vectors for our model and CoCoOp. Here, each point corresponds to the prompt representation vector of an image. For our model and the CoCoOp baseline, the prompt representation vectors of different images are different. However, for Tip-Adapter-F, which has relatively good performance among the baselines, the prompt representation vectors images in the same class are the same. It is meaningless to cluster the prompt representation vectors of Tip-Adapter-F. The other baselines, *i.e.* CoOp, CLIP-Adapter, Tip-Adapter, Tip-Adapter-F, have the same problem, not just Tip-Adapter-F. Therefore, we only demonstrate the clustering results for CoCoOp and our model.

In fact, for different images with the same traffic scene class, each image has its own unique prompt representation vector, which is an important feature of our model. Even for the same road segment, with the same scene, the image taken at different times may have different traffic environment context. Our model can accurately model this difference in its spatio-temporal context-based prompt representation, so as to obtain more accurate scene understanding results.

> **[5. Need clarification]** In Figure 7 ... I wonder why it predicted it wrong. Does Tip-Adapter-F also use SCAMP?

Thanks for such detailed question. Tip-Adapter-F does not use SCAMP. For the misclassification result of Tip-Adapter-F in the case study, we attribute it to the following reasons:

- **Lacking contextual information of other aspects:** When we use Tip-Adapter-F to classify an aspect scene of an image, such as [Accessibility], it only considers the information about this aspect. Oppositely, our ST-CLIP model adopts a **multi-aspect prompt** method in SCAMP. In this way, the information of other aspects that can help the model to understand the target aspect can also be used as a part of the prompt. For example, the aspect of [Width] is useful for understanding the [Accessibility] aspect.
- **Lacking geographical spatio-temporal information:** In our model, we incorporate the geographical spatio-temporal information into prompt representation vectors. This information is very helpful for traffic scene understanding. For example, the traffic speed in our time-varying properties can help model to understand the [Accessibility] of image. However, the Tip-Adapter-F model lacks this information, and fail to predict the [Accessibility] aspect as a correct class.

References

[1] Liu R, Huang J, Li G, et al. Revisiting temporal modeling for clip-based image-to-video knowledge transferring[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 6555-6564.

[2] Luo H, Ji L, Zhong M, et al. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning[J]. Neurocomputing, 2022, 508: 293-304.

[3] Fang H, Xiong P, Xu L, et al. Clip2video: Mastering video-text retrieval via image clip[J]. arXiv preprint arXiv:2106.11097, 2021.

[4] Wasim S T, Naseer M, Khan S, et al. Vita-CLIP: Video and text adaptive CLIP via Multimodal Prompting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23034-23044.

[5] Weng Z, Yang X, Li A, et al. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization[C]//International Conference on Machine Learning. PMLR, 2023: 36978-36989.

[6] Atluri G, Karpatne A, Kumar V. Spatio-temporal data mining: A survey of problems and methods[J]. ACM Computing Surveys (CSUR), 2018, 51(4): 1-41.

[7] 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.

[8] 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.

[9] Zhao P, Luo A, Liu Y, et al. Where to go next: A spatio-temporal gated network for next poi recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(5): 2512-2524.

[10] Wang Y, Jing C, Huang W, et al. Adaptive spatiotemporal inceptionnet for traffic flow forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3882-3907.