PEDESTRIANQA: A Benchmark for Vision-Language Models on Pedestrian Intention and Trajectory Prediction

Naman Mishra[1], Shankar Gangisetty[1], C.V. Jawahar[1]

***Abstract*— Pedestrian intention and trajectory prediction are critical for the safe deployment of autonomous driving systems, directly influencing navigation decisions in complex traffic environments. Recent advances in large vision–language models offer a powerful new paradigm for these tasks by combining high-capacity visual understanding with flexible natural language reasoning. In this work, we introduce PedestrianQA, a large-scale video-based dataset that formulates pedestrian intention and trajectory prediction as question–answering tasks augmented with structured rationales. PedestrianQA expresses richly annotated pedestrian sequences, in natural language, enabling VLMs to learn from visual dynamics, contextual cues, and interactions among traffic agents while generating concise explanations of their predictions without needing specialized architectures tailored for each task. Empirical evaluations across PIE, JAAD, TITAN, and IDD-PeD show that finetuning state-of-the-art VLMs on PedestrianQA significantly improves intention classification, trajectory forecasting accuracy, and the quality of explanatory rationales, demonstrating the strong potential of VLMs as a unified and explainable framework for safety-critical pedestrian behavior modeling. Dataset and models are available at https://github.com/botmahn/PedestrianQA**

I. INTRODUCTION

Ensuring pedestrian safety remains one of the most critical challenges in deploying autonomous vehicles (AVs). Pedestrians are highly unpredictable, often exhibiting complex behaviors shaped by dynamic environments and interactions with other agents. Accurately anticipating whether a pedestrian intends to cross and forecasting their future trajectory are essential for safe navigation in structured (e.g., signalized crosswalks) and unstructured settings (e.g., dense urban roads without explicit rules). Failures in these capabilities have repeatedly been linked to AV disengagements and hazardous incidents.

The risk is amplified in unstructured environments [1], where pedestrian movement and behavior are especially erratic. Consider as illustrated in Fig. 1, where a pedestrian may attempt to cross a road segment lacking marked crosswalks and traffic lights, suddenly stepping from between parked vehicles without signaling or checking for oncoming traffic, forcing the ego-vehicle to brake abruptly to avoid collision. Motivated by such scenarios, we aim to advance Advanced Driver Assistance Systems (ADAS) with natural-language explanations that enhance explainability in both structured and unstructured traffic scenarios, build user trust, and support scalable deployment.

Within this context, we focus on two fundamental tasks: Pedestrian Intention Prediction (PIP) [2], [3], [1], [4], [5], which predicts whether a pedestrian *intends* to cross, and Pedestrian Trajectory Prediction (PTP) [6], [7], [3], [2], [1], which forecasts the pedestrian's spatio-temporal path. Both are short-horizon prediction problems that demand rapid responses from traffic agents to prevent accidents. Despite steady progress, existing methods remain limited in two key respects. First, they underutilize multimodal reasoning, failing to integrate the rich semantics of traffic context, agent interactions, and pedestrian attributes within a unified framework. Second, they treat prediction as a black-box classification or regression problem, offering little to no explanation of their predictions, and thereby hindering user trust and large-scale deployment.

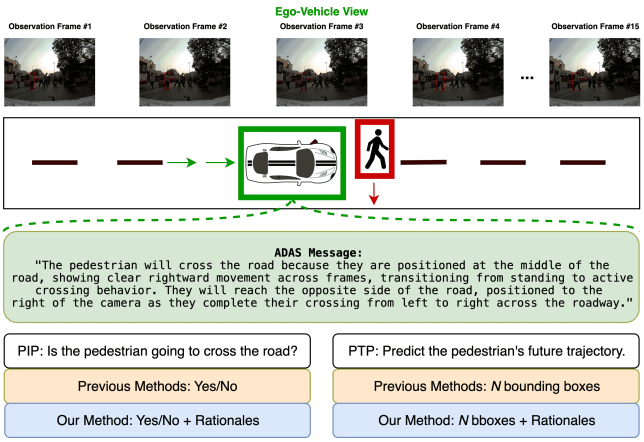

Fig. 1: An illustration of an unstructured-traffic scenario where a pedestrian stands in the middle of the road in front of the ego-vehicle, attempting to cross the road. Unlike prior approaches that provide only predictions, our method predicts the intention and trajectory and generates supporting rationales.

To address these limitations, we explore large-scale vision–language models (VLMs) [8], [9], [10], [11], [12] that have revolutionized the potential use-cases of fusing vision and language modalities. Trained on internet-scale corpora of paired images, videos, and text, they perform strongly in tasks such as visual question answering (VQA), generating detailed scene description, optical character recognition (OCR), and object-grounded spatio-temporal reasoning. With the advent of methods that increase large models efficiency [13], [14], VLMs have successfully been applied to several real-time tasks [15], including autonomous driving [16], [17]. Yet, their application to pedestrian-level safety-critical tasks like PIP and PTP remains limited. Existing pedestrian benchmarks [3], [2], [6], [1] provide rich

[1]CVIT, IIIT-Hyderabad, India. {naman.mishra@research., shankar.gangisetty@ihub-data., jawahar@}iiit.ac.in

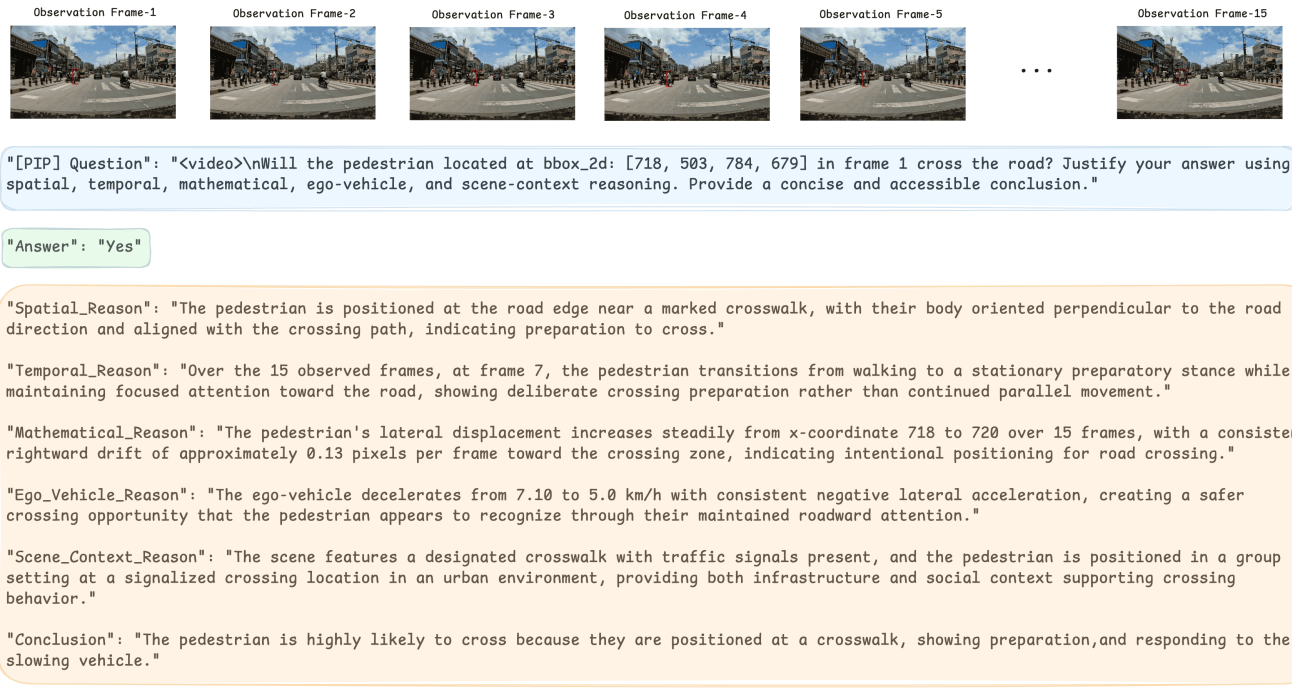

Fig. 2: PedestrianQA Dataset PIP Sample. The top row shows the observation frames. The "Question" and "Answer" are followed by 5 types of rationales and a conclusion.

spatial–temporal data, behavioral categories, environmental context, and ego-vehicle signals that can be reformulated in natural language. Integrating these textual representations with visual inputs can enable VLMs to model pedestrian behavior comprehensively and in a unified manner.

To this end, we introduce **PedestrianQA**, a multimodal video-based question–answering dataset specifically designed for pedestrian intention and trajectory prediction. Unlike prior datasets, PedestrianQA frames pedestrian behavior understanding as a video question–answering–explanation task, enabling VLMs to predict crossing intention and forecast future trajectories while simultaneously generating explanatory natural-language rationales for their decisions. Each pedestrian sequence is associated with a question, an answer, and five structured rationales that capture complementary aspects of the scene. These include spatial information (e.g., body pose, positioning, stride), temporal motion cues (e.g., acceleration, sudden stops, changes in walking speed), scene-level context (e.g., presence of crosswalk markings, traffic signals, traffic density, visibility conditions), and ego-vehicle interactions (e.g., decelerating, yielding, or assertive). For example, a rationale might describe a pedestrian leaning forward with an extended stride (spatial), increasing walking pace over the past two seconds (temporal), approaching an unsignalized intersection with moderate vehicle traffic (scene context), while the ego-vehicle slows and flashes hazard lights (interaction), collectively supporting the prediction that the pedestrian will cross imminently. Overall, this design requires models not only to predict *what* will happen (intention and trajectory) but also to explain *why* those outcomes are plausible. One such sample is depicted in Fig. 2. Further, we show that state-of-the-art VLMs finetuned on PedestrianQA achieve substantial gains in predictive accuracy and reasoning quality. Finally, PedestrianQA enables evaluation metrics that jointly assess prediction correctness and explanation quality, advancing transparent, safety-critical decision-making in autonomous driving.

In summary, our contributions are:

- **PedestrianQA**: a novel multimodal dataset that unifies pedestrian intention and trajectory prediction with fine-grained, rationale-based question–answer annotations.
- **Baseline**: a strong prediction and reasoning baseline established by finetuning a state-of-the-art VLM on PedestrianQA, demonstrating how multimodal models can be adapted to pedestrian intention and trajectory prediction.
- **Benchmark**: comprehensive experiments showing that VLMs finetuned on PedestrianQA outperform existing baselines in intention and trajectory prediction, as well as rationale quality.

II. Related Works

A. Pedestrian Intention and Trajectory Prediction

Several datasets [3], [2], [6], [1], [4], [5], [18], [19], [20] support this line of work. JAAD [2] pioneered PIP using behavioral and contextual cues but lacked scale, ego-vehicle data, and interacting vehicles localization. PIE [3] addresses these gaps by introducing vehicle trajectories and ego-vehicle odometry, while TITAN [6] offers fine-grained action priors and detailed scene dynamics. Finally, IDD-PeD [1] targets

unstructured traffic scenes with expanded pedestrian behavioral attributes. We refer readers to Table I for details. While these datasets offer diverse information, they lack inherent integration or shared structure. Further, they do not provide explicit rationales capturing these underlying relationships.

To address this gap, PIE++ [4] augments PIE with manually annotated sentence-level rationales, and PedVLM [5] introduces PedPrompt, a binary classification question–answering corpus. These datasets have drawbacks in that PIE++ remains limited in scale and annotation diversity, and PedPrompt reduces intention to binary labels, overlooking the richer information in the TRANS [21] dataset and failing to unify it into a descriptive natural language resource. In contrast, our approach automatically generates questions and answers along with diverse, fine-grained, object-grounded natural language rationales that integrate cues from all scene elements for explainable pedestrian behavior prediction.

Parallely, several video question–answer datasets [22], [23], [24], [25], [26], [27], [28], [29], [30] exist in the domain of autonomous driving. We highlight RoadSocial [25], and Rank2Tell [26] due to similarities with our own work. RoadSocial leverages multi-view videos sourced from `x.com` and generates a large-scale question-answer-reasoning dataset using a state-of-the-art LLM [31]. Rank2Tell provides multimodal urban-intersection videos with synchronized RGB, LiDAR, and CAN data, annotated for important-agent localization, importance ranking, and natural language explanations. In contrast, our work is explicitly pedestrian-centric, emphasizing predictions with object-grounded spatio-temporal reasoning.

Methodologically: Early approaches fused visual cues with SVMs [2], LSTM encoders [3], or I3D-based action models combined with ego-motion [6]. In the line of spatio-temporal fusion, PCPA [32], MaskPCPA [33], and PIP-Net [18] apply temporal attention over multimodal streams. In traditional vision-language methods, PIE++ [4] employs MINDREAD, a cross-modal encoder that fuses visual features with textual rationales. PedVLM [5] combines visual tokens via CLIP [34] and text tokens, for generating binary crossing intention labels from T5 [35]. Rank2Tell integrates 2D CNNs and 3D point clouds in a relational graph for joint ranking and captioning. Models such as LG-Traj [7] and IntentFormer [36] jointly encode visual and trajectory cues, and ClipCross [37] aligns CLIP image–text embeddings for intention classification. Utilizing large VLMs, among recent work [38], [39], GPT-4V [38] and VLMs guided by hierarchical prompts [39] explore zero-shot performances on PIP. In contrast to these methods, we use a single open-source VLM for intention prediction, trajectory prediction, and rationale generation, all without any architectural modifications.

B. Vision-Language Models

Early VLMs extended pre-trained vision encoders with instruction-tuned LLMs, for example, InstructBLIP [8] builds on BLIP-2 [40] with instruction tuning on open-source datasets, while LLaVA [11] combines vision transformers [41], [34] with the Vicuna [42] for conversational tasks. Recent models [43], [9], [10], [44] can scale to multi-image and long-video inputs. LLaVA-NeXT [43] generalizes to videos using AnyRes frame packing, linear RoPE [45] scaling for long sequences, and video supervised finetuning. Qwen2.5-VL [9] introduces a novel dynamic-resolution Vision Transformer and absolute time encoding for fine-grained grounding, and long-video comprehension. InternVL3 [10] employs native multimodal pre-training with variable visual position encoding, advanced post-training, and test-time scaling. Building on advances in reasoning for LLMs, recent work adds explicit reasoning capability to VLMs [46]. Kwai Keye-VL [44], for instance, targets short-form video with a multi-stage training recipe and a dedicated *thinking* mode for enhanced reasoning.

Several VLM methods have also been adapted to autonomous driving. Dolphins [47] adapts OpenFlamingo [48] with Grounded Chain-of-Thought (CoT) reasoning and driving-specific instruction tuning. OmniDrive [49] builds on LLaVA with two complementary designs: Omni-L and Omni-Q for joint vision-language alignment and driving-scene reasoning. DriveLM [22] extends BLIP-2 by chaining perception, prediction, and planning QAs in a graph, using context from preceding nodes to reason about the scene and generate a natural-language behavior description. In contrast, we prioritize fine-grained modeling of pedestrian dynamics and spatio-temporally grounding traffic agents to predict intentions, forecast trajectories, and provide explanatory rationales, all without making any architectural changes.

III. PedestrianQA Dataset

In this section, we describe the generation process of our PedestrianQA dataset, with the full pipeline illustrated in Fig. 3 and a PIP example of our dataset in Fig. 2

A. Data Collection

We construct our PedestrianQA corpus from four publicly available pedestrian datasets: 3 structured (JAAD [2], PIE [3], TITAN [6]), and 1 unstructured (IDD-PeD [1]). We selected these datasets for their comprehensive annotations, including localization across frames for target pedestrians, interacting pedestrians and vehicles, and traffic control elements like crosswalks and traffic lights, detailed labels of pedestrian and ego-vehicle behaviors. We refer readers to Table I for details. Each instance in our dataset features a unique target pedestrian observed from the ego-vehicle, along with surrounding interacting pedestrians, non-ego interacting vehicles, and traffic-control elements such as crosswalks and traffic lights. In subsequent sequences, the target pedestrian from an earlier sequence can act as an interacting pedestrian.

B. Sequence Sampling

Following [32], we sample 0.5 s observational frames from JAAD [2], PIE [3], and IDD-PeD [1], each of which is recorded at approximately 30 fps. A Time-To-Event (TTE) between 1–2 s is defined such that the final frame

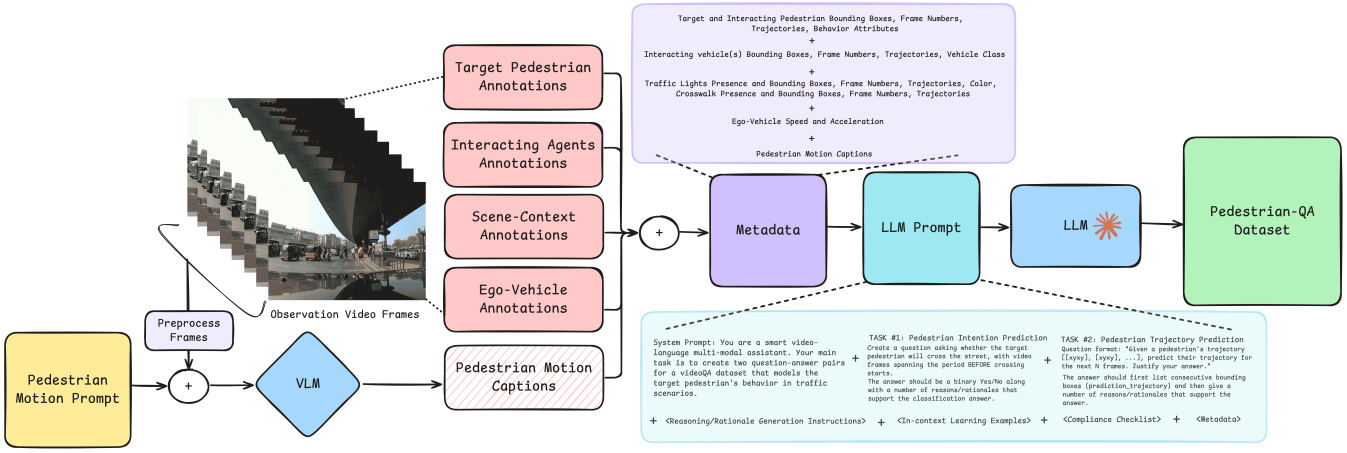

Fig. 3: Data generation pipeline: We first aggregate all ground-truth, human-annotated annotations from the constituent datasets into a unified metadata schema. We use generated VLM captions to enrich motion semantics using carefully designed pedestrian-motion prompts that target fine-grained cues. These captions are validated for format and appended to the metadata. We then construct a single instruction package containing: (i) a system prompt, (ii) task definitions for PIP and PTP, (iii) step-by-step guidance for producing structured, fine-grained rationales, (iv) a small set of in-context exemplars, (v) a compliance checklist for high-quality rationale generation, and (vi) the sequence-level metadata tables. This package is provided to `claude-sonnet-4-20250514` LLM-API to generate triplets of questions, answers, and rationales.

of each target pedestrian sequence falls within this range. For example, if the annotated crossing point occurs at frame 100, the sequence ends between frames 40 and 70 for a video recorded at 30 fps. For trajectory prediction in these datasets, we retain only those sequences that contain at least 1.5 s of video after the observation period. In TITAN [6], we extract observation sequences guided by the `"Simple Context"` annotation. Specifically, we sample frames with the TTE constraint preceding a change in context to either `"crossing a street at pedestrian crossing"` or `"jaywalking (illegally crossing NOT at pedestrian crossing)"`. Furthermore, we leverage the attribute `"waiting to cross street"` as an additional indicator of positive crossing intention. Following details in [6], TITAN is sampled at 10 fps, with an observation window of 1 s and prediction up to 2 s beyond it. We adopt the same TTE sampling protocol for TITAN, as we did with JAAD [2], PIE [3], and IDD-PeD [1].

TABLE I: Annotation availability across datasets. "✓" indicates availability of the corresponding annotation type.

Category	Annotation	IDD-PeD [1]	JAAD [2]	PIE [3]	TITAN [6]
Target Pedestrian	B.Boxes	✓	✓	✓	✓
	Trajectory	✓	✓	✓	✓
Interacting Pedestrians	B.Boxes	✓	✓	✓	✓
	Trajectory	✓	✓	✓	✓
Pedestrian Behavior	–	✓	✓	✓	✓
Interacting Vehicles	B.Boxes	✓	✗	✓	✓
	Trajectory	✓	✗	✓	✓
	Class	✓	✗	✓	✓
Ego Vehicle	Speed	✓	✗	✓	✓
	Acceleration	✓	✓	✓	✓
Scene Context	Object B.Boxes	✓	✗	✓	✗
	Object Class	✓	✓	✓	✗
Motion Captions	–	✓	✓	✓	✓

C. Metadata Construction

We begin by collecting and consolidating all ground-truth, human-annotated labels provided by the constituent datasets, including pedestrian and interacting vehicle bounding boxes across frames, behavioral labels, ego-vehicle motion, and scene context (see Table I). This metadata is normalized into a unified tabular schema to ensure consistency across datasets.

D. Pedestrian Motion Captioning

To enrich the annotations with finer behavioral cues, we generate supplementary captions using a state-of-the-art VLM [9] prompted with carefully designed pedestrian-motion prompts. These prompts are tailored to elicit fine-grained motion descriptions (see Table II). Only the observation frames are used for this purpose. Each frame is preprocessed by extracting a 448×448 crop centered on the pedestrian and overlaying a red bounding box around the target individual across all crops, ensuring that the model consistently attends to the correct pedestrian and their immediate surroundings. Empirically, we found that feeding crops instead of full frames elicits more accurate responses from the VLM. The resulting captions are integrated with the original metadata.

E. Prompt Design and Quality Assurance

The metadata is transformed into a structured TSV format, where the rows correspond to frame indices and columns correspond to various attributes. Next, we construct a comprehensive prompt for LLM-based QA generation. In the

TABLE II: Prompts for generating pedestrian motion captions. Each row describes one category and its corresponding prompt.

Category	Prompt
Initial State and Readiness	What is the pedestrian doing at the start? Are they stationary, adjusting posture, shifting weight, or initiating movement?
Motion Trajectory and Dynamics	Is there consistent movement, acceleration, or deceleration? How does their position change over time?
Walking Direction	Are they walking perpendicular to the road (toward/across) or parallel (alongside)? Justify using their movement path and orientation.
Body Language and Orientation	Describe torso, leg, and arm alignment. Are they oriented toward the road or the crossing path?
Head and Gaze Behavior	Is their head facing towards the road, oncoming traffic, or scanning the scene, indicating awareness or intent to cross?
Interaction with Vehicles and Other Agents	Do they pause, slow down, or move in response to surrounding vehicles or pedestrians?
Signs of Hesitation or Yielding	Is there any visible hesitation or cautious behavior during the observation window?
Surface and Environment	What are they walking on? footpath, crosswalk, road? Do nearby elements support the idea of crossing intent?
Behavioral Transition	How does their behavior evolve across the N frames, e.g., from stationary to walking, or from observing to initiating a step?
Risk Awareness and Goal Inference	Are there signs of caution, urgency, or confidence that suggest a deliberate crossing decision (without guessing mental state)?
Future Behavior	Based on the visible behavior, what will the pedestrian most likely do in the next X seconds?
Scene Context	Are there road edges, traffic signals, or markings that help contextualize or influence the pedestrian's movement?
Final Position and Commitment	By the end of the observation window, has the pedestrian committed to crossing (e.g., stepped off curb, entered road)?

prompt, we additionally incorporate a "compliance checklist" to enforce rules for rationale generation, requiring the LLM to satisfy specific constraints for high-quality outputs and to regenerate whenever any constraint is violated. This checklist ensures that the rationales reflect information from all ground-truth annotation sources and motion captions. It also ensures that the rationales are not just *parroted* attributes present in the metadata.

The composite prompt now includes: (i) a system prompt establishing the task domain and constraints, (ii) explicit definitions of PIP and PTP tasks, (iii) detailed instructions for producing structured rationales across five reasoning categories (spatial, temporal, mathematical, ego-vehicle, and scene context), a final destinaton prediction, and a simple conclusion for everyday users, (iv) a curated set of in-context examples to guide reasoning style and output structure, (v) the compliance checklist, and (vi) the full sequence metadata.

This prompt is fed into the `claude-sonnet-4` LLM-API [31], which outputs question–answer–rationale triplets for each pedestrian sequence. Our pipeline yields a scalable method for producing multimodal reasoning annotations. We record $10,251$ samples in our training set and $4,059$ in our test set.

F. Rationale Generation

To ensure that intention and trajectory predictions are explainable, PedestrianQA requires models to generate structured rationales spanning multiple scene dimensions. The rationales are categorized and described as follows:

1) **Spatial Reasoning** captures the pedestrian's physical state in the environment, including their pose, body orientation (e.g., parallel or perpendicular to the road), and exact spatial placement (e.g., standing on a curb or walking along a lane).
2) **Temporal Reasoning** accounts for how motion evolves over time, such as the initiation of walking, acceleration or deceleration, and pauses before crossing. This category of rationale refers to frame numbers as timestamps.
3) **Mathematical Reasoning** introduces a quantitative perspective, incorporating pedestrian–vehicle distance, trajectory angle, pedestrian velocity estimates, and displacement.
4) **Ego-Vehicle Reasoning** links the pedestrian's decisions to the behavior of the ego-vehicle: changes in speed, acceleration/deceleration, and braking that may enable or discourage crossing.
5) **Scene-Context Reasoning** situates the pedestrian within a broader environment, drawing on cues such as traffic lights, crosswalks, road infrastructure, illumination, and interactions with other agents.
6) **Final Destination Prediction** requires the model to infer the pedestrian's likely endpoint within the scene, at the end of their trajectory. For example, reaching the opposite curb, halting midway, or continuing along the same side of the road.
7) **Conclusion** summarizes the overall judgment in simple, accessible terms, providing a binary decision for intention prediction (e.g., cross or not cross) or a concise description of the forecasted trajectory.

IV. Baseline Model

We finetune the `Qwen2.5-VL-3B-Instruct` model [9] on the PedestrianQA dataset using the official implementation[1]. Both PIP and PTP tasks are framed as QAs, with training data drawn from the official train splits (excluding validation) of all constituent datasets. We instruction-finetune the model using parameter-efficient LoRA [50] adapters ($rank = 8$, $\alpha = 16$) to prevent overfitting. Empirically, we observe consistent gains on intention classification, trajectory prediction, and reasoning tasks, suggesting that the dataset size is sufficient for targeted capability adaptation. For JAAD [2], PIE [3], and IDD-PeD [1], we finetune the model with 15 input frames (corresponding to $0.5s$ of observation time), whereas, for TITAN [6], we use 10 input frames (corresponding to $1s$ of observation time). All finetuning runs are performed on the original video resolution, with default hyperparameters, for 3 epochs on `2×Nvidia RTX-A6000` GPUs.

V. Experiments

A. Experimental Settings

Our benchmark includes a representative set of widely used VLMs: Dolphins [47], LLaVA-NeXt [43], and InternVL3 [10], Qwen2.5-VL [9], together with Kwai-Keye [44], a more recent rationale-generation model. Each model is prompted with task-specific instructions tailored to itself, and optimized for prediction and rationale generation, for both PIP and PTP tasks. We evaluate on the default test sets of all constituent datasets. For PIP, we formulate the task as binary classification, report accuracy and F_1-score. For PTP, we compute trajectory forecasting metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE). To assess caption quality, we adapt the CLAIR

[1] https://github.com/QwenLM/Qwen2.5-VL/

TABLE III: PIP Results. Performance comparison of **Driving-specific VLM**, **Large-scale General-Purpose VLMs**, **Small-scale General-Purpose VLMs**, and our finetuned **Qwen2.5-VL-3b** [9] models. Finetune indicates whether the model was finetuned (✓) or evaluated in a zero-shot (✗) setting. **Bold** and underline is for best and second-best results.

Model	Params	Finetune	PIE [3]		JAAD [2]		TITAN [6]		IDD-PeD [1]		Overall	
			Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Dolphins-9B [47]	9B	✗	0.667	0.800	0.554	0.713	0.072	0.134	0.084	0.155	0.239	0.386
InternVL3-8B-Instruct [10]	8B	✗	0.466	0.451	0.509	0.460	0.749	0.130	0.663	0.130	0.635	0.298
Kwai-Keye-8B [44]	8B	✗	0.603	0.723	0.576	0.704	0.226	0.127	0.297	0.144	0.361	0.370
LLaVA-NeXT-Video-7B [43]	7B	✗	0.411	0.301	0.446	0.039	0.578	0.127	**0.910**	0.032	0.675	0.193
Qwen2.5-VL-7B-Instruct [9]	7B	✗	0.449	0.323	**0.627**	0.548	**0.926**	0.029	0.876	0.087	0.780	0.291
Qwen2.5-VL-3B-Instruct [9]	3B	✗	0.667	0.800	0.554	0.713	0.072	0.134	0.084	0.155	0.239	0.386
InternVL3-2B-Instruct [10]	2B	✗	0.557	0.635	0.475	0.513	0.531	0.145	0.488	0.137	0.515	0.352
Ours (PIE)	3B	✓	0.667	**0.800**	0.554	0.705	0.073	0.134	0.102	0.158	0.247	0.389
Ours (JAAD)	3B	✓	0.667	0.795	0.554	**0.713**	0.072	0.134	0.084	0.155	0.239	0.386
Ours (TITAN)	3B	✓	**0.670**	0.792	0.576	0.664	0.673	0.200	0.572	0.188	0.624	0.514
Ours (IDD-PeD)	3B	✓	0.665	0.754	0.588	0.633	0.733	**0.203**	0.867	0.204	0.767	**0.568**
Ours (All Datasets)	3B	✓	0.633	0.709	0.531	0.497	0.808	0.201	0.880	**0.205**	**0.783**	0.542

TABLE IV: Performance comparison of PTP results. ✓ indicates finetuned models, ✗ zero-shot. **Bold** marks best (lowest) values, underline second-best. Dolphins [47] was excluded from evaluation as it cannot generate trajectory bounding boxes required for this analysis. ADE and FDE are measured in pixels, in the image coordinate space.

Model	Params	Finetune	PIE [3]		JAAD [2]		TITAN [6]		IDD-PeD [1]		Overall	
			ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
InternVL3-8B-Instruct [10]	8B	✗	5937	19831	6118	15619	657	782	279	302	28414	8052
Kwai-Keye-8B [44]	8B	✗	54	104	61	124	**30**	**56**	66	123	44	85
LLaVA-NeXT-Video-7B [43]	7B	✗	190	350	143	204	135	243	141	158	1522	261
Qwen2.5-VL-7B-Instruct [9]	7B	✗	**33**	**52**	42	**84**	37	57	56	98	38	**63**
Qwen2.5-VL-3B-Instruct [9]	3B	✗	96	176	54	101	43	66	61	106	61	105
InternVL3-2B-Instruct [10]	2B	✗	827	1552	435	751	293	464	350	586	4632	811
Ours (PIE)	3B	✓	51	87	43	89	36	58	54	94	43	74
Ours (JAAD)	3B	✓	93	167	56	105	43	69	60	104	60	104
Ours (TITAN)	3B	✓	55	94	43	89	35	64	49	87	43	78
Ours (IDD-PeD)	3B	✓	71	127	44	88	37	58	50	88	49	84
Ours (All Datasets)	3B	✓	41	70	**41**	87	31	60	**46**	**80**	**37**	68

TABLE V: Rationale evaluation on the combined dataset, with Claude-Sonnet-4. Average scores (0–100) for Spatial Reasoning (SR), Temporal Reasoning (TR), Mathematical Reasoning (MR), Ego-Vehicle Reasoning (EVR), Scene-Context Reasoning (SCR), Final Destination Prediction (FDP), and Conclusion (C). ✓ indicates finetuned models, ✗ zero-shot. **Bold** shows best score per column; underline marks the second-best. Dolphins [47] generates only a brief conclusion and does not generate category-specific rationales.

Model	Params	Finetune	SR	TR	MR	EVR	SCR	FDP	C
Dolphins-9B [47]	9B	✗	0.00	0.00	0.00	0.00	0.00	0.00	12.05
InternVL3-8B-Instruct [10]	8B	✗	37.01	37.60	26.36	29.17	37.79	29.63	45.05
Kwai-Keye-8B [44]	8B	✗	33.15	30.74	28.10	36.38	37.09	**38.03**	28.70
LLaVA-NeXT-Video-7B [43]	7B	✗	46.39	33.96	39.52	40.06	28.77	17.22	50.33
Qwen2.5-VL-7B-Instruct [9]	7B	✗	40.15	41.90	31.74	38.08	38.04	12.76	50.89
Qwen2.5-VL-3B-Instruct [9]	3B	✗	22.05	19.92	17.25	27.11	28.80	10.48	14.80
InternVL3-2B-Instruct [10]	2B	✗	31.71	32.30	23.02	27.47	34.25	25.95	37.09
Ours (PIE)	3B	✓	25.52	24.48	22.00	36.70	30.43	14.58	18.28
Ours (JAAD)	3B	✓	21.85	19.84	17.03	26.99	28.46	10.95	15.00
Ours (TITAN)	3B	✓	46.26	41.71	39.77	45.25	47.38	22.80	43.06
Ours (IDD-PeD)	3B	✓	56.33	51.35	46.83	57.77	55.63	26.53	55.37
Ours (All Datasets)	3B	✓	**58.36**	**54.84**	**51.68**	**61.51**	**59.72**	32.24	**60.25**

framework [51] to our domain and employ Claude-Sonnet-4 [31] as the evaluator.

B. Pedestrian Intention Prediction Results

In Table III, we observe that `Ours (All Datasets)` model achieves the highest overall accuracy of 78.3%, narrowly outperforming a much larger `Qwen2.5-VL-7B-Instruct`, and comfortably outperforming other 2b/3b-parameter VLMs. These include `Qwen2.5-VL-3B-Instruct` and `InternVL3-2B-Instruct`, which our finetuned model outperforms by 54.4% and 26.8% on overall accuracy, respectively. We also highlight that while `Qwen2.5-VL-7B-Instruct` achieves impressive performance, our finetuned model outperforms other models of similar scale: `Dolphins-9B`, `InternVL3-8B-Instruct`, `Kwai-Keye-8B`, and `LLaVA-NeXT-Video-7B` by 54.4%, 14.8%, 42.2%, and 10.8%, respectively. We attribute the performance gains to the finetuned model's ability to recognize subtle cues of crossing behavior before the action occurs. The model attends to pedestrian orientation, pose, and surrounding traffic in structured environments, while the unstructured datasets contribute edge-case examples of erratic pedestrian behavior. This effect is evident in our IDD-PeD-only and TITAN-only fine-tuned models, which achieve strong

results on the PIE and JAAD datasets. The high accuracy and low F1-score stem from the strong class imbalance between positive and negative crossing samples, with the negative class being far more prevalent. `Ours (All Datasets)` distinguishes the classes more effectively, achieving an overall F1-score that is 25.1% higher than that of `Qwen2.5-VL-7B-Instruct`.

C. Pedestrian Trajectory Prediction Results

In Table IV, we observe that our finetuned model `Ours (All Datasets)` achieves the lowest overall ADE, while narrowly losing out to a much larger `Qwen2.5-VL-7B` model in the FDE metric, by just 5 pixels. The same finetuned model achieves the best ADE in JAAD and IDD-PeD. Our model loses out to `Qwen2.5-VL-7B` in the PIE test set by just 9 pixels. Once again, we observe that our finetuned model `Ours (All Datasets)` achieves a much lower displacement error compared to other models of similar size: `InternVL3-2B` by 426 pixels ADE and 743 pixels FDE, and `Qwen2.5-VL-3B` by 24 pixels ADE and 37 pixels FDE on the overall result. This performance gain arises from jointly training the model on PIP and PTP tasks, enhancing its ability to infer crossing intentions accurately. Pedestrians oriented perpendicular to the road are considerably more likely to cross and exhibit trajectories distinct from those parallel to it. We notice that `Kwai-Keye-8B` achieves the lowest error on the TITAN dataset, narrowly beating our `Ours (All Datasets)` by just 1 pixel ADE and 4 pixels FDE. The rest of our finetuned models remain comparable to `Qwen2.5-VL-7B` and `Kwai-Keye-8B`.

D. Rationale Generation Results

In rationale evaluation (Table V), on the combined dataset, our finetuned `Ours (All Datasets)` achieves the highest score in 6 out of 7 types of captions, losing out to `Kwai-Keye-8B` in the Final Destination Prediction category by nearly 6%. In other categories, it outperforms `Qwen2.5-VL-7B` by as much as 23.43% (Ego-Vehicle Reasoning). While we use `claude-sonnet-4` for both generation and evaluation, scores remain well below saturation ($\leq$ 62%). The evaluation process penalizes hallucinations and weak visual grounding, thus preventing circular bias.

E. Qualitative Analysis

We establish a qualitative analysis between our strongest model: `Ours (All Datasets)` and the strongest default `Qwen-2.5-VL-7b` model in Fig. 4. Our model generates rationales with more accurate object grounding (e.g., detecting crosswalks and traffic lights in PIE) and richer context from visual cues (e.g., identifying the curb as the destination in TITAN). The rationales are also more detailed and less prone to hallucinations. For example, in the IDD-PeD example, our model correctly recognizes a pedestrian standing in the middle of the road. Whereas the `Qwen2.5-VL-7b` baseline incorrectly assumes the starting position is on the edge of the road due to its higher statistical bias of a positive crossing intention.

Fig. 4: Qualitative analysis between `Qwen2.5-VL-7b` and `Ours (All Datasets)` on the compositional datasets (clockwise: IDD-PeD [1], JAAD [2], TITAN [6], PIE [3]). We identify samples having correct predicate predictions for a fair comparison among rationales.

F. Effect of Training on Unstructured Traffic Data

Interestingly, our model trained only on IDD-PeD [1] maintains a comparable performance with our model trained on all datasets (1 to 5% lower rationale scores). This pattern can also be noticed in PTP tasks in the JAAD-ADE (44 vs 41 pixels) and the TITAN-ADE (37 vs 31 pixels) metrics. In PIP, this behavior is observed in PIE accuracy (higher at 66.5% compared to 63.3%), JAAD accuracy (higher at 58.8% compared to 53.1%), comparable performance in TITAN accuracy, and narrowly losing out by 1.6% in the overall accuracy. Based on these observations, we conclude that sufficiently powerful models like [9], if trained only on IDD-PeD [1], a dataset containing only chaotic unstructured traffic scenarios and a much higher density of traffic agents, can have comparable or better performance over similar models trained on structured datasets.

VI. Conclusion

This work reframes pedestrian behavior understanding as a multimodal reasoning problem. It introduces PedestrianQA, a video-based QA corpus that jointly evaluates: (i) PIP, (ii) PTP, and (iii) fine-grained, object-grounded spatio-temporal rationales. By coupling short-horizon visual evidence with structured explanations, PedestrianQA encourages models to predict *what* will happen and to justify *why* it is plausible, advancing explainability for safety-critical autonomous driving. Empirically, finetuning state-of-the-art VLMs on PedestrianQA yields consistent gains across structured (PIE [3], JAAD [2], TITAN [6]), and unstructured (IDD-PeD [1])

datasets, improving both decision accuracy (PIP) and motion forecasting quality (PTP), while producing more informative and category-aligned rationales. Notably, compact models finetuned on our corpus approach or surpass the performance of substantially larger baselines, suggesting that task-specific multimodal supervision is a powerful complement to internet-scale pretraining. The strong transfer observed from IDD-PeD-only training to other datasets indicates that dense, unstructured scenarios provide rich supervisory signals for generalization.

Acknowledgment. This project was supported by iHub-Data and Mobility at IIIT Hyderabad.

REFERENCES

[1] R. Bokkasam, S. Gangisetty, A. H. A. Hafez, and C. V. Jawahar, "Pedestrian intention and trajectory prediction in unstructured traffic using idd-ped," in *ICRA*, 2025.

[2] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017.

[3] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019.

[4] V. Khindkar, V. Balasubramanian, C. Arora, A. Subramanian, and C. Jawahar, "Can reasons help improve pedestrian intent estimation? a cross-modal approach," in *IROS*, 2024.

[5] F. Munir, S. Azam, T. Mihaylova, V. Kyrki, and T. P. Kucner, "Pedestrian vision language model for intentions prediction," *IEEE Open Journal of Intelligent Transportation Systems*, 2025.

[6] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *CVPR*, 2020.

[7] P. S. Chib and P. Singh, "Lg-traj: Llm guided pedestrian trajectory prediction," *arXiv preprint arXiv:2403.08032*, 2024.

[8] W. Dai *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *NeurIPS*, 2023.

[9] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[10] J. Zhu *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[12] G. Team, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[13] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "GPT3.int8(): 8-bit matrix multiplication for transformers at scale," in *NeurIPS*, 2022.

[14] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," in *NeurIPS*, 2022.

[15] A. Sharshar, L. U. Khan, W. Ullah, and M. Guizani, "Vision-language models for edge networks: A comprehensive survey," *IEEE Internet of Things Journal*, 2025.

[16] X. Zhou *et al.*, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[17] B. Jiang, S. Chen, Q. Zhang, W. Liu, and X. Wang, "Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning," *arXiv preprint arXiv:2503.07608*, 2025.

[18] M. Azarmi, M. Rezaei, and H. Wang, "Pip-net: Pedestrian intention prediction in the wild," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[19] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE RA-L*, 2020.

[20] W. Kim *et al.*, "Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections," *IEEE RA-L*, 2019.

[21] D. Guo, T. Mordan, and A. Alahi, "Pedestrian stop and go forecasting with hybrid feature fusion," in *ICRA*, 2022.

[22] C. Sima *et al.*, "Drivelm: Driving with graph visual question answering," in *ECCV*, 2023.

[23] X. Tian *et al.*, "DriveVLM: The convergence of autonomous driving and large vision-language models," in *CoRL*, 2024.

[24] J. Fang *et al.*, "Abductive ego-view accident video understanding for safe driving perception," in *CVPR*, 2024.

[25] C. Parikh, D. Rawat, R. R. T., T. Ghosh, and R. K. Sarvadevabhatla, "Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives," in *CVPR*, 2025.

[26] E. Sachdeva *et al.*, "Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning," in *WACV*, 2024.

[27] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," *AAAI*, 2024.

[28] A. Marcu *et al.*, "Lingoqa: Visual question answering for autonomous driving," in *ECCV*, 2024.

[29] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *WACV*, 2023.

[30] D. Wu, W. Han, Y. Liu, T. Wang, C. zhong Xu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," in *AAAI*, 2025.

[31] Anthropic, "System card: Claude opus 4 & claude sonnet 4," 2025. [Online]. Available: www.anthropic.com/claude/sonnet

[32] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *WACV*, 2021.

[33] D. Yang, H. Zhang, E. Yurtsever, K. Redmill, and U. Ozguner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, 2022.

[34] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[35] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, 2020.

[36] N. Sharma, C. Dhiman, and S. Indu, "Predicting pedestrian intentions with multimodal intentformer: A co-learning approach," *Pattern Recogn.*, 2025.

[37] R. Uziel and O. Bialer, "Optimizing vision-language model for road crossing intention estimation," in *WACV*, 2025.

[38] J. Huang, P. Jiang, A. Gautam, and S. Saripalli, "Gpt-4v takes the wheel: Promises and challenges for pedestrian behavior prediction," *AAAI*, 2024.

[39] M. Azarmi, M. Rezaei, and H. Wang, "Pedestrian intention prediction via vision-language foundation models," *arXiv preprint arXiv:2507.04141*, 2025.

[40] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[41] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[42] W. Chiang *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[43] Y. Zhang *et al.*, "Llava-next: A strong zero-shot video understanding model," 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[44] K. K. Team, "Kwai keye-vl technical report," *arXiv preprint arXiv:2507.01949*, 2025.

[45] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomput.*, 2024.

[46] Y. Li *et al.*, "Perception, reason, think, and plan: A survey on large multimodal reasoning models," *arXiv preprint arXiv:2505.04921*, 2025.

[47] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," in *ECCV*, 2024.

[48] A. Awadalla *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[49] S. Wang *et al.*, "OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," in *CVPR*, 2025.

[50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[51] D. Chan, S. Petryk, J. Gonzalez, T. Darrell, and J. Canny, "CLAIR: Evaluating image captions with large language models," in *"EMNLP"*, 2023.