Supplementary Material for HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction

1. Additional Implementation Details

Preprocessing of Trajectory Data. Each data sample in the Argoverse dataset has 20 past time steps and 30 future time steps. An agent i's geometric attribute at time step t is given by the difference between the agent's positions at time step t and $t-1$, *i.e.*, $\mathbf{p}_i^t - \mathbf{p}_i^{t-1}$. Due to the occlusion and the limited range of perception, an agent may be invisible at some time steps. If $\mathbf{p}_i^t$ is invisible, the geometric attribute of agent i at time step t is padded with zero. To deal with the missing data, we introduce a learnable "padding token" for each of the historical time steps in the temporal learning module, which replaces the spatial embedding if the agent's geometric attribute is padded with zero at the corresponding time step. In the agent-agent interaction module, we further introduce learnable tokens for each time step to handle the case in which $\mathbf{p}_i^t$ is visible while $\mathbf{p}_i^{t-1}$ is invisible.

Preprocessing of Map Data. Each lane in the Argoverse dataset consists of 10 points and is split into 9 lane segments. For a lane segment ξ, the geometric attribute is given by $\mathbf{p}_\xi^1 - \mathbf{p}_\xi^0$, where $\mathbf{p}_\xi^0 \in \mathbb{R}^2$ and $\mathbf{p}_\xi^1 \in \mathbb{R}^2$ are the starting and the ending coordinates of ξ. The semantic attributes of a lane segment include the information about whether it is at an intersection, whether it belongs to a left-turn lane or a right-turn lane, and whether it has traffic control.

Network Architecture. All MLPs for embedding the input vectors have 3 layers. The MLP blocks in the Transformer encoders have 2 layers, which first increase the dimension from d_h to $4 \times d_h$ and then reduce the dimension back to d_h. The embedding $\mathbf{e}_{ij}$ in the global interaction module is shared across all layers. The activation function of the intermediate layers is ReLU. Besides, the decoder uses $\mathrm{ELU}(\cdot) + 1 + \epsilon$ as the activation function to produce positive uncertainty values, where ϵ is set to $1e-3$.

Classification Loss. We use the cross-entropy loss as the classification loss to optimize the mixing coefficients, where the ground-truth coefficient of component f for agent i is empirically defined by

$$\boldsymbol{\pi}_{if} = \frac{\exp\left(-\frac{1}{H}\sum_{t=T+1}^{T+H}\left\|\mathbf{R}_i^\top(\mathbf{p}_i^t - \mathbf{p}_i^T) - \boldsymbol{\mu}_{i,f}^t\right\|_2\right)}{\sum_{f'=1}^{F}\exp\left(-\frac{1}{H}\sum_{t=T+1}^{T+H}\left\|\mathbf{R}_i^\top(\mathbf{p}_i^t - \mathbf{p}_i^T) - \boldsymbol{\mu}_{i,f'}^t\right\|_2\right)} \tag{1}$$

without any trial and error.

2. Additional Qualitative Results

Figure 1 displays more qualitative results of HiVT-128 on the Argoverse validation set. Overall, the predictions produced by our model are map-compliant. In some simple scenarios, our model reasonably predicts trajectories with different speed profiles. In the cases where there are intersections, our model can produce multimodal predictions that cover different possible intentions of the agents. In the scenarios where complex interactions exist between agents, the predicted trajectories of multiple agents are scene-consistent and non-overlapping. These results demonstrate the effectiveness of HiVT.

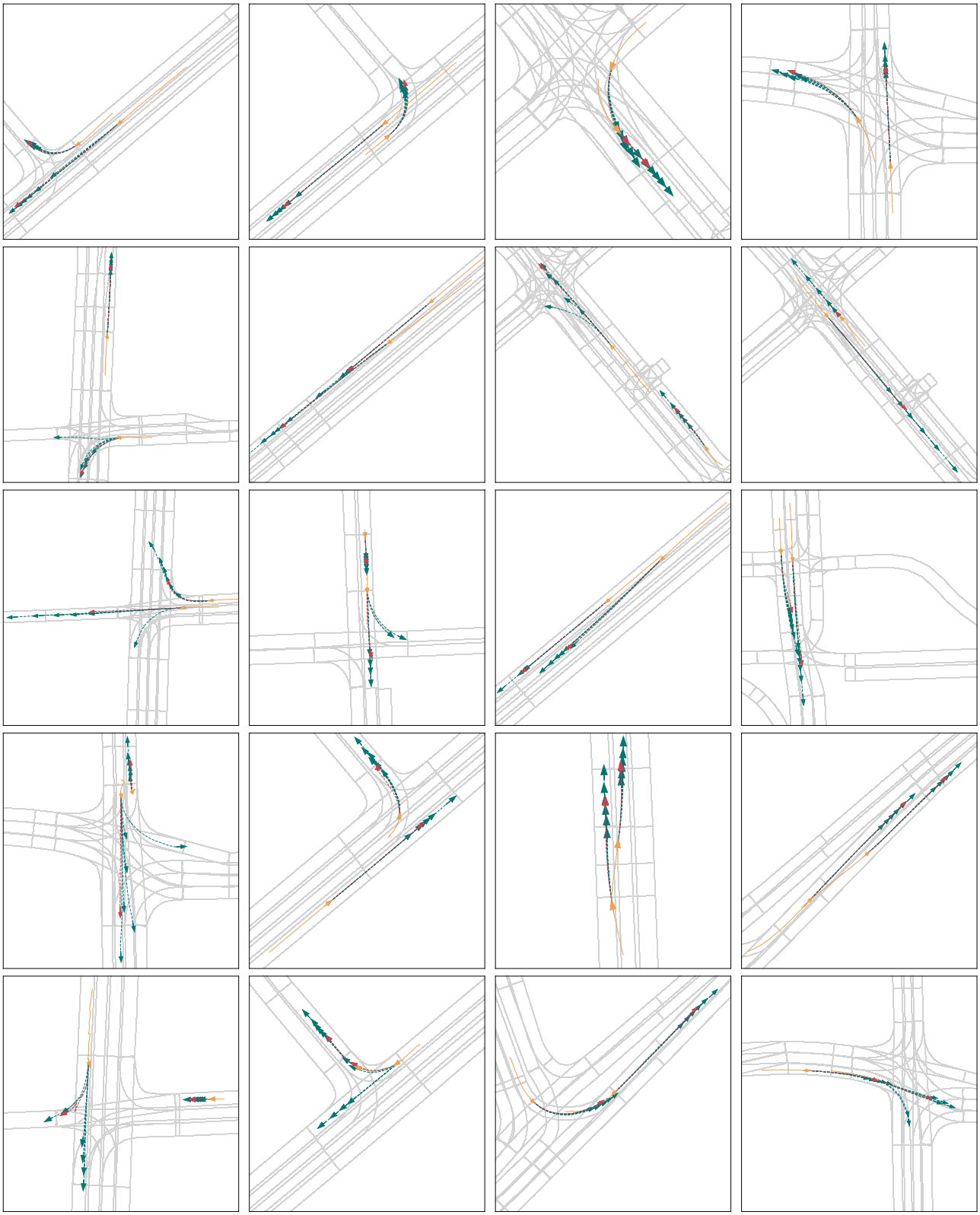

Figure 1. Qualitative results of HiVT-128. The past trajectories are shown in yellow, the ground-truth trajectories are shown in red, and the predicted trajectories are shown in green.