

# Unified Multi-Agent Trajectory Modeling with Masked Trajectory Diffusion

Songru Yang, Zhenwei Shi, Zhengxia Zou\*

Department of Aerospace Intelligent Science and Technology,  
School of Astronautics, Beihang University, Beijing, China  
{songruiyang2001, shizhenwei, zhengxiazou}@buaa.edu.cn

## Abstract

*Understanding movements in multi-agent scenarios is a fundamental problem in intelligent systems. Previous research assumes complete and synchronized observations. However, real-world partial observation caused by occlusions leads to inevitable model failure, which demands a unified framework for coexisting trajectory prediction, imputation, and recovery. Unlike previous attempts that handled observed and unobserved behaviors in a coupled manner, we explore a decoupled denoising diffusion modeling paradigm with a unidirectional information valve to separate the interference from uncertain behaviors. Building on this, we proposed a Unified Masked Trajectory Diffusion model (UniMTD) for arbitrary levels of missing observations. We designed a unidirectional attention as a valve unit to control the direction of information flow between the observed and masked areas, gradually refining the missing observations toward a real-world distribution. We construct it into a unidirectional MoE structure to handle varying proportions of missing observations. A Cached Diffusion model is further designed to improve generation quality while reducing computation and time overhead. Our method has achieved a great leap across human motions and vehicle traffic. UniMTD efficiently achieves 74% improvement in  $\min ADE_{20}$  and reaches SOTA with advantages of 91%, 66%, 69%, and 58% across 4 fidelity metrics on out-of-boundary, velocity, and trajectory length.*

## 1. Introduction

Understanding multi-agent behavior is a critical foundation in various domains, including autonomous driving [35], action analysis [15, 26, 74], video surveillance [79, 83], and scene generation [69]. Trajectory modeling, which analyzes and generates the position sequence of multiple agents within the same scene, is particularly straightforward and effective for understanding agents' behavior.

Existing methods have made significant strides in trajectory prediction [4, 59, 72, 76, 86, 87], imputation [46, 84, 85], and recovery [20, 22] across multiple fields, such as human sports and traffic trajectory. However, existing approaches cannot handle multiple tasks simultaneously, leaving a gap between methods and real-world scenarios where all three tasks coexist, as shown in Fig.1 (a). Recently, increasing attempts have been made to break the task boundaries. [36, 40, 53, 73] employed multi-task frameworks to combine trajectory imputation and prediction. [24, 57, 62, 88] distilled complete trajectory knowledge into partially observed models. Lately, [71] first exploited mask formats as conditional input to unify all three tasks in a BERT-SSM structure.

Even though significant efforts have been devoted, existing approaches handled observed and unobserved behaviors in a coupled manner and encountered challenges in generating realistic multi-agent trajectories. Their suboptimal performance is inevitable for 2 reasons. First, they failed to realize that coupling the accurate representations and the uncertain ones with significant noise is deviating its latent space from the real-world behavior distribution. Second, more than an attached condition, the observation masks are precise boundaries between accuracy and uncertainty that need to be explicitly incorporated into the decoupling.

In this work, we reexamine unified trajectory modeling from coupled to decoupled manner. We introduce UniMTD, a unified framework for trajectory modeling in a decoupled denoising diffusion modeling paradigm, as a 'panacea' for arbitrary levels of missing observations in real practice. To achieve decoupling according to observation masks, a unidirectional attention as the valve unit is proposed. By controlling one-way information flow between observed and masked tokens, it can separate the uncertain noise, maintain the accuracy of the latent space, and consistently project the masked representations into the preserved space. For decoding, we consider the pre-completed encoder output as the deterministic component of the trajectory pattern and directly reuse it as intermediate parts in the diffusion model to efficiently improve trajectory generation quality. Main

\*Corresponding author

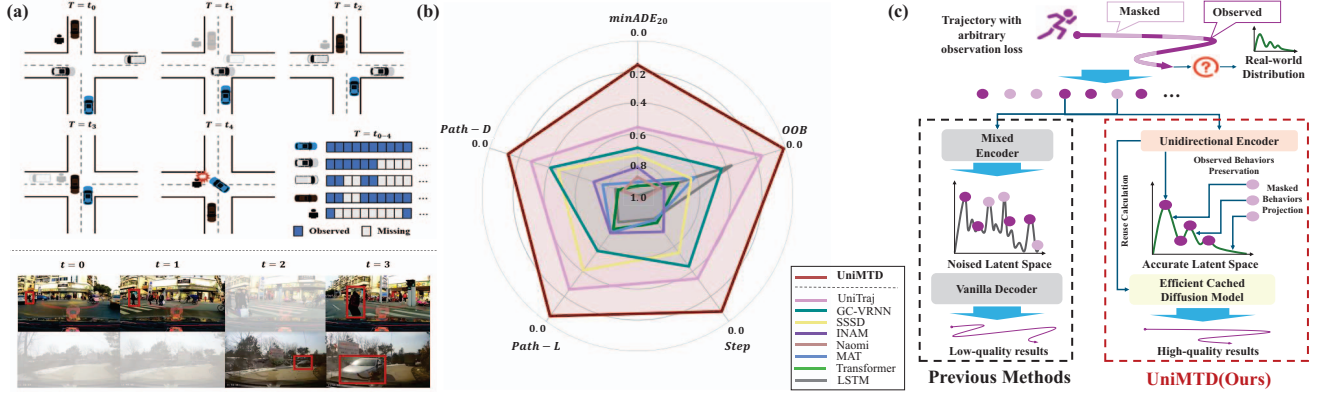


Figure 1. (a) **Above** is a diagram of a traffic accident caused by the blue vehicle’s observation failure of the pedestrian due to the obstruction. From the perspective of multi-agent trajectory modeling, the blue car is fully observed. The white car/van and brown car require prediction and imputation. The pedestrian with severe observation missing needs trajectory recovery. **Below** are real-world cases. In the first row, the pedestrian crashed due to being blocked. In the second row, the vehicle suddenly entered the view, resulting in an accident. (b) UniMTD (Ours) at the outermost side represents the minimum errors (averaged normalized error across three human datasets) on all 5 metrics. (c) Diagram of main differences and advantages of our method compared with previous methods during the encoding and decoding processes.

differences and improvements are shown in Fig. 1 (c)

Specifically, in the encoder, we propose an unidirectional attention (UA) as the valve unit. The UA is composed of a unidirectional masking (UM) and a progressive removal (PR) process. The UM allows one-way attention from observed to unobserved tokens, forcing the model to focus on the accurate information and confining the unobserved behaviors to learn from them. The PR removes the UM on the well-learned unobserved behaviors to further project it into the accurate distribution. The UA is built in a Unidirectional MoE (UMoE) with a local-global experts (LGE) division and a complex adaptive router (CAR) for multi-level observation missing. In the decoder, we design a cached diffusion model (CDiff) reusing the pre-calculated deterministic coordinates and features to conduct a truncated diffusion process, which enables us to improve the quality of trajectory modeling at an extremely low cost in terms of model size ( $< 5\%$ ) and time ( $< 5\%$ ) compared to the original DDPM.

We construct our datasets in both multi-agent vehicle traffic and human behavior scenarios. In both qualitative and quantitative results, UniMTD significantly surpasses existing methods in similar tasks across all 5 evaluation metrics with undamaged efficiency as shown in Fig. 1 (b). It achieves **74%** advantage in accuracy and **91%**, **66%**, **69%**, and **58%** advantages across 4 fidelity metrics on out-of-boundary, velocity, and trajectory length. In summary, the main contributions of our work are as follows:

- We introduce UniMTD, a novel unified trajectory modeling framework, transforming various trajectory-related tasks into an observed-unobserved decoupled masked diffusion paradigm.
- We design a unidirectional attention as a valve unit to decouple accurate and uncertain latent space and further construct a unidirectional MoE apt for arbitrary levels of

missing observations. A cached diffusion model is also designed to generate high-quality trajectories with low time and computation burden.

- We conduct extensive experiments on our framework across multiple human movement and traffic datasets to validate our consistent and outstanding performance with further analysis of our design.

## 2. Related Works

### 2.1. Prediction, Imputation and Recovery

Trajectory prediction predicts future behaviors based on historical observations. In the early studies, kinematics [38, 42] and filter-based [37, 82] methods were widely used. Started by Social-LSTM[1], neural networks such as RNN, LSTM, and GRU [8, 9, 11, 51, 67, 81] are designed to handle sequential data and agent interaction in trajectory prediction. With the popularity of generative methods, more recent studies have adopted generative frameworks such as GAN [3, 21, 28, 39], VAE [13, 29, 41, 70], and Diffusion models [27, 45, 47, 65] as future trajectory generators, enabling models to generate more realistic behaviors. The generative decoders fundamentally rely on a rich representation extracted from the encoder, which has led to the rise of MAE-based pre-training methods such as Traj-mae [14, 66, 68, 75], which are applied to a wide range of downstream tasks through fine-tuning.

Imputation typically serves as a preprocessing method for slight data missing. Traditional statistical methods include interpolation [49], regression [6], and EM algorithms [55]. Statistical methods are efficient but struggle with complex sequences. Deep learning frameworks use RNNs [7, 77], autoregressive models [5, 16], or generative techniques such as GAN, VAE, and Diffusion Model

[19, 48, 52, 78] to generate reconstructed sequences. Existing research on trajectory imputation in multi-agent scenarios is scarce: NAOMI [46] introduces a non-autoregressive imputation method, while Graph Imputer [50] utilized bi-directional features to impute soccer player trajectories.

Trajectory recovery is similar to imputation but differs in the observation ratio. When less than 20% observations are left caused by severe data loss or insufficient sampling, trajectory recovery as a conditional generation task needs to reconstruct the entire trajectory’s motion pattern based on sparse information. Current research involves recovering trajectory data from GPS points [18, 64], but research on multi-agent scenarios is still blank.

Although promising results have been achieved, the above methods have trouble generalizing across tasks. To mitigate cascade errors, the latest research has focused on trajectory prediction with slightly noised data, [24, 36, 40, 53, 57, 62, 73, 88], but these efforts are meeting challenges in generalizing to related tasks and obtaining accurate and realistic prediction results.

## 2.2. Mask Modeling

Mask modeling, since its inception in [23], has been extensively applied to pre-training tasks in trajectory prediction [14, 66, 68, 75] but is still incapable of handling multi-agent scenes with observation missing. [10] first extended mask modeling to generative tasks, iteratively predicting masked tokens, which is naturally adaptable to generation, inpainting, editing, etc. Various tasks can be considered as masking strategies within a unified framework.

Compared to the successful vision and language mask modeling [10, 23, 30, 34], trajectory modeling has higher mask proportion ( $> 50\%$ ), which limits the learning of the real distribution [62]. Besides, the random mask makes it impossible to use the MAE strategy [89] to batch-separate the noise. Moreover, unlike [10], lacking a powerful domain-pretrained tokenizer and decoder like VQGAN [25] or CLIP [54] prevents trajectory modeling in the latent space, and efficiency burdens are making a large number of iterative inferences not preferred.

## 2.3. Diffusion Models

Diffusion models [31] capable of generating high-quality samples through an iterative denoising process, which have recently achieved remarkable results in the fields of image and video generation [32, 63], time series forecast [43, 56], and remote sensing [12, 32, 44], etc. For trajectory modeling, MID [27] is the first diffusion-based method modeling the ambiguous walkable regions to desired trajectories. For the main concerns on computational and time cost, LeapFrog [47] distilled major denoising steps into an initializer to reduce the time costs in pre-trained MID. IDM [45] revealed that introducing goal information reduces de-

noising steps.

Above methods require a full-parameter diffusion model with time and computational redundancies, and CDiff greatly reduces both of them.

## 3. UniMTD

### 3.1. Overall Framework

The overall framework of UniMTD is depicted in Fig.2 (a). Take incomplete trajectories with arbitrary masks as input, the unidirectional encoder composed of UMoE and the diffusion generator with CDiff collaborate to generate completed versions of the input.

### 3.2. Unidirectional Encoder

The unidirectional encoder is composed of UMoE blocks, which include UA, LGE, and CAR.

#### 3.2.1. Twin Structure

The entire encoder is designed in a twin structure  $E_M$  and  $E_U$  with shared weights, inspired by [17] to project masked and unmasked behaviors into the same latent space and provide an accurate reference  $\mathbf{f}_{\text{ref}}$  for the progressive removal. During the training phase,  $E_M$  and  $E_U$  encode masked and complete trajectories to  $\mathbf{f}_M$  and  $\mathbf{f}_U$  respectively. We aligned them in the latent space by MSE loss on masked positions. In the testing phase, only  $E_M$  is used to obtain  $\mathbf{f}_M = \mathbf{f}_{\text{st}}$ .

#### 3.2.2. Unidirectional Attention

The unidirectional attention (UA) is proposed to preserve accurate latent space according to the clean observations. It shoulders 2 tasks: separate the noise from significant observation missing; project the masked tokens into the accurate latent space and extract complete spatiotemporal trajectory features. Thus, unidirectional masking and progressive removal are designed to meet the demands, respectively.

**Unidirectional masking** (UM) is designed to control the information transformation to operate only from the observed data to themselves and the missing data, making the encoding process unidirectional for masked trajectory modeling. By explicitly isolating the observed and the missing data, we can ensure that even with only highly sparse clean observations, they are free from noise interference.

Specifically, similar to the causal mask [61], UM sets the attention weights corresponding to the positions of missing data to 0, meaning that all tokens can only perform attention calculations with tokens that have not been masked. For temporal masking, assuming  $i$ th trajectory  $\mathbf{x}_i \in \mathbb{R}^{T \times 2}$  of length  $T$  and the corresponding mask  $\mathbf{m}_i \in \mathbb{R}^T$ , the temporal  $\text{UM}_i \in \mathbb{R}^{T \times T}$  can be obtained as follows:

$$\begin{aligned} \text{UM}_i &= [\text{UM}_i^0, \text{UM}_i^1, \dots, \text{UM}_i^T]^\top, \\ \text{UM}_i^j &= \begin{cases} 0, & \text{where } \mathbf{m}_i = 1 \\ -\text{Inf}, & \text{where } \mathbf{m}_i = 0 \end{cases}, j \in [0, T] \end{aligned} \quad (1)$$

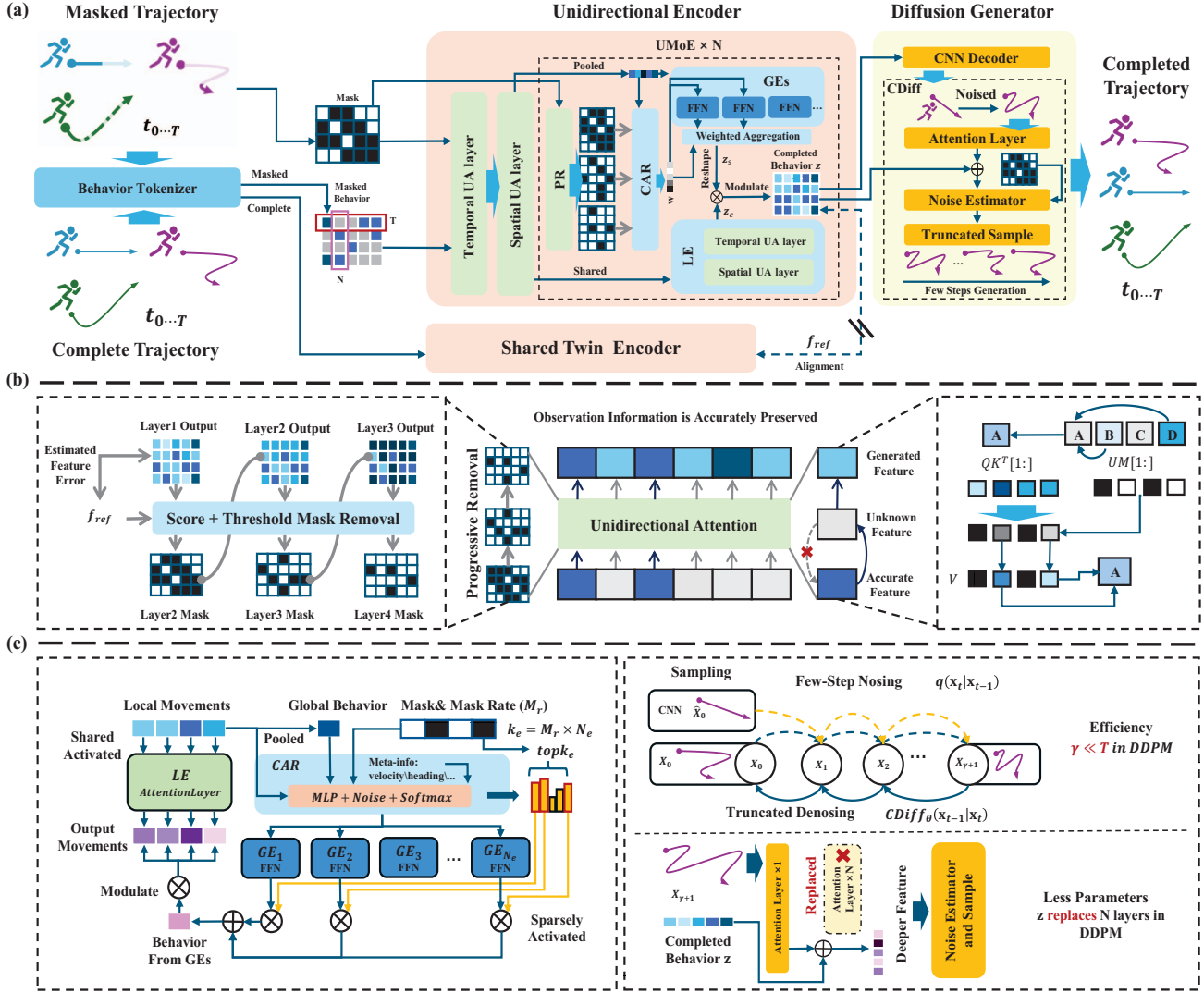


Figure 2. (a) The overall framework of UniMTD, including unidirectional encoder with UMoE, and diffusion generator with CDiff. (b) The detailed calculation of UA, including PR (left) and UM (right). (c) The detailed design of UMoE (left) and CDiff (right).

For social masking,  $N$  multi-agent trajectories at timestep  $t$  is  $\mathbf{x}^t \in \mathbb{R}^{N \times 2}$  and the corresponding mask  $\mathbf{m}^t \in \mathbb{R}^N$ , the social  $\mathbf{UM}_t \in \mathbb{R}^{N \times N}$  can be obtained as same as temporal masking.

Through the above masking strategy, UM allows the information exchange among the clean tokens to form an accurate latent space, and the noisy tokens are confined to gathering information from it.

**Progressive Removal (PR)** is implemented as a refinement process progressively removing the UM mask to project the masked representations into the preserved latent space. Although the continuous existence of UM ensures an accurate latent space, the isolated tokens are not conducive to the extraction of consistent and complete trajectory features. Therefore, when the masked tokens conform to the

accurate latent distribution, the mask on them should be removed. The process is depicted in Fig.2 (b-left).

Applied layer by layer, PR employs a 3-layer LSTM Score( $\cdot$ ) to estimate the confidence  $\mathbf{p}_l$  of the  $l$  layer's masked features conforming to the reference accurate latent features. For each layer,  $k$  masked tokens with the most confidence are unmasked ( $m_k^l$  from 0  $\rightarrow$  1) and used as a condition for predicting other masked tokens in the next layer. The  $l$ th layer mask can be formulated as:

$$\begin{aligned} \mathbf{f}_{\text{st}}^l &= \text{Layer}^{l-1}(\mathbf{f}_{\text{st}}^{(l-1)}, \mathbf{m}^{l-1}), \\ \mathbf{p}_l &= \text{softmax}\left(\frac{1}{\text{Score}(\mathbf{f}_{\text{st}}^l, \mathbf{f}_{\text{ref}}^l)}\right), \\ m_{(n,t)}^l &= \begin{cases} 1, & \text{if } p_{(n,t)}^l \in \text{Topk}(\mathbf{p}_l) \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (2)$$

where  $\mathbf{f}_{st}^l$  is spatiotemporal features of  $l$ th layer, and  $\mathbf{p}_l$  is the confidence. The Score( $\cdot$ ) network can be trained as follows:

$$\mathcal{L} = \frac{1}{L} \sum_{l=1}^L \text{MSE}(\text{LSTM}(\mathbf{f}_{st}^l, l), \|\mathbf{f}_{st}^l - \mathbf{f}_{\text{ref}}^l\|_2) \quad (3)$$

The overall calculation of unidirectional attention is depicted in Fig.2 (b-right), and can be formulated as follows:

$$\text{UA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top \cdot \text{PR}(\text{UM})}{\sqrt{d_k}}\right)\mathbf{V} \quad (4)$$

### 3.2.3. Unidirectional MoE

The unidirectional MoE (UMoE) includes local/global experts (LE/GE) and a complex adaptive router (CAR). Detailed structures are shown in Fig.2 (c-left).

As for trajectory, each coordinate lacks specific meaning, and the behavior is mainly described in a global view. Routing different experts for each time step leads to inconsistent features and weak expert specificity. To address this, we divide LE/GE and use a UA layer as a shared expert LE( $\cdot$ ) activates for all local movement tokens to extract common knowledge  $\mathbf{z}_c$  and maintains the motion stability. Meanwhile, the unmasked tokens of each agent are pooled into global features and routed to a group of sparsely activated experts GEs( $\cdot$ ) for specific behavior knowledge  $\mathbf{z}_s$ , emphasizing the expert specialization among the unique behaviors across multiple agents. We also design a complexity adaptive router to allocate a dynamic number of experts  $\mathbf{k}_e$  according to mask ratios to handle the multi-level complexity inherent in arbitrary mask formats. The calculation in UMoE can be formulated as follows:

$$\begin{aligned} \mathbf{z}_c &= \text{LE}(\mathbf{f}_{st}, \text{PR}(\text{UM})) \in \mathbb{R}^{B \times N \times T \times D}, \\ \mathbf{k}_e &= \text{Normalize}(1 - \text{PR}(\text{UM})) \cdot N_e, \\ \mathbf{w}_e &= \text{Topk}(\mathbf{f}_{\text{meta}}, \mathbf{f}_{st}, \mathbf{k}_e) \in \mathbb{R}^{B \times N \times k_e}, \\ \mathbf{z}_s &= \sum_{i=0}^{k_e} \mathbf{w}_e^i \cdot \text{GE}_i(\text{pool}(\mathbf{f}_{st})) \in \mathbb{R}^{B \times N \times D}, \\ \mathbf{z}_s &\xrightarrow{\text{repeat}} \mathbf{z}_s \in \mathbb{R}^{B \times N \times T \times D}, \\ \mathbf{z} &= \mathbf{z}_c \cdot \mathbf{z}_s \end{aligned} \quad (5)$$

### 3.3. Diffusion Generator

The Diffusion Generator comprises a cached diffusion model and a tiny CNN decoder.

#### 3.3.1. Cached Diffusion Model

The MAE decoding lacks diversity, and the iterative paradigm incurs high computational and time costs. Combine the strengths of the two and make up for their weaknesses, as shown in Fig.2 (c-right), we propose a cache mechanism to replace the deeper features and a large number of preceding inverse diffusion steps using the deterministic feature  $\mathbf{z} \in \mathbb{R}^{(B \cdot l) \times T \times d}$  and coordinate  $\hat{\mathbf{X}}_{\gamma+1} \in$

$\mathbb{R}^{(B \cdot l) \times T \times 2}$  of the encoder. In this way, we have reduced the noise estimator  $f_\epsilon$  with 6 attention layers and 200 steps to only a single layer and 5 steps. The denoising process can be revised as follows:

$$\begin{aligned} \hat{\mathbf{X}}_0 &= \text{CNN}(\mathbf{z}), \\ \hat{\mathbf{X}}_{\gamma+1} &= \sqrt{\bar{\alpha}_\gamma} \hat{\mathbf{X}}_0 + \sqrt{1 - \bar{\alpha}_\gamma} (\hat{\mathbf{X}}_0 \cdot \mathbf{M} + \epsilon \cdot (1 - \mathbf{M})), \\ \mathbf{z}_\gamma &= \text{AttentionLayer}(\hat{\mathbf{X}}_{\gamma+1}) + \mathbf{z}, \\ \epsilon_\theta^\gamma &= f_\epsilon(\hat{\mathbf{X}}_0 \cdot \mathbf{M} + \hat{\mathbf{X}}_{\gamma+1} \cdot (1 - \mathbf{M}), \mathbf{z}_\gamma), \\ \hat{\mathbf{X}}_\gamma &= \frac{1}{\sqrt{\alpha_\gamma}} \left( \hat{\mathbf{X}}_{\gamma+1} - \frac{1 - \alpha_\gamma}{\sqrt{1 - \bar{\alpha}_\gamma}} \epsilon_\theta^\gamma \right) + \sqrt{1 - \alpha_\gamma} \cdot \epsilon \end{aligned} \quad (6)$$

where  $\alpha_\gamma$  and  $\bar{\alpha}_\gamma = \prod_{i=1}^\gamma \alpha_i$  are parameters at denoising step  $\gamma$ .  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  are gaussian noises.

## 4. Experiments

### 4.1. Benchmarks and Setup

**Datasets:** We applied diverse trajectory datasets to verify the applicability of UniMTD on various fields, including three human motion datasets and one vehicle motion dataset with three subsets of different difficulty: *Basketball-U\** [80], *Soccer-U\** [58], *Football-U\**, and *Vehicle-U* (Easy/Ordinary/Hard)\* [60] constructed following [73]. The mask strategy is consistent with [71] for comparison.

**Evaluation Protocol:** We generate a total of  $k = 20$  candidate trajectories based on the masked trajectory input and use the  $\text{minADE}_{20}$  as the accuracy evaluation metric. Furthermore, four additional metrics to reflect the generation quality of agent movements' velocity, range, and behavior differences are employed to comprehensively assess the generation quality. **(1) minADE<sub>20</sub>:** Minimum average displacement error between the generated trajectories and the ground truth across the 20 generated trajectories, **(2) Out-of-Boundary (OOB):** The percentage of generated locations that fall outside the predefined court boundaries, **(3) Step:** The average change in step size across the generated trajectories, **(4) Path-L:** The average length of the trajectories for each agent, **(5) Path-D:** The maximum difference in trajectory lengths among all agents. The calculations are consistent with [71].

### 4.2. Quantitative Results

Quantitative results are recorded in Tab.1 on the *Basketball-U*, *Soccer-U*, and the *Football-U* and Tab.2 on the *Vehicle-U*. Across all datasets, compared with mathematical interpolation methods, general deep learning frameworks, and specialized methods for relevant tasks, our UniMTD

\*<https://github.com/linouk23/NBA-Player-Movements>

\*<https://github.com/AtomScott/SportsLabKit>

\*<https://github.com/nfl-football-ops/Big-Data-Bowl>

\*<https://github.com/shijieS/OmniMOTdataset>

Table 1. Quantitative results on datasets *Basketball-U* (in feet), *Soccer-U* (in pixels), and *Football-U* (in yards)

Dataset	<i>Basketball-U</i> (In Feet)					<i>Soccer-U</i> (In Pixels)					<i>Football-U</i> (In Yards)				
	minADE↓	OOB↓	Step	Path-L	Path-D	minADE↓	OOB↓	Step	Path-L	Path-D	minADE↓	OOB↓	Step	Path-L	Path-D
Mean	14.58	<b>0</b>	0.99	52.39	737.58	417.68	<b>0</b>	4.32	213.05	8022.51	14.18	<b>0</b>	0.52	25.06	606.07
Medium	14.56	<b>0</b>	0.98	51.80	743.36	418.06	<b>0</b>	4.39	214.55	8041.98	14.23	<b>0</b>	0.52	24.96	600.22
Linear Fit	13.54	4.47e-03	0.56	42.86	453.38	398.34	<b>0</b>	<b>0.70</b>	<b>112.34</b>	<b>2047.19</b>	12.66	1.49e-04	<u>0.17</u>	<u>15.83</u>	207.57
LSTM[33]	7.10	9.02e-04	0.76	58.48	449.58	186.93	4.74e-05	7.50	652.98	4542.78	7.20	2.24e-04	0.43	34.06	228.13
Transformer[61]	6.71	2.38e-03	0.79	59.34	517.54	170.94	6.59e-05	6.66	566.14	4269.08	6.84	5.68e-04	0.42	33.01	202.10
MAT[80]	6.68	1.36e-03	0.88	58.83	483.46	170.46	7.56e-05	6.45	562.44	3953.34	6.36	4.57e-04	0.40	31.32	186.11
Naomi[46]	6.52	2.02e-03	0.81	69.10	450.66	145.20	8.78e-05	7.47	649.62	4414.99	6.77	7.66e-04	0.67	42.74	259.11
INAM[52]	6.53	3.16e-03	0.70	58.54	439.87	134.86	4.04e-05	6.37	547.02	4102.37	5.80	8.30e-04	0.39	32.10	177.04
SSSD[2]	6.18	1.82e-03	0.47	46.87	393.12	118.71	4.51e-05	5.11	425.98	3252.66	5.08	6.81e-04	0.39	23.10	122.42
GC-VRNN[73]	5.81	9.28e-04	0.37	28.08	235.99	105.87	1.29e-05	4.92	506.32	3463.26	4.95	7.12e-04	0.29	32.48	149.87
UniTraj[71]	<u>4.77</u>	6.12e-04	<u>0.27</u>	<u>34.25</u>	<u>240.83</u>	<u>94.59</u>	<u>1.29e-05</u>	4.52	349.73	2805.79	<u>3.55</u>	1.12e-04	0.23	19.26	<u>114.58</u>
Ground Truth	0	0	0.17	37.61	269.49	0	0	0.52	112.92	951.00	0	0	0.03	12.56	76.68
UniMTD (Ours)	<b>1.88</b>	<u>9.61e-06</u>	<b>0.23</b>	<b>35.07</b>	<b>270.50</b>	<b>13.65</b>	<b>0</b>	<u>0.95</u>	<u>124.39</u>	<b>1688.11</b>	<b>0.91</b>	<u>2.84e-05</u>	<b>0.09</b>	<b>13.45</b>	<b>107.38</b>

Table 2. Quantitative results on datasets *Vehicle-U* with *Easy*, *Ordinary*, and *Hard* scenarios (in pixels)

Dataset	<i>Vehicle-U Easy</i> (In Pixels)					<i>Vehicle-U Ordinary</i> (In Pixels)					<i>Vehicle-U Hard</i> (In Pixels)				
	minADE↓	OOB↓	Step	Path-L	Path-D	minADE↓	OOB↓	Step	Path-L	Path-D	minADE↓	OOB↓	Step	Path-L	Path-D
Transformer [61]	107.93	5.61e-03	7.12	457.18	5913.21	92.47	4.77e-03	6.79	376.59	5723.45	108.76	1.07e-02	7.04	472.83	5835.42
SSSD [2]	<u>30.99</u>	<u>1.82e-03</u>	<b>0.47</b>	228.30	<b>1611.98</b>	<b>28.90</b>	<u>9.51e-04</u>	<b>0.41</b>	<u>119.57</u>	<u>1829.07</u>	<u>30.65</u>	<u>5.92e-03</u>	<b>0.50</b>	<u>131.65</u>	<u>2160.53</u>
UniTraj [71]	77.65	3.94e-03	2.42	<u>165.28</u>	4036.13	66.05	2.11e-03	4.01	257.17	5350.69	79.39	7.48e-03	4.70	255.12	5204.53
Ground Truth	0	0	0.06	43.30	1574.39	0	0	0.04	28.51	2030.48	0	0	0.05	28.92	1940.34
UniMTD(Ours)	<b>13.12</b>	<b>9.04e-04</b>	<u>0.75</u>	<b>79.15</b>	<u>1726.57</u>	<b>10.62</b>	<b>3.28e-04</b>	<u>0.65</u>	<b>62.37</b>	<b>2079.77</b>	<b>15.40</b>	<b>2.97e-03</b>	<u>0.86</u>	<b>75.19</b>	<b>2100.29</b>

achieves the SOTA performance, leading by a large margin across all 5 metrics.

Specifically, for datasets *Basketball-U*, *Soccer-U*, and *Football-U* depicting complex human behavior, UniMTD outperforms the current best-performing UniTraj by averaged 74% and surpasses the Diffusion model-based SSSD by averaged 80% on the minADE<sub>20</sub> metric, indicating a great leap in accuracy of behavior modeling. For OOB, Step, Path-L, and Path-D, UniMTD has achieved 91%, 66%, 69%, and 58% averaged advantages across data-driven methods, indicating that our generated trajectories align more closely with real motion patterns. For datasets *Vehicle-U*, UniMTD achieves SOTA performance, with averaged 56%, 33%, 44%, 15% advantages on minADE<sub>20</sub>, OOB, Path-L, Path-D across all three subsets, and the step metric is the second-best. Different from human sports in limited scenes, the vehicle dataset exhibits a significant amount of behaviors entering and leaving the scene, involving more large-scale, high-dynamic scenarios. This may lead to an increase in OOB and give advantages to the SSSD method based on a 200-step diffusion model with strong abilities to express detailed and local behaviors, such as step-by-step velocity. Our SOTA performance on the majority of metrics across all datasets demonstrates the model’s versatility in different scenarios.

Regarding the inference time, for a scenario containing 11 agents, UniTraj requires **0.16s**, UniMTD requires **0.18s**, while SSSD, which also uses the diffusion model, requires more than **10s**. This is because our CDiff reduces the number of denoising steps from 200 to 5 and simplifies the cal-

culaton, thus improving the inference efficiency.

### 4.3. Qualitative Results

In Fig.3, we qualitatively compared the results between the prediction model, UniTraj, SSSD, UniMTD, and the UniMTD without CDiff or UA on the *Basketball-U* dataset. The results in the seventh column illustrate that our UniMTD generates behaviors with superior accuracy and fidelity, especially in the recovery of arc (purple) and circle (green, blue) trajectories. In the second column, models trained solely on prediction naturally cannot adapt to imputation and recovery but are unexpectedly not satisfactory on prediction. This result indicates that the unified modeling may benefit its subtasks. In the third column, UniTraj can generate behaviors with misconnection and even failure in recovery. In the fourth column, SSSD obtains inaccurate but detailed behavior sequences in prediction and imputation and also fails in the recovery task. In the fifth column, the woCDiff model generates averaged results with significant connection errors and overly smooth behaviors. In the sixth column, the woUA model discloses fluctuations and distortions due to the unseparated uncertain mask noise.

The qualitative evaluation results confirm the significant superiority of our method in quantitative results and also demonstrate our correct insights into UA and CDiff designs.

### 4.4. Model Analysis

#### 4.4.1. Ablation Study

Tab.3 demonstrates the results of our ablation studies conducted on the *Basketball-U* dataset. The upper part records

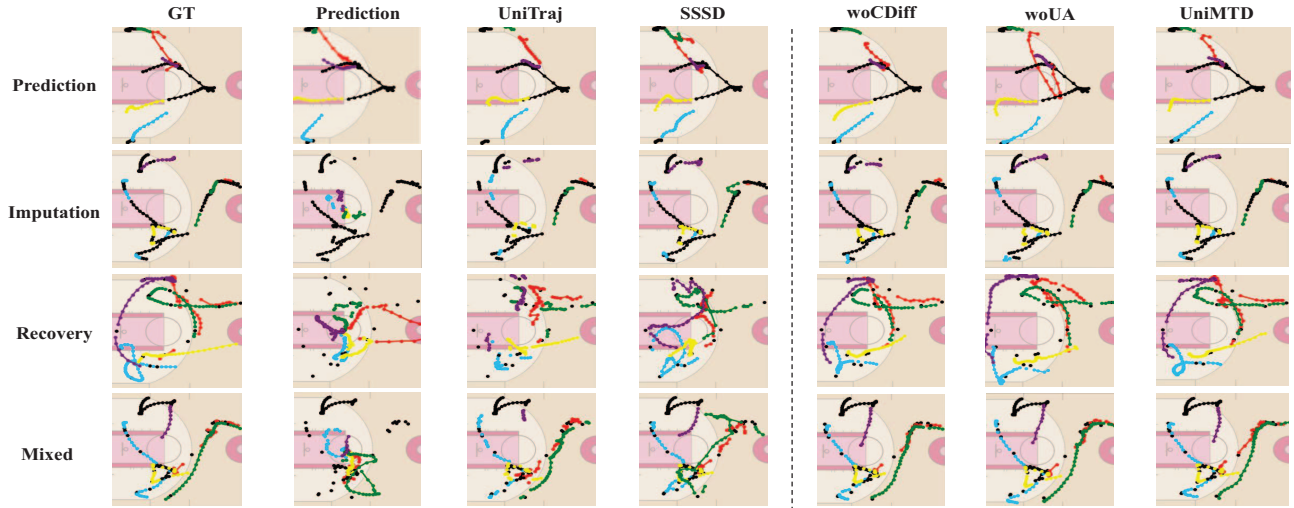


Figure 3. The qualitative comparison of different methods under specific unified mask format across prediction, imputation, recovery, and a mixed scenario of these three. The ground truth is on the left, and our method is shown on the right side. Players 1-5 are in different colors, and the observed area is black.

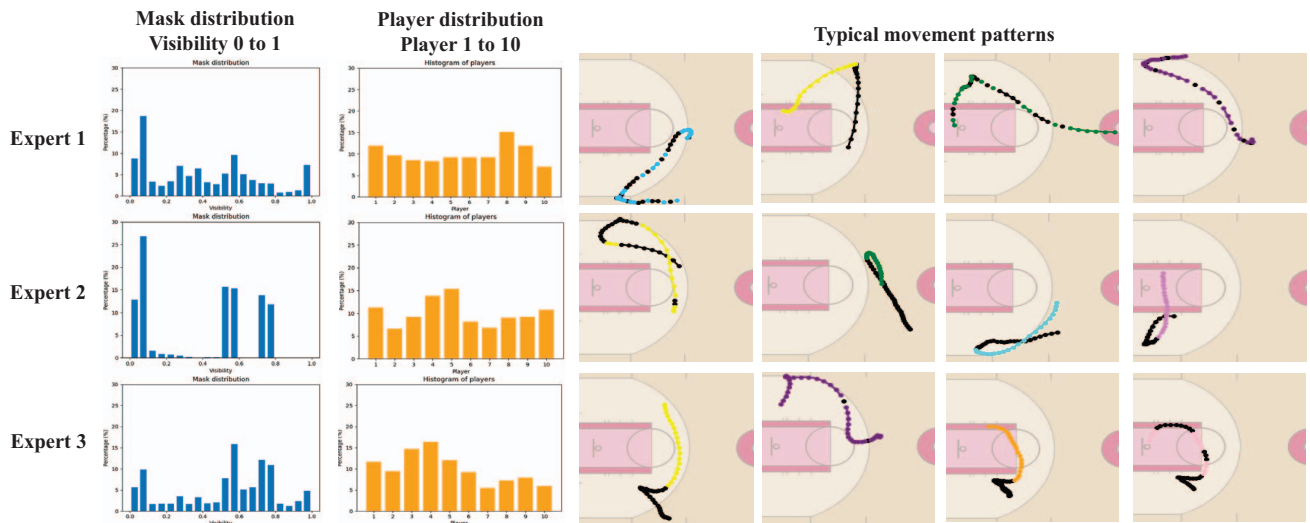


Figure 4. The statistical analysis of the distribution of player and masking ratios across experts 1-3. The typical movement patterns of each expert are given on the right.

the performance variations when key parameters and model capacity are modified, while the lower part documents the performance variations when CDiff, PR, CAR (replaced by Top2 router), UMoEs (replaced by UA layers), and UM (UA to full attention) are successively removed.

For the upper part, when the number of experts increases ( $N_e = 16$ ) or decreases ( $N_e = 4$ ), only the OOB and Path-L metrics improve, while all other metrics deteriorate. When the model capacity is directly increased, the minADE and step improve, while other metrics deteriorate. From the overall trend, as the model becomes more complex, both accuracy and fidelity metrics show the potential for contin-

ued improvement, but this improvement may come at the cost of deterioration in other metrics.

For the lower part, **w/o CDiff** indicates that CDiff improved the accuracy with only a cost of  $<5\%$  size inflation. **w/o UMoE** indicates that replacing UMoE with UA layers leads to severe degradations on both accuracy and fidelity. **w/o CAR** indicates that CAR can better handle the multi-level complexity inherent in arbitrary observation loss than commonly used top-k routing. **w/o PR** and **w/o UM** demonstrate a performance nosedive when observed and unobserved behaviors are coupled, indicating that UA contributes to both accuracy and fidelity with almost no ad-

Table 3. Ablation experiment on *Basketball-U*

Variants	minADE↓	OOB↓	Step	Path-L	Path-D	Size (M)
UniMTD	1.88	9.61e-06	0.23	35.07	270.50	37.19
$N_e = 4$	2.13	5.40e-06	0.25	36.28	335.40	27.71
$N_e = 16$	1.90	1.32e-06	0.24	35.88	202.67	41.93
6UMoEs	1.85	1.18e-05	0.21	35.45	215.08	53.18
w/o CDiff	1.95	1.32e-06	0.26	37.13	245.75	35.86
w/o PR	2.12	1.32e-07	0.27	38.44	203.72	35.54
w/o CAR	2.22	6.32e-06	0.26	37.18	319.15	35.54
w/o UMoE	2.92	9.21e-07	0.37	39.70	331.99	17.33
w/o UM	4.44	5.69e-5	0.44	51.19	398.76	17.33

Table 4. Ablation experiment of **CDiff** efficiency on *Basketball-U*

Variants	minADE↓	OOB↓	Step	Path-L	Path-D	Time (s)
Step=2	1.99	3.68e-06	0.25	37.31	216.41	0.17
Step=5	1.88	9.61e-06	0.23	35.07	270.50	0.18
Step=15	2.14	6.28e-05	0.28	37.41	203.65	0.51
Step=50	2.73	2.25e-03	0.33	42.46	319.55	1.85
Step=100	15.19	6.83e-02	3.22	207.11	5165.04	3.13
Naive DDPM	2.69	3.11e-5	0.28	39.19	304.65	> 10s

ditional learnable parameters.

In the following sections, we further conduct validation on CDiff, UMoE, and UA according to the original design insights.

#### 4.4.2. CDiff Efficiency Analysis

Tab.4 records the impact of truncated denoising steps on trajectory modeling quality and efficiency.

In the top 5 rows, when  $\text{step} > 15$ , as steps increased, the inference time surged while the performance decreased. Suppose the input is too noisy compared to  $\hat{X}_0$ . The single Attention Layer cannot bridge the noisy input and accurate deeper features, leading to contamination and performance damage. In the last row, we compared the standard diffusion model in [27], demonstrating improved efficiency and performance. This is because the  $\hat{X}_0$  ensures the accuracy of CDiff modeling, while DDPM directly samples from the noise, which is prone to generating high-frequency tremors in highly-dynamic scenes.

#### 4.4.3. UMoE Specialization Analysis

To validate the specialization in UMoE, we statistically analyzed the distribution of player and masking ratios corresponding to each expert (experts 1-3 as examples) on the left of Fig.4 and visualized the specialized movement patterns of each expert on the right.

Compared across experts, the specialization in mask distribution is concentrated in several patterns. For instance, expert 2 specializes in visibility at around 0.1, 0.5, and 0.7. Experts are different in their preference for players, such as expert 1 prefers players 1, 8, 9. We also find specialized behavior patterns in the samples assigned to each expert, such as expert 3 processes most trajectories with arc movement after a sharp turn.

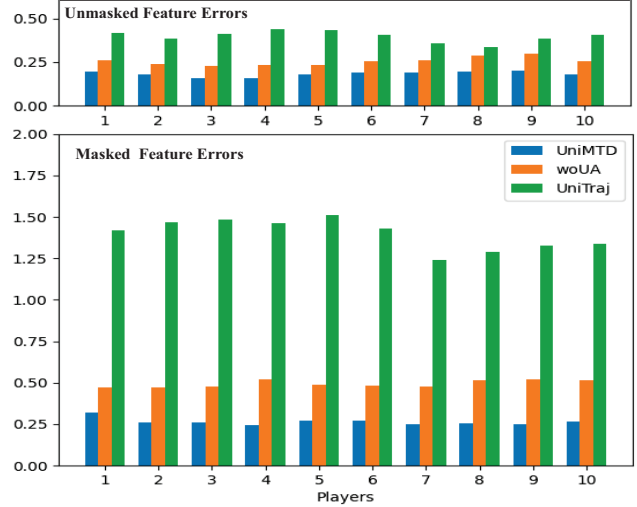


Figure 5. The comparison of the feature errors on masked and unmasked tokens. Different methods are in different colors.

#### 4.4.4. UA Feature Robustness Analysis

Masked trajectory modeling can be regarded as removing masked noise to restore the complete trajectory. We compared the denoising capabilities on the features of the encoder, that is, the feature robustness to mask noise. Specifically, we input the complete and masked trajectories successively and compare the features'  $L1$  distances. The results of masked and unmasked tokens are in Fig. 5.

Compared with UniTraj and the woUA version, our method achieves the least error. The gap between the errors of unmasked tokens is greatly magnified in the distance of masked tokens prediction, indicating the importance of maintaining accurate, clean tokens for the overall quality of features and the effectiveness of the proposed UA.

## 5. Conclusion

In this work, we propose UniMTD, a novel unified multi-agent trajectory modeling framework with an observed-unobserved decoupled masked diffusion paradigm. Essentially, we decouple observed-unobserved behavior to separate the interference from the uncertain latent space. To achieve this, we design a unidirectional attention as valve units to control the information flow between the accurate and the uncertain latent spaces. We further construct a unidirectional MoE and a cached diffusion model to handle the inherent multi-complexity and the efficiency requirements in trajectory modeling under arbitrary mask levels. Through extensive experiments and comprehensive evaluations, UniMTD consistently outperforms existing SOTA methods across trajectory prediction, completion, and recovery tasks with an averaged > 50% advantage on accuracy and fidelity. Establishing itself as a SOTA solution with significant potential for further exploration in the embodied foundation model.

## 6. Acknowledgements

The work was supported by the National Natural Science Foundation of China under Grant 62471014, 62125102, and U24B20177, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities.

## References

- [1] Alexandre Alahi, Kratharth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 2
- [2] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022. 6
- [3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2
- [4] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17890–17901, 2024. 1
- [5] Faraj Bashir and Hua-Liang Wei. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. *Neurocomputing*, 276:23–30, 2018. 2
- [6] Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076, 2010. 2
- [7] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018. 2
- [8] Samuele Capobianco, Leonardo M Millefiori, Nicola Forti, Paolo Braca, and Peter Willett. Deep learning methods for vessel trajectory prediction based on recurrent neural networks. *IEEE Transactions on Aerospace and Electronic Systems*, 57(6):4329–4346, 2021. 2
- [9] Samuele Capobianco, Nicola Forti, Leonardo Maria Millefiori, Paolo Braca, and Peter Willett. Recurrent encoder-decoder networks for vessel trajectory prediction with uncertainty estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 59(3):2554–2565, 2022. 2
- [10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. 3
- [11] 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/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 2
- [12] Bowen Chen, Liqin Liu, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Spectral-cascaded diffusion model for remote sensing image spectral super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [13] Guangyi Chen, Junlong Li, Nuoxing Zhou, Liangliang Ren, and Jiwen Lu. Personalized trajectory prediction via distribution discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15580–15589, 2021. 2
- [14] Hao Chen, Jiase Wang, Kun Shao, Furui Liu, Jianye Hao, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. Traj-mae: Masked autoencoders for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8362, 2023. 2, 3
- [15] Jianqi Chen, Panwen Hu, Xiaojun Chang, Zhenwei Shi, Michael Christian Kampffmeyer, and Xiaodan Liang. Sitcom-crafter: A plot-driven human motion generation system in 3d scenes. *arXiv preprint arXiv:2410.10790*, 2024. 1
- [16] Xinyu Chen, Mengying Lei, Nicolas Saunier, and Lijun Sun. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12301–12310, 2021. 2
- [17] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024. 3
- [18] Yuqi Chen, Hanyuan Zhang, Weiwei Sun, and Baihua Zheng. Rntrajrec: Road network enhanced trajectory recovery with spatial-temporal transformer. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 829–842. IEEE, 2023. 3
- [19] Yakun Chen, Kaize Shi, Zhangkai Wu, Juan Chen, Xianzhi Wang, Julian McAuley, Guandong Xu, and Shui Yu. A temporally disentangled contrastive diffusion model for spatiotemporal imputation. *arXiv preprint arXiv:2402.11558*, 2024. 3
- [20] Zebin Chen, Xiaolin Xiao, Yue-Jiao Gong, Jun Fang, Nan Ma, Hua Chai, and Zhiguang Cao. Interpreting trajectories from multiple views: A hierarchical self-attention network for estimating the time of arrival. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2771–2779, 2022. 1
- [21] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mgan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13158–13167, 2021. 2
- [22] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3767–3776, 2021. 1

- [23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3
- [24] Patrick Ebel, Ibrahim Emre Göl, Christoph Lingenfelder, and Andreas Vogelsang. Destination prediction based on partial trajectory data. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1149–1155, 2020. 1, 3
- [25] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3
- [26] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022. 1
- [27] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17113–17122, 2022. 2, 3, 8
- [28] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 2
- [29] Marah Halawa, Olaf Hellwich, and Pia Bideau. Action-based contrastive learning for trajectory prediction. In *European conference on computer vision*, pages 143–159. Springer, 2022. 2
- [30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3
- [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3
- [32] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3
- [33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6
- [34] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *arXiv preprint arXiv:2408.06687*, 2024. 3
- [35] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022. 1
- [36] Zhe Huang, Ruohua Li, Kazuki Shin, and Katherine Driggs-Campbell. Learning sparse interaction graphs of partially detected pedestrians for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):1198–1205, 2022. 1, 3
- [37] Biao Jin, Bo Jiu, Tao Su, Hongwei Liu, and Gaofeng Liu. Switched kalman filter-interacting multiple model algorithm based on optimal autoregressive model for manoeuvring target tracking. *IET Radar, Sonar & Navigation*, 9(2):199–209, 2015. 2
- [38] Nico Kaempchen, Bruno Schiele, and Klaus Dietmayer. Situation assessment of an autonomous emergency brake for arbitrary vehicle-to-vehicle collision scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 10(4):678–687, 2009. 2
- [39] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Reza Tofighi, and Silvio Savarese. Socialbigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, 32, 2019. 2
- [40] Bernard Lange, Jiachen Li, and Mykel J Kochenderfer. Scene informer: Anchor-based occlusion inference and trajectory prediction in partially observable environments. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14138–14145. IEEE, 2024. 1, 3
- [41] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 336–345, 2017. 2
- [42] Chiu-Feng Lin, A Galip Ulsoy, and David J LeBlanc. Vehicle dynamics and external disturbance estimation for vehicle path prediction. *IEEE Transactions on Control Systems Technology*, 8(3):508–518, 2000. 2
- [43] Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):19–41, 2024. 3
- [44] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 3
- [45] Chen Liu, Shibo He, Haoyu Liu, and Jiming Chen. Intention-aware denoising diffusion model for trajectory prediction. *arXiv preprint arXiv:2403.09190*, 2024. 2, 3
- [46] Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. Naomi: Non-autoregressive multiresolution sequence imputation. *Advances in neural information processing systems*, 32, 2019. 1, 3, 6
- [47] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023. 2, 3
- [48] Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8983–8991, 2021. 3
- [49] Norazian Mohamed Noor, Mohd Mustafa Al Bakri Abdullah, Ahmad Shukri Yahaya, and Nor Azam Ramli. Compar-

- ison of linear interpolation method and mean method to replace the missing values in environmental data set. In *Materials science forum*, pages 278–281. Trans Tech Publ, 2015. 2
- [50] Shayegan Omidshafiei, Daniel Hennes, Marta Garnelo, Eugene Tarassov, Zhe Wang, Romuald Elie, Jerome T Connor, Paul Muller, Ian Graham, William Spearman, et al. Time-series imputation of temporally-occluded multiagent trajectories. *arXiv preprint arXiv:2106.04219*, 2021. 3
- [51] Derek J Phillips, Tim A Wheeler, and Mykel J Kochenderfer. Generalizable intention prediction of human drivers at intersections. In *2017 IEEE intelligent vehicles symposium (IV)*, pages 1665–1670. IEEE, 2017. 2
- [52] Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. Imitative non-autoregressive modeling for trajectory forecasting and imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2020. 3, 6
- [53] Kyle K Qin, Yongli Ren, Wei Shao, Brennan Lake, Filippo Privitera, and Flora D Salim. Multiple-level point embedding for solving human trajectory imputation with prediction. *ACM Transactions on Spatial Algorithms and Systems*, 9(2):1–22, 2023. 1, 3
- [54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [55] Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using a fuzzy clustering-based em approach. *Knowledge and Information Systems*, 46(2):389–422, 2016. 2
- [56] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, pages 8857–8868. PMLR, 2021. 3
- [57] Philip Schörrer, Lars Tötzel, Jens Doll, and J. Marius Zöllner. Predictive trajectory planning in situations with hidden road users using partially observable markov decision processes. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2299–2306, 2019. 1, 3
- [58] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3579, 2022. 5
- [59] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024. 1
- [60] ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Simultaneous detection and tracking with motion modelling for multiple object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 626–643. Springer, 2020. 5
- [61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 6
- [62] Sheng Wang, Yingbing Chen, Jie Cheng, Xiaodong Mei, Ren Xin, Yongkang Song, and Ming Liu. Improving autonomous driving safety with pop: A framework for accurate partially observed trajectory predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14450–14456. IEEE, 2024. 1, 3
- [63] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yanan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 3
- [64] Tonglong Wei, Youfang Lin, Yan Lin, Shengnan Guo, Lan Zhang, and Huaiyu Wan. Micro-macro spatial-temporal graph-based encoder-decoder for map-constrained trajectory recovery. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3
- [65] Keshu Wu, Yang Zhou, Haotian Shi, Xiaopeng Li, and Bin Ran. Graph-based interaction-aware multimodal 2d vehicle trajectory prediction using diffusion graph convolutional networks. *IEEE Transactions on Intelligent Vehicles*, 9(2): 3630–3643, 2023. 2
- [66] Chongjun Xia, Yang Peng, and Dong Qu. A pre-trained model specialized for ship trajectory prediction. In *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1857–1860. IEEE, 2024. 2, 3
- [67] Yang Xing, Chen Lv, and Dongpu Cao. Personalized vehicle trajectory prediction based on joint time-series modeling for connected vehicles. *IEEE Transactions on Vehicular Technology*, 69(2):1341–1352, 2019. 2
- [68] Chenfeng Xu, Tian Li, Chen Tang, Lingfeng Sun, Kurt Keutzer, Masayoshi Tomizuka, Alireza Fathi, and Wei Zhan. Pretram: Self-supervised pre-training via connecting trajectory and map. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 2, 3
- [69] Liangyu Xu, Wanxuan Lu, Hongfeng Yu, Yongqiang Mao, Hanbo Bi, Chenglong Liu, Xian Sun, and Kun Fu. Taformer: A unified target-aware transformer for video and motion joint prediction in aerial scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [70] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022. 2
- [71] Yi Xu and Yun Fu. Sports-traj: A unified trajectory generation model for multi-agent movement in sports. In *ICLR 2025*. 1, 5, 6
- [72] Yi Xu, Jing Yang, and Shaoyi Du. Cf-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12541–12548, 2020. 1

- [73] Yi Xu, Armin Bazarjani, Hyung-gun Chi, Chiho Choi, and Yun Fu. Uncovering the missing pattern: Unified framework towards trajectory imputation and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9643, 2023. 1, 3, 5, 6
- [74] Zhuo Xu, Rui Zhou, Yida Yin, Huidong Gao, Masayoshi Tomizuka, and Jiachen Li. Matrix: Multi-agent trajectory generation with diverse contexts. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12650–12657. IEEE, 2024. 1
- [75] Bingqi Yan, Geng Zhao, Lexue Song, Yanwei Yu, and Junyu Dong. PreIn: Pretrained-based contrastive learning network for vehicle trajectory prediction. *World Wide Web*, 26(4): 1853–1875, 2023. 2, 3
- [76] Biao Yang, Fucheng Fan, Rongrong Ni, Hai Wang, Ammar Jafaripournimchahi, and Hongyu Hu. A multi-task learning network with a collision-aware graph transformer for traffic-agents trajectory prediction. *IEEE Transactions on Intelligent Transportation Systems*, 25(7):6677–6690, 2024. 1
- [77] Jinsung Yoon, William R Zame, and Mihaela Van Der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, 2018. 2
- [78] Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024. 3
- [79] Won Joon Yun, Soohyun Park, Joongheon Kim, MyungJae Shin, Soyi Jung, David A Mohaisen, and Jae-Hyun Kim. Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-uav control. *IEEE Transactions on Industrial Informatics*, 18(10):7086–7096, 2022. 1
- [80] Eric Zhan, Stephan Zheng, Yisong Yue, Long Sha, and Patrick Lucey. Generating multi-agent trajectories using programmatic weak supervision. *arXiv preprint arXiv:1803.07612*, 2018. 5, 6
- [81] Jiandong Zhang, Zhuoyong Shi, Anli Zhang, Qiming Yang, Guoqing Shi, and Yong Wu. Uav trajectory prediction based on flight state recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 60(3):2629–2641, 2023. 2
- [82] Ruifeng Zhang, Libo Cao, Shan Bao, and Jianjie Tan. A method for connected vehicle trajectory prediction and collision warning algorithm based on v2v communication. *International Journal of Crashworthiness*, 22(1):15–25, 2017. 2
- [83] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1
- [84] Yimei Zhang, Xiangjie Kong, Wenfeng Zhou, Jin Liu, Yanjie Fu, and Guojiang Shen. A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 1
- [85] Cong Zhao, Andi Song, Yuchuan Du, and Biao Yang. Trajgat: A map-embedded graph attention network for real-time vehicle trajectory imputation of roadside perception. *Transportation research part C: emerging technologies*, 142: 103787, 2022. 1
- [86] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 1
- [87] Zikang Zhou, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Qcnxt: A next-generation framework for joint multi-agent trajectory prediction. *arXiv preprint arXiv:2306.10508*, 2023. 1
- [88] Pengfei Zhu, Peng Shu, Mengshi Qi, Liang Liu, and Huadong Ma. Target-driven self-distillation for partial observed trajectories forecasting. *arXiv preprint arXiv:2501.16767*, 2025. 1, 3
- [89] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Xuetao Wei, and Yuxuan Liang. Unitraj: Universal human trajectory modeling from billion-scale worldwide traces. *arXiv preprint arXiv:2411.03859*, 2024. 3