Optimizing Vision-Language Model for Road Crossing Intention Estimation Supplementary Material

Roy Uziel* and Oded Bialer
General Motors, Technical Center Israel
uzielr@post.bgu.ac.il, oded.bialer8@gmail.com

1. *ClipCross* Implementation Details

We employed 38 predefined sentence suffixes for Ω_{pre} in conjunction with 15 predefined prefixes, as elaborated in Section 2. The quantity of optimized embedding vectors in Ω_z was optimized and set to 13, as elucidated in Section 3. Five token vectors were used to represent each suffix, with a token vector size of 512. Both the CLIP image and text embedding vectors had a size of 512. We employed CLIP version OpenAI ViT-B/32, Adam optimizer with learning rate of 0.01, a batch size of 1024, loss weight factor $\lambda = 0.013$, and $\epsilon = 0.35$ in the similarity loss. The input images had a frame rate of 30 frames per second. Each image was cropped around the pedestrian's bounding box, with the crop size 15 times larger than the bounding box. The pedestrian-centric CLIP image embedding per frame was obtained by reducing the attention weights outside the pedestrian's bounding box by a factor of 0.1 in the last attention layer of the CLS token in the CLIP image encoder. The temporal combination, resulting in vector $\boldsymbol{x}$, used a sequence of 15 frames of CLIP image embeddings spanning 0.5 seconds, summed with learnable scale factors $\rho_1, .., \rho_{15}$. The MLP that processes the cross-model feature vector $\boldsymbol{\eta}$ consists of two layers and produces two outputs, representing the probabilities of each class. A leaky ReLU activation function was applied to the output of the first layer, followed by a softmax operation on the output of the second layer.

2. Predefined Embedding - Ω_{pre}

Table 1 showcases the 38 predefined suffixes that were used to generate the 38 embedding vectors in Ω_{pre}. These suffixes comprise human descriptions of positive and negative crossing intentions characteristics. The text embedding vectors within Ω_{pre} were derived by averaging the embedding vectors produced using each suffix listed in Table 1 alongside various prefixes, as outlined in Equation 1 of the main paper. The list of utilized prefixes is provided in Table 2, which is sourced from the recommended prefixes in the official OpenAI git repository.

Subsequently, we assess how the choice of the number of predefined suffixes and prefixes, used to generate Ω_{pre}, impacts the performance of *ClipCross*. Figs. 1 and 2 display accuracy metrics on PIE dataset vs. the number of randomly selected predefined prefixes and suffixes from Tables 1,2. The solid lines represent the average results across ten different runs, while the line width represents variance. These figures illustrate that, as the number of predefined suffixes and prefixes increases, the average accuracy improves until it reaches saturation, coupled with a reduction in variance. In this paper, we used 38 suffixes and 15 prefixes. Figs. 1 and 2 indicate that further increasing these numbers results in negligible performance improvement.

3. Insight on Optimized Embedding - Ω_z

In this section, we evaluate the effect of the selection of the number optimized text emebdding vectors, Ω_z, on the performance of *ClipCross*, and how the optimized embeddings are related to the predefined embedding vectors Ω_{pre}. Fig. 3 shows accuracy metrics as a function of the number of optimized embedding vectors. It reveals that optimal performance is achieved at around 13 embedding vectors, as employed in the paper.

The text emebdding optimization process in *ClipCross* can be interpreted as pulling each optimized embedding in Ω_z towards an embedding of a predefined sentence in Ω_{pre} and further refining it. We identify the associated embedding vector in Ω_{pre} for each optimized embedding vector in Ω_z by choosing the vector with the highest cosine similarity. In the second column of Table 1, you can see the number of vectors in Ω_z linked to each suffix vector in Ω_{pre}. The table reveals that the optimization process yielded a reasonably diverse association between the optimized embeddings and the predefined sentences. The 13 optimized embeddings in Ω_z were linked to nine distinct sentences in Ω_{pre}. Additionally, the first two predefined sentences in Table 1 were each transformed into three refined variations.

[*] Roy Uziel is with General Motors Technical Center Israel and Ben-Gurion University of the Negev.

Index	Suffix	# associated vectors from Ω_z
1	away from the road.	3
2	off the road.	3
3	facing a building.	1
4	leaning against a light post.	1
5	sitting on a bench near the sidewalk.	1
6	entering a car.	1
7	walking across a zebra crossing.	1
8	observing a storefront.	1
9	using a crosswalk to cross.	1
10	remaining on the sidewalk.	0
11	staying on the sidewalk.	0
12	along the sidewalk.	0
13	dashing across the street	0
14	sitting next to the road.	0
15	along the side of the road.	0
16	walking parallel to the street.	0
17	crossing a bridge over the road.	0
18	pausing at a street corner.	0
19	crossing the road.	0
20	jaywalking.	0
21	crossing legally.	0
22	stepping onto the street.	0
23	crossing the street at a pedestrian signal.	0
24	navigating through an intersection.	0
25	walking through a pedestrian tunnel under the road.	0
26	intending to cross the road.	0
27	about to cross the road.	0
28	planning to cross the road.	0
29	waiting for green traffic light.	0
30	looking both ways before crossing.	0
31	positioning oneself near a crosswalk.	0
32	checking for oncoming traffic.	0
33	hesitating before crossing.	0
34	walking towards a crosswalk.	0
35	approaching an intersection.	0
36	waiting for a break in traffic.	0
37	signaling their intention to cross.	0
38	preparing to step off the curb.	0

Table 1. List of suffixes used to generate Ω_{pre} and the count of associated optimized text embeddings per suffix.

In Section 4 of the main paper, we conducted an ablation study to evaluate *ClipSimMin*, a variant of *ClipCross* that substitutes the similarity loss in Equation 3 of the main paper with the minimum cosine distance between the vector in Ω_z and any vector in Ω_{pre}. With *ClipSimMin*, we noticed that eight vectors in Ω_z were linked to the first suffix in Table 1, while five vectors were linked to the second suffix. This suggests that the use of the similarity loss in *ClipCross* results in a greater disparity among the optimized vectors in Ω_z compared to *ClipSimMin*. As shown in Table 4 in the main paper, this variation leads to enhanced performance.

4. PIE, PSI and JAAD Datasets

We evaluated *ClipCross* using three publicly available datasets: PIE [3], PSI [1] and JAAD [4]. PIE dataset consists of 6 hours of urban street camera footage captured at 30 fps. Each sequence is 2 seconds long, with 0.6 seconds overlap, and 15 human annotators rated whether the pedestrian intended to cross on a 5-point scale. The ratings were aggregated, and a threshold of 0.5 was used to determine the ground truth label. The PIE training set contains 1020 non-crossing and 1674 crossing intention cases,

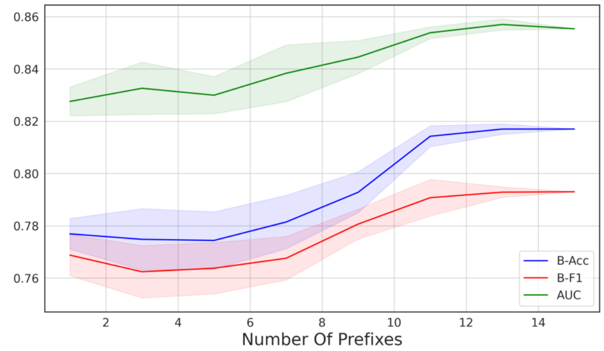

Figure 1. Performance dependence on the number of prefixes.

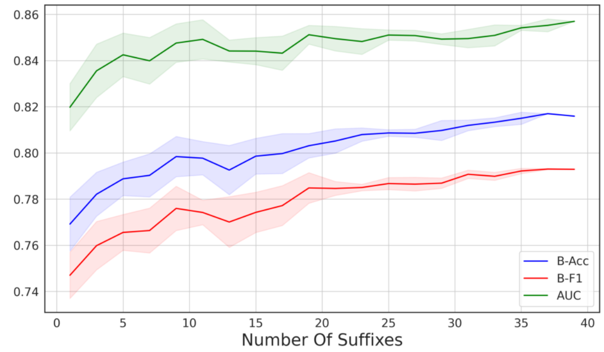

Figure 2. Performance dependence on the number of suffixes.

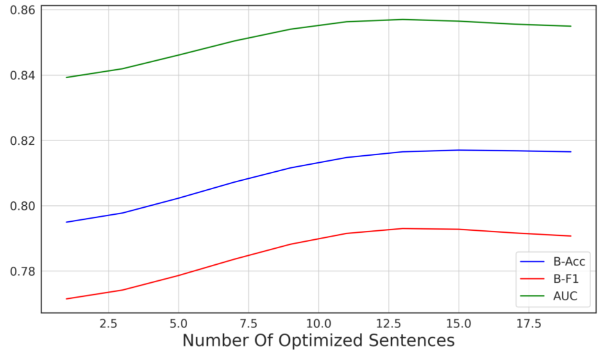

Figure 3. Performance dependence on the number of optimized sentences.

Prefix
a photo of a person
a video of a person
an example of person
a photo of the person
a video of the person
an example of the person
a demonstration of the person
a photo of a person during
a video of a person during
an example of a person during
a demonstration of a person during
a photo of the person during
a video of the person during
an example of the person during
a demonstration of the person

Table 2. List of prefixes

while the test set contains 414 non-crossing and 1386 crossing intention cases. In the PSI dataset, video footage was captured while driving in urban settings at a frame rate of 30 fps. Sequences of 0.5 seconds were extracted from 110 video clips with an overlap ratio of 0.8 and annotated by 24 human drivers to determine whether the pedestrian intended to cross or not. We used a threshold of 0.5 to obtain the ground truth label. The train set includes 2100 examples of non-crossing intentions and 3941 examples of crossing intentions, while the test set has 807 examples of non-crossing intentions and 1975 examples of crossing intentions. The JAAD dataset consists of video footage captured at a frame rate of 30 fps, featuring pedestrians in urban environments from various countries. Our evaluation employed the JAAD$_{\text{all}}$ version [2], encompassing 686 scenarios depicting positive crossing intentions of pedestrians either crossing or about to cross, along with 2100 scenarios depicting instances where pedestrians were far from the road without any intention to cross. The division into training and testing sets was conducted in accordance with the methodology outlined in [2].

5. Evaluation Metrics Details

In this section, we outline the exact calculations of the evaluation metrics utilized in the paper. We initiate by introducing the definitions of the following core classification outcomes:

- **True Positives (TP):** The number of positive instances correctly classified as positive.
- **False Positives (FP):** The number of negative instances incorrectly classified as positive.
- **True Negatives (TN):** The number of negative instances correctly classified as negative.
- **False Negatives (FN):** The number of positive instances incorrectly classified as negative.

The datasets PIE, JAAD and PSI, used for evaluating the performance in this paper, exhibit imbalanced class distributions. This imbalance introduces a significant bias in the accuracy metric when only one class is considered. To mitigate this bias, we calculate accuracy metrics separately for each class and utilize a balanced accuracy metric, along with balanced F1 and the Matthews Correlation Coefficient (MCC). By applying the aforementioned definitions, we computed the subsequent classification metrics:

- **Class 0 accuracy:** The proportion of correctly classified negative instances, calculated as

$$Acc_0 = \frac{TN}{TN + FP}. \tag{1}$$

- **Class 1 accuracy:** The proportion of correctly classified positive instances, calculated as

$$Acc_1 = \frac{TP}{TP + FN}. \tag{2}$$

- **Accuracy of both classes:** The ratio of correctly classified positive and negative instances to all instances, calculated as

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}. \tag{3}$$

- **Balanced accuracy:** The average of Class 0 and Class 1 accuracies, calculated as

$$B_{Acc} = \frac{Acc_0 + Acc_1}{2}. \tag{4}$$

- **F1 score of Class 0:** The harmonic mean of precision and recall for Class 0, calculated as

$$F1_0 = 2 \cdot \frac{Precision_0 \cdot Recall_0}{Precision_0 + Recall_0}, \tag{5}$$

  where $Precision_0 = \frac{TN}{TN+FN}$ and $Recall_0 = \frac{TN}{TN+FP}$.

- **F1 score of Class 1:** The harmonic mean of precision and recall for Class 1, calculated as

$$F1_1 = 2 \cdot \frac{Precision_1 \cdot Recall_1}{Precision_1 + Recall_1}, \tag{6}$$

  where $Precision_1 = \frac{TP}{TP+FP}$ and $Recall_1 = \frac{TP}{TP+FN}$.

- **Balanced F1 score:** The average of F1 scores for Class 0 and Class 1, calculated as

$$B_{F1} = \frac{F1_0 + F1_1}{2}. \tag{7}$$

- **Matthews Correlation Coefficient (MCC):** A metric that provides a balanced measure of classification performance for imbalanced datasets. MCC is calculated using the formula:

$$MCC = \beta\sqrt{Precision_1 \cdot Recall_1 \cdot Precision_0 \cdot Recall_0} \tag{8}$$

  where

$$\beta = \frac{TP \cdot TN - FP \cdot FN}{TP \cdot TN}. \tag{9}$$

  The MCC ranges from -1 to 1, where -1 indicates complete disagreement between predictions and true labels, 0 signifies performance no better than random chance, and 1 represents perfect agreement. Due to its consideration of all four components of the confusion matrix (TP, TN, FP, and FN), MCC is particularly suitable for imbalanced datasets.

- **AUC:** The area under the precision recall curve. The precision and recall refer to $Precision_1$, and $Recall_1$, respectively, which are defined above. These values are computed at different thresholds on the crossing intention classification score that determine positive crossing intention.

References

[1] Tina Chen, Taotao Jing, Renran Tian, Yaobin Chen, Joshua Domeyer, Heishiro Toyoda, Rini Sherony, and Zhengming Ding. Psi: A pedestrian behavior dataset for socially intelligent autonomous car. *arXiv preprint arXiv:2112.02604*, 2021. 2

[2] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1258–1268, 2021. 3

[3] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019. 2

[4] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017. 2