Supplementary Material for Adapting to Length Shift: FlexiLength Network for Trajectory Prediction

Yi Xu Yun Fu
Northeastern University, USA
xu.yi@northeastern.edu, yunfu@ece.neu.edu

1. Implementation Details

We employ the official codes of the AgentFormer model [4], as found in [1], and the HiVT [6] model, as found in [2], to evaluate our proposed framework. We utilize the provided pre-trained models to assess the performance across different observation lengths. Specifically, for the HiVT model, we opt for its smaller variant, HiVT-64, and use the Argoverse 1 validation set. All models are trained on NVIDIA Tesla V100 GPUs, adhering to the same hyperparameters as specified in their respective official implementations.

Training Loss. In our work, we introduce the FlexiLength Network (FLN) framework, designed for easy integration with Transformer-based trajectory prediction models. We demonstrate its application by evaluating it using two models: AgentFormer and HiVT. The AgentFormer model is a two-stage generative model. In both stages, we utilize the output $Y^L \sim \mathcal{D}(\psi^L)$, following their original loss function. Additionally, we incorporate our temporal distillation loss $\mathcal{L}_{kl}$, setting the balance hyperparameter λ to 1. For the HiVT model, which is trained using a combined regression and classification loss function, we also apply the output $Y^L \sim \mathcal{D}(\psi^L)$ in line with its original loss function. Our framework further extends this with our temporal distillation loss $\mathcal{L}_{kl}$, and sets the balance hyperparameter λ to 1.

2. Specialized Layer Normalization Study

In our analysis, we classify the components of typical Transformer-based models for trajectory prediction into several key parts: a Spatial Encoder for extracting spatial features, a Positional Encoder (PE) for embedding positional information, a Transformer Encoder for temporal dependency modeling, and a Trajectory Decoder for generating predicted trajectories. While various designs differ among these components, including Layer Normalization (LN) layers beyond the Transformer Encoder, we investigate these LN layers in different model parts. Our experiments show that the LN shift in the Transformer Encoder is the main cause of the performance drop.

Regarding the AgentFormer model, it incorporates two LN layers in its Transformer Encoder and three in the Trajectory Decoder. We train the AgentFormer model separately for observation lengths of 2, 6, and 8 timesteps (Isolated Training) on the Eth dataset [2, 3]. We then use the same trajectory to pass through these three trained models independently and analyze the LN statistics in the first LN layer of the Trajectory Decoder. The input feature of this layer has a dimension of 20×256, and we chart these values across the 20 dimensions depicted in Fig. 1, as the LN affects the last dimension. Our observations reveal minimal statistical variance at different observation lengths, indicating that feature representations extracted from the same trajectory at varying observation lengths have a very similar statistical structure for subsequent decoding (prediction). We conduct further experiments that apply three additional specialized LN layers in the Trajectory Decoder on the nuScenes dataset [5], detailed in Tab. 1. Applying specialized LN to the first LN layer results in almost the same performance. The addition of two or three specialized LN layers within the Trajectory Decoder shows minimal improvement. Consequently, we decide to implement specialized LN layers only in the Transformer Encoder to balance performance with model complexity.

The HiVT model uses several LN layers in each of its components. Yet, our observations indicate that a notable statistical discrepancy arises primarily in the Transformer Encoder. We conduct additional experiments that involve the use of additional specialized LN layers in different components on the Argoverse 1 [1] validation set, as detailed in Tab. 2. The findings are consistent with our observations from the AgentFormer model. Consequently, we decide to implement only two specialized LN layers in the Temporal Transformer Encoder.

In conclusion, it becomes evident that normalization shifts typically occur within the Transformer Encoder (Temporal Modeling) when the observation lengths are different. This shift is also one of the reasons for the perfor-

[1] https://github.com/Khrylx/AgentFormer
[2] https://github.com/ZikangZhou/HiVT

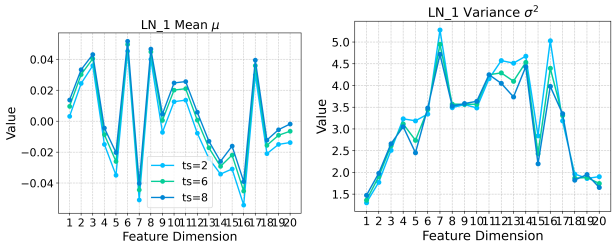

Figure 1. Layer Normalization statistics for the first LN layer of the Trajectory Decoder in the AgentFormer model, trained (Isolated Training) on the Eth dataset with observation lengths of 2, 6, and 8 timesteps separately

Method	Note	$ADE_5/FDE_5 \downarrow$ K = 5 Samples		
		2Ts	3Ts	4Ts
AFormer-IT [4]	-	2.02/4.23	1.93/3.97	1.86/3.89
AFormer-FLN	-	1.92/3.91	1.88/3.89	1.83/3.78
AFormer-FLN	+1 SLN	1.92/3.92	1.87/3.89	1.83/3.78
AFormer-FLN	+2 SLN	1.91/3.91	1.87/3.88	1.83/3.77
AFormer-FLN	+3 SLN	1.91/3.90	1.86/3.88	1.82/3.76

Table 1. Specialized Layer Normalization study in the AgentFormer model on the nuScenes dataset. The term +l SLN refers to applying l extra Specialized Layer Normalization layers within the Trajectory Decoder.

Method	Note	$ADE_6/FDE_6 \downarrow$ K = 6 Samples		
		10Ts	20Ts	30Ts
HiVT-64-IT [6]	-	0.92/1.43	0.81/1.17	0.69/1.04
HiVT-64-FLN		0.81/1.25	0.72/1.08	0.65/0.98
HiVT-64-FLN	+2 SLN-SE	0.80/1.24	0.72/1.08	0.65/0.98
HiVT-64-FLN	+4 SLN-TD	0.81/1.24	0.71/1.07	0.64/0.97

Table 2. Specialized Layer Normalization study in the HiVT-64 model using the Argoverse 1 validation set. The term +2 SLN-SE denotes the addition of 2 Specialized Layer Normalizations in the Agent-Agent Interaction module within the HiVT's Spatial Encoder, while +4 SLN-TD refers to the addition of 4 additional Specialized Layer Normalizations in the Trajectory Decoder.

mance drop. This finding is consistent with the empirical results discussed in the main section of the paper.

3. Validation of Normalization Shift

To further validate that our FlexiLength Network (FLN) can alleviate the Normalization Shift problem, we analyze the Layer Normalization (LN) statistics between Isolated Training (IT) and FLN, both trained using the AgentFormer model on the nuScenes dataset. We use an identical trajectory with 3 agents over an observation period of 4 timesteps and process it through both IT and FLN models separately. The first LN layer in these models receives an intermediate input feature with dimensions of 12×256. Since LN operates along the last dimension, the statistical outcomes,

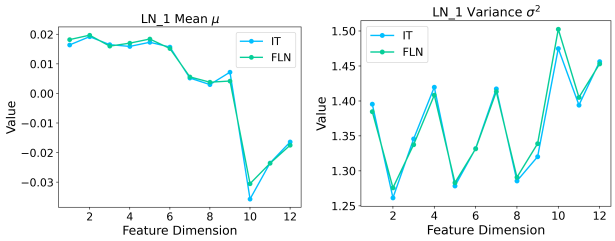

Figure 2. Layer Normalization statistics for the first layer of the Transformer Encoder in the AgentFormer model, with IT trained at the observation length of 4 timesteps. Both IT and FLN are evaluated at the same observation length of 4 timesteps.

Method	Dataset	$ADE_{20}/FDE_{20} \downarrow$ K = 20 Samples		
		2 Ts	6 Ts	8 Ts
AFormer [4]	Eth	0.661/0.966	0.640/0.946	0.451/0.748
AFormer-IT		0.467/0.757	0.452/0.757	0.451/0.748
AFormer-FLN		**0.450/0.742**	**0.432/0.730**	**0.411/0.721**
AFormer [4]	Hotel	0.225/0.349	0.166/0.277	0.142/0.225
AFormer-IT		0.161/0.276	0.148/0.242	0.142/0.225
AFormer-FLN		**0.153/0.248**	**0.138/0.232**	**0.124/0.210**
AFormer [4]	Univ	0.341/0.538	0.275/0.475	0.254/0.454
AFormer-IT		0.263/0.478	0.251/0.458	0.254/0.454
AFormer-FLN		**0.251/0.457**	**0.244/0.447**	**0.232/0.430**
AFormer [4]	Zara1	0.250/0.412	0.212/0.347	0.177/0.304
AFormer-IT		0.184/0.319	0.179/0.310	0.177/0.304
AFormer-FLN		**0.178/0.308**	**0.162/0.300**	**0.160/0.288**
AFormer [4]	Zara2	0.190/0.312	0.178/0.286	0.140/0.236
AFormer-IT		0.140/0.238	0.142/0.241	0.140/0.236
AFormer-FLN		**0.131/0.230**	**0.131/0.221**	**0.128/0.217**

Table 3. Comparison with baseline models on the ETH/UCY dataset, evaluated using the ADE_{20}/FDE_{20} metric. The best results are highlighted in bold.

depicted in Fig. 2, are presented along the 12-dimensional axis. The alignment of the two curves indicates that the statistical values are quite similar, demonstrating the effectiveness of FLN in mitigating the normalization shift problem.

4. Quantitative Results

In the Experiments section, we present the performance of our proposed FlexiLength Network (FLN) on the ETH/UCY dataset through various figures. Additionally, we include corresponding quantitative results in Tab. 3 for further reference. It is evident that our FLN consistently surpasses the performance of Isolated Training (IT) across different observation lengths in all five datasets.

5. Visualizations

In Fig. 3, we present trajectory prediction visualizations using the AgentFormer model on the nuScenes dataset. These visualizations showcase the same trajectory but with different observation lengths. We focus on a single agent and

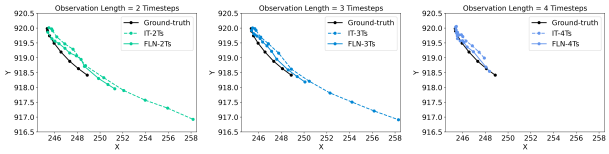

Figure 3. Visualizations of trajectory predicted by Isolated Training (IT) and Our FlexiLength Network (FLN).

maintain the figure size for easier comparison. These visualizations clearly demonstrate that our FlexiLength Network (FLN) outperforms Isolated Training (IT) across various observation lengths, confirming the effectiveness of FLN in handling inputs with differing observation lengths.

References

[1] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 1

[2] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, pages 655–664. Wiley Online Library, 2007. 1

[3] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc J. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–268, 2009. 1

[4] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9813–9823, 2021. 1, 2

[5] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 1

[6] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 1, 2