2024 SAIL seminar

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

CVPR 2024

2024. 11. 14

김병훈

순천향대학교 일반대학원 석사과정

Contents

I Introduction

II Related works

III Methodology

IV Experiments

V Conclusion

1

Introduction

I Introduction

Background

Importance of Pedestrian Trajectory Prediction

- Forecasting pedestrian trajectories is essential for systems such as route planning, social robots and others
- Conventional methods are operated by taking in pedestrian coordinates and estimating future paths
- Specifically, additional information about social interactions can significantly improve performance
  - Social interactions: distance between pedestrians, similarity of behavior, etc.

Development of Language Models(LM)

- Language models have recently advanced, demonstrating the ability to provide contextual understanding and condition generation in a variety of tasks
- In particular, it can describe social reasoning beyond physics-based interactions
- Therefore, introducing language models for trajectory prediction can improve interaction modeling

I Introduction

Problem Statement

Problems of introducing LM into trajectory prediction

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

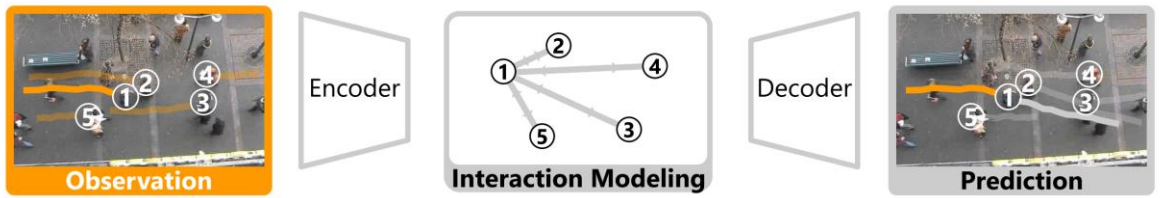
Research Objective

LMTraj

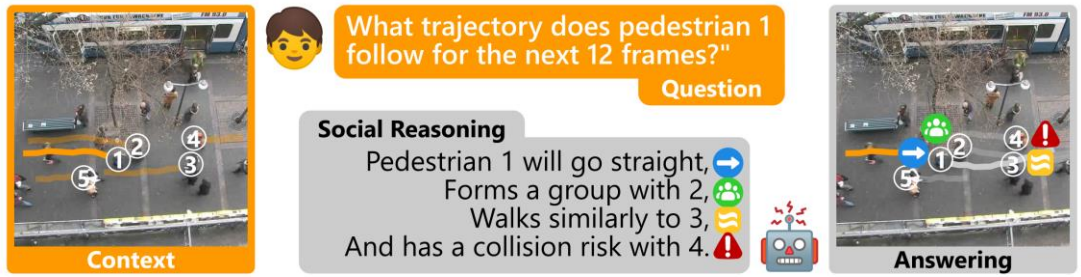
(Language-based Trajectory Prediction)

- A novel NLP model is propose to predict the future trajectory of pedestrians
- The proposed model effectively solves the existing problem

Model Operation Comparison



(a) Previous trajectory prediction

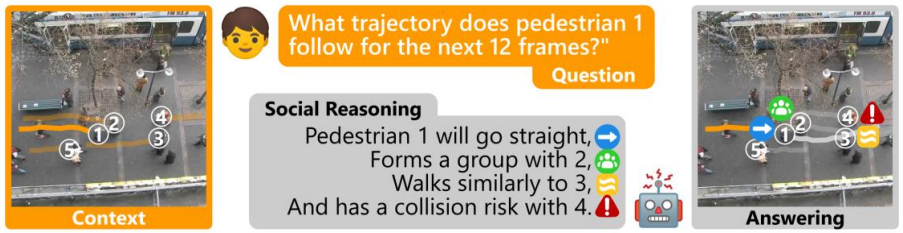


(b) Our language-based trajectory prediction

I Introduction

Proposed Method

- **LMTraj consists of four steps:**



(b) Our language-based trajectory prediction

1. Convert raw trajectory coordinates and scenes to text prompts and integrate into QA templates
2. Introduce additional information to predict future trajectories reflecting social relationships
3. Optimized tokenization to clearly split numerical and text to learn sequential correlations
4. Incorporated beam search and temperature tuning techniques to generate the most probable multi-modal trajectories

Contributions

- Proposed the LMTraj to interpret and predict spatio-temporal numerical information in trajectory data
- By converting trajectory prediction into a QA task, we demonstrated accurate predictability of zero-shot and supervised learning methods through LMTraj-ZERO and LMTraj-SUP
- Enhanced social reasoning by applying language model optimization techniques such as tokenization optimization and beam search

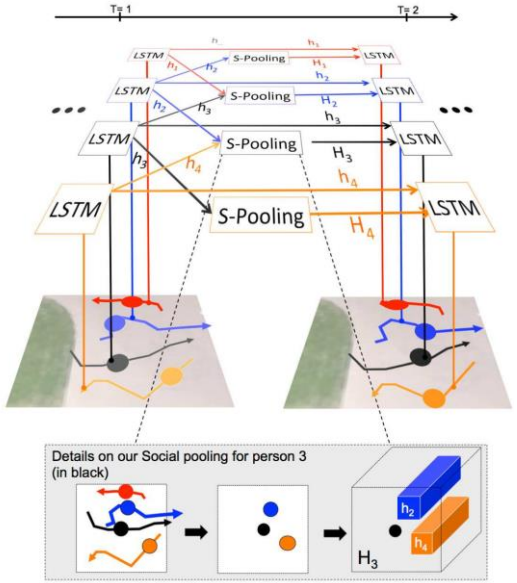
2

Related works

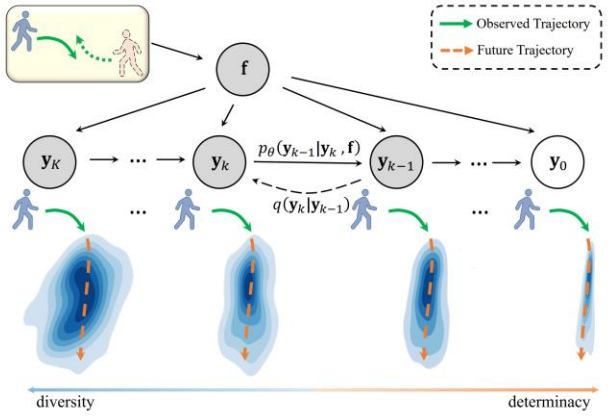
II Related works

Pedestrian Trajectory Prediction

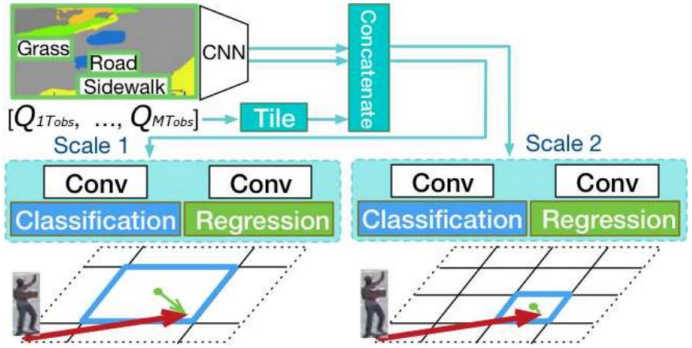
- Numerical-based prediction
  - Social-LSTM
    - Social interaction between neighboring agents is modeled by aggregating hidden states via a pooling mechanism
  - Attention, GCN, GAT, Transformer
    - Directly model mutual influences among agents
- Numerical + Environmental information
- Probabilistic trajectory generation
  - Bivariate Gaussian distribution
  - Generative Adversarial Networks(GANs)
  - Conditional Variational AutoEncoder (CVAE)
  - Diffusion models
- Heatmap-based prediction
  - Pixel-level path prediction
- Discretized(Manhattan) grid-based prediction → Social norms are not considered
  - Simplified path prediction



Social-LSTM



Diffusion model



Discretized Grid

II Related works

Language-Based Reasoning and Prediction

- Transformer architectures
  - BERT: Masked language modeling (MLM)
  - GPT-2: Causal language modeling (CLM)
  - T5: Sequence-to-sequence (Seq2Seq) modeling
- LM-based time-series forecasting
  - ForecastQA
    - Proposes a QA benchmark
    - Verify forecasting ability for future events
  - PromptCast
    - Predictions on weather temperature, energy consumption, and customer flow
- Using linguistic intermediate representations
  - Kuo et al.(2022) - Most relevant work
    - Using linguistic intermediate representations for trajectory prediction
    - Solving action-related reasoning through language priors

Language foundation model

Text generative tasks
- Machine translation
- Text generation
- Question-answering

Other tasks
- Vision-language tasks
- Mathematical problem

Limitation

- Using tokenizers pre-trained on text data, language models are not sufficiently utilized
- In particular, these approaches are not well-suited for trajectory prediction tasks due to inconsistent analysis of numerical data
- In addition, it inhibits the understanding of high-level features such as social interactions

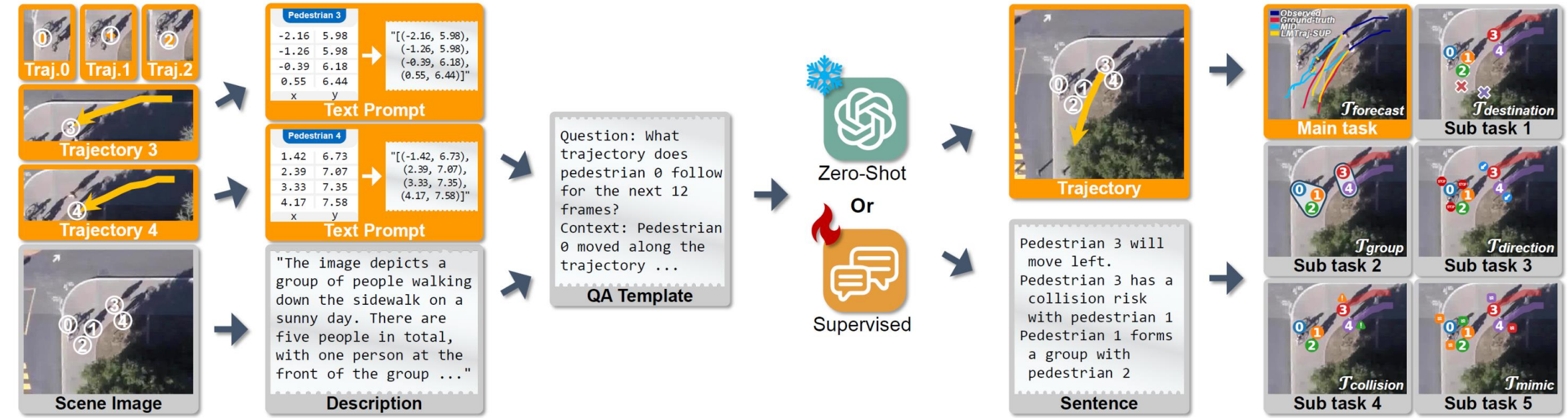
3

Methodology

Methodology

LMTraj Framework

- **Prompt-based trajectory prediction**
  - Recasting trajectory prediction tasks in a sentence-by-sentence manner
    - → Applying language models for numerical prediction



Methodology

Problem Definition

Trajectory prediction problems

- Coordinate sequence for each agent → future coordinates (sequence-to-sequence problem)

$$\mathcal{S}_{n,obs} = \{(x_n^t, y_n^t) \in \mathbb{R}^2 \mid t \in [1, ..., T_{obs}]\} \Rightarrow \mathcal{S}_{n,pred} = \{(x_n^t, y_n^t) \in \mathbb{R}^2 \mid t \in [T_{obs}+1, ..., T_{obs}+T_{pred}]\}$$

- The model takes as input both the past trajectory $\mathcal{S}_{obs}$ and the image $\mathcal{I}$
  - Deterministic prediction: predicting the most-likely trajectory
  - Stochastic prediction: predicting possible multi-modal future trajectories

Methodology

Data Space Conversion to Prompt

- For input to the LM, pedestrian trajectories and image data must be converted into text prompts
- Next, the converted data is aggregated into language sentences using QA templates for input and output of the language model

Converting trajectory coordinates into the prompt

- Convert all float coordinate values to text strings represented in decimal
  - Decimal representation is more compatible with natural language
- Then, combine the 2D coordinate sequences using commas, brackets, etc.
- This process converts the past and future trajectories $\mathcal{S}_{n,obs}$, $\mathcal{S}_{n,pred}$ of all N pedestrians in the scene into a text prompt $\mathcal{P}_{\mathcal{S}_{n,obs}}$, $\mathcal{P}_{\mathcal{S}_{n,pred}}$

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.

Methodology

Data Space Conversion to Prompt

Converting image data into the prompt

- Convert scene image $\mathcal{I}$ to prompt $\mathcal{P}_{\mathcal{I}}$
- Using the BLIP-2 model to extract a textual description of the scene in which the agent is moving
- This allows it to learn various environmental details and identify patterns of movement speed and behavior similar to traditional map encoding
  - Placement of buildings and vehicles, density of people, pedestrian flow, etc.

Converting predictions into the prompt

- Numeric coordinate prompts and scene description prompts are pre-processed and input into LMTraj
- Provide Question and Answer templates for trajectory prediction $\mathcal{T}_{forecast} = \{\mathcal{P}_{\mathcal{C}}, \mathcal{P}_{\mathcal{Q}}, \mathcal{P}_{\mathcal{A}}\}$
- Where, denote the historical coordinates of all agents as context $\mathcal{P}_{\mathcal{C}}$, the question $\mathcal{P}_{\mathcal{Q}}$ about predicting the future trajectory of a particular pedestrian n, and the expected answer $\mathcal{P}_{\mathcal{A}}$

Methodology

Domain Shift to Sentence Generation

Optimizing the tokenizer for numeric data

- Using pre-trained tokenizers to convert QA prompts into a form that LMTraj-SUP can understand
- However, as shown in Figure 2(b), numbers are split irregularly into tokens and special characters such as periods and commas are sometimes grouped together
  - This can disturb the training
- To solve this problem, we train a new tokenizer on numeric data

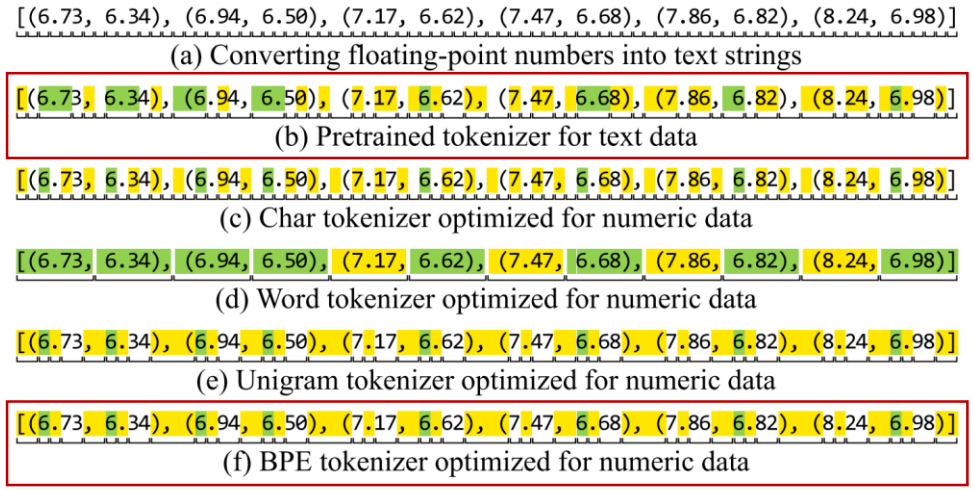


Figure 2. Comparison of the text-pretrained tokenizer and our numeric data-optimized tokenizer. Under brackets with yellow or white highlight colors indicate that the corresponding letters have been tokenized. The green color highlights that the token contains an integer with 6.

Methodology

Domain Shift to Sentence Generation

Multi-task training for social relation reasoning

- In trajectory prediction tasks, modeling the interactions between agents is the most important aspect
- Introduce auxiliary tasks to fully utilize reasoning and understanding of scene context and social dynamics
- Types of auxiliary tasks
  - Destination suggestion
  - Moving direction prediction
  - Similar pattern search
  - Group member prediction
  - Collision possibility assessment
- Extract common features of agent behavior and leverage social knowledge learned from each auxiliary task (e.g., group walking and collision avoidance) to improve the accuracy of the main prediction task

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."

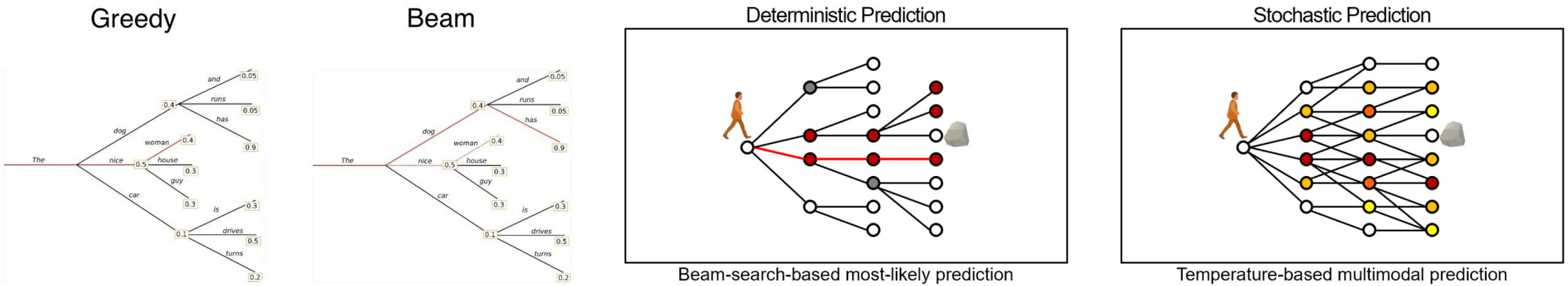
Table 1. QA templates to convert raw trajectory data into prompts.

Methodology

Domain Shift to Sentence Generation

- **Generating most-likely and multimodal outputs**

  - Generate all possible multipath $\widehat{\mathcal{S}}_{pred}^{k}$ or the most likely single path $\widehat{\mathcal{S}}_{pred}$
    - Deterministic prediction (Single path)
      - Using beam search to predict the path $\widehat{\mathcal{P}}_{\mathcal{A}}$ with the highest probability search (depth d)
    - Stochastic prediction (Multipath)
      - Using a temperature parameter τ to modulate the token probabilities to generate different outputs $\widehat{\mathcal{P}}_{\mathcal{A}}^{k}$



Methodology

Forecasting With the Language Model

LMTraj-ZERO: Zero-shot prediction in the language foundation model

- Prompt tuning: fine-tuning language models to optimize input prompts to produce the expected output
- Using GPT-3.5 and GPT-4, which are large pre-trained language models not trained for trajectory prediction
  1. Provide QA prompts to teach LMTraj-ZERO the desired output
  2. Transform the output $\widehat{\mathcal{P}}_{\mathcal{A}}^{k}$ into numerical coordinates $\widehat{\mathcal{S}}_{pred}^{k}$ for evaluation
- Throughout the process, the language model is fixed and not trained or fine-tuned

LMTraj-SUP: Supervision of language-based predictor

- Optimal language model selection: Select the T5 (Seq2Seq) model, which is an encoder-decoder language model
  - MLM cannot be used for trajectory prediction tasks
  - CLM causes accumulative errors
  - T5 encodes the input sequence at once and makes predictions based on it, which solves the above problems and performs better [4, 64, 73]

[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
[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
[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

Methodology

Implementation Details

LMTraj-ZERO

- Using GPT-3.5 and GPT-4 as the fundamental language models for prompt engineering
- Multi-process by creating a pool of 1,000 units of threads to evaluate all paths in the dataset
- If the response does not match the desired answer format, retry

LMTraj-SUP

- Using the BPE model for tokenizer
- Using T5 as the backbone language model
- T5 trains end-to-end using cross-entropy loss between generated output and tokenized ground truth answers

Hyperparameter

- Beam search depth $d = 2$, temperature parameter $\tau = 0.7$
- Optimizer: AdamW, Batch size: 512, Learning rate: 1e-4, Epochs: 200
- NVIDIA 4090 GPU x 8

4

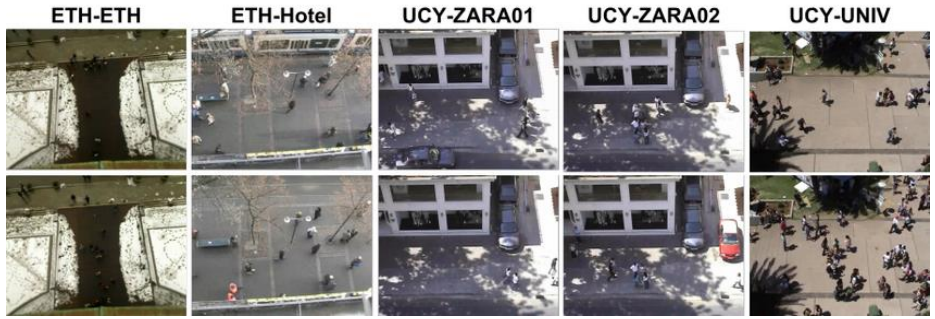
Experiments

IV Experiments

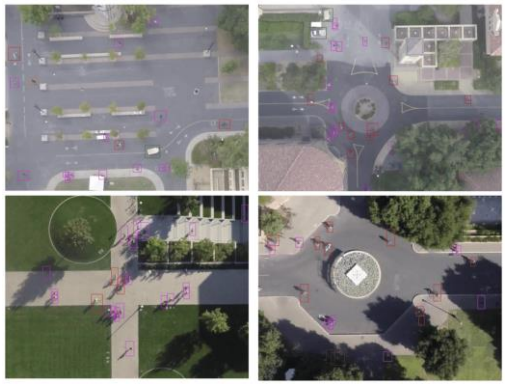
Experimental Setup

Datasets

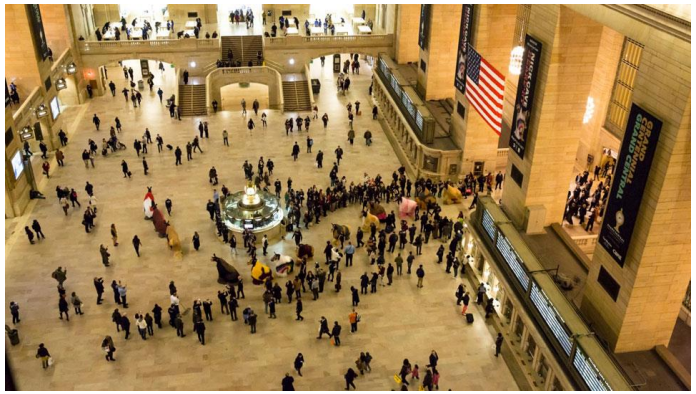
- ETH/UCY Dataset:
  - Different scenes from 5 CCTV cameras
  - 1,536 pedestrians



- Stanford Drone Dataset (SDD):
  - Drone videos from 8 different college campuses
  - 6 agent categories: pedestrian, car, bicycle, etc.
  - 5,232 trajectories



- Grand Central Station (GCS) Dataset:
  - Video of pedestrians swarming the terminal exit during extremely crowded conditions
  - 12,684 pedestrians



Using the first 3.2 seconds (T_{obs} = 8 frames) as observation and the next 4.8 seconds (T_{pred} = 12 frames) as prediction

IV Experiments

Experimental Setup

Evaluating tokenization

- To evaluate similarity, using Recall-Oriented Understudy for Gisting Evaluation-1 (ROUGE-1) scores
  - Specifically, ROUGE-1 checks the overlap ratio of each word between the source and target sentence

$$\textbf{ROUGE-1} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_1 \in S} Count_{match}(gram_1)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_1 \in S} Count(gram_1)}$$

정답문장: "한화는 10 년 안에 우승 할 것이다."
생성문장: "두산은 3 년 안에 우승 할 것이다."

$$N_{\text{정답문장}} = 7$$

$$N_{\text{년,안,우승,할,것이다}} = 5$$

$$ROUGE-1 = \frac{5}{7}$$

Evaluating trajectory prediction accuracy

- To measure accuracy, using Average Displacement Error (ADE) and Final Displacement Error (FDE)
  - Generate K=20 for stochastic prediction

$$\text{ADE} = \frac{1}{T}\sum_{t=1}^{T} \|p_T^{pred} - p_T^{gt}\| \qquad \text{FDE} = \|p_T^{pred} - p_T^{gt}\|$$

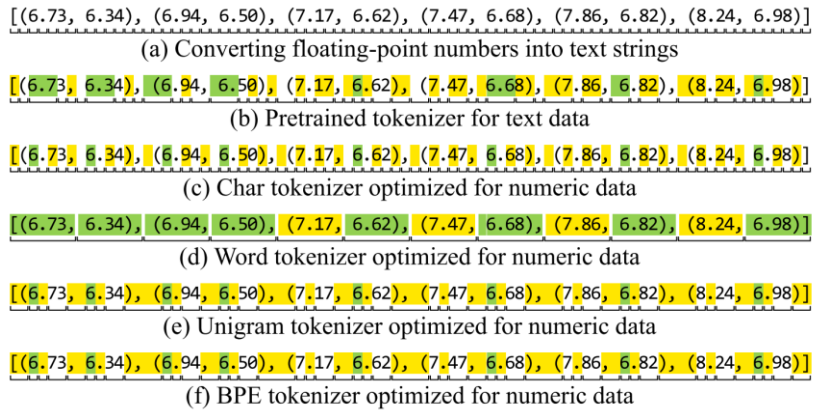
IV Experiments

Evaluation Results

Evaluation of the numerical tokenizer

- Char: breaks the text down into individual characters
- Word: splits the text into words, which are separated by whitespace
- Unigram: Tokenizes sentences by lexicalizing byte pairs based on probability values for neighbor characters
- Byte Pair Encoding (BPE): Tokenizes sentences by iteratively merging the most frequent pairs of character sequences in a vocabulary

"Selected a BPE tokenizer that can represent sentences with a reduced number of tokens"



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.

IV Experiments

Evaluation Results

Evaluation of the zero-shot approach

- Quantitative evaluation
  - Achieved the best performance of all zero-shot methods
  - LMTraj-ZERO with GPT-4 shows similar performance to the supervised model Social-STGCNN

Zero-shot	Stop	Linear	Kalman filter	AutoTraj-ectory [62]	LMTraj-ZERO -GPT-3.5	LMTraj-ZERO -GPT-4	Social-STGCNN[73]
ETH	2.84/4.82	1.00/2.23	0.94/2.13	N/A	1.07/1.82	**0.80/1.64**	0.65 / 1.10
HOTEL	1.15/2.09	0.32/0.62	0.26/0.50	N/A	0.42/0.65	**0.20/0.37**	0.50 / 0.86
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**	0.44 / 0.80
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**	0.34 / 0.53
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**	0.31 / 0.48
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**	0.45 / 0.75
SDD	64.0/116.7	18.8/38.0	16.6/33.9	N/A	17.8/29.1	**10.9/21.0**	20.8 / 33.2
GCS	76.0/138.8	18.9/40.7	18.3/39.4	N/A	27.7/44.8	**12.7/25.5**	14.7 / 23.9

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.

- Qualitative evaluation

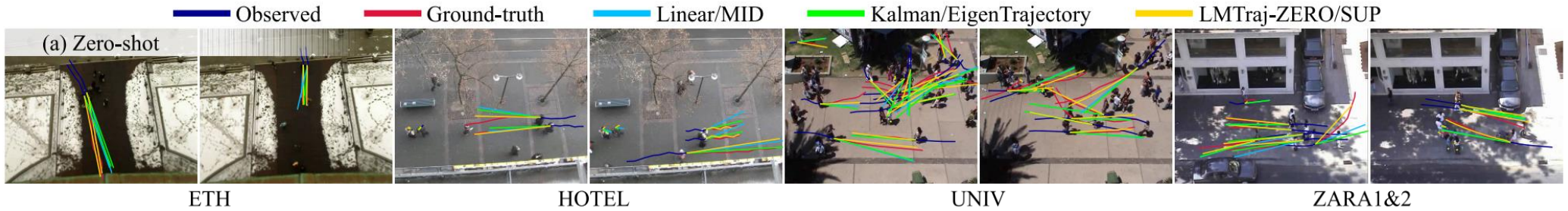


Figure 3. Visualization of prediction results on (a) zero-shot and two supervised trajectory prediction benchmarks: (b) deterministic and (c) stochastic approach. To aid visualization for the stochastic approach, we report one best trajectory of $K = 20$ samples each.

IV Experiments

Evaluation Results

- **Evaluation of the supervised approach**
  - Quantitative evaluation

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]	EigenTraj ectory [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]	AgentFor mer [140]	MID† [31]	GP-Graph [5]	NPSN [6]	SocialVAE [126]	EqMotion [125]	EigenTraj ectory [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.

IV Experiments

Evaluation Results

- **Evaluation of the supervised approach**
  - Qualitative evaluation

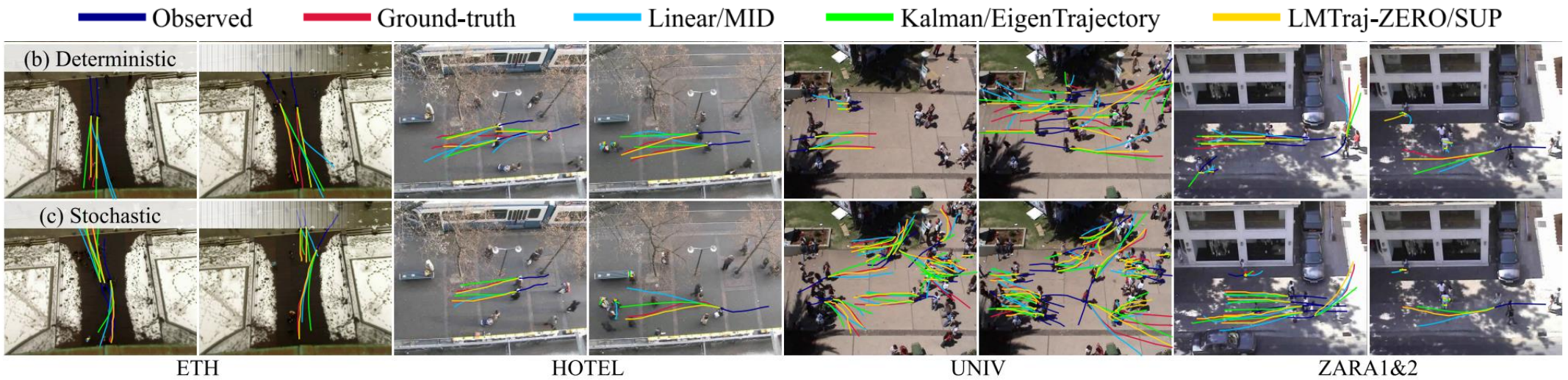


Figure 3. Visualization of prediction results on (a) zero-shot and two supervised trajectory prediction benchmarks: (b) deterministic and (c) stochastic approach. To aid visualization for the stochastic approach, we report one best trajectory of $K = 20$ samples each.

IV Experiments

Ablation Studies

Effectiveness of the numerical tokenizer

- Comparison of the effectiveness of text-based pre-trained tokenizers vs. numeric tokenizers
- Using a numeric tokenizer, outperforms pre-trained tokenizers in deterministic prediction accuracy

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**

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

Model size

- Comparison of performance differences based on the size of backbone Seq2Seq models
- Performance increases slightly as model size increases, but the increase is not significant
- As a result, we select the lightweight model for real-time prediction

Model		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
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

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

IV Experiments

Ablation Studies

Multi-task training strategy

- Multi-task learning strategy outperforms single-task learning strategy
- By incorporating domain knowledge, the model better understands group behavior and conflict avoidance, suggesting that it helps with key predictive tasks

Model		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
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**

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

Beam-search and temperature analysis

- Validate that using beam search with depth $d = 2$ and temperature tuning of $\tau = 0.7$ produces the best performance

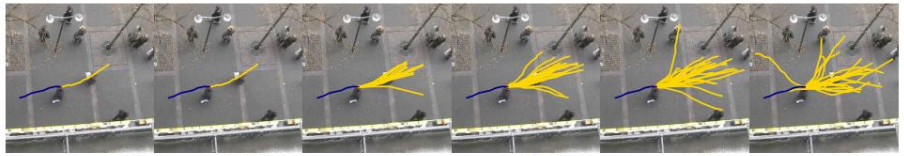


Figure 5. Visualization of the most-likely and multimodal trajectory generation capability of our LMTraj-SUP (τ: temperature).

Model		ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
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

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

IV Experiments

Ablation Studies

Computational cost

- Inference times are slightly slower than the fastest models because of the structural nature of the language model, which predicts the next token sequentially
- However, it produces promising results with reasonable GPU memory consumption as well as real-time inference

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.

5

Conclusion

V Conclusion

Summary

Predict trajectories with a question-and-answer approach

- By transforming the trajectory prediction task into a question-answer format, we demonstrate that past trajectory data can be used to successfully predict future trajectories

Demonstrate prediction performance with zero-shot and supervised learning

- Through a variety of experiments, we demonstrate the accuracy of future trajectory prediction in both zero-shot and supervised ways using LMTraj-ZERO and LMTraj-SUP

Demonstrate the effectiveness of large language model optimization techniques

- Ablation study demonstrates that language model-specific techniques such as tokenization optimization and multi-task learning improve the model's social reasoning capabilities and trajectory prediction performance

감사합니다

Thank you for listening