StreamMOTP: Streaming and Unified Framework for Joint 3D Multi-Object Tracking and Trajectory Prediction

Jiaheng Zhuang[1] , Guoan Wang[2], Siyu Zhang[2], Xiyang Wang[2] ,
Hangning Zhou[2] , Ziyao Xu[2], Chi Zhang[2], and Zhiheng Li[1,3] ⋆

[1]Tsinghua University [2]Mach Drive
[3] Tsinghua Innovation Center in Zhuhai

Abstract. 3D multi-object tracking and trajectory prediction are two crucial modules in autonomous driving systems. Generally, the two tasks are handled separately in traditional paradigms and a few methods have started to explore modeling these two tasks in a joint manner recently. However, these approaches suffer from the limitations of single-frame training and inconsistent coordinate representations between tracking and prediction tasks. In this paper, we propose a **stream**ing and unified framework for joint 3D **M**ulti-**O**bject **T**racking and trajectory **P**rediction (**StreamMOTP**) to address the above challenges. Firstly, we construct the model in a streaming manner and exploit a memory bank to preserve and leverage the long-term latent features for tracked objects more effectively. Secondly, a relative spatio-temporal positional encoding strategy is introduced to bridge the gap of coordinate representations between the two tasks and maintain the pose-invariance for Trajectory prediction. Thirdly, we further improve the quality and consistency of predicted trajectories with a dual-stream predictor. Extensive experiments conducted on the popular nuScenes dataset show the effectiveness and superiority of StreamMOTP, which outperforms previous methods significantly on both tasks. Furthermore, the proposed framework has great potential and advantages in actual applications of autonomous driving.

Keywords: 3D MOT · Trajectory Prediction · Streaming Framework

1 Introduction

In autonomous driving systems, 3D Multi-Object Tracking (MOT) [2,12,19,22, 26,27,34] and trajectory prediction [6,8–10,14,17,18,36–38] are two crucial tasks which play a vital role in ensuring the driving performance of ego-vehicle. Obviously, high-precision tracking can provide a more solid foundation for prediction, and in turn, accurate predictions can enhance the effectiveness of tracking. As depicted in Fig.1 (a), the two tasks are executed one after another in current mainstream pipelines of autonomous driving. Although this paradigm has

---
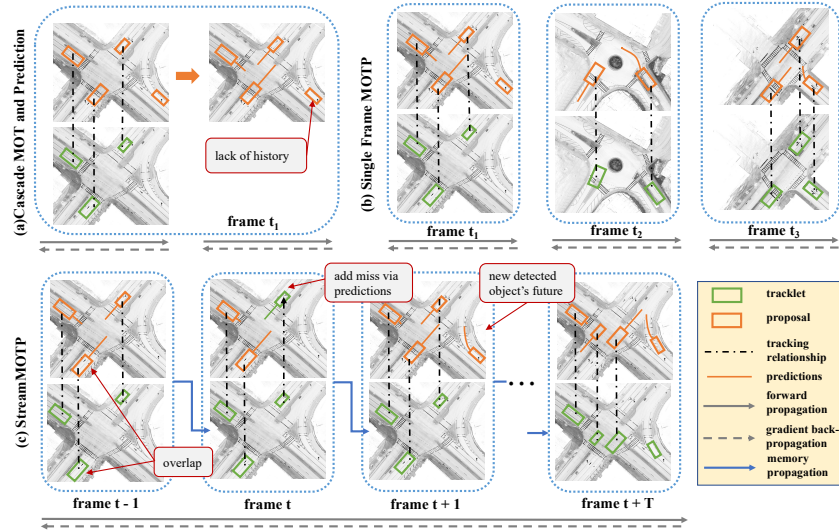⋆ Corresponding author: zhhli@mail.tsinghua.edu.cn

Fig. 1: Different pipelines for the tasks of multi-object tracking and trajectory prediction in autonomous driving. **(a)** Cascade paradigm, where the two tasks are performed separately with non-differentiable transitions. **(b)** Joint single-frame paradigm, where the two tasks are performed jointly in a parallelized framework per frame. **(c)** The proposed StreamMOTP, where the memory, feature, and gradient are propagated across consecutive frames to enhance the long-term modeling ability and temporal coherence.

achieved some success, the separated processing flow can not fully exploit the potential complementarity between the tasks of tracking and prediction, since it suffers from information loss, feature misalignment, and error accumulation across modules [25]. Despite some methods [24, 28, 33] attempt to integrate the two tasks as shown in Fig.1 (b), some limitations and problems have still not been well explored: (1) the tasks of multi-object tracking and trajectory prediction are both executed in a streaming manner in actual deployments, while the training procedure of most previous methods is conducted in a snap-shot pattern, where the length of historical window is fixed and the long-term information can not be fully exploited efficiently. (2) In general, the coordinates representation of objects for tracking and prediction are different, where a unified coordinate system is needed in MOT for optimal association while most prediction methods adopt the agent-centric coordinate representation for each object to ensure pose-invariance. (3) Most methods focus on predicting the future trajectories of objects visible in current frame, inadvertently overlooking those lost because of either occlusions or miss from upstream perception, which may result in adversely affecting downstream tasks.

In this paper, we introduce **StreamMOTP**, a streaming framework for joint multi-object tracking and trajectory prediction as depicted in Fig.1 (c), where the tasks of MOT and trajectory prediction are jointly performed on successive frames. Specifically, we associate the newly perceived objects with historical

tracklets and predict their future trajectories simultaneously. Different from previous works, the extracted latent features of objects are sequentially utilized in StreamMOTP as part of the representation for the subsequent tracked objects during the forward propagation phase. As for the back-propagation, the gradients are not confined to a single frame but are propagated through multiple frames, which greatly narrows the gap between training and online inference, allowing for a more comprehensive learning process by accounting for temporal dependencies across the entire sequence.

Concretely, we extend the pattern of training from single-frame to multi-frame and introduce a memory bank to maintain and update long-term latent features for tracked objects, thereby improving the model's capability for long-term sequence modeling. Aiming to address the coordinate system discrepancy between the tasks of tracking and prediction, we propose a relative Spatio-Temporal Positional Encoding (STPE) strategy, which is applied to realize the compromise and unification of the different agent- and ego-centric representation in the two tasks. At the same time, based on the observation that there is an obvious overlap between the predicted trajectories of objects in consecutive adjacent frames as depicted in Fig.1 (c, frame $t-1$), we apply dual-stream predictor to effortlessly and elegantly generate future trajectories for both tracked and new-come objects simultaneously, which benefits to both tasks of MOT and trajectory prediction.

It should be pointed out that, with the design of the streaming and unified framework, StreamMOTP obtains the potential and advantages to handle more complex driving scenarios in actual applications. On the one hand, the predicted trajectories for tracked objects could help deal with the problem of occlusions at the current moment by marking the possible positions of obscured targets in the current frame, as shown in Fig.1 (c, frame t). On the other hand, for the objects newly perceived in the current frame, StreamMOTP maintains the capability to predict their future trajectories by leveraging social interactions and contextual features stored in the memory bank while traditional prediction methods may fail due to the lack of historical information about them, as shown in Fig.1 (c, frame $t+1$).

The core contributions are summarized as follows:

- We propose StreamMOTP, a joint MOT and trajectory Prediction model based on a streaming framework to bridge the gap between training and real-world deployment. A memory bank for tracked objects is introduced in this framework for utilizing long-term features more effectively.
- We introduce a spatio-temporal positional encoding strategy to construct the relative relationship between objects in different frames, which reaches the compromise and unification of inconsistent coordinate representation in tracking and prediction.
- We design a dual-stream predictor to simultaneously predict object trajectories in the current and previous frames. The predicted trajectory from the previous frame can further assist in predicting newly perceived objects' trajectories, improving temporal consistency in trajectory prediction.

- We get better performance for 3D MOT and trajectory prediction on nuScenes, improving AMOTA / MOTA by 3.84% / 8.44% and reducing minADE / minFDE by 0.220 / 0.141.

2   Related Work

2.1   3D Multi-Object Tracking

Existing multi-object tracking paradigms, such as tracking-by-detection (Deep-SORT [29], AB3DMOT [26]), Joint Detection and Embedding learning (Fair-MOT [35], JDE [23]), and joint detection and tracking (Tracktor++ [4], YONDT-MOT [22]), typically rely on Kalman filters(KFs) to predict the positions of tracked objects for better-association. Yet, KFs require fine-tuning of parameters and struggle with occlusions (PC3TMOT [30], DeepFusionMOT [21]). In contrast, dedicated prediction tasks can provide superior short-term prediction results for tracking, especially in handling complex scenarios such as occlusions. Therefore, combining the two tasks of multi-object-tracking and trajectory prediction can effectively improve the overall performance of multi-object tracking. This combination not only reduces the dependence on traditional methods like KFs but also enhances the robustness and adaptability of the tracking methods.

2.2   Trajectory Prediction

There has been significant progress in trajectory prediction recently. With the use of pooling [10], graph convolution [14], attention mechanism [17] [38], vector-based methods [8] can efficiently aggregate sparse information in traffic scenes. As the future is uncertain, some works (Multipath++ [18], HiVT [39], Wayformer [15]) predict multimodal future distribution by decoding a set of trajectories from scene context while others (DenseTNT [9], MTR [17], PiH [11]) generate multimodal prediction by leveraging anchors. Though these methods greatly improve trajectory prediction, most of them use GT past trajectories as input for training and testing, neglecting tracking error accumulation with imperfect inputs. Therefore, we handle the tasks of tracking and prediction jointly with no need for GT trajectories as predictor's inputs to provide more robust predictions based on practical detectors in the real world.

2.3   Joint Tracking and Prediction

Recent years have seen increasing interest in joint tracking and prediction. For example, [32] refine the inputs for the predictor through a re-tracking module, MTP [25] propose multi-hypothesis data association to create multiple track sets for predictor simultaneously. Besides polishing the input tracklets for the prediction module, some studies combine the tasks of tracking and prediction with joint optimization. PTP [28] and PnPNet [13] uses the shared feature representation to address both tasks. AffinPred [24], TTFD [33] use affinity matrices instead
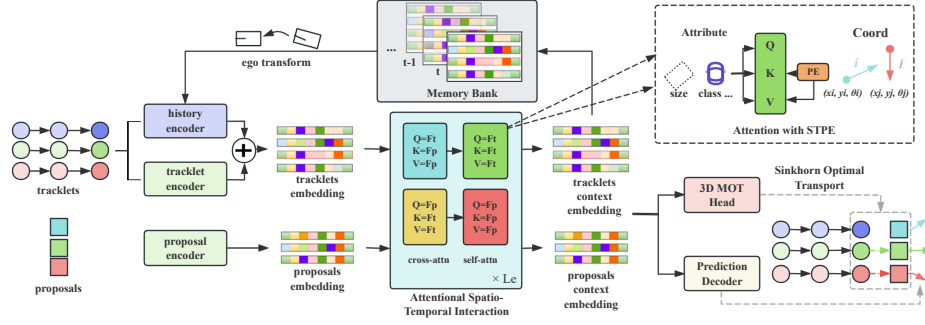
Fig. 2: Overview of StreamMOTP. Tracklets and proposals denote the previous frame trajectories and the current frame detections respectively. The model first performs Attentional Spatio-Temporal Interaction, which is based on attention with STPE, to get context features. The tasks of tracking and prediction are then performed based on those context features. Current context features and tracking results are updated in memories at each time step.

of tracklets as inputs of the prediction module to enhance forecasting, but they sacrifice the capability to provide tracking results explicitly. However, all of these methods are performed in a snap-shot form and neglect the misalignment issue between tracking and prediction. Compared to those approaches, our method uses a streaming framework and a unified spatio-temporal positional encoding strategy to address the above problems.

3   Approach

3.1   Streaming Framework

Simply, let $\mathcal{D} = \{d_1, \ldots, d_N\}$ represent the set of objects perceived in the current frame from a 3D object detector, where N denotes the number of objects. Concretely, each object at frame t is represented as $d_i^t = [d_i^{\text{pos},t}, d_i^{\text{size},t}, d_i^{\text{head},t}, d_i^{\text{class},t}, d_i^{\text{score},t}]$ where each element denotes the position, size, heading angle, class and confidence score from the module of detection, respectively. In this paper, the goal of joint 3D multi-object tracking and trajectory prediction includes two parts, to obtain the association of multiple obstacles in adjacent frames by assigning a unique track ID to each object, and meanwhile to predict the future trajectories $\mathcal{F} = \{f_1, \ldots, f_N\}$ for all agents in the current frame, with each element of a trajectory specified by a two-dimensional coordinate (x, y).

Based on the observation that the actual physical world is continuous and long-term history is essential for a safer autonomous driving system, we model the task of joint 3D multi-object tracking and trajectory prediction in a streaming manner (shown as Fig. 2). First of all, we extend the pattern of training from single-frame to multi-frame so as to narrow the gap between training and actual deployment. To be more specific, we introduce a Memory Bank for tracked objects to maintain long-term latent features for utilizing the long-term information

more effectively, where the latent features are maintained through consecutive frames and could further benefit the performance of both tasks, including not only multi-object tracking but also trajectory prediction.

Specifically, the memory bank consists of $F \times N$ latent features where F is the length of the memory bank and N is the number of objects stored per frame. At each time, the latent feature of those tracked objects that have been associated with the new perceived objects in the current frame would be saved into the memory bank. These features are then utilized in subsequent frames to enhance features for tracked objects, detailed in Sec. 3.2. The entrance and exit of the memory bank follow the first-in, first-out rule.

3.2   Spatio-Temporal Encoder

Feature Extraction. To capture the semantic and motion information of the obstacles in the driving scenario efficiently and adequately, we conduct feature extraction for the tracked and new-come objects separately. For the perceived objects from adjacent frames, we use $d \in \mathbb{R}^{N_p \times C}$ and $\tau \in \mathbb{R}^{N_t \times C}$ to represent the semantic features, where N_p and N_t denote the number of objects at current frame t (named as *proposals*) and previous frame $t - 1$ (named as *tracklets*), respectively. At the same time, the historical trajectories of last T_h frames for N_t tracked objects are represented with $H \in \mathbb{R}^{N_t \times T_h \times C}$. Simply and effectively, we deploy the Multi-Layer Perceptron (MLP) to encode the semantic information into high-dimension features and fuse the historical data H to *tracklets* τ through a Multi-Head Cross Attention (MHCA) as:

$$F_d = \mathrm{MLP}(d), \tilde{F}_t = \mathrm{MLP}(\tau) + \mathrm{MHCA}(\mathrm{MLP}(H)) \tag{1}$$

where $\tilde{F}_t \in \mathbb{R}^{N_t \times D}$, $F_p \in \mathbb{R}^{N_p \times D}$, and C, D correspond to the dimension of the semantic and latent high-dimension features respectively.

Additionally, to equip our model with long-term temporal modeling capability, we exploit the latent features saved in the memory bank. Inspired by dynamic weight learning [1] [20] , an ego transformation is applied to ensure the temporal alignment and effective feature usage across frames:

$$\begin{aligned} \alpha, \beta &= \mathrm{MLP}(E_t - E_s) \\ M &= \alpha \, \mathrm{LN}(\tilde{M}) + \beta \end{aligned} \tag{2}$$

where Eq.2 is an affine transformation that converts latent feature from past frame $\tilde{M}$ to current time M. The parameters α, β are derived from the ego difference between two frames. E_s, E_t denotes the frames where the latent feature is saved and utilized. LN stands for layer normalization. Then we apply temporal aggregation of long-term latent memory maintained in the memory bank for each tracked object with MHCA and then fuse the latent memory feature with the extracted feature of *tracklets* as follows:

$$F_t = \tilde{F}_t + \mathrm{MHCA}(M) \tag{3}$$

Spatio-Temporal Positional Encoding. For the task of tracking, aligning all features within a unified coordinate system is essential for feature association. In contrast, for prediction tasks, previous research [18] [39] have demonstrated the advantages of agent-centric representations, which normalize various trajectories to local coordinate systems centered on the selected agent. To bridge the gap between coordinate representation between tracking and prediction, we propose a relative **S**patio-**T**emporal **P**ositional **E**ncoding (**STPE**) strategy. This approach differentiates between coordinate-independent and dependent features, using the former as query tokens for attention mechanism during feature interaction, while the latter is incorporated into attention through relative positional encoding. To be specific, we encode the relative spatio-temporal position between object i in previous frame t (*tracklet frame*) and object j in current frame p (*proposal frame*) as follows:

$$\delta_{ij}^{tp} = \text{MLP}([s_j^p - s_i^t, \theta_j^p - \theta_i^t]) \tag{4}$$

where s denotes the coordinate (x, y) of the object, and θ denotes the heading rate, both of which are coordinate-dependent features. We calculate the relative difference between two objects at different frames and embed it into δ_{ij}^{tp}.

Attentional Spatio-Temporal Interaction. Based on relative embedding from the spatio-temporal positional encoding strategy, we fuse the features of *proposals* and *tracklets* with cross-attention and self-attention iteratively. Take the proposal branch as an example, we use query-centric attention with a spatio-temporal positional encoding strategy, incorporating the relative positional embedding into the key/value of the attention mechanism.

$$\tilde{F}_i^p = \text{MHCA}\left(\mathbf{Q} = F_i^p, \mathbf{K}/\mathbf{V} = \{F_j^t + \delta_{ij}^{tp}\}_{j \in N_i}\right)$$
$$F_i^{p\prime} = \text{MHSA}\left(\mathbf{Q} = \tilde{F}_i^p, \mathbf{K}/\mathbf{V} = \{\tilde{F}_j^p + \delta_{ij}^{p\prime}\}_{j \in N_i}\right) \tag{5}$$

As shown in Eq. 5, F_i^p denotes the coordinate independent embedding of object i in current frame p, δ_{ij}^{tp} refers to the relative embedding calculated in Eq. 4. We first employ cross-attention to fuse the tracked objects' information from previous frames into new-come objects from current frames. Subsequently, self-attention is utilized within the current frame to foster awareness among detected objects in this frame. This process enables the features of newly perceived objects to incrementally assimilate comprehensive information, enriching their contextual awareness. We denote the result of this branch as *proposals* context feature $F_i^{p\prime}$. Similarly, The tracklet branch undergoes the same propagation and gets *tracklets* context feature $F_i^{t\prime}$ in parallel as Eq. 6:

$$\tilde{F}_i^t = \text{MHCA}\left(\mathbf{Q} = F_i^t, \mathbf{K}/\mathbf{V} = \{F_j^p + \delta_{ij}^{pt}\}_{j \in N_i}\right)$$
$$F_i^{t\prime} = \text{MHSA}\left(\mathbf{Q} = \tilde{F}_i^t, \mathbf{K}/\mathbf{V} = \{\tilde{F}_j^t + \delta_{ij}^t\}_{j \in N_i}\right) \tag{6}$$

3.3   MOT Head

Association with Optimal Transport. The core purpose of the MOT head for StreamMOTP is to associate the M-tracked objects in the previous frame and the N-perceived objects in the current frame. To find the association relationship, we learn an affinity matrix $A^{(\text{tp})} \in \mathbb{R}^{N_t \times N_p}$ based on the *tracklets* context feature and *proposals* context feature after feature interaction. We use the dot product to calculate the similarity score, where each entry $A_{ij}^{(\text{tp})}$ represents the similarity between tracked object i and detected object j.

$$A_{ij}^{(\text{tp})} = \frac{\langle F_i^{t'}, F_j^{p'} \rangle}{\sqrt{D}}, \forall (i,j) \in N_t \times N_p \tag{7}$$

where D is the dimension of the context feature.

Given the affinity matrix, we get the optimal affinity matrix $A^{(\text{opt})} \in \mathbb{R}^{(N_t+1) \times (N_p+1)}$ through **log sinkhorn algorithm** as SuperGlue [16], which performs differentiable optimal transport in log-space for stability. Under our streaming framework, the use of the log sinkhorn algorithm allows the model to modify the model parameters of previous frames while optimizing subsequent frames for continuous tracking and prediction. The last row and the last column of $A^{(\text{opt})}$ respectively represent newly appeared objects and tracklets without corresponding matched objects.

Tracking Loss. We supervise the output affinity matrix $A^{(\text{opt})}$ with the ground truth (GT) relationship represented by the matrix $A^{(\text{g})} \in \mathbb{R}^{(N_t+1) \times (N_p+1)}$. The accuracy of $A^{(\text{opt})}$ is judged by how closely its high-value elements align with the ones in $A^{(\text{g})}$. Therefore, we use the following loss:

$$\mathcal{L}_{\text{tracking}} = -\frac{1}{N_m} \cdot (A^{(\text{opt})} e^{-U} + U) \cdot A^{(\text{g})} \tag{8}$$

where the uncertainty matrix $U \in \mathbb{R}^{(N_t+1) \times (N_p+1)}$ is derived from *tracklets* and *proposals* feature using MLP and max operation, ensuring the robustness of training [?]. e^{-U} denotes the exponentiation of each element in $-U$. N_m is the number of matching pairs in $A^{(g)}$. Finally, we get association relationship A from $A^{(\text{opt})}$.

3.4   Dual-Stream Predictor

The predictor predicts all agents' multi-modal future trajectories. The detail of the predictor is shown in Fig. 3.

Single Frame Prediction. To jointly predict all future trajectories for perceived objects in the current frame, we utilize a transformer-based decoder that incorporates the previous encoded context feature by learnable intention queries. To combine the advantages of the prior acceleration of convergence provided by the anchor-based model [17] and the high flexibility of the anchor-free model [18], we combine learnable tokens and anchors to form the query:
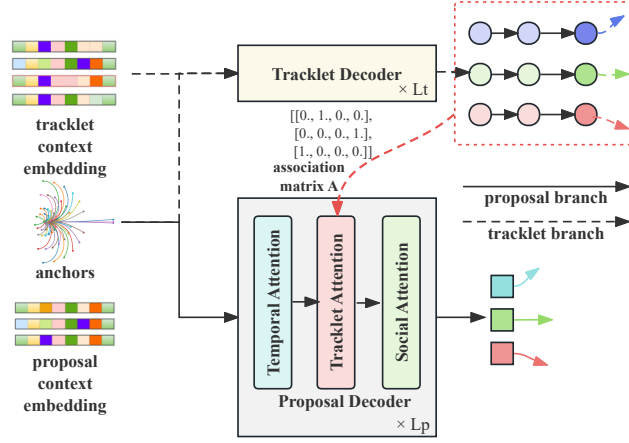
Fig. 3: Overview of dual-stream predictor. Two branches predict the previous frame trajectories and the current frame detections simultaneously, The streaming connection between consecutive frames smooth the predicted trajectories.

$$Q_p^l = I + \phi(A_T) + \phi(\hat{x}_T^{l-1}) \tag{9}$$

where $Q_p^l \in \mathbb{R}^{N_p \times K \times D}$ is the query input at the current frame and layer l decoder, which is combined from a learnable embedding I, the endpoints of the anchors A_T, and the predicted endpoints of previous layer $\hat{x}_T^{l-1}$, which are fused through ϕ (a sinusoidal position encoding followed by an MLP). Next, to aggregate features from context embedding, we perform attention mechanism on the temporal and social dimensions to get multi-modal prediction output.

Dual-Stream Predictor. Predictions for previously tracked and currently perceived objects largely overlap on matched objects. As shown in Fig.4, the T_f+1 predictions from frame t-1 should be consistent with the T_f predictions from current frame t. Besides, it's much more feasible to generate consecutive output trajectories with the streaming nature of the proposed framework of StreamMOTP. Based on the observations, we propose a dual-stream predictor. The predictor comprises two branches: a primary branch focuses on making predictions for the detected objects in current frame and a supportive auxiliary branch for the previous tracked objects. The primary branch follows *Single Frame Prediction* to predict from the context features of *proposals*, while the auxiliary branch leverages the context features of *tracklets* to generate K adaptive predictions $\hat{Y}_t \in \mathbb{R}^{N_t \times K \times (T_f+1) \times 2}$ specific to the tracked objects. Since the prediction result $\hat{Y}_t$ from the *tracklet frame* and $\hat{Y}_p$ from the *proposal frame* have T_f overlapping, using $\hat{Y}_t$ to guide the prediction of $\hat{Y}_p$ enhances both accuracy and temporal coherence of the predicted trajectory. Specifically, we encode and map the overlapping T_f frame of $\hat{Y}_t$ to yield auxiliary features $F_{Y_t} \in \mathbb{R}^{N_p \times K \times T_f \times D}$:

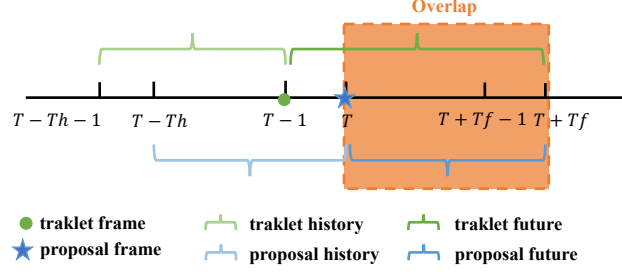$$F_{Y_t} = \mathrm{MLP}(\mathrm{PE}(A^T \hat{Y}_t)) \tag{10}$$

Fig. 4: The idea of temporal consistency between consecutive frames, where the consistency of the overlap is beneficial for aligning trajectories for continuity and stability.

where $\mathrm{PE}(\cdot)$ denotes sinusoidal position encoding, $A \in \mathbb{R}^{N_t \times N_p}$ denotes the association matrix given by MOT head.

In addition to *Single Frame Prediction*, auxiliary features and anchor queries from the current frame are aggregated together in our dual-stream predictor. We adopt cross attention, taking current anchor embedding as query, and the prediction features from the auxiliary tracklet branch as key and value:

$$Q = \mathrm{MHCA}(\mathbf{Q} = Q, \mathbf{K}/\mathbf{V} = F_{Y_t}) \tag{11}$$

We place Eq. 11 after the interaction between queries and *proposals* context features, while before the self-attention of the queries, making the queries interact sequentially with historical features, future features, and the social context.

Multi-modal Prediction Loss with Gaussian Mixture Model. As the future behaviors of the agents are highly multi-modal, we follow [18] to represent the distribution of predicted trajectories with Gaussian Mixture Model (GMM):

$$f\left(\{\mathbf{Y}_i^t\}_{t=1}^{T_f}\right) = \sum_{k=1}^{K} p_{i,k} \prod_{t=1}^{T_f} \mathrm{GMM}\ \left(\mathbf{Y}_i^t \mid \boldsymbol{\mu}_{i,k}^t, \sigma_{i,k}^t\right) \tag{12}$$

where $\{p_{i,k}\}_{k=1}^{K}$ is the probability distribution between K modes, and the $k-$th mixture component's Gaussian density for agent i at time step t is parameterized by $\mu_{i,k}^t$ and $\sigma_{i,k}^t$. Given Eq. 12 for all predicted steps, we adopt negative log-likelihood loss and supervised predictions for new-come objects in the current frame and for the tracked objects simultaneously. Loss can be formulated as:

$$\mathcal{L}_{\mathrm{prediction}} = -\log f(\hat{Y}_p) - \log f(\hat{Y}_t) \tag{13}$$

Then, the final loss of our model is denoted as:

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{tracking}} + \mathcal{L}_{\mathrm{prediction}} \tag{14}$$

where $\lambda \in \mathbb{R}_{>0}$ is the weight for tracking loss to balance the joint optimization of the two tasks.

Table 1: Comparison with existing approaches (on nuScenes). All results are based on detections from Megvii and use only Lidar as input.

(a) 3D MOT Result			**(b)** One Step MOTP Result			**(c)** Multi Step MOTP Result		
Methods	AMOTA ↑	MOTA ↑	Methods	minADE ↓	minFDE ↓	Methods	minADE ↓	minFDE ↓
mmMOT [34]	23.93	19.82	Social-GAN [10]	1.794	2.850	PTP [28]	2.320	3.819
GNN3DMOT [27]	29.84	23.53	TraPHic [6]	1.827	2.760	MTP(S=10) [28]	1.585	2.512
AB3DMOT [26]	39.90	31.40	Graph-LSTM [7]	1.646	2.445	MTP(S=200)	1.325	1.979
PTP [28]	42.36	32.06	PTP [28]	1.017	1.527	AffinPred [24]	0.977	1.628
StreamMOTP	**46.30**	**40.50**	StreamMOTP	**0.810**	**1.481**	StreamMOTP	**0.757**	**1.487**

4   Experiments

4.1   Experimental Setup and Implementation Details

Dataset and Metrics. The proposed method is evaluated using the widely recognized nuScenes dataset. Following the standard practices of nuSences dataset [5], we predict trajectories for objects perceived in the current frame and use the distance threshold of 2m to match them with GT future trajectories. In the task of trajectory prediction, the models predict future trajectories for 3s and 6s, with a time interval of 0.5s, based on 2s historical data, to align with other works. As for the task of MOT, We employ the commonly used AMOTA, MOTA, and AMOTP for evaluation. Standard minADE and minFDE metrics are used to evaluate the prediction performance. Moreover, we design the metric of '**tc**' to evaluate the temporal consistency, which is calculated as the ADE in $T_f - 1$ overlapping frames between predictions from T to $T + T_f$ and predictions from $T - 1$ to $T + T_f - 1$.

 Inputs. In StreamMOTP, input data is formatted in a sequential format. During training, we split the streaming video into training slices and use a sliding window to sequentially get the inputs at each timestamp. To address detector noise, we incorporate the detected results and employ the ground truth (GT) matching relationships up to the $(t-1)$-th frame to create history tracks. Newly perceived objects without association in the current frame serve as *proposals*. In online inference, the model takes raw detections as input to perform tracking and prediction jointly.

 Training. To avoid poor latent memories which may impede the training procedure in early stages, **scheduled sampling** [3] is applied to the memory bank. We train our model for 180 epochs. Specifically, features in the memory bank are selected through sampling, and the sampling rate starts to increase at epoch 30, following a sigmoid curve.

4.2   Comparison with Related Work

Table 1 compares StreamMOTP with other methods in tracking and prediction, using the same Megvii [40] detector for fairness. For MOT, we evaluate all categories, while for trajectory prediction, we adopt two settings from prior

Table 2: Ablation study on the components of StreamMOTP.

Memory Bank	STPE	Stream Predictor	AMOTA	AMOTP	MOTA	minADE	minFDE	MR	tc
			0.523	0.781	0.426	0.572	0.942	0.113	-
	✓	✓	0.556	0.770	0.466	0.384	0.594	0.075	-
✓		✓	0.528	0.782	0.431	0.524	0.838	0.103	-
✓	✓		0.544	**0.768**	0.456	0.488	0.776	0.098	2.081
✓	✓	✓	**0.556**	0.779	**0.472**	**0.377**	**0.586**	**0.072**	**1.942**

studies: (1) **Setting1: One Step MOTP**. In Setting1, we follow a single-step tracking and 3s prediction, similar to PTP [28]. The model uses GT past trajectories $t \in \{T_c - T_h, \cdots, T_c - 1\}$ and GT detections in the current frame T_c, conducts MOT at the current frame, and forecasts future trajectories in frames $t \in \{T_c + 1, \cdots, T_c + T_f\}$. Results for all classes from the nuScenes Prediction Challenge are reported. This setting is more suitable for Vehicle-to-Vehicle (V2V) scenario. (2) **Setting2: Multi Step MOTP**. In setting2, we perform standard tracking and 6s prediction for detected objects in T_c, based on their tracked histories, and evaluate prediction results on all vehicle classes. This setting aligns more closely with the current stage of autonomous driving and is more widely adopted in industry deployments.

Our model surpasses previous related work in both tasks of multi-object tracking and trajectory prediction. In MOT performance, shown in Table 1a, our model not only achieves gains over PTP baseline [28] with improvements of 3.94% in AMOTA and 8.44% in MOTA, but also surpasses several competing trackers. Table 1b shows the prediction comparison for one-step MOTP. Our model reaches the lowest minADE of 0.810 and minFDE of 1.481, which outperforms PTP [28] by 0.207 on minADE and 0.046. Moreover, Table 1c offers a comparison of multi-step MOTP's predictions, where our model attains state-of-the-art performance with a minADE of 0.757 and a minFDE of 1.487, outperforming AffinPred [24] by 0.220 and 0.141, respectively. The improvements in Table 1c are more obvious than in Table 1b for the reason that trajectory prediction in setting1 is more saturated than in setting2, indicating the larger growth potential for prediction based on tracked trajectory.

4.3   Ablation Studies

We evaluated the impact of each module within our StreamMOTP framework, as summarized in Table 2, where the bottom row represents the full implementation of our method. All models are experimented on Setting2, except that the detector is switched to CenterPoint [31] and 3s prediction metrics are computed on True Positive detections at a recall rate of 0.6. The Megvii detector, being an older model, exhibits subpar detection capabilities. Therefore, we switch to a detector with relatively moderate performance to better measure each module's efficacy.

Effects of each module. Firstly, upon removing the memory bank, we observed a slight decline in performance for both tracking and prediction tasks. We will further explore it later. Secondly, we remove the spatio-temporal positional

Table 3: The effect of slice length (abbreviated as "Slice") and the use of a memory bank (abbreviated as "Mem") on tracking and prediction.

Slice	Mem	AMOTA	MOTA	minADE	minFDE	MR
3		0.570	0.490	0.633	0.953	0.137
5		0.560	0.478	0.402	0.621	0.075
10		0.557	0.466	0.384	0.594	0.075
3	✓	0.570	0.486	0.537	0.813	0.119
5	✓	0.564	0.478	0.392	0.602	0.072
10	✓	0.556	0.472	**0.377**	**0.586**	**0.072**

Table 4: Ablation study of Memory Bank when training slice=3.

Memory Length F	AMOTA	MOTA	minADE	minFDE	MR
0	0.570	0.490	0.633	0.953	0.137
1	0.569	0.487	0.603	0.921	0.135
2	0.570	0.486	**0.537**	**0.813**	**0.120**

encoding in the spatio-temporal interaction module and encode the absolute co-ordinate feature in the same way as the attribute feature. There is a significant drop in performance for both tasks of tracking and prediction, which shows that spatio-temporal positional encoding maintains the pose-invariance for trajectory predictions and effectively addresses the issue of inconsistent coordinate representations. Thirdly, we replace the dual streaming predictor with a single frame predictor. The second-last row shows that the dual-stream predictor plays a vital role in advancing prediction performance. The modest decrease in tracking further corroborates that augmenting prediction capabilities also benefits tracking results. Notably, the tc metric also drops when the dual-stream predictor is eliminated, which indicates that the dual-stream predictor enhances the trajectory predictions' quality and consistency, since it leverages predictions from previous frames as a valuable prior for predicting current perceived objects' trajectories, yielding more viable and steady outcomes.

Effects of streaming framework. The effectiveness of the streaming framework and the memory bank is explored by adjusting the lengths of training segments. In Table 3, tracking performance remains stable, but prediction accuracy improves significantly with longer segments due to reliance on extensive sequential information. This finding stems from the gap that our models are trained in split slices (multi-frame sequences of length k) but evaluated in streaming video (the average length is 40 in nuScenes, $k \ll 40$). This gap constrains the effectiveness of approaches, especially for previous snap-shot methods. Our streaming framework addresses this gap by utilizing temporal information across frames, boosting prediction. Moreover, the integration of the memory bank, particularly with shorter slices, markedly boosts prediction accuracy by the retention and utilization of long-term latent features in the memory bank, therefore improving
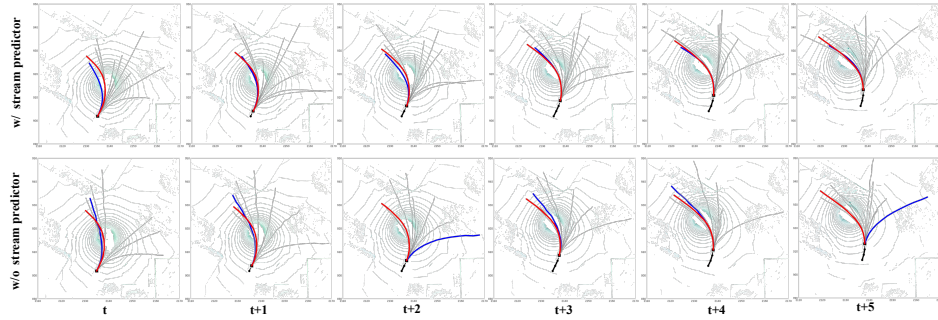
Fig. 5: Qualitative results of StreamMOTP on the nuScenes validation set during consecutive frames. The tracked history and detection are shown in black, models' best score prediction and ground-truth trajectories are drawn in blue and red. The predictions of other modes are drawn in gray. The top row shows the results given by the dual-stream predictor while the bottom row shows the results with a base predictor.

the model's capability for long-term sequence modeling. This is crucial under resource constraints that limit slice length and temporal receptive field. Furthermore, Table 4 shows that as the length of the memory bank expands, the model's performance grows, which further demonstrates the impact of the memory bank.

4.4   Qualitative Results

We provide some qualitative results in Fig. 5 to show our predictions. There is a brand new object without historical trajectory perceived at frame t. StreamMOTP successfully predicts its future trajectory with social interactions. Moreover, by comparing the two rows, we can see that all mode predictions in the top row are smoother and more precise, and the highest score of the predictions fluctuates less. More visualization can be found in our supplementary video.

5   Conclusion

In this paper, we introduce StreamMOTP, a streaming and unified framework for joint multi-object tracking and trajectory prediction. With the design of the memory bank, spatio-temporal positional encoding strategy, and dual-stream predictor, streamMOTP bridges the gap between training and actual deployment, as well as maintains better capability and great potential for both tasks of MOT and prediction. The experiments on nuSences demonstrate the effectiveness and superiority of the proposed framework. We hope this work could further offer insights into the multi-task end-to-end autonomous driving systems.

Acknowledgments

References

1. Aydemir, G., Akan, A.K., Güney, F.: Adapt: Efficient multi-agent trajectory prediction with adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8295–8305 (2023) 6
2. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022) 1
3. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. Advances in neural information processing systems **28** (2015) 11
4. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 941–951 (2019) 4
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 11
6. Chandra, R., Bhattacharya, U., Bera, A., Manocha, D.: Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8483–8492 (2019) 1, 11
7. Chandra, R., Guan, T., Panuganti, S., Mittal, T., Bhattacharya, U., Bera, A., Manocha, D.: Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. IEEE Robotics and Automation Letters **5**(3), 4882–4890 (2020) 11
8. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11525–11533 (2020) 1, 4
9. Gu, J., Sun, C., Zhao, H.: Densetnt: End-to-end trajectory prediction from dense goal sets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15303–15312 (2021) 1, 4
10. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2255–2264 (2018) 1, 4, 11
11. Li, D., Zhang, Q., Xia, Z., Zheng, Y., Zhang, K., Yi, M., Jin, W., Zhao, D.: Planning-inspired hierarchical trajectory prediction via lateral-longitudinal decomposition for autonomous driving. IEEE Transactions on Intelligent Vehicles (2023) 4
12. Li, X., Xie, T., Liu, D., Gao, J., Dai, K., Jiang, Z., Zhao, L., Wang, K.: Polymot: A polyhedral framework for 3d multi-object tracking. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9391–9398. IEEE (2023) 1
13. Liang, M., Yang, B., Zeng, W., Chen, Y., Hu, R., Casas, S., Urtasun, R.: Pnpnet: End-to-end perception and prediction with tracking in the loop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11553–11562 (2020) 4

14. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph lstm. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 125–143. Springer (2016) 1, 4

15. Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K.S., Sapp, B.: Wayformer: Motion forecasting via simple & efficient attention networks. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2980–2987. IEEE (2023) 4

16. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020) 8

17. Shi, S., Jiang, L., Dai, D., Schiele, B.: Motion transformer with global intention localization and local movement refinement. Advances in Neural Information Processing Systems **35**, 6531–6543 (2022) 1, 4, 8

18. Varadarajan, B., Hefny, A., Srivastava, A., Refaat, K.S., Nayakanti, N., Cornman, A., Chen, K., Douillard, B., Lam, C.P., Anguelov, D., et al.: Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 7814–7821. IEEE (2022) 1, 4, 7, 8, 10

19. Wang, L., Zhang, X., Qin, W., Li, X., Gao, J., Yang, L., Li, Z., Li, J., Zhu, L., Wang, H., et al.: Camo-mot: Combined appearance-motion optimization for 3d multi-object tracking with camera-lidar fusion. IEEE Transactions on Intelligent Transportation Systems (2023) 1

20. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. arXiv preprint arXiv:2303.11926 (2023) 6

21. Wang, X., Fu, C., Li, Z., Lai, Y., He, J.: Deepfusionmot: A 3d multi-object tracking framework based on camera-lidar fusion with deep association. IEEE Robotics and Automation Letters **7**(3), 8260–8267 (2022) 4

22. Wang, X., He, J., Fu, C., Meng, T., Huang, M.: You only need two detectors to achieve multi-modal 3d multi-object tracking. arXiv preprint arXiv:2304.08709 (2023) 1, 4

23. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: European conference on computer vision. pp. 107–122. Springer (2020) 4

24. Weng, X., Ivanovic, B., Kitani, K., Pavone, M.: Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6573–6582 (2022) 2, 4, 11, 12

25. Weng, X., Ivanovic, B., Pavone, M.: Mtp: Multi-hypothesis tracking and prediction for reduced error propagation. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 1218–1225. IEEE (2022) 2, 4

26. Weng, X., Wang, J., Held, D., Kitani, K.: Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. arXiv preprint arXiv:2008.08063 (2020) 1, 4, 11

27. Weng, X., Wang, Y., Man, Y., Kitani, K.M.: Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6499–6508 (2020) 1, 11

28. Weng, X., Yuan, Y., Kitani, K.: Ptp: Parallelized tracking and prediction with graph neural networks and diversity sampling. IEEE Robotics and Automation Letters **6**(3), 4640–4647 (2021) 2, 4, 11, 12
29. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 4
30. Wu, H., Han, W., Wen, C., Li, X., Wang, C.: 3d multi-object tracking in point clouds based on prediction confidence-guided data association. IEEE Transactions on Intelligent Transportation Systems **23**(6), 5668–5677 (2021) 4
31. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021) 12
32. Yu, R., Zhou, Z.: Towards robust human trajectory prediction in raw videos. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8059–8066. IEEE (2021) 4
33. Zhang, P., Bai, L., Wang, Y., Fang, J., Xue, J., Zheng, N., Ouyang, W.: Towards trajectory forecasting from detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 2, 4
34. Zhang, W., Zhou, H., Sun, S., Wang, Z., Shi, J., Loy, C.C.: Robust multi-modality multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2365–2374 (2019) 1, 11
35. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**, 3069–3087 (2021) 4
36. Zhao, J., Zhuang, J., Zhou, Q., Ban, T., Xu, Z., Zhou, H., Wang, J., Wang, G., Li, Z., Li, B.: Kigras: Kinematic-driven generative model for realistic agent simulation. arXiv preprint arXiv:2407.12940 (2024) 1
37. Zhou, Y., Shao, H., Wang, L., Waslander, S.L., Li, H., Liu, Y.: Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15281–15290 (2024) 1
38. Zhou, Z., Wang, J., Li, Y.H., Huang, Y.K.: Query-centric trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17863–17873 (2023) 1, 4
39. Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.: Hivt: Hierarchical vector transformer for multi-agent motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8823–8833 (2022) 4, 7
40. Zhu, B., Jiang, Z., Zhou, X., Li, Z., Yu, G.: Class-balanced grouping and sampling for point cloud 3d object detection. arXiv preprint arXiv:1908.09492 (2019) 11