Thank you very much for your recognition of our work in utilizing spatio-temporal information to address the TSU problem. Below, we provide a detailed explanation of your feedback and questions. If there are any unclear points or if you have any other questions, please do not hesitate to let us know. We are committed to addressing these issues promptly and comprehensively.

> **[W1]** The framework lacks sufficient explanation and analysis – it's a classification task, as the generated descriptions only have a few possibilities (like "The width is normal/narrow/extremely narrow/unknown"). The author used a CLIP-based contrastive learning method. However, by using learnable embeddings instead of actual words, which harms the generalization ability of the CLIP model. The author should compare it to a simpler baseline treating it as a classification task.

Thanks for your insightful suggestions. In the form of the problems, the traffic scene understanding task in our paper is indeed a classification problem. However, unlike traditional classification tasks with a fixed number of sample categories, our task may involve more complex and variable traffic scene categories. Traditional image classification models usually require a large amount of labeled data for training, with high demands on both the quantity and quality of data. Additionally, these models often lack scalability and require separate training when adding a new classification category. However, in TSU task, new categories of scenarios may emerge, often with only a small number of annotated samples available. In such cases, traditional classification models may fall short. Therefore, leveraging the generalization ability of CLIP and fine-tuning it through few-shot learning is a possible solution. In other words, CLIP is just a tool for our model to handle new category and few sample challenges in TSU tasks. As shown in section "4.3 Few-shot Experiments" and Figure 3, even with 8 training samples, our model can also achieve satisfactory performance. The complete few-shot results can be found at **figures/few-shot_results.png**.

As you suggested, we additionally conduct comparative experiments with the classification baseline models (**ResNet-50** and **ViT-B/32**) here. The experiments contain two setups, few-shot and large-scale. In **Few-shot (M1)** setup, we fine-tune the baseline model using the same 16-shot dataset as our model. In **Large-scale (M2)** setup, we expand the Beijing training set to 5,000 images to train our model and baselines. The experimental results are shown in the table below:

Model	Beijing Scene	Beijing Surface	Beijing Width	Beijing Accessibility
ResNet-50 (M1)	0.487	0.831	0.336	0.638
ViT-B/32 (M1)	0.490	0.824	0.350	0.645
ResNet-50 (M2)	0.534	0.839	0.387	0.648
ViT-B/32 (M2)	0.551	0.830	0.399	0.661
ST-CLIP	**0.757**	**0.859**	**0.596**	**0.799**

The experimental results indicate that the classification baseline can not achieve satisfactory performance on few-shot dataset (**M1**). Even with the training set expanded to 5,000 images, the classification results of the classification models still fall far behind those of ST-CLIP (**M2**), which demonstrates the effectiveness of our approach.

In addition, we demonstrate the performance of our model with new classification categories. We extend three additional class-specific words for the "Scene" aspect:

- **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 adopt 16 samples for each new scene class to fine-turn our model. The experiment results are reported in the following table.

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.

Therefore, we leverage the generalization capability of CLIP to combine spatio-temporal data with learnable prompt for addressing the TSU task, rather than solely attributing it to an image classification task. We believe this framework has the potential to be extended to other “image-and-spatio-temporal-data” related multimodal data applications.

Thank you very much for your valuable feedback, which help us to understand our work more deeply. We will add the experiment with classification baselines in the supplementary materials of the final version, and emphasize our contribution in the introduction.

> **[W2]** This paper is not well-motivated; it's hard to understand why the trajectories of the vehicle would be helpful in determining features like road width or surface.

Thanks for this constructive remark. In the real world, the behaviors of vehicle have close relations with traffic scenes, such as road widths and surfaces. For instance, a narrow and rugged road is likely to present more navigation challenges than a broad and smooth road, resulting in fewer trajectories and lower speeds. To further demonstrate the role of trajectory behaviors in determining the characteristics of roads, we compare the mean speed of the traffic scenes with soil, broken, and normal categories in the [Surface] aspect.

Road Type	Mean Speed
Soil	6.7
Broken	15.4
Normal	30.2

The comparative results indicate significant differences in trajectory characteristics among roads in different surface categories.

In our model, we construct Time-varying Properties to capture the information in trajectories (see Section “3.1.2 Time-varying Representations for Segments.”) and incorporate them into prompts of the CLIP model for high performance traffic scene understanding. The experiments compared with the baselines also verified the effectiveness of our model.

> **[W3]** This paper uses data from within the company (Didi Rider), which may not be replicable, and the significance of the research question is also small, making it unsuitable for the research track.

Thanks for your thoughtful concern. The relevant code has been provided with the link.

In terms of the significance of our research question, as mentioned in the Abstract and Section "1. Introduction", the aim of our paper is to provide a tool for processing a type of emerging multimodal data: **image data with spatio-temporal information**, which involves the image modality data and the spatio-temporal modality data. This type of data is gathered in substantial volumes through navigation and ride-sharing apps. In recent years, there has been many papers in KDD dedicated to addressing the challenges posed by this type of data [1, 2]. The ST-CLIP model proposed by our paper focuses on the traffic scene understanding, which is a core technology for analyzing such images associated with spatio-temporal information data. Our work integrates spatio-temporal data into multimodal models. Spatio-temporal data mining has been a long-standing area of interest for the KDD community. Meanwhile, research on pre-trained multimodal models is currently thriving in KDD research track [3, 4]. To the best of our knowledge, this is the first attempt to integrate spatio-temporal information into pre-trained multimodal models. I believe that our research could offer valuable insights to researchers in the KDD community working on spatio-temporal data and multimodal data processing.

> **[W4]** The paper has some typos, for example, line 293: the corresponding entry of A is ‘1’; line 111: Spatio-Temporal Data Enhanced ‘CILP’ model.

Thanks for pointing out the writing issues. The two typos will be rectified in the revisions, and our manuscript will be meticulously reviewed.

Finally, we extend our sincere gratitude to reviewer RZdj. Your valuable feedback has prompted us to gain a deeper insight into the motivation and contribution of our paper. We will incorporate these insights into the revision. Regardless of the outcome of this article's acceptance, we are grateful for your assistance in enhancing our research.

References

[1] Yang J, Ye X, Wu B, et al. DuARE: Automatic road extraction with aerial images and trajectory data at Baidu maps[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 4321-4331.

[2] Yu F, Ao W, Yan H, et al. Spatio-temporal vehicle trajectory recovery on road network based on traffic camera video data[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 4413-4421.

[3] Wang D, Salamatian K, Xia Y, et al. BERT4CTR: An Efficient Framework to Combine Pre-trained Language Model with Non-textual Features for CTR Prediction[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 5039-5050.

[4] Huang J, Wang H, Sun Y, et al. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 3029-3039.