TrajPrompt: Aligning Color Trajectory with Vision-Language Representations

Li-Wu Tsao[1], Hao-Tang Tsui[1], Yu-Rou Tuan[1], Pei-Chi Chen[1], Kuan-Lin Wang[1], Jhih-Ciang Wu[2], Hong-Han Shuai[1], and Wen-Huang Cheng[2]

[1] National Yang Ming Chiao Tung University, Taiwan
{lwtsao.ee09,hhshuai}@nycu.edu.tw
[2] National Taiwan University, Taiwan
wenhuang@csie.ntu.edu.tw

Supplementary Materials

We demonstrate more detailed discussions on the shape of trajectory points on the BEV scene. Furthermore, we provide the sensitivity analysis on the selection of top-k candidates in trajectory sampling with quantitative and qualitative perspectives. Finally, we visualize and compare the differences in scene between SDD and DroneCrowd under diverse scene situations.

A Detailed Discussion on Shape of Points

Table 8: Effect of different thicknesses while drawing the line. We position each trajectory point as the center and control the thickness based on the pixel-unit configuration.

Configuration	Deterministic		Stochastic	
Square (Width × Height)	ADE	FDE	ADE	FDE
3×3	46.51	54.27	9.29	15.41
5×5	14.64	21.96	7.33	12.67
7×7	20.84	28.92	9.43	16.20
Circle (Radius)				
$r = 1$	12.48	23.17	7.32	12.24
$r = 2$	8.61	15.99	4.28	7.60
$r = 4$	8.66	16.05	6.07	10.29
Gaussian (Standard Deviation)				
$\sigma = 1$	10.64	19.77	4.80	8.57
$\sigma = 2$	**7.73**	**14.38**	**3.78**	**6.81**
$\sigma = 4$	8.46	15.74	3.98	7.23
$\sigma = 8$	10.72	19.89	4.65	8.29

As drawing lines on the BEV scene includes the fundamental element of thickness, we conduct the sensitivity analysis to investigate how the choice of this hyperparameter affects the performance. In Tab. 8, we gradually increase the range of coloring pixels based on each dependent configuration, that is, the length of width, height, radius, and standard deviation. The quantitative results indicate that the moderate thickness brings effective predictions for all shapes. From the qualitative perspective shown in Fig. 7, larger strokes ($\sigma = 8$) make the CLIP

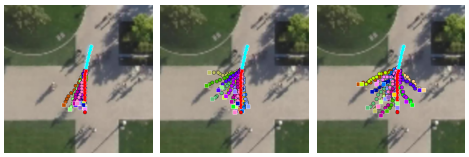
(a) $\sigma = 2$ (b) $\sigma = 8$ (c) CLIP results

Fig. 7: Visualize the effect of different thicknesses on the boundary of multiple surroundings using Gaussian configuration.

model confuse due to delivering multiple surrounding information. According to Fig. 7c, the illustration in Fig. 7b displays a lower probability on pavement and a higher probability on buildings compared to Fig. 7a. This is due to the unclear boundary around the thick line. Based on the observation, the improper thicknesses can hinder the model's understanding and lead to poor performance. Therefore, we select the appropriate stroke size ($\sigma = 2$) to keep the line's precise location and the completeness of background information.

Table 9: Comparing the results obtained from different ranges of top-k candidates (hyperparameter k) in stochastic setting.

Top-k sampling	ADE	FDE
k = 2	4.57	9.33
k = 5	**3.78**	**6.81**
k = 8	5.18	7.94



(a) k = 2 (b) k = 5 (c) k = 8

Fig. 8: Visualize the results by sampling top-k candidates for each prediction step.

B Sensitivity Analysis on Trajectory Sampling

Number of k for Trajectory Sampling. According to the success of our retrieval process, we deliver another perspective on trajectory sampling since most of the trajectory prediction results care much about diversity. Different from retrieval that takes the closest pixel embedding in the dictionary, the sampling process considers the top-k relative features as candidates and builds up the sampling probability based on the embedding distance. Tab. 9 shows the sensitivity test on different k, and the best choice of k is 5 in TrajPrompt. Comparing the results in Fig. 8, the case for k = 5 generates more natural and reasonable diverse trajectory sampling as shown in Fig. 8b. Increasing k involves selecting more locations from a larger set, including k pixels with similar embeddings. However, a main disadvantage occurs when the number of pixels is too large to be cluttered and unrealistic, such as Fig. 8c. In contrast, smaller k can only select the nearby pixels with conservative decisions, such as Fig. 8a. Thus, selecting a proper k is essential to fit diverse scene situations by our sampling strategy.

(a) (b) (c) (d)

Fig. 9: Visualize the results on SDD dataset. (a) and (b) discuss the effect while encountering the boundary of two surrounding types. (c) and (d) illustrate the difference of diversity control on the restricted walkable area.

(a) (b) (c) (d)

Fig. 10: Visualize the results on DroneCrowd dataset. Since there are more restrictions on the city street environments, the walking patterns on this dataset appears to follow the road structure precisely.

C More Qualitative Results

We demonstrate more qualitative results in Fig. 9 and Fig. 10 to discuss the performance of TrajPrompt on various BEV scene scenarios. In Fig. 9a, TrajPrompt demonstrates the ability to act like human, avoiding occupied on road and rapidly enter the pavement considering safety issues. Similarly, the prediction of TrajPrompt prevents the agent from entering the meadow that follows the natural behavior of pedestrians, where the predicted paths are distributed on the pavement as shown in Fig. 9b. These human-like decisions have shown TrajPrompt realizes the correlation of different surroundings well. Fig. 9c and Fig. 9d show the general arc-shaped pattern that appears commonly in the trajectory prediction model, since these cases are all located on the pavement without obstacles or turns on their moving direction. Considering Fig. 9c and Fig. 9d in detail, TrajPrompt can make precise control on the narrow pavement as shown in Fig. 9d, in which the agent is walking down the stairs within a constrained area compared to the wide area in Fig. 9c. Different from the Fig. 9 scenarios in campus, Fig. 10 shows the effectiveness of TrajPrompt on city street with diverse illumination condition. Since there are more building structures on the street, it might cause a little difference on SDD and DroneCrowd dataset. Despite challenging illumination conditions on DroneCrowd dataset, TrajPrompt performs well in Fig. 10, covering all reasonable decisions without violating the understanding of the BEV scene.