GUIDE-CoT: Goal-driven and User-Informed Dynamic Estimation for Pedestrian Trajectory using Chain-of-Thought

Sungsik Kim
Kookmin University
Seoul, Korea

Janghyun Baek
Korea University
Seoul, Korea

Jinkyu Kim
Korea University
Seoul, Korea
jinkyukim@korea.ac.kr

Jaekoo Lee
Kookmin University
Seoul, Korea
jaekoo@kookmin.ac.kr

ABSTRACT

While Large Language Models (LLMs) have recently shown impressive results in reasoning tasks, their application to pedestrian trajectory prediction remains challenging due to two key limitations: insufficient use of visual information and the difficulty of predicting entire trajectories. To address these challenges, we propose *Goal-driven and User-Informed Dynamic Estimation for pedestrian trajectory using Chain-of-Thought (GUIDE-CoT)*. Our approach integrates two innovative modules: (1) a goal-oriented visual prompt, which enhances goal prediction accuracy combining visual prompts with a pretrained visual encoder, and (2) a chain-of-thought (CoT) LLM for trajectory generation, which generates realistic trajectories toward the predicted goal. Moreover, our method introduces controllable trajectory generation, allowing for flexible and user-guided modifications to the predicted paths. Through extensive experiments on the ETH/UCY benchmark datasets, our method achieves state-of-the-art performance, delivering both high accuracy and greater adaptability in pedestrian trajectory prediction. Our code is publicly available at https://github.com/ai-kmu/GUIDE-CoT.

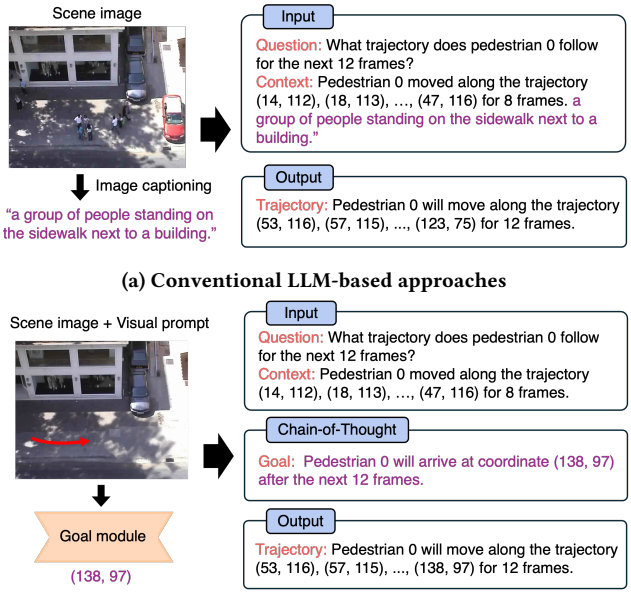
KEYWORDS

LLM-based Pedestrian Trajectory Prediction, Chain-of-Thought (CoT) Reasoning, Visual Prompting

ACM Reference Format:
Sungsik Kim, Janghyun Baek, Jinkyu Kim, and Jaekoo Lee. 2025. GUIDE-CoT: Goal-driven and User-Informed Dynamic Estimation for Pedestrian Trajectory using Chain-of-Thought. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

Pedestrian trajectory prediction has emerged as a critical task in various applications, including autonomous driving and urban planning [40]. In autonomous driving, vehicles or robots must dynamically adjust their paths by predicting and responding to the movements of nearby pedestrians [33]. The inability of an autonomous agent to accurately predict future pedestrian trajectories can lead to severe consequences, such as collisions, posing significant safety risks [20]. Furthermore, in the context of urban planning, particularly in smart cities, accurately understanding and predicting pedestrian flow is crucial. This insight enables the optimization of transportation networks and infrastructure, leading to smoother traffic management and greater overall efficiency [17].



(a) Conventional LLM-based approaches

(b) Our proposed approach (GUIDE-CoT)

Figure 1: Comparison between conventional LLM-based methods and our approach for pedestrian trajectory prediction. (a) Conventional approaches often leverage LLM's reasoning capability to predict pedestrians' future trajectories conditioned on textual contexts, which contain their past observations and scene descriptions (from an off-the-shelf image captioning model). (b) Our proposed method, called GUIDE-CoT, further improves the model's prediction performance by predicting pedestrians' final goal position given the scene image with overlaid visual prompts (i.e., a red arrow). Such predicted goal position is then augmented into an LLM in a similar manner to Chain-of-Thought (CoT), offering rich intermediate reasoning contexts.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Deep learning has led to substantial breakthroughs in pedestrian trajectory prediction, with recent studies showing that deep learning models trained on large-scale datasets can accurately predict a wide range of pedestrian behaviors [46]. While existing approaches have focused on modeling pedestrian interactions and environmental factors, there has been growing interest in goal-based pedestrian trajectory prediction [10, 13, 30, 31]. These methods aim to predict a pedestrian's future destination based on their past trajectory and map information. Since the accurate prediction of a pedestrian's goal is crucial for trajectory estimation, improving goal prediction accuracy has become a central focus in trajectory prediction research.

In this work, we propose a novel approach by leveraging large language models (LLMs), which have recently been applied to a variety of reasoning and predicting tasks across domains [27, 28, 48]. Building on prior work, which has shown that converting a pedestrian's past trajectory into a natural language format and feeding it into an LLM can result in state-of-the-art (SOTA) trajectory prediction performance [4], we propose a method that significantly enhances LLM-based methods by incorporating goal prediction as a key step in the chain-of-thought (CoT) reasoning process. As illustrated in Figure 1, unlike existing LLM-based methods that predict trajectories directly from historical trajectories and scene image caption, our approach first predicts CoT reasoning of the pedestrian's goal to guide the final trajectory generation.

In contrast to existing methods that rely on semantic segmentation maps with trajectory heatmaps [10, 30] or dynamic features [13] for goal prediction, our approach introduces a goal-oriented visual prompt that leverages visual cues more effectively. To be specific, we enhance goal prediction by incorporating visual prompts, represented as directional arrows on RGB images to indicate pedestrian movement. These visual prompts are processed by a pretrained model, and rather than solely depending on the model's output for direct goal prediction, we combine the visual features with semantic map-based goal prediction. This approach enables more accurate and context-aware predictions by utilizing both spatial and visual information. The predicted goals are then integrated into the CoT reasoning process, which provides context for the LLM, guiding it to generate more precise future trajectories through structured reasoning.

We propose a controllable trajectory generation method by manipulating the CoT reasoning process. Specifically, we explore how modifying the goal context during inference can adjust the predicted trajectory, enabling user-guided control over future path predictions. Experiments on the ETH/UCY benchmark datasets show that our method delivers results competitive with state-of-the-art approaches, while introducing new capabilities in controllability and reasoning.

Our main contributions are summarized as follows:

- We introduce a goal-oriented visual prompt that effectively integrates spatial and visual information for superior goal prediction, enhancing trajectory prediction performance.
- We propose a CoT reasoning prompt that leverages goal predictions as intermediate steps, enabling the LLM to predict future trajectories more accurately.
- We enable controllable trajectory generation, allowing users to modify goal context during inference, providing dynamic control over the predicted trajectories.

2 RELATED WORK

2.1 Pedestrian Trajectory Prediction

Pedestrian trajectory prediction has been extensively studied, with early approaches relying on simple physics-based models [22]. As deep learning advanced, data-driven models that learn pedestrian movement patterns and predict future trajectories become the dominant methodology. A notable example is Social-LSTM [2], which introduced LSTM-based modeling for pedestrian paths. Following this, a wide variety of learning approaches, including GANs [3, 12, 20, 42], CVAEs [43, 50], Diffusion models [11, 19, 32], GCNs [34, 39, 45], and Transformers [52, 53], have been explored to further improve prediction accuracy.

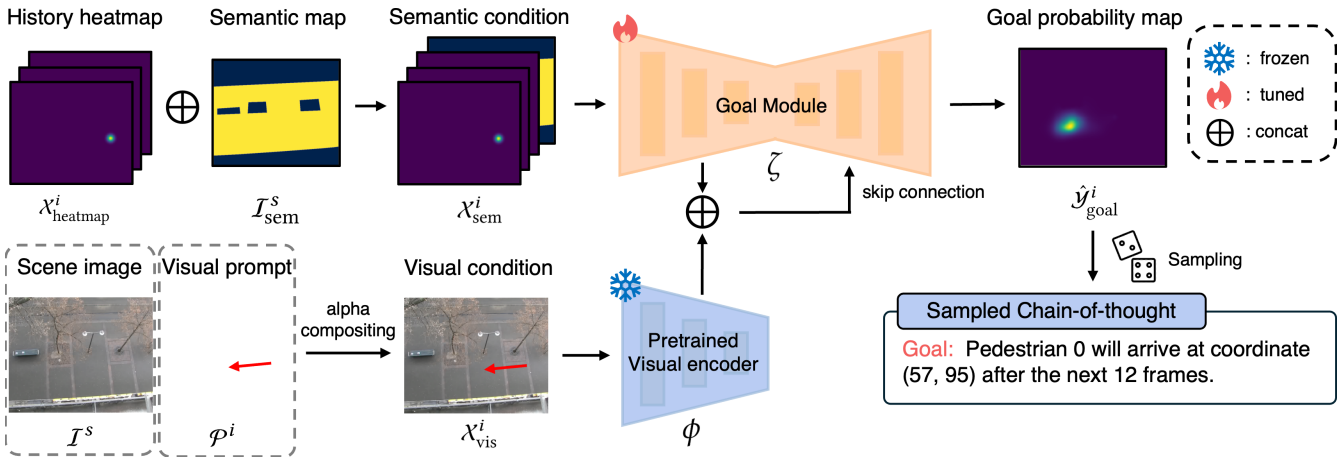
While these approaches have made significant strides, many focus solely on predicting the immediate future path based on historical data and interactions with the environment. However, goal-based pedestrian trajectory prediction has recently emerged as an important alternative. This approach assumes that pedestrian trajectories are heavily influenced by their intended destination, rather than just local interactions. For instance, Y-Net [30] leverages history trajectories and environmental context (in the form of heatmaps and semantic segmentation map) and uses U-Net [41] to predict a goal probability map. This goal-based approach is critical because predicting the final destination can significantly enhance trajectory accuracy.

Most existing methods still struggle to effectively incorporate information of multiple modalities and handle the complexity of trajectory prediction as a reasoning process. Recent advancements in LLMs have demonstrated strong reasoning capabilities across various domains, including pedestrian trajectory prediction. As a LLM-based SOTA model, the LMTraj model [4] reframed trajectory prediction as a question-answering task, utilizing LLMs to infer pedestrian movements, and achieved superior performance over alternatives. However, LLM-based models [4] have two key limitations: insufficient integration of visual information and difficulty in predicting the full trajectory.

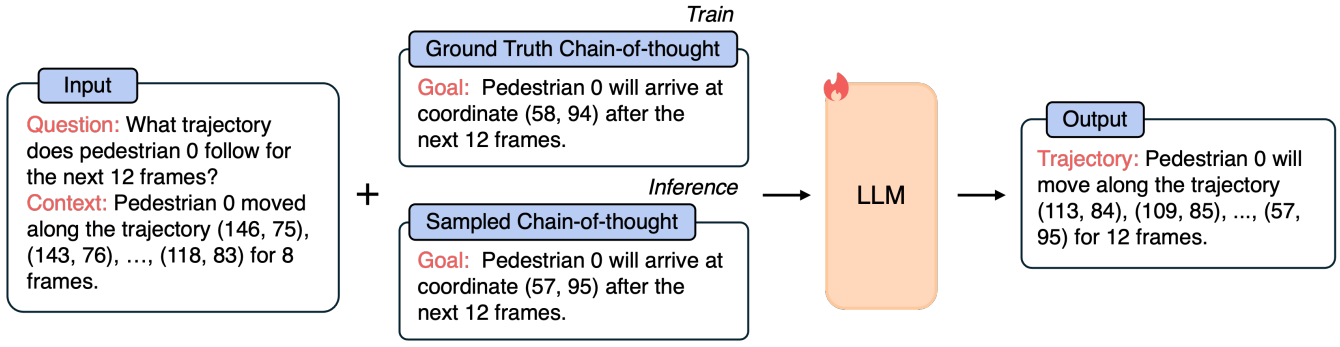
2.2 Large Language Models for Sequence Prediction

Large Language Models (LLMs) have driven significant advances in natural language processing by leveraging large datasets and transformer architectures. These models have achieved impressive results across a wide range of applications, such as text generation, translation, and summarization [15, 37, 38]. Larger LLMs [1, 8] have shown an enhanced ability to generalize to unseen data and tasks, demonstrating robust performance even in few-shot and zero-shot learning scenarios, where limited task-specific examples are available.

LLMs have been widely utilized in various sequence prediction tasks to enhance inference performance. For example, Time-LLM [24] leverages a pretrained LLM for time series forecasting and has demonstrated high performance in both zero-shot and few-shot settings. Furthermore, recent studies have proposed the use of LLMs



(a) Goal predictor with visual prompt



(b) LLM-based trajectory prediction with Chain-of-Thought (CoT) reasoning

Figure 2: An overview of our proposed approach, called GUIDE-CoT. (a) Our model first predicts each pedestrian's final goal position given (i) pedestrians' past observations, (ii) semantic BEV map, and (iii) top-down view scene image with visual prompt (i.e., a red arrow). Our model generates a sentence describing their final positions, such as "Pedestrian 0 will arrive at coordinate (57, 95) after the next 12 frames." (b) Such a generated goal description is then augmented into the LLM in a similar way to Chain-of-Thought reasoning, generating the final trajectory of each pedestrian.

to predict complex biological sequences [25], such as DNA patterns [54] and protein structures [18]. Specifically, DNAGPT [54] introduced a generalized pretrained LLM for modeling and predicting DNA sequences, while ProtGPT2 [18] presented a GPT-2 [37]–based protein sequence generation model that samples novel protein sequences. Similarly, studies have emerged exploring the use of LLMs to predict pedestrian trajectories–a sequential prediction task.

2.3 Reasoning via Visual Prompt

To enhance the reasoning capabilities of Large Multimodal Models (LMMs), additional visual information, known as visual prompts, has been introduced [9]. Visual prompts provide crucial visual cues that assist models in making more accurate inferences. One common approach is to integrate learnable parameters directly into the visual data. For instance, VPT [23] proposes adding learnable prompts to the input patches of each layer in the Vision Transformer [16]. Another method involves padding image borders with learnable parameters embedding additional information around the edges of the image to guide the model's attention [7].

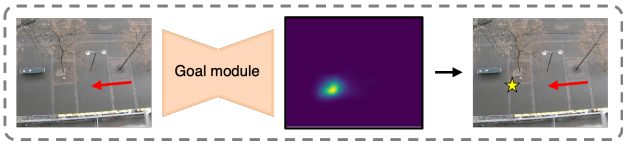
In addition, visual markers are often employed to highlight specific objects or regions within an image. For example, CPT [51] uses unique colors to distinguish objects in region proposals, while Shtedritski *et al.* [44] employs red circles to mark objects, leveraging CLIP [36] for zero-shot reasoning. ViP-LLaVA [9] expands visual prompts by incorporating arrows, points, and triangles for more diverse reasoning tasks.

Inspired by these approaches, we propose a novel method to represent pedestrian trajectories using visual prompts. Our approach guides the trajectory of pedestrians, allowing the model to better understand movement patterns and predict future paths more effectively.

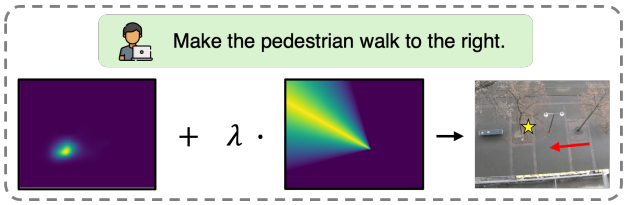
3 METHOD

3.1 Problem Definition

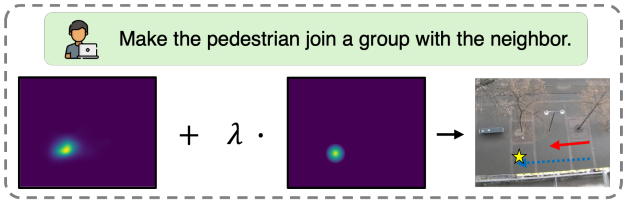
Pedestrian trajectory prediction task aims to forecast future trajectories of pedestrians based on their past movements. For each pedestrian i at time step t, the observed past trajectory is composed



(a) without user-provided guidance



(b) with directional user guidance



(c) with positional user guidance

Figure 3: Our goal-conditioned model further allows the user to provide text-driven guidance, controlling the model's trajectory prediction process. Such guidance may contain (b) a directional guide (e.g., "make the pedestrian walk to the right") or (c) a positional guide (e.g., "make the pedestrian join a group with the neighbor"). Compare results with and without user-provided guidance, i.e., (a) vs. (b) and (c).

of 2D position coordinates p spanning a time window of length τ_{obs}. This trajectory is represented as:

$$\mathcal{X}_t^i = \left\{ p_{t-\tau_{\text{obs}}+1}^i, \ \ldots, \ p_t^i \right\} \tag{1}$$

The input to the model consists of the past trajectories of N pedestrians in the scene, denoted as:

$$\mathcal{X}_t = \left(\mathcal{X}_t^0, \ \ldots, \ \mathcal{X}_t^{N-1} \right) \in \mathbb{R}^{N \times \tau_{\text{obs}} \times 2} \tag{2}$$

Additionally, the model is provided with a scene image $\mathcal{I}^s$ and a corresponding semantic map $\mathcal{I}_{\text{sem}}^s$ for the scene s.

The goal of task is to predict the future trajectories of these N pedestrians over a prediction window of length τ_{pred}. The predicted future trajectories are represented as:

$$\hat{\mathcal{Y}}_t = \left(\mathcal{Y}_t^0, \ \ldots, \ \mathcal{Y}_t^{N-1} \right) \in \mathbb{R}^{N \times \tau_{\text{pred}} \times 2} \tag{3}$$

The objective of learning is to minimize the error between the predicted future trajectories $\hat{\mathcal{Y}}_t$ and the ground-truth future trajectories $\mathcal{Y}_t$.

3.2 Goal-oriented Visual Prompt for Trajectory

Building on prior work that demonstrates improvements in reasoning tasks through the visual knowledge of pretrained vision models [47], we propose a novel approach to effectively leverage visual information from scene images for the goal prediction task.

A naïve approach might consider using top-view video frames of pedestrians for trajectory prediction via a video encoder. However, the naïve approach faces three key challenges: (1) obtaining top-view pedestrian videos is impractical, (2) distinguishing pedestrians within the scene is challenging, and (3) video processing is computationally inefficient compared to handling individual images.

To address these challenges while still utilizing the rich visual information from scene images, we introduce visual prompts—such as circles, arrows, and points [9, 44]—to designate pedestrian positions and movements for goal prediction. Specifically, our approach uses arrow-shaped visual prompts to guide the goal prediction, allowing the model to better interpret and predict future trajectories.

As shown in Figure 2a, given the past trajectory $\mathcal{X}^i$ of the i-th pedestrian as input, we generate a visual prompt $\mathcal{P}^i = Draw(\mathcal{X}^i)$ by connecting the 2D coordinates with arrows in the process referred to as $Draw$. Next, the visual condition $\mathcal{X}_{\text{vis}}^i$ is constructed through alpha compositing between the scene image $\mathcal{I}^s$ and the visual prompt.

$$\mathcal{X}_{\text{vis}}^i = (1-\alpha) \cdot \mathcal{I}^s + \alpha \cdot \mathcal{P}^i \tag{4}$$

Additionally, the past trajectory is represented as a heatmap, denoted as the history heatmap $\mathcal{X}_{\text{heatmap}}^i$. The heatmap is concatenated with the semantic map $\mathcal{X}_{\text{sem}}^i$ to form the semantic condition $\mathcal{X}_{\text{sem}}^i$ as follows:

$$\mathcal{X}_{\text{sem}}^i = \mathcal{X}_{\text{heatmap}}^i \oplus \mathcal{I}_{\text{sem}}^s \tag{5}$$

where $\oplus$ denotes channel-wise concatenation.

Finally, we compute the goal probability map $\hat{\mathcal{Y}}_{\text{goal}}^i$, from which the 2D coordinates of the goal are sampled. The $\hat{\mathcal{Y}}_{\text{goal}}^i$ is defined as:

$$\hat{\mathcal{Y}}_{\text{goal}}^i = \sigma\left[\zeta\left(\phi(\mathcal{X}_{\text{vis}}^i), \ \mathcal{X}_{\text{sem}}^i \right) \right] \tag{6}$$

Here, ζ represents the U-Net-based goal module [41], ϕ refers to the pretrained visual encoder, and σ is the sigmoid function. Similar to prior work [10, 30], we adopt a U-Net-based encoder-decoder architecture [41] for the goal module. However, unlike the original designs, we modify the skip connections to concatenate the encoded features from both $\mathcal{X}_{\text{sem}}^i$ and $\mathcal{X}_{\text{vis}}^i$ before passing them to the decoder.

To predict pedestrian trajectories using LLMs (Figure 2b), prompt engineering is needed to convert the past trajectory $\mathcal{X}_t^i$ from numeric form into a natural language prompt. A tokenization process is then required to input this prompt into the LLM. For both prompt generation and tokenization, we adopted the method proposed by LMTraj [4]. Unlike previous studies, we incorporate the CoT goal context as input for the LLM. Rather than merely predicting the goal, CoT extracts a goal prompt by leveraging past trajectory data to represent the pedestrian's intended goal. The LLM then focuses solely on generating a realistic trajectory toward the goal without performing additional goal prediction. The CoT-based approach allows the model to better understand underlying movement patterns, enabling more accurate and context-aware future trajectory predictions. By structuring the trajectory generation process around a well-defined goal, this method significantly enhances the realism of the predicted paths.

3.3 User-guided Trajectory Generation

Unlike existing LLM-based models such as SOTA LMTraj [4], our approach enables comprehensive trajectory modeling, not limited to simple trajectory prediction. Notably, our method allows users to provide high-level guidance, such as direction or group behavior, without the need to specify exact 2D goal coordinates. This makes our model more interpretable and adaptable to user input. Figure 3 illustrates examples of user-guided trajectory generation.

To achieve this, we adjust the goal logit map $\zeta\left(\phi(\mathcal{X}^i_{\text{vis}}),\ \mathcal{X}^i_{\text{sem}}\right)$ by incorporating a guidance function $\mathcal{G}(\mathcal{X}^i_t)$ as a regularization term, allowing for user-guided trajectory adjustments. The modified goal probability map is computed as follows:

$$\hat{\mathcal{Y}}^i_{\text{goal}} = \sigma\left[\zeta\left(\phi(\mathcal{X}^i_{\text{vis}}),\ \mathcal{X}^i_{\text{sem}}\right) + \lambda \cdot \mathcal{G}(\mathcal{X}^i_t)\right] \tag{7}$$

We define guidance functions for two scenarios: direction guidance and group behavior. For direction (Figure 3b), logits, derived by $\mathcal{G}_{\text{direction}}(\mathcal{X}^i_t) = \max\left(0,\ 1 - \frac{|\theta-\theta_p|}{\theta_{\max}}\right)$, are added to regions rotated by a user-defined angle relative to the pedestrian's current heading. θ is the target angle for the pedestrian, θ_p is the angle between each position and the current position, and $\theta_{\max}$ is the maximum angle for adding logits. For group behavior (Figure 3c), logits, derived by $\mathcal{G}_{\text{group}}(\mathcal{X}^i_t) = \max\left(0,\ 1 - \frac{d_p}{d_{\max}}\right)$, are added around the predicted goals of neighboring pedestrian. d_p represents the distance to the neighboring pedestrian's future position, and $d_{\max}$ is the maximum distance. The level of influence from this guidance is controlled via the parameter λ, which adjusts the strength of the user input.

3.4 Training and Inference Scheme

The proposed approaches are learned through the following loss functions. First, for a goal-oriented visual prompt (Figure 2a), model learns to align the predicted goal probability map $\hat{\mathcal{Y}}_{\text{goal}}$ with the ground truth probability map $\mathcal{Y}_{\text{goal}}$, which is generated from the final 2D coordinates $p_{t+\tau_{\text{pred}}}$ of the ground truth trajectory $\mathcal{Y}_t$. The $\mathcal{L}_{\text{goal}}$ is computed using binary cross entropy (BCE) as follows: $\mathcal{L}_{\text{goal}} = BCE\left(\hat{\mathcal{Y}}_{\text{goal}},\ \mathcal{Y}_{\text{goal}}\right)$ Next, for the chain-of-thought (CoT) LLM for trajectory generation (Figure 2b), the predicted text-based trajectory $\hat{\mathcal{Y}}_{\text{text}}$ is compared with the ground truth trajectory $\mathcal{Y}_{\text{text}}$, represented in text form, using cross entropy (CE) to compute the $\mathcal{L}_{\text{text}}$. As a result, the $\mathcal{L}_{\text{text}}$ is expressed as follows: $\mathcal{L}_{\text{text}} = CE\left(\hat{\mathcal{Y}}_{\text{text}},\ \mathcal{Y}_{\text{text}}\right)$

During training, the LLM uses the ground truth chain-of-thought from the dataset instead of the one generated by the goal-oriented visual prompt. This ensures that the goal-oriented visual prompt and LLM are trained independently, without influencing each other. The overall training process of the goal-oriented visual prompt and the CoT LLM is shown in Algorithm 1. At inference time, the chain-of-thought sampled from the goal-oriented visual prompt is passed to the LLM to predict the final trajectory.

Algorithm 1 GUIDE-CoT Training Scheme

(Step 1) Goal-oriented visual prompt training scheme
INPUT: Past trajectories of pedestrians $\mathcal{X}$, Future trajectories of pedestrians $\mathcal{Y}$, Scene image $\mathcal{I}^s$, Semantic map $\mathcal{I}^s_{\text{sem}}$, Goal module ζ, Pretrained visual encoder ϕ, Batch size N
OUTPUT: Trained goal module ζ

for batch in $\mathcal{X}$ **do**
  for $i = 1$ to N **do**
    $\mathcal{P}^i \leftarrow Draw(\mathcal{X}^i)$
    $\mathcal{X}^i_{\text{vis}} \leftarrow (1-\alpha)\cdot\mathcal{I}^s + \alpha\cdot\mathcal{P}^i$
    $\mathcal{X}^i_{\text{sem}} \leftarrow \mathcal{X}^i_{\text{heatmap}} \oplus \mathcal{I}^s_{\text{sem}}$
    $\hat{\mathcal{Y}}^i_{\text{goal}} \leftarrow \sigma\left[\zeta(\phi(\mathcal{X}^i_{\text{vis}}),\ \mathcal{X}^i_{\text{sem}})\right]$
  end for
  $\mathcal{L}_{\text{goal}} \leftarrow \text{BCE}(\hat{\mathcal{Y}}_{\text{goal}},\ \mathcal{Y}_{\text{goal}})$
  Backprop and update ζ
end for
Return: Trained goal module ζ

(Step 2) CoT LLM training scheme
INPUT: Past trajectories of pedestrians $\mathcal{X}$, Future trajectories of pedestrians $\mathcal{Y}$, LLM, Batch size N
OUTPUT: Trained LLM

for batch in $\mathcal{X}$ **do**
  for $i = 1$ to N **do**
    CoT prompt ← from ground-truth $\mathcal{Y}^i$
    $\hat{\mathcal{Y}}^i_{\text{text}} \leftarrow$ LLM(prompt)
  end for
  $\mathcal{L}_{\text{text}} \leftarrow \text{CE}(\hat{\mathcal{Y}}_{\text{text}},\ \mathcal{Y}_{\text{text}})$
  Backprop and update LLM
end for
Return: Trained LLM

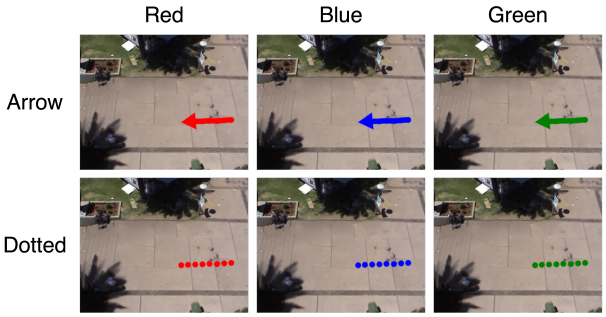


Figure 4: Examples of variants of our used visual prompts with different colors (i.e., red, blue, and green) and shapes (i.e., arrow and points).

4 EXPERIMENTAL RESULTS

4.1 Implementation and Evaluation Details

Datasets. We evaluated the trajectory prediction performance using the widely used ETH [35] and UCY [26] datasets. Following prior works, we adopted the leave-one-out cross-validation setting

Table 1: Comparison of ADE/FDE across different models on the ETH/UCY datasets. The 1st/2nd best performances are indicated in boldface and underline. †: Each model is evaluated under the common pedestrian trajectory prediction setting, i.e., consistent dataset splits and evaluation units (meters). ‡: Results for these models are reproduced.

Model	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
	ADE (↓) / FDE (↓)	ADE (↓) / FDE (↓)	ADE (↓) / FDE (↓)	ADE (↓) /FDE (↓)	ADE (↓) / FDE (↓)	ADE (↓) / FDE (↓)
Social-GAN (18' CVPR) [20]	0.77 / 1.40	0.43 / 0.88	0.75 / 1.50	0.35 / 0.69	0.36 / 0.72	0.53 / 1.04
PECNet (20' ECCV)[1†] [31]	0.61 / 1.07	0.22 / 0.39	0.34 / 0.56	0.25 / 0.45	0.19 / 0.33	0.32 / 0.56
Trajectron++ (20' ECCV)[†] [43]	0.61 / 1.03	0.20 / 0.28	0.30 / 0.55	0.24 / 0.41	0.18 / 0.32	0.31 / 0.52
AgentFormer (21' ICCV)[†] [53]	0.46 / 0.80	0.14 / 0.22	0.25 / 0.45	0.18 / 0.30	0.14 / 0.24	0.23 / 0.40
MID (22' CVPR)[†] [19]	0.57 / 0.93	0.21 / 0.33	0.29 / 0.55	0.28 / 0.50	0.20 / 0.37	0.31 / 0.54
Goal-SAR (22' CVPRW)[1†‡] [10]	0.60 / 0.81	0.19 / 0.18	0.85 / 0.59	0.20 / 0.30	0.21 / 0.27	0.41 / 0.43
NPSN (22' CVPR) [6]	**0.36** / 0.59	0.16 / 0.25	0.23 / 0.39	0.18 / 0.32	0.14 / 0.25	**0.21** / 0.36
SocialVAE (22' ECCV) [50]	0.41 / 0.58	0.13 / 0.19	**0.21** / 0.36	**0.17** / 0.29	**0.13** / 0.22	**0.21** / 0.33
EqMotion (23' CVPR) [49]	0.40 / 0.61	0.12 / 0.18	0.23 / 0.43	0.18 / 0.32	**0.13** / 0.23	**0.21** / 0.35
EigenTrajectory (23' ICCV) [5]	**0.36** / 0.53	0.12 / 0.19	0.24 / 0.43	0.19 / 0.33	0.14 / 0.24	**0.21** / 0.34
LED (23' CVPR) [32]	0.39 / 0.58	**0.11** / 0.17	0.26 / 0.43	0.18 / **0.26**	**0.13** / 0.22	**0.21** / 0.33
LMTraj-SUP (24' CVPR)[2] [4]	0.41 / 0.51	0.12 / 0.16	0.22 / **0.34**	0.20 / 0.32	0.17 / 0.27	0.22 / 0.32
GUIDE-CoT (Ours)[1,2]	0.38 / **0.43**	0.13 / **0.15**	0.34 / 0.48	0.19 / 0.29	0.16 / **0.21**	0.24 / **0.31**

1: Goal-based methods. 2: LLM-based methods

Table 2: Ablation study with (a) different types of visual prompts, (b) different combinations of visual conditions, and (c) different pretraining strategies (with the same ResNet-50 [21] backbone). We use ETH/UCY datasets and report FDE scores (lower is better).

(a) Effect on types of visual prompts

	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg.
Blue points	0.46	0.17	0.50	**0.28**	0.22	0.33
Blue arrow	0.46	0.17	0.50	**0.28**	0.22	0.33
Green points	**0.43**	0.16	0.50	0.29	**0.21**	0.32
Green arrow	0.44	0.16	0.50	**0.28**	0.22	0.32
Red points	0.46	0.17	0.50	**0.28**	0.22	0.33
Red arrow	**0.43**	**0.15**	**0.48**	0.29	**0.21**	**0.31**

(b) Effect on combinations of visual prompts

$\mathcal{X}_{\text{sem}}$	$\mathcal{X}_{\text{vis}}$	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg.
✓		0.66	0.22	0.64	0.36	0.24	0.42
	✓	0.66	0.32	1.22	0.60	0.33	0.51
✓	✓	**0.43**	**0.15**	**0.48**	**0.29**	**0.21**	**0.31**

(c) Effect on pretraining strategies

	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg.
ImageNet-1K [14]	0.44	0.18	0.49	0.30	0.22	0.33
Remote-CLIP [29]	**0.41**	0.17	0.49	**0.28**	0.22	**0.31**
CLIP [36]	0.43	**0.15**	**0.48**	0.29	**0.21**	**0.31**

for our experiments. We also followed the previous experimental settings [4–6], i.e., dataset splits.

Evaluation metrics. For quantitative evaluation, we used well-known metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE). The model predicted 20 possible future trajectories, from which the one with the smallest error was selected. All results were reported in meters.

Implementation details. For goal-oriented visual prompt, we employed the CLIP [36] visual encoder based on the ResNet-50 [21]. The goal module utilized a CNN-based U-Net encoder-decoder architecture. We trained the goal-oriented visual prompt for 50 epochs, with a batch size of 64, on a single RTX-3090Ti GPU, which took approximately 7 hours. For the LLM, we used the T5 small [38] model, consistent with LMTraj [4]. The prompt used to transfer Goal CoT to the LLM is: "Pedestrian i will arrive at coordinate $p^i_{t+\tau_{\text{pred}}}$ after the next τ_{pred} frames.". The experimental code can be found on our GitHub at https://github.com/ai-kmu/GUIDE-CoT

4.2 Trajectory Prediction Performance Comparison with Existing Approaches

The results of pedestrian trajectory prediction on the ETH/UCY datasets are presented in Table 1. The best performance is highlighted in bold, and the second-best is underlined. Our proposed method achieves ADE comparable to existing models and demonstrates the best performance in terms of average FDE across the datasets. This demonstrates that the goal-oriented visual prompt effectively contributes to improving goal prediction performance.

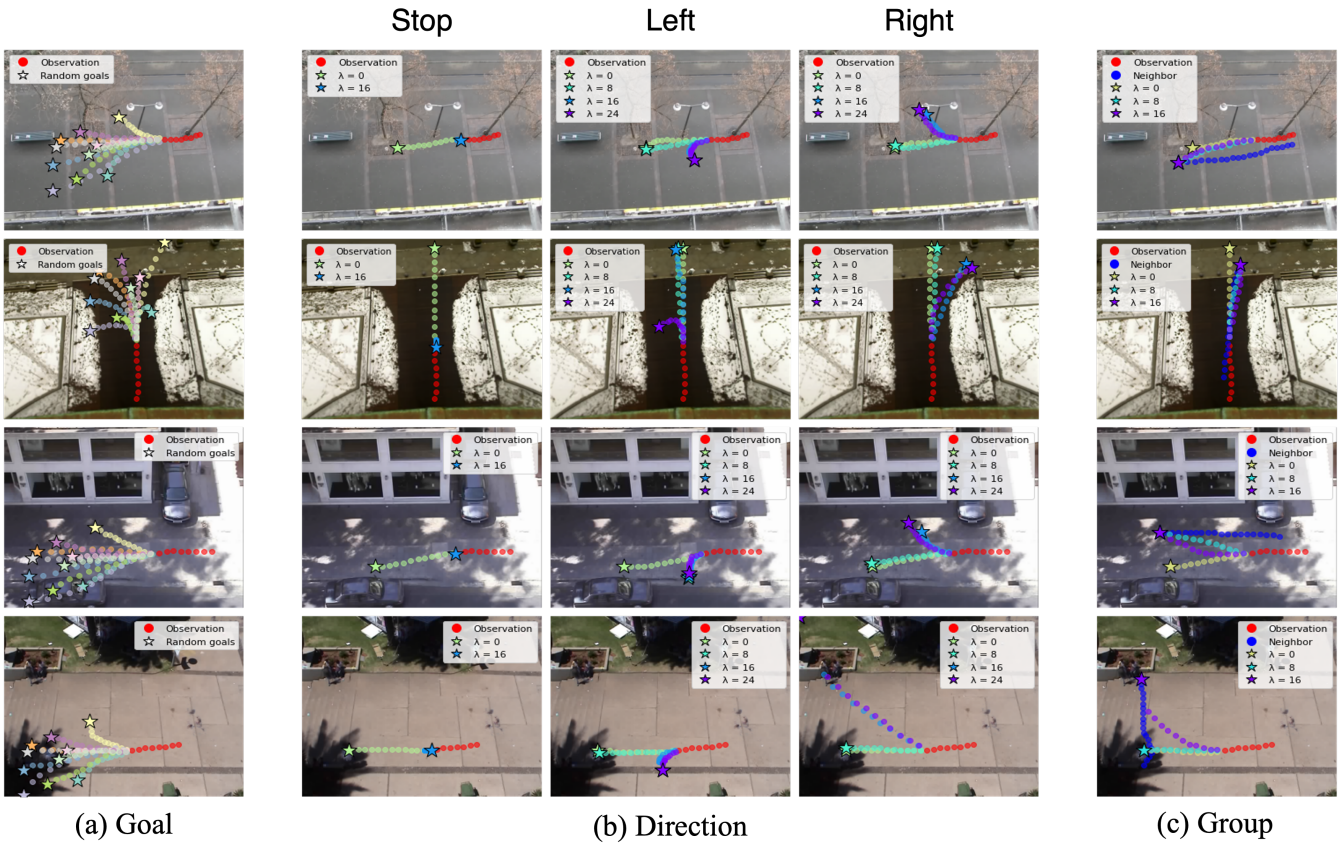


Figure 5: Visualization of controllable trajectory generation based on user guidance. The red points represent the pedestrian's observed trajectory, and the stars indicate the generated goals according to the goal-oriented visual prompt. Points in other colors show predicted trajectories from goal adjustments, illustrating variations like stopping, turning, and grouping with nearby pedestrians. These trajectories demonstrate the model's adaptability to user-guided trajectory generation.

Comparison with LLM-based models. Our method surpasses LMTraj-SUP [4], which also utilizes an LLM. This demonstrates that providing the LLM with appropriate context, rather than merely tasking it with inferring the full trajectory, enables more effective reasoning and leads to enhanced predictive performance. Furthermore, the incorporation of goal-oriented visual prompts not only enhances trajectory estimation but also allows the model to leverage spatial and visual information more effectively. This leads to a deeper understanding of pedestrian intent, resulting in more accurate predictions across diverse environments.

Our method provides flexibility through user-guided trajectory generation, allowing high-level guidance without precise goal coordinates. This adaptability allows precise handling of complex pedestrian behaviors and dynamic environments, outperforming previous models. Its ability to generalize across datasets demonstrates robustness, ensuring reliable performance under diverse conditions. Integrating vision-based and semantic map-based goal prediction creates a synergistic effect, supporting more context-aware predictions. By leveraging multiple modalities instead of relying solely on historical data, our approach ensures realistic and adaptable trajectories suited to complex urban environments. These strengths enable state-of-the-art performance while introducing new capabilities for user-informed trajectory generation.

Comparison with goal-based models. GUIDE-CoT shares similarities with existing goal-based pedestrian trajectory prediction methods, as it predicts both the goal and the trajectory leading toward it. However, our method achieves superior performance over the representative goal-based approach, Goal-SAR [10], in terms of goal prediction accuracy, as measured by FDE. Although both GUIDE-CoT and Goal-SAR [10] employ the same U-Net-based goal module [41] to process semantic conditions, our results demonstrate that the introduction of goal-based visual prompts plays a crucial role in enhancing performance.

These performance improvements offer two key insights. First, incorporating well-designed visual prompts enables the pretrained visual encoder to extract more meaningful features. Second, these enriched features help the model effectively utilize environmental information for goal prediction, which significantly contributes to improved overall performance.

4.3 Effect of Goal-oriented Visual Prompt

We present the first approach to represent pedestrian trajectories as visual prompts for reasoning. To further analyze the effectiveness of our approach, we conduct experiments to evaluate the impact of different combinations of visual prompts and pretrained visual encoders on model performance.

For visual prompts. As shown in Figure 4, we test six types of visual prompts, combining three colors—red, blue, and green—with two shapes: arrows connecting the pedestrian's starting and ending points, and dotted points representing the movement path. Most visual prompts yield performance comparable to leading methods, with the red arrow achieving the remarkable results as shown in Table 2a. These findings suggest that the CLIP [36] visual encoder effectively interprets a wide range of visual prompts.

For conditions. Table 2b presents the performance evaluation from ablation experiments, which test the effect of using different input conditions for the goal module when generating the goal probability map. In the experiments where only the visual condition was applied—consisting of scene images and visual prompts—we combined a pretrained visual encoder with a CNN-based decoder, structured similarly to U-Net [41], and trained only the decoder for evaluation. The results indicate that the best performance was achieved when both conditions (visual and semantic) were applied together. This suggests that integrating features extracted by the pretrained visual encoder into existing goal modules [10, 30] significantly improves the model's ability to predict accurate goals.

For pretraining strategies. Table 2c presents a comparison of different visual encoder pretraining strategies. We evaluated the performance of the ImageNet-1K [14], CLIP [36], and Remote-CLIP [29] visual encoders. The results show that CLIP consistently outperforms ImageNet-1K across all datasets, highlighting its effectiveness in extracting relevant features for trajectory prediction. This superior performance can be attributed to CLIP's ability to leverage multimodal data and align visual features with contextual information more effectively than traditional pretraining methods.

As a result, leveraging visual prompts along with a pretrained visual encoder has a clear positive impact on goal prediction by enhancing the model's contextual understanding of pedestrian intent. Our findings align with prior research using visual prompts for reasoning tasks [9, 44], indicating that goal prediction shares key similarities with general reasoning tasks. This suggests that advancements in reasoning research can further improve goal prediction performance.

4.4 Effect of User-guided Trajectory Generation

We perform a qualitative analysis to evaluate user-guided trajectory generation using three types of high-level guidance: goal, direction, and group. The visual results are shown in Figure 5. For goal guidance, a random goal was assigned for the pedestrian. For direction guidance, we tested three scenarios: stop, left, and right. For group guidance, a random neighboring pedestrian was selected, and the model was guided to group the target pedestrian with them. In Figure 5a, the randomly assigned goal led the LLM to generate realistic trajectories that followed the given goal accurately.

For direction guidance (Figure 5b), we tested three commands: stop, left, and right. Without guidance ($\lambda = 0$), the predicted trajectory followed a straight path to the nearest goal. As λ increased, the trajectory adjusted to the specified direction. Importantly, the model avoided assigning goals to non-traversable areas like obstacles or off-road locations. This behavior stems from our method, where the guidance function $\mathcal{G}(\mathcal{X}_t^i)$ modifies the goal logit map, fine-tuning the path rather than replacing it.

For group guidance (Figure 5c), we selected a random neighboring pedestrian and guided the target pedestrian to form a group with them. When $\lambda = 0$, the red and blue pedestrians were predicted to remain independent. However, as λ increased, the trajectories adjusted to bring them closer, successfully forming a group.

5 CONCLUSION

We propose GUIDE-CoT, a novel framework that enhances pedestrian trajectory prediction by integrating goal-oriented visual prompts with a chain-of-thought LLM. Our method leverages visual prompts and pretrained visual encoders to deliver precise goal context to the LLM, improving both prediction accuracy and adaptability. Unlike existing LLM-based trajectory prediction methods, GUIDE-CoT uses the goal context as an input, enabling controllable trajectory generation by adjusting the goal dynamically. Extensive experiments on the ETH/UCY datasets demonstrate the effectiveness of our approach, showing that GUIDE-CoT not only outperforms current methods but also introduces new capabilities for user-guided, context-aware trajectory predictions.

One key limitation of this study is that the effectiveness of visual information can vary with the environment. In areas with distinct obstacles, such as trees and buildings, pedestrian future goals can be predicted with relatively high accuracy. However, in open spaces like UNIV, where there are few obstacles, predicting these goals becomes more challenging (see Table 1). To address this issue, future work should explore integrating additional contextual cues from the surrounding environment–beyond just visual information–into the LLM.

ACKNOWLEDGMENTS

This research was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(IITP-2025-RS-2022-00156295; Information Technology Research Center), the grant (IITP-2024-RS-2024-00397085; Leading Generative AI Human Resources Development, and IITP-2024-RS-2024-00417958; Global Research Support Program in the Digital Field program), and the National Research Foundation of Korea(NRF) grant (No.RS-2023-00212484; xAI for Motion Prediction in Complex, Real-World Driving Environment) funded by the Korea government(MSIT)

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.

[3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. 2019. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 0–0.

[4] Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. 2024. Can Language Beat Numerical Regression? Language-Based Multimodal Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[5] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. 2023. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10017–10029.

[6] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. 2022. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6477–6487.

[7] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274* (2022).

[8] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[9] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12914–12923.

[10] Luigi Filippo Chiara, Pasquale Coscia, Sourav Das, Simone Calderara, Rita Cucchiara, and Lamberto Ballan. 2022. Goal-driven self-attentive recurrent networks for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2518–2527.

[11] Younwoo Choi, Ray Coden Mercurius, Soheil Mohamad Alizadeh Shabestary, and Amir Rasouli. 2024. Dice: Diverse diffusion model with scoring for trajectory prediction. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 3023–3029.

[12] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. 2021. 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*. 13158–13167.

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

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[15] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[17] Zhuangyuan Fan and Becky PY Loo. 2021. Street life and pedestrian activities in smart cities: opportunities and challenges for computational urban science. *Computational urban science* 1 (2021), 1–17.

[18] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* 13, 1 (2022), 4348.

[19] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. 2022. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17113–17122.

[20] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2255–2264.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.

[24] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.

[25] Inkyung Kim, Juwan Lim, and Jaekoo Lee. 2024. Human activity recognition via temporal fusion contrastive learning. *IEEE Access* 12 (2024), 20854–20866.

[26] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. 2007. Crowds by example. In *Computer graphics forum*, Vol. 26. Wiley Online Library, 655–664.

[27] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns* 5, 3 (2024).

[28] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.

[29] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[30] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. 2021. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15233–15242.

[31] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. 2020. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 759–776.

[32] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. 2023. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5517–5526.

[33] Zhiwei Meng, Jiaming Wu, Sumin Zhang, Rui He, and Bing Ge. 2023. Interaction-aware trajectory prediction for autonomous vehicle based on LSTM-MLP model. In *Proceedings of KES-STS International Symposium*. Springer, 91–99.

[34] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. 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*. 14424–14432.

[35] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*. IEEE, 261–268.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[39] Ben A Rainbow, Qianhui Men, and Hubert PH Shum. 2021. Semantics-STGCNN: A semantics-guided spatial-temporal graph convolutional network for multi-class trajectory prediction. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2959–2966.

[40] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13756–13766.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[42] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. 2019. 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*. 1349–1358.

[43] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 683–700.

[44] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11987–11997.

[45] Bogdan Ilie Sighencea, Ion Rareș Stanciu, and Cătălin Daniel Căleanu. 2023. D-STGCN: Dynamic Pedestrian Trajectory Prediction Using Spatio-Temporal Graph Convolutional Networks. *Electronics* 12, 3 (2023), 611.

[46] Bogdan Ilie Sighencea, Rareș Ion Stanciu, and Cătălin Daniel Căleanu. 2021. A review of deep learning-based methods for pedestrian trajectory prediction. *Sensors* 21, 22 (2021), 7543.

[47] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2023. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562* (2023).

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[49] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. 2023. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1410–1420.

[50] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. 2022. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*. Springer, 511–528.
[51] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open* 5 (2024), 30–38.
[52] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 507–523.
[53] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9813–9823.
[54] Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. 2023. DNAGPT: a generalized pretrained tool for multiple DNA sequence analysis tasks. *bioRxiv* (2023), 2023–07.