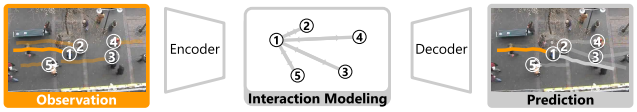
Can Language Beat Numerical Regression? Language-Based Multimodal Trajectory Prediction

Inhwan Bae[1], Junoh Lee[2] and Hae-Gon Jeon[1,2*]
[1]AI Graduate School, [2]School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology, Gwangju, South Korea
{inhwanbae, juno}@gm.gist.ac.kr, haegonj@gist.ac.kr

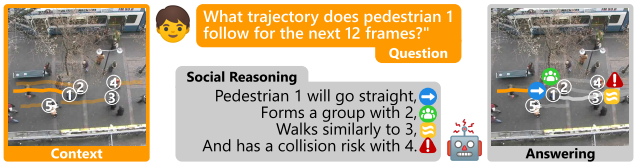
Abstract

Language models have demonstrated impressive ability in context understanding and generative performance. Inspired by the recent success of language foundation models, in this paper, we propose LMTraj (Language-based Multimodal Trajectory predictor), which recasts the trajectory prediction task into a sort of question-answering problem. Departing from traditional numerical regression models, which treat the trajectory coordinate sequence as continuous signals, we consider them as discrete signals like text prompts. Specially, we first transform an input space for the trajectory coordinate into the natural language space. Here, the entire time-series trajectories of pedestrians are converted into a text prompt, and scene images are described as text information through image captioning. The transformed numerical and image data are then wrapped into the question-answering template for use in a language model. Next, to guide the language model in understanding and reasoning high-level knowledge, such as scene context and social relationships between pedestrians, we introduce an auxiliary multi-task question and answering. We then train a numerical tokenizer with the prompt data. We encourage the tokenizer to separate the integer and decimal parts well, and leverage it to capture correlations between the consecutive numbers in the language model. Lastly, we train the language model using the numerical tokenizer and all of the question-answer prompts. Here, we propose a beam-search-based most-likely prediction and a temperature-based multimodal prediction to implement both deterministic and stochastic inferences. Applying our LMTraj, we show that the language-based model can be a powerful pedestrian trajectory predictor, and outperforms existing numerical-based predictor methods. Extensive experiments show that our LMTraj can successfully understand social relationships and accurately extrapolate the multimodal futures on the public pedestrian trajectory prediction benchmark. Code is publicly available at https://github.com/inhwanbae/LMTrajectory.

*Corresponding author



(a) Previous trajectory prediction



(b) Our language-based trajectory prediction

Figure 1. **Traditional vs. Our language-based trajectory prediction, LMTraj**. Given each observation data, (a) traditional predictors directly use the numerical values; (b) the proposed method converts the raw trajectory data to the linguistic prompt, and then captures reasoning social relations to predict a socially acceptable future with the question-answering template.

1. Introduction

Forecasting pedestrian trajectories in crowded environments is essential for path planning, social robots and autonomous maneuvering systems. The mainstream models used for this task take the position of pedestrians in world-coordinates as input, and infer their possible future paths by regressing a set of coordinate sequences [1, 3, 6, 14, 27, 32, 41, 49, 50, 64, 73, 96, 99, 102, 106, 124, 139]. Capturing social relations between pedestrians, based on their distance and motion similarity, has resulted in impressive performance improvements [73, 99, 115, 116, 139, 140].

Meanwhile, recent advances in language models have demonstrated their ability to provide context understanding and conditional generation across a spectrum of tasks [30, 59, 92]. The language models also offer accurate results when solving mathematical problems [104, 149]. This is because the language models can provide higher-level connotations [45, 131, 134], and benefit from tokenizers [44]

and extensive knowledge embedded in large pretrained models [87, 88]. The beauty of the language models is to account for social reasoning [113] beyond the physics-based interactions. As a result, we can intuitively expect improvement in interaction modeling when a language model is introduced in trajectory prediction. However, there are remaining challenges before they can be practically applied: (1) since it is trained on text data, the text tokenizer often does not work on numerical data; (2) it does not consider numerical data with decimal precision; (3) it does not attempt to extrapolate the time-series data using the numerical data itself.

In this paper, we investigate the feasibility of using natural language processing (NLP) to infer the future trajectories of pedestrians. We attempt to bridge the gap between traditional trajectory predictors and the capability of contemporary language models, offering a holistic solution for forecasting in crowded scenarios. Here, we introduce the Language-based Multimodal Trajectory predictor (LMTraj), which reevaluates language models from their foundational levels for numerical forecasting, as illustrated in Fig. 1. Our LMTraj consists of four steps: (1) We convert the raw trajectory coordinates and scene images into textual prompts. The raw coordinates are transformed into a set of decimal notations, and the images are converted into natural language through an image captioning model. Both prompts are then integrated into the question-answering (QA) template as context information. (2) We introduce supplementary tasks to push the model to learn a higher level of context understanding. Auxiliary questions about the number of group members and collision occurrences drive the model to consider social relationships when forecasting future trajectories. (3) We conduct an in-depth analysis of tokenizers, which have been largely overlooked by existing models. Our numerical tokenizer, optimized using the trajectory prompt, splits text and numbers clearly so that the model can learn correlations between sequential natures. (4) We enable the language model to infer future trajectories in both deterministic and stochastic manners. To generate the most likely and multimodal trajectories, we incorporate beam-search and temperature-tuning techniques.

Lastly, we evaluate the language model as a numerical regressor through both *zero-shot* and *supervised* approaches. We perform a zero-shot evaluation using prompt engineering on the two language foundation models. To take full advantage of the language model, we integrate all the proposed components into our LMTraj model. By effectively incorporating the proposed methods, our model achieves state-of-the-art results using a variety of public pedestrian trajectory prediction benchmarks, which are commonly regarded as the area of numerical regressors.

2. Related Works

2.1. Pedestrian Trajectory Prediction

Beginning with physics-based mathematical formulation methods [33, 71, 81, 135], trajectory forecasting has significantly improved under the numerical-based prediction paradigm. Following advances in convolutional neural networks (CNNs) and recurrent neural networks (RNNs), trajectory prediction become capable of inferring socially-acceptable paths using social interactions and motion modeling. One pioneering work is Social-LSTM [1], which recurrently predicts future coordinates using a long short-term memory (LSTM), while the social interaction between neighboring agents is modeled by aggregating hidden states via a pooling mechanism. Employing methods such as attention mechanisms [28, 36, 96, 117], graph convolutional networks (GCNs) [3, 40, 73, 107], graph attention networks (GATs) [6, 34, 41, 56, 57, 99, 116], or transformers [5, 17, 20, 31, 75, 85, 101, 120, 121, 139, 140, 148] allows us to directly model mutual influences among agents. Plus, additional environmental information can lead to better prediction results [21–23, 41, 55, 65, 66, 68, 69, 95, 98, 106, 107, 112, 114, 132, 141, 145]. Subsequent works take either recurrent [1, 11, 15, 16, 26, 31, 32, 42, 46, 60, 62, 63, 70, 77, 84, 96, 109, 119, 128, 130, 142, 143] or simultaneous approaches [2–4, 7, 35, 52, 73, 79, 99, 100] to extrapolate the future trajectories. Recent works combine probabilistic inferences with the bivariate Gaussian distribution [1, 3, 13, 50, 73, 74, 90, 99, 102, 103, 129, 137, 139], Generative Adversarial Network (GAN) [22, 32, 34, 41, 54, 58, 95, 106, 110, 145], Conditional Variational AutoEncoder (CVAE) [10, 15, 36, 46, 47, 49, 64, 96, 108, 118, 123, 126, 144] and diffusion [31, 37, 67, 91] for multi-modal trajectory generation.

Departing from the mainstream methods, works in [19, 65] predict heatmaps at the pixel level in images for possible future paths. Like the classification task, some works [23, 55, 78] have output classified positions on a discretized (Manhattan) grid. Unfortunately, they reach a limit because the trajectory prediction task requires forecasting accurate pathways based on social norms.

2.2. Language-Based Reasoning and Prediction

Transformer architectures and their training schemes have led to the notable development of language foundation models in the NLP field. In particular, BERT [39] employs a masked language modeling (MLM), which randomly masks a certain percentage of words and trains the model to predict them. GPT-2 [87] uses a causal language modeling (CLM), an autoregressive method for predicting the next token. T5 [89] involves sequence-to-sequence (Seq2Seq) modeling, using an encoder-decoder architecture to generate the output sequence. These unique models stand out in various genera-

tive tasks, including machine translation [12, 127, 136], text generation [30, 59, 92], and question-answering [8, 25, 86].

Beyond the NLP field, language foundational models have also exhibited superior performance in vision-language tasks and solving mathematical problems. This includes classification [88, 146], generation [24, 147], and problem-solving [104, 149]. These works explore the application of foundational language models, with the goal of extending the scope of the pre-training/fine-tuning paradigm.

Most recently, there have been attempts to incorporate language priors into time-series forecasting [9, 76, 105]. For instance, ForecastQA [38] proposes a QA benchmark with timestamp constraints to verify its forecasting ability regarding future events. Xue *et al*. [134] study mobility prediction, inferring how people move in cities. Inspired by chatbot applications, PromptCast [131] has made predictions on weather temperature, energy consumption, and customer flow. The most relevant work to ours [45] uses linguistic intermediate representations for trajectory prediction, solving action-related reasoning through language priors. However, they cannot fully take advantage of the linguistic model, in that they inherently use pre-trained tokenizers learned from text data. In particular, their approach is not suitable for trajectory prediction tasks because of the inconsistent analysis of numerical data. Furthermore, when dealing with coordinate sequences, existing numerical regressors are directly utilized as auxiliary modules to language models, inhibiting a higher level of understanding like social interactions.

3. Methodology

Our approach shifts the paradigm from conventional trajectory prediction to a prompt-based perspective. We recast the trajectory prediction task in a sentence-to-sentence manner, which uses the numerical input and output as a prompt and applies a language model for the purpose of numerical forecasting. In this work, we propose a language-based trajectory prediction framework, LMTraj, consisting of LMTraj-ZERO and LMTraj-SUP, using both zero-shot and supervised approaches, respectively.

We start with a numerical definition of the trajectory forecasting task in Sec. 3.1. We then describe our considerations in converting the numerical trajectories and images into text prompts and designing prompt templates to obtain desirable responses from language models for both LMTraj-ZERO and LMTraj-SUP in Sec. 3.2. Using these text prompts, we obtain the best performance with the language model-based trajectory predictor, LMTraj-SUP, as described in Sec. 3.3. In Sec. 3.4, we introduce how to build the language models, whose implementation details can be found in Sec. 3.5.

3.1. Problem Definition

The problem of trajectory prediction involves forecasting the time-series future coordinates of each agent from their historical coordinate sequences. This task can be regarded as a sequence-to-sequence problem. Formally, given a scene image $\mathcal{I}$ and a past observation trajectory with length T_{obs}, it can be denoted as $\mathcal{S}_{n,obs} = \{(x_n^t, y_n^t) \in \mathbb{R}^2 \mid t \in [1, ..., T_{obs}]\}$, where (x_n^t, y_n^t) is the 2D coordinate of a specific pedestrian n at time t. In the same way, a ground truth future trajectory for the prediction length T_{pred} can be written as $\mathcal{S}_{n,pred} = \{(x_n^t, y_n^t) \in \mathbb{R}^2 \mid t \in [T_{obs}+1, ..., T_{obs}+T_{pred}]\}$. The prediction model takes both $\mathcal{S}_{obs}$ and $\mathcal{I}$ as input. It either predicts one most-likely path $\widehat{\mathcal{S}}_{pred}$ or generates K possible multi-modal future trajectories $\widehat{\mathcal{S}}_{pred}^k$, which are called deterministic and stochastic predictions, respectively.

3.2. Data Space Conversion to Prompt

To make predictions using a language model, we first need to convert the raw data into text prompts. The most common data used in trajectory prediction is a numerical coordinate sequence and top-down view images of a scene. In this section, we start by transforming the pedestrian trajectory and environmental data. The converted data are then aggregated into linguistic sentences using a QA template for the input and output of the language model.

Converting trajectory coordinates into the prompt. We convert the entire float-type coordinate value to a text string with decimal representation. Compared to the binary numerical system, which is commonly used for network input, decimal representation is more compatible with natural language. In this process, we round the continuous values to discrete values with two decimal places for the word coordinate system in order to efficiently use the prompt. We leave it as an integer value if the trajectory is in the pixel coordinate system. Second, to represent the sequence of 2D coordinates, we concatenate the x_n^t and y_n^t coordinates using a comma separator and round bracket, and combine the time-series coordinates $\{(x_n^t, y_n^t) \mid t\}$ using the square bracket, as shown in Tab. 1. By using the different bracket symbols, it becomes easier to parse the spatio-temporal information. With this trick, we transform both the history and future trajectories $\mathcal{S}_{n,obs}$, $\mathcal{S}_{n,pred}$ into text prompts $\mathcal{P}_{\mathcal{S}_{n,obs}}$, $\mathcal{P}_{\mathcal{S}_{n,pred}}$, and repeat the process for all N pedestrians in the scene.

Converting image data into the prompt. We convert the scene image $\mathcal{I}$ into prompts $\mathcal{P}_{\mathcal{I}}$ as well. Inspired by image captioning, we employ the BLIP-2 model [51], trained on ImageNet [94], to extract text descriptions that depict the agent-moving scene. Taking the image description prompt as input, the model is able to learn various environmental details, such as the placement of buildings and vehicles, the density of people, and the flow of pedestrians. This helps the model to determine moving speeds and behavior patterns, similar to a traditional map encoding using pretrained segmentation models [65].

Converting predictions into the prompt. Next, the numerical coordinate prompt and the scene description prompt

Prompt	Type	Field	Template
$\mathcal{P}_{\mathcal{S}_n,obs}$	-	-	"[($\{x_n^1\}$,$\{y_n^1\}$),($\{x_n^2\}$,$\{y_n^2\}$),...,($\{x_n^{T_{obs}}\}$,$\{y_n^{T_{obs}}\}$)]"
$\mathcal{T}_{\mathcal{S}_n,obs}$	-	-	"Pedestrian_$\{n\}$_moved_along_the_trajectory_$\{\mathcal{P}_{\mathcal{S}_n,obs}\}$_for_$\{T_{obs}\}$_frames."
$\mathcal{T}_{forecast}$	Input	Question	"What_trajectory_does_pedestrian_$\{n\}$_follow_for_the_next_$\{T_{obs}\}$_frames?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	"Pedestrian_$\{n\}$_will_move_along_the_trajectory_$\{\mathcal{S}_{n,pred}\}$_for_the_next_$\{T_{pred}\}$_frames."
$\mathcal{T}_{dest}$	Input	Question	"At_which_coordinates_does_pedestrian_$\{n\}$_arrive_after_the_next_$\{T_{pred}\}$_frames?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	"Pedestrian_$\{n\}$_will_arrive_at_coordinate ($\{x_n^{T_{obs}+T_{pred}}\}$,$\{y_n^{T_{obs}+T_{pred}}\}$)_after_the_next_$\{T_{pred}\}$_frames."
$\mathcal{T}_{dir}$	Input	Question	"In_which_direction_will_pedestrian_$\{n\}$_move_in_the_future?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	"Pedestrian_$\{n\}$_will {move_forward|move_backward|move_left|move_right|stop}."
$\mathcal{T}_{mimic}$	Input	Question	"Which_pedestrian_seems_to_walk_similarly_to_pedestrian_$\{n\}$?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	Case 1: "Pedestrian_$\{n\}$_walks_similarly_to_pedestrian_$\{k\}$." Case 2: "Pedestrian_$\{n\}$_will_walk_alone."
$\mathcal{T}_{group}$	Input	Question	"With_which_pedestrians_does_pedestrian_$\{n\}$_form_a_group?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	Case 1: "Pedestrian_$\{n\}$_forms_a_group_with_pedestrian_$\{k\}$." Case 2: "Pedestrian_$\{n\}$_will_walk_alone."
$\mathcal{T}_{col}$	Input	Question	"With_which_pedestrian_does_pedestrian_$\{n\}$_have_a_collision_risk?"
		Context	"$\{\mathcal{P}_{\mathcal{I}}\}$_$\{\mathcal{T}_{\mathcal{S}_1,obs}\}$_$\{\mathcal{T}_{\mathcal{S}_2,obs}\}$_..._$\{\mathcal{T}_{\mathcal{S}_N,obs}\}$"
	Output	Answer	Case 1: "Pedestrian_$\{n\}$_has_a_collision_risk_with_pedestrian_$\{k\}$." Case 2: "Pedestrian_$\{n\}$_has_no_collision_risk."

Table 1. QA templates to convert raw trajectory data into prompts.

are preprocessed before being fed into LMTraj. For trajectory prediction with a language model, we need to make it suitable for the NLP task. Note that the QA task gives context information to a language model and asks questions to ensure the correct answers. We introduce a question-answering template $\mathcal{T}_{forecast} = \{\mathcal{P}_{\mathcal{C}}, \mathcal{P}_{\mathcal{Q}}, \mathcal{P}_{\mathcal{A}}\}$ for trajectory forecasting. We provide the history coordinates of all agents in a scene as context $\mathcal{P}_{\mathcal{C}}$ and ask the model to predict the future trajectory for a specific pedestrian n using $\mathcal{P}_{\mathcal{Q}}$. The answer we expect is $\mathcal{P}_{\mathcal{A}}$. This template-based description can effectively transform the data into text [133].

3.3. Domain Shift to Sentence Generation

After the conversion process, we revisit each component of the conventional NLP pipelines, and introduce a domain adaptation for LMTraj-SUP.

Optimizing the tokenizer for numeric data. The first thing that we revisit is the tokenizer. A tokenizer is an essential component that breaks text down into smaller units called tokens, which are used as the preliminary step in conventional NLP models to parse and understand languages [43, 44, 72, 82].

Following a conventional NLP pipeline [80, 122], we use a tokenizer to convert the QA prompt into a form that the LMTraj-SUP can understand. In this step, existing studies directly employ pretrained tokenizers [72, 82]. However, we figure out that when they are optimized for text data, they often fail to properly represent the numerical data. When using this tokenizer, numbers are irregularly split into tokens, and occasionally, special characters like periods and commas

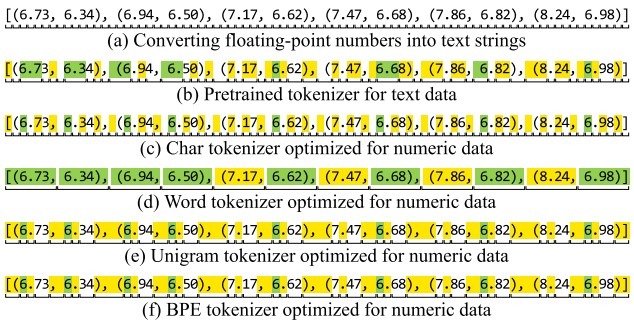


Figure 2. Comparison of the text-pretrained tokenizer and our numeric data-optimized tokenizer. Under brackets with yellow or white highlight colors indicate that the corresponding letters have been tokenized. The green color highlights that the token contains an integer with 6.

are grouped together, as shown in Fig. 2(b). This can disturb the training for consecutiveness and associations between adjacent numbers.

To address this issue, we train a new tokenizer for the numerical data using our QA prompts $\mathcal{P}_{\mathcal{C}}$, $\mathcal{P}_{\mathcal{Q}}$ and $\mathcal{P}_{\mathcal{A}}$ consisting of numerical coordinates and image description prompts. As demonstrated in Fig. 2(f), our numerical tokenizer clearly breaks down words, integers and decimal parts well. In addition, because the total number of tokens for the same sentence is reduced by removing the unnecessary splitting problem, LMTraj-SUP can become lighter and faster.

Multi-task training for social relation reasoning. In trajectory prediction tasks, the most crucial component is modeling interactions between agents. To enhance the reasoning capacity with social relations, we develop a training scheme for our LMTraj-SUP. It is a widely known technique which allows language models to achieve high-level knowledge understanding through multi-task learning [39, 87, 89]. While LMTraj-SUP can learn to perform prediction tasks only with the forecasting QA prompt, we introduce auxiliary tasks to fully take advantage of its understanding and reasoning ability for both scene context and social dynamics.

The five auxiliary tasks are as follows: destination suggestion, moving direction prediction, similar pattern search, group member prediction, and collision possibility assessment. We implement these synthetic tasks using pseudo labels and QA templates $\mathcal{T}_{dest}$, $\mathcal{T}_{dir}$, $\mathcal{T}_{mimic}$, $\mathcal{T}_{group}$ and $\mathcal{T}_{col}$ in the same way as the forecasting task. Table 1 lists up the prompt templates, and LMTraj-SUP yields the six types of outputs. By explicitly teaching various social relations, LMTraj-SUP can better capture, understand, and use social norms. Among the outputs, LMTraj-SUP can extract common features for agent motions, and leverage social knowledge (e.g., group walking and collision avoidance) learned from each auxiliary task to enhance the fidelity of the main forecasting task.

Tokenizer	Summary				Input sentence		Output sentence	
	# Vocab	# Mixed	Clarity	Cover	# Token	Rouge	# Token	Rouge
Pretrained	32000	504	98.43	1.00	566.25	1.00	44.47	1.00
Char	**59**	**0**	**100**	1.00	952.80	1.00	77.48	1.00
Word	13586	13497	0.655	1.00	**142.21**	1.00	**12.10**	1.00
Unigram	1113	**0**	**100**	1.00	421.63	1.00	27.46	1.00
BPE	1224	**0**	**100**	1.00	402.73	1.00	27.46	1.00

Table 2. Evaluation of the tokenizer characteristics. # Vocab: the total number of unique words in the tokenizer, # Mixed: The number of unique entries that contain both characters and numerals, Clarity: Percentage of non-mixed cases in vocab, Cover: Coverage of the tokenizer that can cover all sentences in the dataset, # Token: The average number of tokens per sentence. Rouge: ROUGE-1 score between the original sentences and their reconstructed ones after tokenization.

Zero-shot	Stop	Linear	Kalman filter	AutoTrajectory [62]	LMTraj-ZERO	
					-GPT-3.5	-GPT-4
ETH	2.84/4.82	1.00/2.23	0.94/2.13	N/A	1.07/1.82	**0.80/1.64**
HOTEL	1.15/2.09	0.32/0.62	0.26/0.50	N/A	0.42/0.65	**0.20/0.37**
UNIV	1.36/2.47	0.52/1.17	0.55/1.20	0.89/1.45	0.56/0.98	**0.37/0.77**
ZARA1	2.51/4.61	0.43/0.96	0.45/0.98	0.48/0.91	0.47/0.91	**0.33/0.66**
ZARA2	1.38/2.53	0.33/0.73	0.34/0.75	0.50/1.03	0.39/0.71	**0.24/0.50**
AVG	1.85/3.31	0.52/1.14	0.51/1.11	0.62/1.13	0.58/1.01	**0.39/0.79**
SDD	64.0/116.7	18.8/38.0	16.6/33.9	N/A	17.8/29.1	**10.9/21.0**
GCS	76.0/138.8	18.9/40.7	18.3/39.4	N/A	27.7/44.8	**12.7/25.5**

Table 3. Comparison of LMTraj-ZERO methods with other zero-shot methods (ADE/FDE, Unit: meter for ETH/UCY and pixel for SDD/GCS). **Bold**: Best, Underline: Second best.

Generating most-likely and multimodal outputs. Lastly, we tune the text generation stage, where LMTraj-SUP infers the output. In trajectory prediction, it is crucial to generate all possible multiple paths $\widehat{\mathcal{S}}^k_{pred}$ or the single most likely path $\widehat{\mathcal{S}}_{pred}$. Numerical regression-based methods predict a deterministic path using encoder-decoder architectures, and extend it to stochastic inferences for diverse multiple paths by introducing a random latent vector as an additional input.

In the same way, to use a language model in this task, it must be able to produce diverse multiple outputs. We assume that if we can effectively leverage the stochasticity of the language model, inherently based on distributional semantics [18], it can function as a probability-based numerical approach. We handle the stochasticity by introducing a text-generation technique. Using beam search [29], LMTraj-SUP can predict the path $\widehat{\mathcal{P}}_{\mathcal{A}}$ with the highest probability search controlled by a hyperparameter on a depth d. On the other hand, the model can generate diverse outputs $\widehat{\mathcal{P}}^k_{\mathcal{A}}$ by modulating the token probability using a temperature parameter τ in LMTraj-SUP. By using these tricks, the language-based model can perform at par with and even potentially replace existing predictor methods.

3.4. Forecasting With the Language Model

Lastly, we incorporate our proposed methods into the trajectory forecasting model. To do this, we adopt two widely-used approaches in computer vision and natural language tasks: (1) conducting zero-shot evaluation through prompt engineering of a pretrained language foundation model, LMTraj-ZERO and (2) an end-to-end supervision, LMTraj-SUP.

LMTraj-ZERO: Zero-shot prediction in the language foundation model. Prompt-tuning is a method that fine-tunes a language model, not by retraining it but by optimizing the input prompt that goes into a frozen pre-trained model to produce a desired output [53, 83]. The advantage of prompt-tuning is that it allows us to leverage the existing/extensive knowledge embedded within large pre-trained models.

In this work, we also use pre-trained large language models, GPT-3.5 and GPT-4, not trained for the purpose of trajectory forecasting. Following [131], we optimize the input prompt with the following steps: (1) We make an initial forecasting QA prompt $\mathcal{P}_{\mathcal{Q}}$ to instruct LMTraj-ZERO on what the desired output should be; (2) The prompts are fed into LMTraj-ZERO; (3) The outputs $\widehat{\mathcal{P}}^k_{\mathcal{A}}$ are evaluated by transforming them back into the numerical coordinates $\widehat{\mathcal{S}}^k_{pred}$, ensuring a fair comparison with the conventional metrics. In all the processes, the language model is frozen and is neither trained nor fine-tuned.

LMTraj-SUP: Supervision of language-based predictor. Next, we evaluate the maximum capacity and performance of the language model through end-to-end training. First of all, we analyze various structures of sentence-to-sentence language models to choose the best model for forecasting. In trajectory prediction, it has been proven that predicting the trajectories through an encoder-decoder architecture is better than using a procedural generation from recurrent models due to the error accumulation issue [4, 64, 73]. Therefore, instead of using a CLM, we choose the Seq2Seq model, an encoder-decoder language model. Similar to the zero-shot predictors, the set of context and question sentences $\{\mathcal{P}_{\mathcal{C}}, \mathcal{P}_{\mathcal{Q}}\}$ are given to the network, and the output sentences $\widehat{\mathcal{P}}_{\mathcal{A}}$ are transformed back into numerical data $\widehat{\mathcal{S}}_{pred}$. The difference between LMTraj-ZERO and LMTraj-SUP lies in the multi-task QA template, whose loss is back-propagated to train LMTraj-SUP.

3.5. Implementation Details

To demonstrate the zero-shot performance of the proposed LMTraj-ZERO, we use GPT-3.5 and GPT-4 as foundational language models for prompt engineering. In this experiment, since the proposed model does not require a training procedure, we exclude the tokenizer optimization and multi-task training methods. Each API call for one trajectory inference takes about 20 seconds, so we carry out a multi-process by creating a thread pool of 1,000 units to evaluate all paths in the datasets. We ensure prediction fidelity from the output sentence by retrying if responses are not aligned with the desired answer format.

Deterministic	Social-LSTM [1]	Social-GAN [32]	SR-LSTM† [142]	STGAT [34]	STAR-D† [139]	Trajectron ++† [96]	PECNet [64]	MID [31]	GP-Graph [5]	SocialVAE [126]	NPSN [6]	EigenTrajectory [7]	**LMTraj-SUP**
ETH	1.09 / 2.35	1.13 / 2.21	1.01 / 1.93	0.88 / 1.66	0.97 / 2.00	1.02 / 2.09	1.20 / 2.73	1.42 / 2.94	0.89 / 1.78	0.97 / 1.93	0.95 / 2.04	0.92 / 2.03	**0.65 / 1.04**
HOTEL	0.79 / 1.76	1.01 / 2.18	0.35 / 0.72	0.56 / 1.15	0.32 / 0.73	0.33 / 0.63	0.68 / 1.51	0.64 / 1.30	0.47 / 1.03	0.40 / 0.78	0.32 / 0.57	0.29 / 0.57	**0.26 / 0.46**
UNIV	0.67 / 1.40	0.60 / 1.28	0.66 / 1.38	**0.52 / 1.13**	0.56 / 1.25	**0.52** / 1.16	0.78 / 1.71	0.76 / 1.62	0.56 / 1.19	0.54 / 1.16	0.59 / 1.23	0.57 / 1.21	0.57 / 1.16
ZARA1	0.47 / 1.00	0.42 / 0.91	0.56 / 1.23	0.41 / 0.91	0.44 / 0.96	0.42 / 0.94	0.82 / 1.85	0.74 / 1.59	**0.40 / 0.87**	0.44 / 0.97	0.42 / 0.89	0.45 / 0.99	0.51 / 1.01
ZARA2	0.56 / 1.17	0.52 / 1.11	0.44 / 0.90	**0.31 / 0.68**	0.35 / 0.77	0.32 / 0.71	0.62 / 1.46	0.60 / 1.31	0.35 / 0.77	0.33 / 0.74	**0.31 / 0.68**	0.34 / 0.75	0.38 / 0.74
AVG	0.72 / 1.54	0.67 / 1.41	0.60 / 1.23	0.54 / 1.11	0.53 / 1.14	0.52 / 1.11	0.82 / 1.85	0.83 / 1.75	0.53 / 1.13	0.54 / 1.12	0.52 / 1.08	**0.51** / 1.11	**0.48 / 0.88**
SDD	31.2 / 57.0	27.3 / 41.4	31.4 / 56.8	28.0 / 41.3	28.8 / 51.4	22.7 / 42.0	29.8 / 65.1	25.2 / 57.6	24.7 / 49.0	24.2 / 49.3	22.1 / 38.0	20.7 / 41.9	**17.5 / 34.5**
GCS	40.2 / 67.2	33.6 / 50.5	31.9 / 48.4	31.8 / 49.3	29.3 / 46.5	16.9 / 35.1	28.3 / 61.2	19.4 / 41.5	16.7 / 34.9	16.6 / 35.0	**16.5** / 36.3	17.6 / 37.2	16.9 / **34.8**

Stochastic	Social-GAN [32]	Social-STGCNN[73]	PECNet† [64]	Trajectron ++† [96]	AgentFormer [140]	MID† [31]	GP-Graph [5]	NPSN [6]	SocialVAE [126]	EqMotion [125]	EigenTrajectory [7]	LED [67]	**LMTraj-SUP**
ETH	0.77 / 1.40	0.65 / 1.10	0.61 / 1.07	0.61 / 1.03	0.46 / 0.80	0.57 / 0.93	0.43 / 0.63	**0.36** / 0.59	0.41 / 0.58	0.40 / 0.61	**0.36** / 0.53	0.39 / 0.58	0.41 / **0.51**
HOTEL	0.43 / 0.88	0.50 / 0.86	0.22 / 0.39	0.20 / 0.28	0.14 / 0.22	0.21 / 0.33	0.18 / 0.30	0.16 / 0.25	0.13 / 0.19	0.12 / 0.18	0.12 / 0.19	**0.11** / 0.17	0.12 / **0.16**
UNIV	0.75 / 1.50	0.44 / 0.80	0.34 / 0.56	0.30 / 0.55	0.25 / 0.45	0.29 / 0.55	0.24 / 0.42	0.23 / 0.39	**0.21** / 0.36	0.23 / 0.43	0.24 / 0.43	0.26 / 0.43	0.22 / **0.34**
ZARA1	0.35 / 0.69	0.34 / 0.53	0.25 / 0.45	0.24 / 0.41	0.18 / 0.30	0.28 / 0.50	0.17 / 0.31	0.18 / 0.32	**0.17 / 0.29**	0.18 / 0.32	0.19 / 0.33	0.18 / 0.26	0.20 / 0.32
ZARA2	0.36 / 0.72	0.31 / 0.48	0.19 / 0.33	0.18 / 0.32	0.14 / 0.24	0.20 / 0.37	0.15 / 0.29	0.14 / 0.25	**0.13 / 0.22**	**0.13** / 0.23	0.14 / 0.24	**0.13 / 0.22**	0.17 / 0.27
AVG	0.53 / 1.04	0.45 / 0.75	0.32 / 0.56	0.31 / 0.52	0.23 / 0.40	0.31 / 0.54	0.23 / 0.39	**0.21** / 0.36	**0.21** / 0.33	**0.21** / 0.35	**0.21** / 0.34	**0.21** / 0.33	0.22 / **0.32**
SDD	13.6 / 24.6	20.8 / 33.2	10.0 / 15.9	11.4 / 20.1	8.7 / 14.9	7.6 / 14.3	9.1 / 13.8	8.6 / 11.9	8.1 / 11.7	7.9 / 11.9	8.1 / 13.1	8.5 / 11.7	**7.8 / 10.1**
GCS	15.9 / 32.6	14.7 / 23.9	17.1 / 29.3	12.8 / 24.2	10.2 / 16.9	10.7 / 18.2	7.8 / 13.7	7.7 / 13.4	7.4 / 11.9	7.6 / 13.1	7.4 / 12.5	N / A	**7.1 / 9.6**

Table 4. Comparison of LMTraj-SUP methods with other state-of-the-art deterministic and stochastic methods (ADE/FDE, Unit: meter for ETH/UCY and pixel for SDD/GCS). †: Issues raised in the authors' GitHubs are fixed, **Bold**: Best, Underline: Second best.

For the supervised training, we leverage the full potential of our LMTraj-SUP model by integrating all proposed techniques in Secs. 3.2 and 3.3. We use the BPE model [97] for the tokenizer, and encoder-decoder, T5 [89], as our backbone language model. The two models are trained using all the multi-task QA templates. T5 is trained using a cross-entropy loss between the generated outputs and the tokenized ground-truth answers in an end-to-end manner. The hyper-parameters d and τ in Sec. 3.3 are empirically set to 2 and 0.7, respectively. AdamW optimizer [61] is used, whose batch size is 512 and learning rate is 1e-4 over 200 epochs. The training time takes about 4 hours, leveraging a distributed data parallel pipeline on a machine of 8 NVIDIA 4090 GPUs.

4. Experiments

In this section, we conduct comprehensive experiments to verify the effectiveness of our language-based approach for trajectory prediction. We first describe the experimental setup in Sec. 4.1. We then provide comparison results with various zero-shot and supervised trajectory prediction methods in Sec. 4.2. We lastly conduct an extensive ablation study to validate the effect of each component in our method in Sec. 4.3.

4.1. Experimental Setup

Datasets. We conduct experiments on four public datasets: ETH [81], UCY [48], Stanford Drone Dataset (SDD) [93], and the Grand Central Station (GCS) [138] dataset to compare our LMTraj model with state-of-the-art trajectory predictors. The ETH and UCY datasets consist of five subset scenes (ETH, Hotel, Univ, Zara1 and Zara2) with 1,536 pedestrians recorded with the surveillance camera. We use the standard train-val-test split and adopt the leave-one-out strategy [1, 32] for the training and evaluation. SDD has 5,232 trajectories of six agent categories, including pedestrians, cars, and bicyclists, in eight different university campus scenes from a top-down drone view. GCS shows a highly congested terminal scene with 12,684 pedestrians streaming to the exit. We follow the standard benchmark protocol in [6, 32, 34, 73, 99] that the first 3.2 seconds ($T_{obs} = 8$ frames) are used as observations, and the following 4.8 seconds ($T_{pred} = 12$ frames) are predicted.

Evaluation protocols. To evaluate the tokenizer for LMTraj, we use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score from the NLP task to measure the text similarity. Specifically, ROUGE-1 checks the overlap ratio of each word between the source and target sentence. In order to measure the prediction accuracy of LMTraj, we use two metrics as accuracy measures: Average Displacement Error (ADE) and Final Displacement Error (FDE). The ADE and FDE compute the Euclidean distance between a predicted and a ground-truth path and their destination, respectively. Following [32], we generate $K = 1$ samples for the most-likely evaluation and $K = 20$ samples for the multimodal trajectory prediction. For the multimodal predictions, we generate multiple outputs and then choose the best path for the performance evaluation.

4.2. Evaluation Results

Evaluation of the numerical tokenizer. First, we check the efficiency of our numerical tokenizers compared to pre-trained tokenizers in [89] trained on texts. To find the most suitable tokenizer type for numerical trajectory data, we test four types: char, word, unigram, and byte pair encoding (BPE) using six forecasting QA prompts. In particular, the Char-based model [111] breaks the text down into individual characters, while the word-based model [82] splits the text into words, which are separated by whitespace. BPE [97] tokenizes sentences by iteratively searching the text and by repeatedly merging the most frequent sequence pairs of

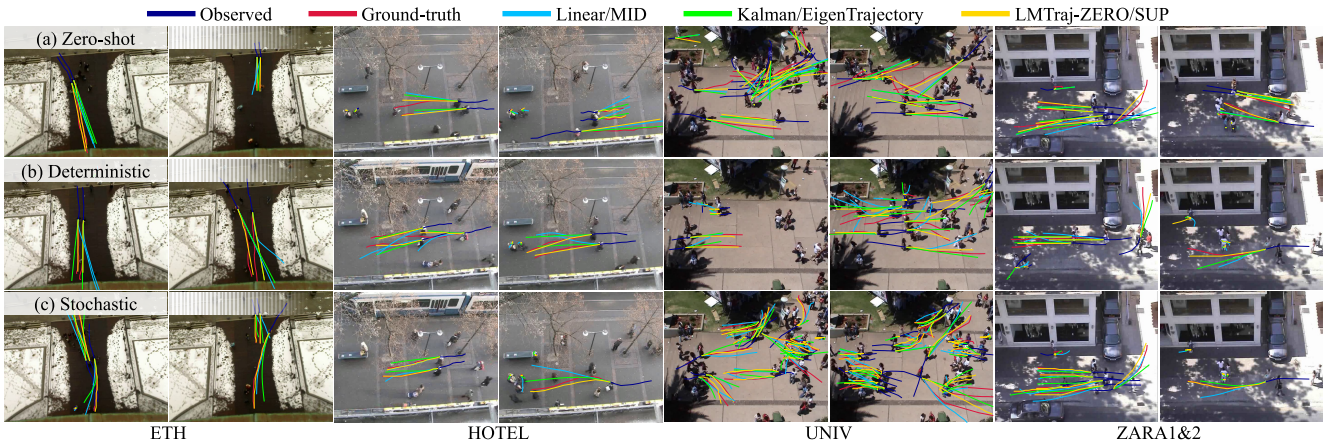


Figure 3. Visualization of prediction results on (a) zero-shot and two supervised trajectory prediction benchmarks: (b) deterministic and (c) stochastic approach. To aid visualization for the stochastic approach, we report one best trajectory of $K = 20$ samples each.

letters in a vocabulary. The Unigram model [43] uses an approach similar way to BPE, but generates a vocabulary by lexicalizing byte pairs based on probability values for neighboring characters. Additionally, the pretrained model employs the unigram model trained on 750GB of web crawled text data [89].

Table 2 shows that all five methods cover all the words in the trajectory prompts well. Since no vocabulary is missing, the input and output sentences are exactly the same as the original sentence after tokenization. However, the pretrained and word tokenizers often have a mixture of letters and numbers. As shown in Fig. 2-(b,d), certain tokens combine letters and numbers, and even multiple tokens are used to represent the integer part of the number 6. This disturbs LMTraj's understanding of the sequential nature of the numbers. While a Char tokenizer is capable of separating numerical and letter notation, it requires too many tokens for tokenization, as in Fig. 2-(c). Both unigram and BPE tokenizers effectively distinguish between letters and numbers while decreasing the average number of tokens by merging multi-digit numbers into a single token in Fig. 2-(e,f). This directly reduces the computational complexity of LMTraj. We select the BPE tokenizer for our LMTraj-SUP model, thanks to its ability to represent a sentence with a smaller number of tokens.

Evaluation of the zero-shot approach. To demonstrate the potential of prompt engineering with language foundation models for trajectory prediction, we conduct a quantitative comparison between LMTraj-ZERO and various zero-shot methods. We provide three algorithmic approaches and one learnable model. The 'Stop' operates by stopping walking at the final observation point without making predictions, while 'Linear' and 'Kalman filter' methods serve as state extrapolation techniques. AutoTrajectory [62] is trained in an unsupervised manner without any ground-truth trajectory label. As shown in Tab. 3 and Fig. 3-(a), we observe that our method achieves the best performance among all the zero-shot methods. Particularly, using LMTraj-ZERO with GPT-4 yields results superior to that of GPT-3.5, indicating the model has better performance when combined with larger language foundation models. LMTraj-ZERO with GPT-4 shows comparable performance to the supervised model, Social-STGCNN [73]. This opens the possible study of zero-shot trajectory prediction.

Evaluation of the supervised approach. Next, to check the maximum performance of the linguistic approach, we compare LMTraj-SUP to both deterministic and stochastic trajectory prediction methods. As shown in Tab. 4, LMTraj-SUP outperforms the deterministic predictions on most datasets, while the other models reached a plateau. This demonstrates the effectiveness of the LMTraj-SUP for reasoning about complex social relationships in Fig. 4 as well as performing beam search based on cumulative probabilities for the most likely path, as visualized in Fig. 3-(b). This provides a significant advantage over the previous works that rely on the graph-based social interaction modeling and the greedy selection of footsteps.

In addition, LMTraj-SUP also shows promising results for stochastic trajectory prediction. By understanding potential future behavior patterns through scene descriptions and social reasoning, LMTraj-SUP, generating sentences of realistic trajectories, achieves better performance than the previous works. As shown in Fig. 3-(c), our LMTraj-SUP successfully generates multimodal trajectories using the temperature tuning technique to diversify the outputs, as in NLP. This means that our approach offers a new potential solution to the limited performance of existing physics-based social relationships.

4.3. Ablation Studies

Effectiveness of the numerical tokenizer. We compare the effectiveness of the text-based pretrained tokenizer with our numerical tokenizer for stochastic trajectory prediction.

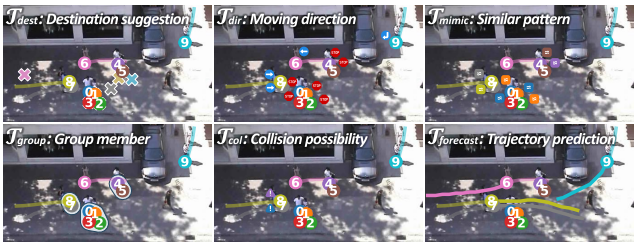


Figure 4. Visualization of the social reasoning using observed paths and the corresponding trajectory prediction results.

Model		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Tokenizer	Pretrained	0.85/1.49	0.46/0.93	0.97/2.00	0.55/1.06	0.43/0.89	0.65/1.28
	Numerical	**0.65/1.04**	**0.26/0.46**	**0.57/1.16**	**0.51/1.01**	**0.38/0.74**	**0.48/0.88**
Size	Small	**0.65/1.04**	0.26/0.46	0.57/1.16	0.51/1.01	**0.38**/0.74	0.48/**0.88**
	Medium	0.68/1.17	0.26/**0.45**	0.57/1.16	0.51/1.02	0.39/0.76	0.48/0.91
	Large	0.71/1.22	0.26/0.46	0.57/1.16	**0.50/1.00**	**0.38/0.73**	0.48/0.91
Multi-task	No	0.74/1.27	0.31/0.59	0.74/1.51	0.53/1.06	0.41/0.79	0.55/1.04
	Yes	**0.66/1.07**	0.26/0.46	0.57/**1.16**	**0.52/1.02**	0.38/**0.74**	0.48/**0.89**
Depth	$d=1$	0.66/1.05	0.26/0.46	0.57/1.17	**0.51/1.00**	0.38/0.75	0.48/0.89
	$d=2$	**0.65/1.04**	0.26/0.46	0.57/**1.16**	**0.51**/1.01	0.38/0.74	0.48/**0.88**
	$d=3$	0.66/1.07	0.26/0.46	0.57/**1.16**	0.52/1.02	0.38/0.74	0.48/0.89
	$d=4$	0.67/1.09	0.26/0.46	0.57/**1.16**	0.52/1.03	0.38/**0.73**	0.48/0.89
	$d=5$	0.67/1.10	0.26/0.46	0.57/**1.16**	0.52/1.03	0.38/0.74	0.48/0.90
Temperature	$\tau=0.1$	0.49/0.70	0.19/0.31	0.41/0.80	0.34/0.64	0.29/0.53	0.34/0.60
	$\tau=0.3$	0.45/0.58	0.15/0.21	0.29/0.52	0.26/0.45	0.22/0.38	0.27/0.43
	$\tau=0.5$	0.42/0.54	0.13/0.17	0.24/0.41	0.21/0.36	0.19/0.31	0.24/0.36
	$\tau=0.7$	**0.41/0.51**	**0.12/0.16**	**0.22/0.34**	**0.20/0.32**	**0.17/0.27**	**0.22/0.32**
	$\tau=0.9$	0.42/0.53	0.13/0.18	**0.22**/0.35	0.22/0.35	0.18/**0.27**	0.23/0.34

Table 5. Ablation studies on each component of LMTraj-SUP (ADE/FDE, meter). **Bold**: Best, Underline: Second best.

In Tab. 5, our LMTraj-SUP with the numerical tokenizer outperforms the pretrained tokenizer in deterministic prediction accuracy. This shows the advantage of our approach for numerical tasks, by allowing the model to better understand numerical information from sentences.

Model size. Next, we vary the sizes of the backbone Seq2Seq model in Tab. 5. As expected, the performance varies slightly with increasing model size, but the gain is marginal. As a result, we choose the smallest and the lightweight model for the real-time prediction.

Multi-task training strategy. To enhance the model's ability to reason about social interactions, we include a multitask training in the forecasting pipeline. As reported in Tab. 5, the multi-task training strategy improves the performance compared to a single-task training strategy. This improvement suggests that integrating domain knowledge pushes the model to better understand group behaviors and collision avoidance, helpful for the main forecasting task.

Beam-search and temperature analysis. We conduct a parameter study for both the deterministic and stochastic predictions. Table 5 validates that using beam search with a depth of $d = 2$ and the temperature-tuning with $\tau = 0.7$ produces the best performance. In addition, adjusting the temperature parameter τ affects the level of uncertainty in the multi-modal generation, allowing for controlled variations within socially acceptable limits in Fig. 5

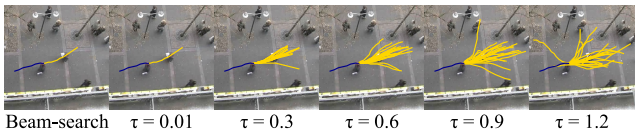


Figure 5. Visualization of the most-likely and multimodal trajectory generation capability of our LMTraj-SUP (τ: temperature).

Model	Accuracy		Complexity		
	ADE	FDE	GPU memory	Training	Inference
PECNet [64]	0.32	0.56	1733 MB	**0.3 h**	57.0 ms
MID [31]	0.31	0.54	2929 MB	6.9 h	35.0 ms
STAR [139]	0.26	0.53	1735 MB	36.3 h	97.0 ms
AgentFormer [140]	0.23	0.40	9639 MB	22.0 h	**8.2 ms**
SocialVAE [126]	**0.21**	0.33	1762 MB	2.1 h	73.0 ms
LMTraj-SUP	0.22	**0.32**	**1401 MB**	3.8 h	18.3 ms

Table 6. Computational complexity analysis of our LMTraj-SUP with other numerical-based trajectory prediction models. 'Inference' measures the average inference time per trajectory.

Computational cost. Lastly, we check the computational efficiency of LMTraj-SUP in Tab. 6. Due to the structural nature of the language model that sequentially predicts the next token, the inference time is a little slower than the fastest model. However, it produces promising results with the reasonable GPU memory consumption as well as real-time inference.

5. Conclusion

This paper demonstrates the ability of language models to understand and extrapolate spatio-temporal numeric information from trajectory data. We shift the domain of the trajectory prediction task to a question-answering task, which provides historical data as context and then forecasts futures when answering the given question templates. The history data, transformed into a text prompt format, can offer rich information for the language model, and capture human dynamics. We show that both the prompt engineering of the language foundation models and their end-to-end training can successfully predict accurate future paths in zero-shot and supervised manners using our LMTraj-ZERO and LMTraj-SUP, respectively. In addition, the specialized techniques for large language models, including tokenizer optimization, multi-task learning, beam-search, and temperature tuning scheme, allow our model to better comprehend high-level social reasoning, and to operate like conventional deterministic and stochastic trajectory predictor models.

Acknowledgement This research was supported by 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT (No.S1602-20-1001), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST) and No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 6

[2] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[3] Inhwan Bae and Hae-Gon Jeon. Disentangled multi-relational graph convolutional network for pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1, 2

[4] Inhwan Bae and Hae-Gon Jeon. A set of control points conditioned pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 5

[5] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6

[6] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6

[7] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. EigenTrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6

[8] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the IEEE International Conference on Machine Learning (PMLR)*, 2020. 3

[9] Taylor Berg-Kirkpatrick and Daniel Spokoyny. An empirical investigation of contextualized number prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 3

[10] Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008*, 2020. 2

[11] Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018. 2

[12] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. Neural machine translation with monolingual translation memory. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 3

[13] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2019. 2

[14] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[15] 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 (ICCV)*, 2021. 2

[16] Guangyi Chen, Zhenhao Chen, Shunxing Fan, and Kun Zhang. Unsupervised sampling promoting for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[17] Hao Chen, Jiaze 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 (ICCV)*, 2023. 2

[18] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 5

[19] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[20] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[21] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixe. Goal-gan: Multimodal trajectory prediction based on goal position estimation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

[22] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[23] Nachiket Deo and Mohan M. Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020. 2

[24] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2021. 3

[25] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2019. 3

[26] Yonghao Dong, Le Wang, Sanping Zhou, and Gang Hua. Sparse instance conditioned multimodal trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[27] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018. 1

[28] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Networks*, 108:466–478, 2018. 2

[29] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 5

[30] Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2020. 1, 3

[31] 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 (CVPR)*, 2022. 2, 6, 8

[32] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6

[33] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[34] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 6

[35] Ronny Hug, Wolfgang Hübner, and Michael Arens. Introducing probabilistic bézier curves for n-step sequence prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[36] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[37] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[38] Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. ForecastQA: a question answering challenge for event forecasting with temporal text data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 3

[39] Jacob Devlin Ming-Wei Chang Kenton and Lee 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 (NAACL-HLT)*, 2019. 2, 4

[40] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[41] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2

[42] Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[43] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 4, 7

[44] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018. 1, 4

[45] Yen-Ling Kuo, Xin Huang, Andrei Barbu, Stephen G McGill, Boris Katz, John J Leonard, and Guy Rosman. Trajectory prediction with linguistic representations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 1, 3

[46] Mihee Lee, Samuel S. Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[47] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[48] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3): 655–664, 2007. 6

[49] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. 1, 2

[50] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with

dynamic relational reasoning. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[51] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 3

[52] Shijie Li, Yanying Zhou, Jinhui Yi, and Juergen Gall. Spatial-temporal consistency network for low-latency trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[53] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 5

[54] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[55] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[56] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[57] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[58] Rongqin Liang, Yuanman Li, Xia Li, Yi Tang, Jiantao Zhou, and Wenbin Zou. Temporal pyramid network for pedestrian trajectory prediction with multi-supervision. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[59] Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1, 3

[60] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 6

[62] Yuexin Ma, Xinge Zhu, Xinjing Cheng, Ruigang Yang, Jiming Liu, and Dinesh Manocha. Autotrajectory: Label-free trajectory extraction and prediction from videos using dynamic points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 7

[63] Takahiro Maeda and Norimichi Ukita. Fast inference and update of probabilistic density estimation on trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[64] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6, 8

[65] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[66] Huynh Manh and Gita Alaghband. Scene-lstm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*, 2018. 2

[67] 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 (CVPR)*, 2023. 2, 6

[68] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Mantra: Memory augmented networks for multiple trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[69] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2

[70] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Smemo: Social memory for trajectory forecasting. *arXiv preprint arXiv:2203.12446*, 2022. 2

[71] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[72] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2013. 4

[73] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, 7

[74] Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[75] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[76] Ki-In Na, Ue-Hwan Kim, and Jong-Hwan Kim. Spubert: Faster human multi-trajectory prediction from socio-physical understanding of bert. *Knowledge-Based Systems*, 2023. 3

[77] Ingrid Navarro and Jean Oh. Social-patternn: Socially-aware trajectory prediction guided by motion patterns. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. 2

[78] Ehsan Pajouheshgar and Christoph H Lampert. Back to square one: probabilistic trajectory forecasting without bells and whistles. In *Proceedings of the Neural Information Processing Systems Workshop (NeurIPSW)*, 2018. 2

[79] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[80] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 4

[81] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2009. 2, 6

[82] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4, 6

[83] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 5

[84] Mark Pfeiffer, Giuseppe Paolo, Hannes Sommer, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2

[85] Mozhgan Pourkeshavarz, Changhe Chen, and Amir Rasouli. Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[86] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. ProphetNet: Predicting future n-gram for sequence-to-SequencePretraining. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 3

[87] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2, 4

[88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2, 3

[89] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 2, 4, 6, 7

[90] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2

[91] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[92] Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 1, 3

[93] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6

[94] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal on Computer Vision (IJCV)*, 2015. 3

[95] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[96] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6

[97] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016. 6

[98] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human trajectory prediction via conditional 3d attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[99] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 6

[100] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. Social interpretable tree for pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2

[101] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[102] Xiaodan Shi, Xiaowei Shao, Zipei Fan, Renhe Jiang, Haoran Zhang, Zhiling Guo, Guangming Wu, Wei Yuan, and Ryosuke Shibasaki. Multimodal interaction-aware trajectory prediction in crowded space. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1, 2

[103] Xiaodan Shi, Xiaowei Shao, Guangming Wu, Haoran Zhang, Zhiling Guo, Renhe Jiang, and Ryosuke Shibasaki. Socialdpf: Socially acceptable distribution prediction of futures. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[104] Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 1, 3

[105] Georgios Spithourakis and Sebastian Riedel. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 3

[106] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[107] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[108] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[109] Jianhua Sun, Yuxuan Li, Liang Chai, Hao-Shu Fang, Yong-Lu Li, and Cewu Lu. Human trajectory prediction with momentary observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[110] Jianhua Sun, Yuxuan Li, Liang Chai, and Cewu Lu. Stimulus verification is a universal and effective sampler in multi-modal human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[111] Ilya Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011. 6

[112] Chaofan Tao, Qinhong Jiang, and Lixin Duan. Dynamic and static context-aware lstm for multi-agent motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[113] Elliot Turiel. *The development of social knowledge: Morality and convention*. Cambridge University Press, 1983. 2

[114] Daksh Varshneya and G. Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. *arXiv preprint arXiv:1705.09436*, 2017. 2

[115] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017. 1

[116] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[117] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2

[118] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 2

[119] Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[120] Song Wen, Hao Wang, and Dimitris Metaxas. Social ode: Multi-agent trajectory forecasting with neural ordinary differential equations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[121] Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[122] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 4

[123] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[124] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), 2022. 1

[125] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[126] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6, 8

[127] Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. Improving multilingual neural machine translation with auxiliary source languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 3

[128] 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 (AAAI)*, 2020. 2

[129] Yi Xu, Lichen Wang, Yizhou Wang, and Yun Fu. Adaptive trajectory prediction via transferable gnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[130] 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 (CVPR)*, 2023. 2

[131] Hao Xue and Flora D Salim. PromptCast: A new prompt-based learning paradigm for time series forecasting. *arXiv preprint arXiv:2210.08964*, 2022. 1, 3, 5

[132] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2

[133] Hao Xue, Flora D Salim, Yongli Ren, and Charles LA Clarke. Translating human mobility forecasting through natural language generation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022. 4

[134] Hao Xue, Bhanu Prakash Voutharoja, and Flora D Salim. Leveraging language foundation models for human mobility forecasting. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, 2022. 1, 3

[135] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[136] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[137] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters (RA-L)*, 2021. 2

[138] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[139] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6, 8

[140] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 8

[141] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[142] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6

[143] He Zhao and Richard P. Wildes. Where are you heading? dynamic trajectory prediction with expert goal examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[144] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, Congcong Li, and Dragomir Anguelov. Tnt: Target-driven trajectory prediction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2020. 2

[145] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[146] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 3

[147] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[148] Yiyao Zhu, Di Luan, and Shaojie Shen. Biff: Bi-level future fusion with polyline-based coordinate for interactive trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[149] Mingyu Zong and Bhaskar Krishnamachari. Solving math word problems concerning systems of equations with gpt-3. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 1, 3