BIO-TAD: BIOMECHANICAL TARGET-ANCHORED DIFFUSION FOR MULTIMODAL PEDESTRIAN TRAJECTORY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intelligent systems in autonomous driving necessitate effective approaches for predicting multimodal pedestrian trajectories, especially during highly stochastic maneuvers characterized by minimal early macro-spatial displacement. We present Bio-TAD, an innovative backbone-agnostic generative framework that utilizes fine-grained biomechanical posture cues to effectively anchor a trajectory diffusion model. Instead of passively concatenating continuous features, our method employs an Orthogonal Prototype Memory Bank to convert kinematic signals into discrete behavioral primitives, thereby constraining the generative space. The generation process is regulated by a target-oriented Adaptive Layer Normalization (AdaLN) engine and optimized through a cohesive Multi-Task Diffusion Objective that aligns spatial diffusion errors and destination anchoring concurrently. Comprehensive assessments across three benchmark datasets demonstrate superior performance, yielding up to a 12% reduction in Final Displacement Error relative to the closest pose-augmented prior work, and an absolute improvement of up to 12.11% in semantic Intent Rate over coordinate-only baselines. Through controlled ablation, we further show that discretizing pose into orthogonally regularized primitives outperforms continuous concatenation under otherwise identical conditions, isolating discretization as the source of the gain. To expose this effect, to which standard minADE/minFDE metrics are structurally blind, we introduce a novel Intent Rate metric that measures semantic mode correctness against a training-decoupled geometric classifier.

1 INTRODUCTION

Human locomotion is not simply a series of movements of the center of mass; it is mostly indicated by body language prior to any significant spatial displacement. Subtle indicators, such as a twisted shoulder or a rotated pelvis, serve as immediate antecedents to abrupt trajectory alterations (Liang et al., 2019; Gao et al., 2025). When models disregard these tiny physiological indications, they must produce future trajectories as unstructured, random statistical fluctuations.

Our empirical diagnostic analysis reveals a pronounced and persistent predictive deficiency in coordinate-only paradigms, particularly in the moments immediately preceding abrupt maneuvers. As illustrated in Figure 1, within a critical multi-frame interval spanning from two frames to one frame prior to a sudden turning event, trajectory-only baselines such as AutoBots (Girgis et al., 2021) exhibit a near-complete failure to anticipate the maneuver. This is evidenced by a substantial escalation in spatial error ($\mathrm{FDE} \geq 2.934\,\mathrm{m}$) and a marked degradation in intent alignment, with accuracy declining to as low as 36.5%. This limitation arises because the center of mass has not yet undergone any significant macro-spatial displacement during this phase.

Despite rapid advancements, state-of-the-art multimodal trajectory prediction models, including those employing Generative Adversarial Networks (GANs) (Dou & Lu, 2024), Variational Autoencoders (VAEs) (Xiang et al., 2024), and contemporary Diffusion models (Bae et al., 2024), predominantly depend on historical (X, Y) coordinate sequences to anticipate movement. By treating pedestrians as point masses, these models structurally exclude the biomechanical signals that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

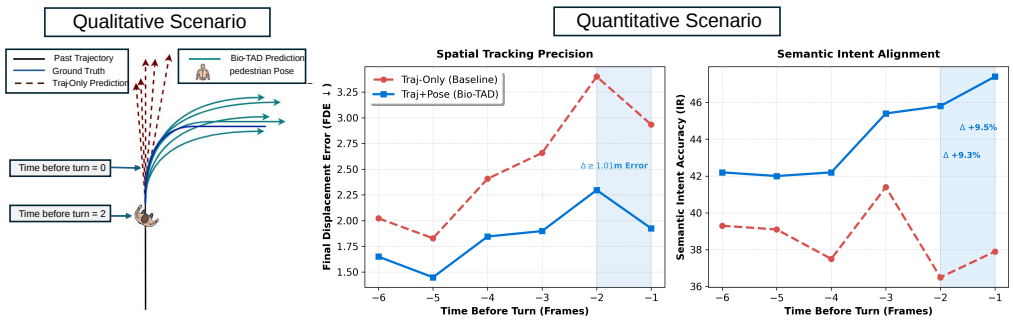


Figure 1: Predictive performance comparison between the coordinate-only baseline (Girgis et al., 2021) and the Bio-TAD framework at critical moments preceding a turn. Left: Schematic illustration of predicted trajectories at two temporal snapshots: two frames and zero frames before a direction change, relative to ground truth. Right: Final Displacement Error (FDE, lower is better) and Intent Rate (IR, higher is better) as a function of time before the turn. The shaded region highlights the critical window one to two frames before the turn, where the coordinate-only baseline fails most severely. Within this window, Bio-TAD reduces FDE by over 1.01 m and improves intent alignment by up to 9.5%.

precede spatial displacement, treating pose as irrelevant rather than as an early predictive signal that coordinates, by definition, cannot carry.

In contrast, by incorporating fine-grained biomechanical posture cues, our framework is able to detect early physiological precursors to motion. This enables significantly improved predictive performance, reducing spatial error by over $1.01\,\mathrm{m}$ (achieving an FDE of $1.924\,\mathrm{m}$) and increasing absolute intent prediction accuracy by up to 9.5% within this critical temporal window. Incorporating these social and biomechanical signals is therefore essential because they serve as a definitive limit on the unpredictable generating space, allowing models to provide accurate multimodal future predictions that better reflect intended behavior.

To address this disparity, we propose Bio-TAD (Biomechanical Target-Anchored Diffusion). While foundational generative models have successfully utilized discrete latent spaces (Chai et al., 2020) and Adaptive Layer Normalization (AdaLN) (Peebles & Xie, 2023) for macroscopic trajectory and image generation, Bio-TAD represents the first systematic study of adapting these mechanisms to organize dense biomechanical pose. Specifically, we introduce an Orthogonal Prototype Memory Bank to quantize continuous kinematic signals into discrete behavioral slots. This establishes a fundamentally different role for pose, rather than acting as a passively concatenated auxiliary feature, our architecture translates it into discrete priors that actively shape the generative manifold of a diffusion model. Bio-TAD fundamentally redesigns multi-future forecasting by separating the prediction pipeline into discrete macroscopic (spatial routing) and microscopic (biomechanical posture) processing streams. In contrast to contemporary methods like social-pose (Gao et al., 2025), which passively concatenate continuous posture embeddings, Bio-TAD employs Orthogonal Prototype Memory Banks to convert these biomechanical signals into discrete behavioral primitives. Concatenation leaves the unconstrained latent space of the backbone to resolve the mapping between pose and multimodal futures, with no structural guarantee that distinct behavioral modes remain separable within the generative manifold. The generating process is regulated by a target-oriented Adaptive Layer Normalization (AdaLN) engine that derives anticipated destination way points and consistently directs the diffusion denoiser towards that particular physical goal. The whole system is optimized end-to-end through a unified Multi-Task Diffusion Objective, a cohesive loss function that simultaneously aligns spatial diffusion error prototype orthogonality, and destination anchoring. The intent classifier used exclusively for evaluation in Intent Rate (IR) is a deterministic geometric rule applied post-hoc to predicted trajectory endpoints and requires no training. The intent classifier shares no parameters with the training pipeline, fully avoiding circularity. We confirm Bio-TAD's model-agnostic behaviour by overlaying it on recent state-of-the-art models with different architectures, demonstrating consistent enhancements across various encoder architectures.

The main technical contributions of the proposed framework can be summarized as follows:

- We introduce Bio-TAD, a backbone-agnostic generative framework that moves beyond continuous feature concatenation by translating biomechanical posture into discrete, orthogonally-regularized behavioral primitives that actively anchor the generative space. Through controlled ablation (A.1), we show this discretization yields measurable gains over continuous concatenation under otherwise identical conditions.
- We introduce a specialized multi-task optimization objective that concurrently enforces spatial diffusion precision, orthogonal memory variety, and explicit target convergence.
- We employ a target-oriented Adaptive Layer Normalization (AdaLN) engine that closely aligns produced pathways with anticipated destination waypoints throughout the denoising process.
- We identify a measurement gap in standard evaluation: minADE/minFDE reward a single well-placed sample and are structurally blind to whether the predicted distribution matches true behavioral intent. We introduce Intent Rate (IR) to fill this gap, measuring mode correctness against an independent geometric classifier that is fully decoupled from training.
- We achieve state-of-the-art results on benchmark datasets across multiple encoder architectures.

2 Related Work

2.1 Multimodal Trajectory Forecasting

Trajectory prediction has transitioned from deterministic forecasting to probabilistic modeling to encapsulate the intrinsic multimodality of human motion. Foundational models such as Social GAN effectively combined sequence-to-sequence architectures with generative adversarial networks to generate socially acceptable and diverse trajectories (Gupta et al., 2018). Simultaneously, techniques such as MultiPath utilized predefined sets of future state-sequence anchors to provide varied probabilistic hypotheses (Chai et al., 2020). In contrast to MultiPath's spatial trajectory discretizations, Bio-TAD introduces discretizations at the microscopic kinematic level, converting pre-movement biomechanical posture cues into discrete behavioral primitives before any spatial displacement occurs.To address the increasing need for intricate spatio-temporal reasoning, sequence representations employing agent-aware Transformers, like AgentFormer, were developed to concurrently describe temporal and social dimensions (Yuan et al., 2021).

In order to produce realistic trajectory assumptions, recent diffusion-based methods iteratively denoise data distributions to overcome extreme stochasticity. This evolution has led to an increasing focus on explicitly quantifying uncertainty (Suk & Kim, 2025), utilizing large language models for historical motion cues (Chib & Singh, 2025), and incorporating pedestrian motion intent to enhance interpretability (Liu et al., 2026). Collectively, these developments demonstrate how the field is changing from passively anticipating geographical pathways to actively anchoring generative processes in behavioral causes that can be explained. However, contemporary intention-aware diffusion models (Liu et al., 2026) continue to obtain intent solely from the spatial domain, which predicts macroscopic endpoints by deriving short-term directional patterns from coordinate histories. By defining intent at the microscopic physiological level and extracting behavioral primitives straight from posture before any spatial displacement, our Bio-TAD framework, in contrast, creates a real biomechanical foundation.

2.2 Integration of Auxiliary Inputs

Multi-task learning has proven effective in capturing the complex intricacies of human movement. Recent studies indicate that traditional coordinate-based trajectory prediction can be enhanced by incorporating additional physical indicators, including semantic intention forecasting (Bouhsain et al., 2020), 3D bounding-box estimation (Saadatnejad et al., 2022), and head pose prediction (Hasan et al., 2021). Prior enhancements systematically employed 2D posture keypoints to directly influence short-term trajectory predictions (Yagi et al., 2018).

The inherent depth ambiguity of two-dimensional keypoints has led to increased interest in employing three-dimensional body pose for trajectory prediction (Gao et al., 2025). Existing frameworks primarily regard 3D biomechanical integration as a passive feature-engineering task. Current

methods integrate pose through training-time knowledge distillation (Jeong et al., 2025), post-hoc physics-based regularization (Taketsugu et al., 2025), or by concatenating continuous pose embeddings into unrestricted latent spaces (Gao et al., 2025). In contrast to these methods, which treat pose as a training signal, a post-generation filter, or a passively concatenated auxiliary feature, Bio-TAD translates biomechanical posture into discrete, backbone-independent behavioral primitives that actively anchor the forward diffusion process.

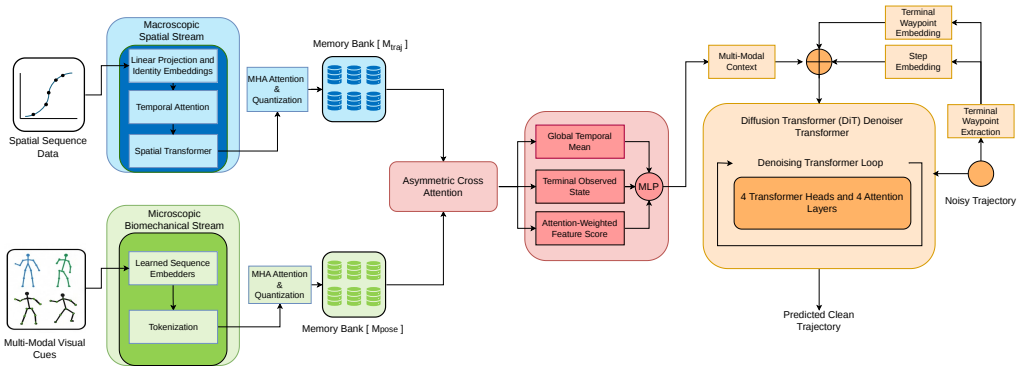


Figure 2: **Overall architecture of the Bio-TAD framework.** A dual-stream encoder processes macroscopic spatial trajectories and microscopic biomechanical cues, translating them into distinct behavioral primitives using orthogonal memory banks. The streams are integrated by an asymmetric cross-attention process and consolidated into a cohesive multi-modal environment. A target-oriented diffusion transformer (DiT) utilizes the aggregated context, step embeddings, and dynamically derived terminal waypoints to iteratively denoise and predict the clean future trajectory.

3 Proposed Method: Bio-TAD

3.1 Formal Problem Formulation

Considering a historical spatial sequence $\mathbf{X}_{traj} = \{x_1, \ldots, x_{t_p}\}$ and its associated multi-modal visual cues, which include 2D/3D bounding boxes and 2D/3D skeleton pose keypoints. Let $\mathbf{X}_{bio} = \{p_1, \ldots, p_{t_p}\}$ be the observations over a historical interval t_p. The goal is to predict K feasible future trajectories $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^K\}$ over the future horizon t_f.

3.2 Hierarchical Dual-Stream Encoding

To ensure backbone-agnostic compatibility, Bio-TAD processes incoming tracking data through a parallel, dual-stream Transformer architecture mapped to a hidden dimension of $d = 256$. As illustrated in Figure 2, the Macroscopic Spatial Stream encodes historical trajectories via linear projections enhanced with spatial and agent identity embeddings, routing features through localized temporal attention and global spatial transformers to yield a spatial condition tensor $\mathbf{H}_{traj}$. Concurrently, the Microscopic Biomechanical Stream maps physical bounding boxes and multi-dimensional joint configurations into a joint-level kinematic tensor $\mathbf{H}_{pose}$.

3.3 Orthogonal Prototype Memory Bank

Continuous embedding spaces include intrinsic variations that collapse into unstructured statistical noise during high-uncertainty pedestrian movements. To mitigate this deterioration, we propose parallel, dual-stream learnable memory matrices for both encoders: $\mathbf{M}_{traj}, \mathbf{M}_{pose} \in \mathbb{R}^{S \times d}$, equipped with $S = 256$ unique behavioral slots. To extract discrete behavioral primitives, we apply a Gumbel-Softmax operation to the Multi-Head Attention (MHA) scores, enforcing a hard routing to the memory slots. The quantized descriptor is formulated as:

$$\hat{H}_* = H_* + \text{Dropout}(\text{MHA}_{\text{Gumbel}}(\text{Norm}(H_*), M_*, M_*)) \tag{1}$$

where $* \in \{\text{traj}, \text{pose}\}$. To guarantee that the codebook slots encompass a wide, non-redundant range of pedestrian actions without mode collapse, we implement structural diversity by minimizing a joint slot orthogonality regularization loss relative to the identity matrix $\mathbf{I}$:

$$\mathcal{L}_{ortho} = \sum_{*\in\{traj,pose\}} ||\bar{\mathbf{M}}_* \bar{\mathbf{M}}_*^\top - \mathbf{I}||_2^2 \tag{2}$$

where $\bar{\mathbf{M}}_*$ denotes row-wise L_2-normalized variant of each memory bank matrix. Unlike standard vector-quantized codebooks, which are optimized for reconstruction fidelity, our prototype bank is optimized for behavioral diversity through the orthogonality constraint in Equation 2. This prevents the slot collapse that would otherwise reduce the discrete primitives to redundant modes, which is essential when the primitives serve as generative anchors rather than reconstruction targets.

3.4 Asymmetric Cross-Attention Modality Fusion

Bio-TAD employs an asymmetric cross-attention gating method to include fundamental physiological priors into the spatial navigation framework. The quantized trajectory representation $\hat{\mathbf{H}}_{traj}$ functions as the query vector, effectively extracting detailed structural information from the keys and values of the memory-enhanced postural tensor $\hat{\mathbf{H}}_{pose}$. We use the trajectory stream as the query because the macroscopic trajectory is the primary prediction target. This formulation allows us to selectively retrieve relevant biomechanical cues and reduce the risk of diluting the spatial signal with high-frequency postural noise.

We utilize a learnt, trajectory-dependent sigmoid fusion gate α to dynamically regulate the infusion of body language vectors in response to structural uncertainty. The fused spatio-temporal tensor $\mathbf{H}_{fused}$ is derived as:

$$\begin{aligned} \mathbf{H}_{fused} = (1-\alpha) \odot \hat{\mathbf{H}}_{traj} \\ + \alpha \odot \text{CA}(\hat{\mathbf{H}}_{traj}, \hat{\mathbf{H}}_{pose}, \hat{\mathbf{H}}_{pose}) \end{aligned} \tag{3}$$

where CA denotes Cross-Attention and $\alpha = \sigma(\text{Linear}(\hat{\mathbf{H}}_{traj}))$.

3.5 Attentive Condition Pooling

Prior to conditioning the downstream denoiser, length-varying temporal states are aggregated into a static-dimensional description using an Attentive Condition Pool. The pooling layer derives three uniform vectors from $\mathbf{H}_{fused}$: the global temporal mean, the final observed state, and an attention-weighted feature score. The three vectors are concatenated and processed through a multilayer perceptron with SiLU activation to get the compressed global context vector. $\mathbf{c}_{fused}$:

$$\mathbf{c}_{fused} = \text{MLP}(\text{SiLU}([\mathbf{h}_{mean} \parallel \mathbf{h}_{last} \parallel \mathbf{h}_{attn}])) \tag{4}$$

This architecture preserves both low-frequency momentum and high-frequency pre-turn biomechanical cues.

3.6 Target-Oriented AdaLN Diffusion Engine

The primary route generating pipeline is regulated by a Diffusion Transformer (DiT) engine (Peebles & Xie, 2023). In the forward diffusion process, parametric Gaussian noise is consistently introduced into the ground-truth trajectory $\mathbf{y}_0$ at step t to produce the latent future state $\mathbf{y}_t$.

While standard DiT architectures heavily rely on Adaptive Layer Normalization (AdaLN) for global class conditioning, Bio-TAD uniquely repurposes this mechanism for dynamic waypoint anchoring. Furthermore, unlike traditional goal-conditioned frameworks (Mangalam et al., 2020; Zhao et al., 2021; Gu et al., 2021; Mangalam et al., 2021) that rely on a separate sub-network to statically predict spatial endpoints upfront, Bio-TAD dynamically extracts the evolving terminal destination waypoint $\mathbf{w}_t = \mathbf{y}_t[-1, :]$ directly from the latent state after each backward denoising iteration. The terminal anchor is linearly integrated and combined with the step embedding $\phi_t(t)$ and the multi-modal context $\mathbf{c}_{fused}$ to create a dense, target-aware conditioning vector: $\mathbf{c}_{total} = \phi_t(t) + \phi_c(\mathbf{c}_{fused}) + \phi_w(\mathbf{w}_t)$. Here ϕ_t, ϕ_c, and ϕ_w are learned multilayer perceptrons (MLPs) with SiLU activations that map the timestep embedding, the fused multimodal context, and the terminal waypoint into a shared conditioning dimension.

3.7 MULTI-TASK DIFFUSION OBJECTIVE

The complete Bio-TAD network is optimized using a composite multi-task objective. While the individual loss components rely on standard, well-established mathematical formulations, their concurrent application is carefully structured to simultaneously enforce spatial diffusion precision, memory bank orthogonality, and explicit target convergence. The diffusion loss, $\mathcal{L}_{diff}$, is a composite loss that reduces noise estimation error $\mathcal{L}_\epsilon = ||\epsilon - \epsilon_\theta||_2^2$, in conjunction with real-space coordinate adjustments ($\mathcal{L}_{x_0}$) and structural drift evaluations ($\mathcal{L}_{ADE}$). It is formulated as:

$$\mathcal{L}_{diff} = \mathcal{L}_\epsilon + \lambda_{x_0}\mathcal{L}_{x_0} + \lambda_{ADE}\mathcal{L}_{ADE} \tag{5}$$

To strictly enhance target-oriented trajectory convergence, we use a specific Smooth L_1 endpoint penalty that solely penalizes the variance of the destination estimate relative to the ground truth:

$$\mathcal{L}_{waypoint} = \text{SmoothL1}(\hat{\mathbf{y}}_0[-1,:], \mathbf{y}_{GT}[-1,:]) \tag{6}$$

The total joint optimization balance is mathematically formalized as:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \lambda_{ortho}\mathcal{L}_{ortho} + \lambda_{waypoint}\mathcal{L}_{waypoint} \tag{7}$$

where λ_{x_0}, λ_{ADE}, λ_{ortho}, and $\lambda_{waypoint}$ are empirically tuned hyper parameters.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We assess our framework using three public datasets that provide extensive visual and biomechanical data: the JTA Dataset(Fabbri et al., 2018), the JRDB Dataset(Vendrow et al., 2023), and the Urban Dataset(Kress et al., 2022). To conduct a thorough evaluation of Bio-TAD's backbone-agnostic capabilities, we benchmark against a varied array of methodological paradigms, including Social-LSTM (Alahi et al., 2016), Social-GAN (Gupta et al., 2018), AutoBots (Girgis et al., 2021), Social-Transmotion (Saadatnejad et al., 2024), VLMTraj (Bae et al., 2025), and SingularTrajectory (Bae et al., 2024). The complete dataset splits, sequence processing rates, and baseline architectural paradigms are elaborated in Appendix B and Appendix C. Appendix D delineates comprehensive architectural requirements, temporal diffusion settings, and multi-task optimization weights.

4.2 EVALUATION METRICS

Standard Geometric Metrics: We adopt the standard Average Displacement Error ($minADE_{20}$) and Final Displacement Error ($minFDE_{20}$) to evaluate geometric prediction accuracy. For the JRDB dataset, the evaluation is performed deterministically ($K = 1$); therefore, standard ADE and FDE are reported.

Novel Target/Semantic Metric:

- **Intent Rate (IR):** Assesses behavioral congruence. A deterministic classifier $\Phi(\cdot)$ assigns a trajectory to one of eight distinct intent categories, outputting a class label c. The assigned class $c = \Phi(y)$ is determined by the previous and subsequent Euclidean displacements (d_{past}, d_{future}) in relation to a movement threshold of $\tau = 0.2$ m, and the relative heading alteration $\theta \in [-\pi, \pi]$:
  - **Idle / Stopping / Starting:** Governed by displacement transitions: Idle ($d_{past} \leq \tau \wedge d_{future} \leq \tau$), Stopping ($d_{past} > \tau \wedge d_{future} \leq \tau$), and Starting ($d_{past} \leq \tau \wedge d_{future} > \tau$).
  - **Straight / Turns:** Active motion ($d_{past}, d_{future} > \tau$) partitioned by angular deviation: Straight ($|\theta| \leq 15°$), Left/Right Turn ($15° < |\theta| \leq 50°$), and Far Left/Right Turn ($|\theta| > 50°$), where the sign of θ determines direction.

The IR quantifies the ratio of the K predicted trajectories that accurately correspond to the ground-truth semantic class c_{GT}:

$$IR = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(\Phi(\hat{y}_k) = c_{GT}) \tag{8}$$

Note that IR is a precision-style measure over the K samples; it rewards each predicted mode for being a behaviorally valid future, not for collapsing all samples onto a single class. It is therefore complementary to the diversity we encourage in Appendix A.1. A strong model produces a spread of distinct, plausible behavioral modes, of which the ground-truth class is one, rather than a spread of geometrically scattered noise. High IR and healthy sample diversity are thus compatible: the former penalizes samples that fall outside any coherent intent, while the latter penalizes deterministic collapse.

Table 1: Intent accuracy at T=12 frames on JTA dataset (Trajectory (T) vs T + 3D Pose (T+P))

Model	**Intent Class**	**T**	**T+P**
AutoBots (Girgis et al., 2021)	Left Turn	31.54%	**42.18%**
	Right Turn	23.66%	**31.08%**
	Far Left Turn	29.12%	**53.12%**
	Far Right Turn	36.13%	**54.97%**
Social-Transmotion (Saadatnejad et al., 2024)	Left Turn	28.53%	**43.48%**
	Right Turn	22.85%	**38.57%**
	Far Left Turn	34.93%	**54.60%**
	Far Right Turn	42.52%	**50.34%**
SingularTrajectory (Bae et al., 2024)	Left Turn	34.19%	**46.34%**
	Right Turn	28.45%	**36.29%**
	Far Left Turn	36.43%	**56.17%**
	Far Right Turn	40.71%	**52.48%**
VLMTraj (Bae et al., 2025)	Left Turn	35.46%	**47.53%**
	Right Turn	29.58%	**37.45%**
	Far Left Turn	37.55%	**57.34%**
	Far Right Turn	41.86%	**53.62%**

4.3 Quantitative Evaluation

This section presents our experimental findings, assessing the Bio-TAD framework across three benchmarks. Table 3 details the spatial displacement errors, whereas Tables 10, 11, and 12 from A.7 present the discrete intent accuracy over different prediction horizons for the JTA, JRDB, and Urban datasets respectively. Table 9 from A.5 additionally ablates input-modality configurations, showing that progressively richer visual prompts monotonically improve accuracy. Table 1 presents intent accuracy at T = 12 on the JTA dataset for recent baselines; the supplementary material reports the full results across all horizons and all baselines.

JTA Dataset The JTA dataset functions as our principal testing ground for dense multi-agent scenarios with high-fidelity 3D biomechanical keypoints. The incorporation of Bio-TAD results in significant decreases in spatial inaccuracy across all foundational paradigms. In particular, when combined with the multimodal LLM baseline (VLMTraj (Bae et al., 2025)), $minADE_{20}$ and $minFDE_{20}$ decrease from 0.59/1.36 to 0.48/1.00. Table 1 and Table 10 (an extended version of Table 1) jointly illustrate that the Bio-TAD-augmented VLMTraj model improved spatial performance and improved the correct assessment of pedestrian intent across all temporal horizons.

This semantic alignment is essential since conventional coordinate-only models inherently fail to predict highly stochastic directional changes. By integrating physiological inputs, our approach establishes unambiguous restrictions on the generative space. The Bio-TAD-augmented AutoBots model demonstrates a $+24.00\%$ absolute increase in Intent Rate for "Far Left Turns" and $+18.84\%$

for "Far Right Turns" at the $T = 12$ horizon relative to its unaugmented baseline. This improvement indicates that, from the $K = 20$ produced multimodal futures, a much greater fraction of the projected trajectories now precisely align with the actual semantic behavior of a sharp turn. By recording pre-movement biomechanical torsion, such as pelvic or shoulder rotation, before significant spatial displacement, Bio-TAD minimizes unstructured statistical noise and facilitates early, accurate target convergence. The continual enhancements in discrete intent accuracy affirm a fundamental architectural necessity. They empirically establish that conceptualizing biomechanical posture as an active generative anchor, rather than simply appending it as a continuous auxiliary feature, is crucial for effectively addressing high-uncertainty motion patterns and mitigating the generation of unstructured statistical noise.

JRDB Dataset: To assess the framework's adaptability to constrained real-world modalities, we analyzed Bio-TAD on the JRDB dataset utilizing just 2D postures. Bio-TAD consistently enhances geometric performance across all baseline models, systematically reducing $minFDE_1$ for all architectures, from Social-LSTM (0.95m to 0.87m) to AutoBots (0.80m to 0.71m). Table 11 from A.7, illustrates that the semantic behavior provides essential architectural insights. Across all paradigms, Intent Rate improvements for sharp turns are roughly halved compared to the 3D JTA dataset, quantifying the effect of Z-axis depth ambiguity. Projecting three-dimensional human kinematics into a two-dimensional pixel space obscures essential torsional dynamics, compelling the diffusion denoiser to function with reduced-fidelity microscopic priors. Nevertheless, the uniform positive improvement across all baseline architectures demonstrates that even simplified 2D limb articulations offer adequate physiological support for the Multi-Task Diffusion Objective to surpass just coordinate-driven baselines.

Urban Road Traffic: The Urban dataset evaluates the framework's efficacy for high-frequency (25 fps) real-world trajectories. Table 12 from A.7, illustrates that, at this high temporal resolution, a conventional macroscopic trajectory array accurately reflects instantaneous momentum changes. This leads to increased base Intent Rates for unaltered models, exemplified by baseline AutoBots forecasting "Far Left Turns" with an accuracy of 44.34% at T=63. As a result, the geometric margins have become more stringent, with AutoBots improving from 0.38/0.68 to 0.35/0.63 in $minADE_{20}/minFDE_{20}$. Even with this temporal coupling, the incorporation of the Bio-TAD microscopic processing stream consistently enhances semantic precision over the whole range of baseline models. The incorporation of explicit biomechanical priors leads to consistent improvements in Intent Rate, irrespective of the underlying architecture, prediction horizon, or specific intent class, which may vary from linear navigation to intricate, sudden directional changes. Especially for complex intent categories with substantial directional deviations, Bio-TAD attains an average absolute improvement of approximately $+11.56\%$ for far-left turns and $+12.11\%$ for far-right turns at extended prediction horizons, highlighting its efficacy in addressing high-uncertainty motion patterns. This verifies that even when large-scale historical routing is remarkably seamless and concentrated, isolating the biomechanical posture explicitly enhances the target-oriented AdaLN engine with superior semantic priors, ensuring resilient multi-future forecasting across varied pedestrian kinematics.

Table 2: Deterministic evaluation ($k = 1$) on the JTA dataset across multi-modal backbones.

Model Backbone	**ADE**	**FDE**
AutoBots (Girgis et al., 2021)	1.20	2.70
AutoBots + Social-Pose (Gao et al., 2025)	0.90	1.91
AutoBots + Bio-TAD (Ours)	**0.84**	**1.68**
Social-Transmotion(Saadatnejad et al., 2024) + Social-Pose (Gao et al., 2025)	0.94	1.94
Social-Transmotion + Bio-TAD (Ours)	**0.88**	**1.71**

We perform a comparative study of Bio-TAD against contemporary pose-augmented architectures utilizing a deterministic inference approach ($k = 1$) on the JTA dataset with multiple backbones. As shown in Table 2, the Social-Pose framework (Gao et al., 2025) attains an ADE/FDE of 0.90/1.91m when its continuous concatenation encoder is integrated into an AutoBots backbone, while the native pose-fused Social-Transmotion architecture achieves 0.94/1.94m. Conversely, our Bio-TAD architecture consistently exhibits enhanced performance relative to these configurations across both backbone families. AutoBots integrated with Bio-TAD attains an ADE/FDE of 0.84/1.68m, whereas Social-Transmotion with Bio-TAD obtains 0.88/1.71m, reflecting a decrease of around 12% in FDE

and 7% in ADE. This consistent geometric advantage illustrates that purposefully guiding the generative diffusion process via a target-focused engine and distinct behavioral primitives is beneficial across several baseline formulations.

Table 3: Quantitative results on the three datasets with ground truth pose (3D pose information for JTA and Urban, 2D pose information for JRDB). Metrics are reported in meters

Model	**Input Modality**	**JTA** $\mathbf{minADE_{20}/minFDE_{20}}$	**JRDB** $\mathbf{minADE_1/minFDE_1}$	**Urban** $\mathbf{minADE_{20}/minFDE_{20}}$
Social-LSTM	Traj	0.58/1.25	0.47/0.95	0.38/0.69
Social-LSTM + Bio-TAD (ours)	Traj + Pose Information	**0.51/1.05**	**0.43/0.87**	**0.36/0.65**
Social-GAN	Traj	0.68/1.51	0.50/0.99	0.39/0.70
Social-GAN + Bio-TAD (ours)	Traj + Pose Information	**0.60/1.22**	**0.45/0.91**	**0.37/0.66**
AutoBots	Traj	0.72/1.50	0.39/0.80	0.38/0.68
AutoBots + Bio-TAD (ours)	Traj + Pose Information	**0.56/1.13**	**0.34/0.71**	**0.35/0.63**
Social-Transmotion	Traj	0.65/1.49	0.40/0.77	0.38/0.68
Social-Transmotion	Traj + Pose	-/-	0.37/0.73	-/-
Social-Transmotion + Bio-TAD (ours)	Traj + Pose Information	**0.49/1.08**	**0.35/0.69**	**0.35/0.62**
VLMTraj	Traj	0.59/1.36	0.38/0.76	0.37/0.67
VLMTraj + Bio-TAD (ours)	Traj + Pose Information	**0.48/1.00**	**0.33/0.67**	**0.34/0.62**
Singular Trajectory	Traj	0.55/1.29	0.38/0.75	0.36/0.65
Singular Trajectory + Bio-TAD (ours)	Traj + Pose Information	**0.51/1.04**	**0.34/0.68**	**0.34/0.61**

4.4 Ablation Studies

To further validate our architectural design choices and isolate the specific drivers of performance, we provide comprehensive ablation studies in the Appendix. Specifically, Appendix A.1 analyzes the sensitivity of the Orthogonal Prototype Memory Bank capacity, empirically demonstrating that discrete behavioral primitives outperform continuous feature concatenation. Appendix A.2 dissects the Multi-Task Diffusion Objective, confirming the complementary necessity of both orthogonality and waypoint constraints for semantic alignment and destination convergence. Furthermore, Appendix A.3 utilizes a capacity-matched control experiment to verify that our predictive gains stem from structural design rather than mere parameter inflation. Finally, Appendix A.4 and Appendix A.5 systematically isolate the contributions of core pipeline components (such as the dual-stream encoder and AdaLN engine) and cross-modal input configurations to confirm the necessity of explicit biomechanical constraints. Appendix A.6 presents a targeted ablation that isolates the pose integration mechanism, confirming that discretizing posture into an orthogonal memory bank significantly outperforms standard continuous feature concatenation.

5 Conclusion

We introduced Bio-TAD, an advanced backbone-agnostic generative system for multimodal pedestrian trajectory prediction. Our method transforms detailed biomechanical posture signals into distinct behavioral primitives, constraining the stochastic generating space to precisely model motion patterns during abrupt movements. We also presented a new evaluation metric, Intent Rate, that assesses semantic behavioral alignment beyond purely geometric displacement, which we anticipate will promote more rigorous benchmarking within the trajectory forecasting community. Despite these advancements, our framework possesses specific limitations that require further investigation. Initially, Bio-TAD achieves inference in 82 ms, adequately supporting standard 2.5 fps tracking pipelines. However, scaling the dual-stream diffusion architecture for ultra-high-frequency environments (e.g., 25 fps) will necessitate additional architectural pruning. Furthermore, the gains in semantic intent are partially affected by the deterministic geometric thresholds applied in our evaluation. Expanding our benchmarks to encompass a wider range of unstructured environments and concurrent baselines is an area for future research.

ETHICS STATEMENT

This study does not engage human participants, identifiable personal information, or confidential private data. The datasets employed in this study (JTA, JRDB, and Urban) are publicly accessible benchmarks designated solely for academic research. We anticipate no imminent adverse social effects, conflicts of interest, or equity issues stemming directly from the methodological innovations introduced in our trajectory prediction framework.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our Bio-TAD framework, we have provided a comprehensive account of the model architectures, dataset splits, evaluation metrics, and multi-task optimization weights in Section 4 and Appendix D. Additionally, thorough ablation studies are presented to confirm our design decisions. The complete source code, trained models, and necessary scripts for replicating the multi-task diffusion objective will be made publicly available following the publication of this work.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks that require disclosure (such as generating synthetic data sets, developing theoretical models, implementing methods, or interpreting results), and all required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, specifically to edit the research paper to improve readability. We have reviewed all AI-assisted work; specifically, the authors critically read, edited, and verified all LLM-refined text to ensure that scientific integrity, accuracy, and the original intent of the research were strictly preserved. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17890–17901, 2024.

Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Social reasoning-aware trajectory prediction via multimodal language model. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

S Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. *arXiv preprint arXiv:2010.10270*, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 86–99. PMLR, 30 Oct–01 Nov 2020. URL https://proceedings.mlr.press/v100/chai20a.html.

Pranav Singh Chib and Pravendra Singh. Lg-traj: Llm guided pedestrian trajectory prediction. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6861–6871. IEEE, 2025.

Wanqing Dou and Lili Lu. Sisgan: A generative adversarial network pedestrian trajectory prediction model combining interaction information and scene information. *Applied Sciences*, 14(20):9537, 2024.

Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 430–446, 2018.

Yang Gao, Saeed Saadatnejad, and Alexandre Alahi. Social-pose: Enhancing trajectory prediction with human body pose. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint arXiv:2104.00563*, 2021.

Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15283–15292. IEEE, 2021.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2255–2264, 2018.

Irtiza Hasan et al. Forecasting people trajectories and head poses by jointly reasoning on tracklets and vislets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1267–1278, 2021.

Jaewoo Jeong, Seohee Lee, Daehee Park, Giwon Lee, and Kuk-Jin Yoon. Multi-modal knowledge distillation-based human trajectory forecasting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24222–24233. IEEE, 2025.

Viktor Kress, Fabian Jeske, Stefan Zernetsch, Konrad Doll, and Bernhard Sick. Pose and semantic map based probabilistic forecast of vulnerable road users' trajectories. *IEEE Transactions on Intelligent Vehicles*, 8(3):2592–2603, 2022.

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5725–5734, 2019.

Yu Liu, Zhijie Liu, Xiao Ren, Youfu Li, and He Kong. Intention-aware diffusion model for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18469–18477, 2026.

Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*, pp. 759–776. Springer, 2020.

Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 15213–15222. IEEE, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Saeed Saadatnejad, Y Ju, and Alexandre Alahi. Pedestrian 3d bounding box prediction. *arXiv preprint arXiv:2206.14195*, 2022.

Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *International conference on learning representations*, volume 2024, pp. 38299–38316, 2024.

Hoseong Suk and Seonghun Kim. Uncertainty-aware multimodal trajectory prediction via a single inference from a single model. *Sensors*, 25(1):217, 2025.

Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, and Norimichi Ukita. Physical plausibility-aware trajectory prediction via locomotion embodiment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12324–12334. IEEE, 2025.

Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4811–4820, 2023.

Wei Xiang, YIN Haoteng, He Wang, and Xiaogang Jin. Socialcvae: Predicting pedestrian trajectory via interaction conditioned latents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6216–6224, 2024.

Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7593–7602, 2018.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M. Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9793–9803, 2021.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on robot learning*, pp. 895–904. PMLR, 2021.

A Ablation Study and Discussion

A.1 Sensitivity Analysis on memory slots

We conduct a sensitivity analysis on the number of discrete memory slots to validate the efficacy of our Orthogonal Prototype Memory Bank. This analysis utilizes the Social-Transmotion baseline integrated with our Bio-TAD framework and is evaluated on the JTA dataset. The memory bank's capacity dictates a crucial trade-off in generative trajectory forecasting: it balances precise semantic alignment with the multimodal diversity of predicted futures. An excessively small memory bank risks deterministic mode collapse, resulting in predictions that follow rigid, repetitive paths. Conversely, an excessively large bank may lead to significant over-fragmentation, causing unstructured spatial dispersion that disregards biomechanical constraints. To quantify this trade-off, we assess the Intent Rate (IR), focusing on the average IR for the most significant spatial deviations, "Far Left Turn" and "Far Right Turn," at the maximum prediction horizon of $T = 12$, in conjunction with Shannon Entropy. In this context, Shannon Entropy serves as a precise measure of distributional diversity among the generated samples. A low entropy score suggests that the network is consistently converging on a single deterministic intent, whereas a higher entropy score indicates a robust, probabilistically diverse exploration of various distinct future actions.

The quantitative results presented in Table 4 illustrate this trade-off and support our architectural configuration of 256 learnable behavioral slots. When memory capacity is limited to 64 slots, the model demonstrates mode collapse, resulting in a low Shannon Entropy of 0.65, indicative of an artificial reduction in generative diversity. This over-constraint subsequently restricts the average Intent Rate for sharp turns to 46.1%. Conversely, increasing the bank to 512 slots leads to excessive fragmentation of the latent intent space. Although this approach maximizes generative diversity (Entropy of 1.78), it results in the indiscriminate scattering of generated samples across improbable paths, leading to a degradation of the Intent Rate to 44.8%. The optimal equilibrium is attained at 256 slots, where the memory bank offers adequate capacity to accurately represent nuanced biomechanical body language without devolving into statistical noise. This configuration achieves the highest Intent Rate (52.47%) for complex turning maneuvers while maintaining a probabilistically accurate multimodal spread (Entropy of 1.35). Crucially, removing the memory bank entirely and reverting to continuous concatenation (0 slots) yields an Intent Rate of only 41.2%, well below the most constrained discrete configuration (64 slots, 46.1%) at an entropy of 1.82. This isolates discretization itself, rather than mere capacity tuning, as the source of Bio-TAD's semantic gains. Importantly, continuous features leave the backbone's unconstrained latent space to resolve the pose-to-future mapping, whereas even a coarse discrete prototype bank imposes structure that the diffusion denoiser can anchor to. This directly substantiates our central claim that biomechanical pose benefits trajectory diffusion only when organized into discrete behavioral primitives.

Table 4: Sensitivity analysis of the Orthogonal Prototype Memory Bank capacity on the Social-Transmotion baseline.

Memory Slots	Avg. Intent Rate (IR) (↑)	Shannon Entropy
0	41.2%	1.82
64	46.1%	0.65
128	50.2%	1.12
256	**52.47%**	1.35
512	44.8%	1.78

Table 5: Ablation study of the Multi-Task Diffusion Objective components on the JTA dataset using the AutoBots baseline.

Configuration	minADE_{20} / minFDE_{20}	**Avg. Sharp Turn IR**
Full Bio-TAD	**0.56 / 1.13**	**54.05%**
w/o Orthogonal Loss	0.61 / 1.27	48.30%
w/o Waypoint Loss	0.64 / 1.38	46.80%
w/o Both	0.68 / 1.45	38.50%

A.2 Impact of Multi-Task Optimization Objectives

We perform an ablation study on the JTA dataset using the AutoBots baseline to systematically isolate the contributions of our specialized Multi-Task Diffusion Objective. We assess the effects of eliminating the Orthogonal Loss, which promotes behavioral diversity within the memory bank, and the Waypoint loss, which directly ties the generative process to expected outcomes. Performance is assessed through standard spatial displacement errors and the average Intent Rate (IR) for sharp directional shifts at the maximum prediction horizon. Table 5 demonstrates that the complete Bio-TAD framework attains optimal performance, successfully constraining the generative space to produce an ADE/FDE of 0.56/1.13m and a sharp-turn IR of 54.05%. Ablating the Orthogonal Loss yields a notable IR reduction (54.05% $\rightarrow$ 48.30%) alongside increased displacement errors (0.61/1.27m). Removing only the Waypoint loss results in even higher spatial errors (0.64/1.38m) and drops the IR to 46.80%. Ultimately, removing both losses drives the sharpest degradation (ADE/FDE $\rightarrow$ 0.68/1.45m, IR $\rightarrow$ 38.50%), confirming the two constraints are complementary: orthogonality preserves distinct behavioral primitives while the waypoint penalty secures destination convergence.

A.3 Capacity-Matched Baseline Analysis

To validate that the predictive enhancements of Bio-TAD arise from our proposed architectural formulations rather than merely an increase in model capacity, we conduct a capacity-matched control experiment. The incorporation of the Bio-TAD framework adds approximately 11.37 million additional parameters to the base architecture. To isolate the impact of parameter volume, we create capacity-matched variants of the Social-LSTM and AutoBots baselines. Specifically, for the RNN-based Social-LSTM, capacity was matched by increasing the hidden state dimensions (up to $d = 1024$) and stacking additional recurrent layers. For the Transformer-based AutoBots, we scaled the architecture by deepening the network with additional attention blocks and widening the multi-head attention and feed-forward dimensions ($d_{model} = 512, d_{ff} = 2048$) until the target parameter budget was reached.

We assess three different configurations for each backbone on the JTA dataset: the original baseline (utilizing only trajectory coordinates), the capacity-matched baseline (employing only trajectory coordinates but scaled by +11.37M parameters), and our proposed model (the baseline augmented with the complete Bio-TAD framework).

The quantitative results presented in Table 6 demonstrate that simply scaling the raw capacity of coordinate-only baselines results in only marginal geometric improvements. Increasing the capacity of the AutoBots model results in a slight reduction of the minFDE_{20} from 1.50m to 1.42m, while the Far Left Turn Intent Rate shows a marginal improvement of +2.68%.

Table 6: Capacity-Matched ablation on the JTA Dataset ($T = 12$). Spatial metrics (minADE$_{20}$/minFDE$_{20}$) are reported in meters (lower is better). Intent Rate (IR) is reported as a percentage (higher is better).

Model Variant	Params	minADE$_{20}$ ↓	minFDE$_{20}$ ↓	Left Turn IR ↑	Right Turn IR ↑	Far Left Turn IR ↑	Far Right Turn IR ↑
Social-LSTM (Base)	∼0.3M	0.58	1.25	28.10%	22.15%	26.40%	31.80%
Social-LSTM (Capacity-Matched)	∼11.71M	0.56	1.21	29.80%	23.40%	28.10%	33.50%
Social-LSTM + Bio-TAD (Ours)	∼11.67M	**0.51**	**1.05**	**39.40%**	**30.20%**	**48.70%**	**51.20%**
AutoBots (Base)	∼1.2M	0.72	1.50	31.54%	23.66%	29.12%	36.13%
AutoBots (Capacity-Matched)	∼12.26M	0.68	1.42	33.20%	25.10%	31.80%	38.40%
AutoBots + Bio-TAD (Ours)	∼12.57M	**0.56**	**1.13**	**42.18%**	**31.08%**	**53.12%**	**54.97%**

Table 7: Component-level ablation of the Bio-TAD architecture on the JTA Dataset (AutoBots Baseline, $T = 12$). Components were systematically ablated during the input processing and output generation phases.

Model Variant	minADE$_{20}$ ↓	minFDE$_{20}$ ↓	Far Left Turn IR ↑	Far Right Turn IR ↑
AutoBots (Baseline)	0.72	1.50	29.12%	36.13%
AutoBots + Bio-TAD (Ours)	**0.56**	**1.13**	**53.12%**	**54.97%**
Input Phase Ablations				
w/o Dual-Stream (Early Fusion)	0.63	1.28	41.50%	43.20%
w/o Asym. Gate α (Symmetric Cross-Attention)	0.60	1.21	44.80%	46.50%
Output Phase Ablations				
w/o AdaLN Engine (Plain Concatenation)	0.65	1.35	39.20%	40.10%
w/o DiT Decoder (MLP Decoder)	0.67	1.40	37.50%	39.00%

Integrating Bio-TAD within the same parameter budget results in significant performance improvements. The Bio-TAD augmented AutoBots framework reduces the minFDE$_{20}$ to 1.13m and raises the Far Left Turn Intent Rate to 53.12%, indicating an absolute improvement of +24.00% compared to the base model and +21.32% relative to the capacity-matched control. This significant divergence confirms that the predictive superiority of Bio-TAD is not a result of parameter inflation. It is the structural imposition of discrete behavioral primitives and the active anchoring of the generative diffusion space that effectively address semantic misalignments during highly stochastic maneuvers.

A.4 Architectural Component Ablations

We conduct a systematic component-level ablation to assess the individual contributions of our core design choices and to isolate the specific drivers of performance within the added parameter budget. We utilize the AutoBots baseline enhanced with Bio-TAD on the JTA dataset ($T = 12$) to conduct an ablation study of four specific structural components during the input and output phases of our pipeline. The quantitative results are presented in Table 7.

Encoding and Fusion Our framework concurrently processes macroscopic trajectory data and microscopic biomechanical data. To validate this separation, we substitute the dual-stream encoder with a single-stream variant that concatenates the spatial and postural inputs before encoding. As indicated in Table 7, the dilution of the biomechanical signal results in a notable decline in intent alignment, with the Far Left Turn IR decreasing from 53.12% to 41.50%. Additionally, we replace the learned, trajectory-dependent sigmoid gate with a standard symmetric cross-attention mechanism, thereby abating the asymmetric cross-attention gate (α). The observed performance decline (minFDE$_{20}$ rises from 1.13m to 1.21m) indicates that it is essential for the framework to selectively retrieve biomechanical cues in response to macroscopic trajectory uncertainty. This structural necessity helps mitigate the impact of high-frequency postural noise on the primary spatial routing.

Conditioning and Generation To validate our generative pipeline, we initially ablate the target-oriented Adaptive Layer Normalization (AdaLN) engine. Removing the dynamic terminal waypoint extraction and reverting to standard global feature concatenation leads to a loss of destination anchoring in the model, causing significant spatial drift (minFDE$_{20}$ increases to 1.35m) and a notable decrease in accurate turn predictions. We ultimately substitute the 4-layer Diffusion Transformer (DiT) denoiser with a conventional deterministic Multi-Layer Perceptron (MLP) decoder. This substitution results in significant degradation across all metrics, with a minimum final displacement error (minFDE$_{20}$) of 1.40m and a Far Left Turn IR of 37.50%. This conclusively demonstrates that

the diffusion process itself, rather than solely the encoded representations, is essential for effectively addressing the multimodal uncertainty present in pedestrian motion.

A.5 Impact of Input Modality Configurations

To systematically investigate the individual and cumulative contributions of heterogeneous visual prompts, we conduct an ablation study on the JTA dataset across two representative transformer-based baselines: AutoBots (Girgis et al., 2021) and Social-Transmotion (Saadatnejad et al., 2024). We evaluate the models under four cross-modal input configurations: (1) purely pixel-space human posture (Traj+2D Pose), (2) pixel-space posture augmented with structural layout (Traj+2D Pose+ 2D BB), (3) the addition of spatial volumetric frames (Traj + 2D Pose + 2D BB + 3D BB), and (4) the combination of spatial layout constraints with high-fidelity physical posture (Traj + 2D BB + 3D BB + 3D Pose). Performance is reported in terms of $minADE_{20}$ and $minFDE_{20}$ to isolate how well each representation helps constrain the unconstrained generative diffusion workspace.

The empirical results compiled in Table 9 reveal distinct architectural trends regarding cross-modality representation alignment. Evaluating the initial configuration (Traj + 2D Pose) shows a solid baseline, but introducing localized bounding box coordinates (Traj + 2D Pose + 2D BB) triggers an immediate performance gain, lowering Social-Transmotion's spatial errors from 0.61/1.24m to 0.58/1.19m. This behavior demonstrates that pixel-space bounding boxes provide critical macroscopic spatial anchoring that complements the localized limb details of a 2D pose. Most notably, shifting from configuration (3) to configuration (4), where low-fidelity 2D postures are substituted with 3D kinematic keypoints alongside multi-dimensional layout boxes, yields the most substantial accuracy optimization across both baselines, bringing AutoBots down to 0.56/1.13m and Social-Transmotion to a state-of-the-art result of 0.49/1.08m. This final improvement confirms our primary hypothesis: while spatial bounding boxes provide essential path boundaries, full 3D body language offers the definitive spatiotemporal constraints needed to steer the target-oriented AdaLN diffusion engine toward accurate future trajectories.

A.6 Ablation on Pose Integration Mechanisms

To validate the structural design of the Orthogonal Prototype Memory Bank, we conducted a targeted ablation to isolate the pose integration mechanism. We assess whether the predictive advantages of Bio-TAD arise from the inclusion of 3D pose data or from our particular approach to its discretization. We utilized the AutoBots baseline on the JTA dataset ($T = 12$) and evaluated three configurations:

1. **Continuous Concatenation:** The 3D pose is embedded and directly concatenated with the trajectory features, completely bypassing the memory bank.
2. **Prototype Memory w/o $\mathcal{L}_{ortho}$:** The continuous pose is routed through the memory bank to form discrete behavioral primitives, but the orthogonality constraint is removed.
3. **Full Bio-TAD:** The complete memory bank with the orthogonality constraint ($\mathcal{L}_{ortho}$) active.

Table 8: Ablation on the Mechanism of Biomechanical Integration (AutoBots Baseline, JTA Dataset, $T = 12$). This isolates the impact of discrete memory versus continuous concatenation, and the necessity of the orthogonal constraint.

Integration Mechanism	minADE$_{20}$	minFDE$_{20}$	Far Left Turn IR	Far Right Turn IR
AutoBots (Coordinate-Only Baseline)	0.72	1.50	29.12%	36.13%
+ Continuous Concatenation (Traj + 3D Pose)	0.66	1.38	39.40%	41.20%
+ Prototype Memory w/o $\mathcal{L}_{ortho}$	0.61	1.27	48.30%	49.80%
+ Full Bio-TAD (Memory + $\mathcal{L}_{ortho}$)	**0.56**	**1.13**	**53.12%**	**54.97%**

The quantitative results in Table 8 empirically identify discretization as the main factor influencing semantic alignment. Appending 3D pose through continuous feature concatenation results in only moderate enhancements compared to the coordinate-only baseline. This arises from the challenges faced by the unconstrained continuous latent space in reliably mapping high-frequency kinematic

cues to multi-modal futures. Transitioning to a discrete Prototype Memory Bank, in the absence of the orthogonality constraint, enhances spatial error and intent alignment by compelling the network to adhere to discrete behavioral representations. However, in the absence of $\mathcal{L}_{ortho}$, the memory slots are prone to partial mode collapse, which restricts the diversity of the generated intents. The complete Bio-TAD framework, which enforces orthogonal slot diversity, attains the lowest displacement errors (0.56/1.13 m) and the highest Intent Rates, demonstrating that biomechanical signals are most effective for diffusion models when converted into distinct, regularized behavioral primitives.

A.7 Comprehensive Intent Accuracy Evaluations

This subsection provides the extended semantic intent evaluations across all comparative baselines and experimental settings. Specifically, Tables 10, 11, and 12 detail the intent alignment accuracy over varying prediction horizons for the JTA, JRDB, and Urban datasets, respectively.

B Datasets

To validate our proposed methodology, we assess it using three publicly accessible datasets that provide rich visual and biomechanical information.

- **JTA Dataset (Fabbri et al., 2018):** A synthetic dataset consisting of 256 training sequences, 128 validation sequences, and 128 test sequences, containing approximately 10 million 3D keypoint annotations. We predict the next 12 time-steps using the prior 9 time-steps at a rate of 2.5 frames per second (fps).
- **JRDB Dataset (Vendrow et al., 2023):** A comprehensive dataset featuring a variety of pedestrian trajectories, accompanied by 2D bounding boxes and postures, encompassing both indoor and outdoor environments. We forecast 12 subsequent time-steps based on 9 preceding time-steps at a rate of 2.5 frames per second.
- **Urban Dataset (Kress et al., 2022):** The data gathered from real urban traffic environments comprises over 2,000 trajectories of pedestrians, including 3D body posture keypoints, amounting to 50,000 test samples. Models examine 1 second of past data to predict the following 2.52 seconds at a frame rate of 25 frames per second.

C Baseline Models

Our baseline suite encompasses many methodological paradigms, including recurrent, generative, attention-based, language-driven, and diffusion-based frameworks, guaranteeing a thorough assessment of Bio-TAD. We benchmark against representative models such as Social-LSTM (Alahi et al., 2016), Social-GAN (Gupta et al., 2018), AutoBots (Girgis et al., 2021), Social-Transmotion (Saadatnejad et al., 2024), VLMTraj (Bae et al., 2025), and SingularTrajectory (Bae et al., 2024).

D Implementation Details

The Bio-TAD framework is implemented in PyTorch. Our dual-stream architecture concurrently processes trajectories and poses, mapping characteristics to a latent dimension of $d = 256$. The orthogonal Prototype Memory Bank comprises 256 trainable behavioral slots.

For the Gumbel-Softmax quantization step, we employ a temperature annealing schedule. The temperature τ is initialized at $\tau = 1.0$ to allow stable gradient flow and is exponentially decayed by a factor of 0.99 per epoch to a minimum floor of $\tau = 0.1$ to enforce strict discrete routing. The AdaLN Temporal Diffusion Denoiser consists of four transformer layers, each including four attention heads and a feed-forward dimension of 1024. The diffusion technique employs a cosine noise schedule over $T = 1000$ training timesteps. During inference, we utilize DDIM sampling with 50 steps ($\eta = 0.0$) to produce $K = 20$ multimodal futures following standard practice for minADE20/minFDE20 evaluation.

The framework is fully optimized using the Adam optimizer with an initial learning rate of 1×10^{-3}. The multi-task objective is calibrated to emphasize target convergence, with $\lambda_{waypoint} =$

Table 9: Ablation study of Multi-Modality configurations on the JTA Dataset. Metrics are reported as $minADE_{20}$ / $minFDE_{20}$ in meters

Input Modality Configuration	AutoBots	Social-Transmotion
Traj	0.72 / 1.50	0.65 / 1.49
Traj + 2D Pose	0.66 / 1.41	0.61 / 1.24
Traj + 2D Pose + 2D BB	0.63 / 1.34	0.58 / 1.19
Traj + 2D Pose + 2D BB + 3D BB	0.61 / 1.28	0.55 / 1.14
Traj + 2D BB + 3D BB + 3D Pose	**0.56 / 1.13**	**0.49 / 1.08**

Table 10: Intent accuracy across prediction horizons on JTA dataset: Trajectory-only (T) vs. Trajectory + 3D Pose (T+P)

Base Model	Intent Class	T=3 Frames		T=6 Frames		T=9 Frames		T=12 Frames	
		T	T+P	T	T+P	T	T+P	T	T+P
Social-LSTM	Left Turn	25.23%	**35.18%**	23.47%	**32.21%**	24.12%	**33.89%**	25.86%	**35.43%**
	Right Turn	20.15%	**28.42%**	21.09%	**27.91%**	20.44%	**26.35%**	21.73%	**27.18%**
	Far Left Turn	28.04%	**46.39%**	27.51%	**45.26%**	28.92%	**46.14%**	30.17%	**47.53%**
	Far Right Turn	32.41%	**42.67%**	33.15%	**41.52%**	34.08%	**43.91%**	35.26%	**44.75%**
Social-GAN	Left Turn	26.14%	**36.42%**	24.38%	**33.71%**	25.26%	**35.15%**	26.79%	**36.54%**
	Right Turn	21.42%	**29.68%**	22.35%	**29.17%**	21.81%	**27.74%**	22.92%	**28.45%**
	Far Left Turn	29.56%	**48.21%**	28.73%	**46.38%**	30.14%	**47.59%**	31.45%	**49.12%**
	Far Right Turn	33.58%	**44.15%**	34.62%	**43.06%**	35.71%	**45.24%**	36.84%	**46.31%**
AutoBots	Left Turn	28.84%	**40.97%**	27.93%	**38.65%**	29.22%	**39.61%**	31.54%	**42.18%**
	Right Turn	23.30%	**34.10%**	24.10%	**32.37%**	23.41%	**30.80%**	23.66%	**31.08%**
	Far Left Turn	29.15%	**53.32%**	28.32%	**52.44%**	29.12%	**52.68%**	29.12%	**53.12%**
	Far Right Turn	34.00%	**54.33%**	35.38%	**52.05%**	35.07%	**52.32%**	36.13%	**54.97%**
Social-Transmotion	Left Turn	26.59%	**45.98%**	26.74%	**41.02%**	27.40%	**41.90%**	28.53%	**43.48%**
	Right Turn	17.70%	**41.39%**	20.74%	**37.85%**	21.55%	**37.16%**	22.85%	**38.57%**
	Far Left Turn	31.03%	**56.50%**	31.29%	**52.53%**	33.41%	**53.82%**	34.93%	**54.60%**
	Far Right Turn	38.65%	**56.38%**	40.37%	**54.09%**	40.06%	**49.43%**	42.52%	**50.34%**
SingularTrajectory	Left Turn	32.48%	**45.61%**	31.15%	**42.72%**	32.63%	**44.18%**	34.19%	**46.34%**
	Right Turn	26.52%	**37.66%**	27.84%	**36.41%**	27.17%	**35.03%**	28.45%	**36.29%**
	Far Left Turn	34.71%	**55.84%**	33.56%	**53.42%**	35.28%	**54.91%**	36.43%	**56.17%**
	Far Right Turn	38.16%	**51.25%**	39.42%	**49.83%**	39.15%	**50.22%**	40.71%	**52.48%**
VLMTraj	Left Turn	33.54%	**46.42%**	32.27%	**43.95%**	33.81%	**45.26%**	35.46%	**47.53%**
	Right Turn	27.63%	**38.81%**	28.95%	**37.64%**	28.32%	**36.19%**	29.58%	**37.45%**
	Far Left Turn	35.82%	**56.97%**	34.69%	**54.58%**	36.44%	**56.02%**	37.55%	**57.34%**
	Far Right Turn	39.27%	**52.16%**	40.58%	**50.91%**	40.34%	**51.45%**	41.86%	**53.62%**

0.4, $\lambda_{ortho} = 0.1$, $\lambda_{x_0} = 0.1$, and $\lambda_{ADE} = 0.05$. All models are trained and assessed on a single NVIDIA H100 GPU. The full Bio-TAD module introduces approximately 11.37M additional parameters relative to the base backbone and runs at 82 ms per inference (K=20, DDIM-50) on a single H100, remaining within the real-time budget of the target deployment frequency.

Table 11: Intent accuracy across prediction horizons on JRDB dataset: Trajectory-only (T) vs. Trajectory + 2D Pose (T+P)

Base Model	Intent Class	T=3 Frames		T=6 Frames		T=9 Frames		T=12 Frames	
		T	T+P	T	T+P	T	T+P	T	T+P
Social-LSTM	Left Turn	24.31%	**29.15%**	23.15%	**27.52%**	24.08%	**28.51%**	25.62%	**30.29%**
	Right Turn	19.85%	**23.36%**	20.42%	**23.85%**	20.15%	**23.58%**	21.28%	**24.84%**
	Far Left Turn	27.55%	**36.62%**	26.83%	**35.79%**	28.41%	**37.52%**	29.67%	**38.89%**
	Far Right Turn	32.14%	**37.28%**	33.02%	**38.07%**	33.85%	**39.02%**	34.91%	**40.21%**
Social-GAN	Left Turn	25.82%	**30.91%**	24.17%	**28.84%**	25.04%	**30.12%**	26.55%	**31.78%**
	Right Turn	21.14%	**25.18%**	22.18%	**25.64%**	21.72%	**24.75%**	22.84%	**25.62%**
	Far Left Turn	29.31%	**38.64%**	28.52%	**37.41%**	29.84%	**38.57%**	31.12%	**39.95%**
	Far Right Turn	33.15%	**38.43%**	34.21%	**38.44%**	35.18%	**39.94%**	36.42%	**41.15%**
AutoBots	Left Turn	28.84%	**34.90%**	27.93%	**33.29%**	29.22%	**34.41%**	31.54%	**36.86%**
	Right Turn	23.30%	**28.70%**	24.10%	**28.24%**	23.41%	**27.10%**	23.66%	**27.37%**
	Far Left Turn	29.15%	**41.23%**	28.32%	**40.38%**	29.12%	**40.90%**	29.12%	**41.12%**
	Far Right Turn	34.00%	**44.16%**	35.38%	**43.71%**	35.07%	**43.69%**	36.13%	**45.55%**
Social-Transmotion	Left Turn	26.59%	**36.28%**	26.74%	**33.88%**	27.40%	**34.65%**	28.53%	**36.00%**
	Right Turn	17.70%	**29.54%**	20.74%	**29.29%**	21.55%	**29.35%**	22.85%	**30.71%**
	Far Left Turn	31.03%	**43.76%**	31.29%	**41.91%**	33.41%	**43.61%**	34.93%	**44.76%**
	Far Right Turn	38.65%	**47.51%**	40.37%	**47.23%**	40.06%	**44.74%**	42.52%	**46.43%**
SingularTrajectory	Left Turn	32.14%	**38.52%**	31.08%	**36.87%**	32.41%	**38.18%**	33.82%	**39.89%**
	Right Turn	26.15%	**31.54%**	27.42%	**31.70%**	26.85%	**30.77%**	28.14%	**32.06%**
	Far Left Turn	34.25%	**44.71%**	33.14%	**43.07%**	34.92%	**44.73%**	36.18%	**46.05%**
	Far Right Turn	37.82%	**44.36%**	39.14%	**44.34%**	38.86%	**44.39%**	40.32%	**46.20%**
VLMTraj	Left Turn	33.41%	**39.95%**	32.16%	**38.00%**	33.52%	**39.24%**	35.18%	**41.21%**
	Right Turn	27.32%	**32.91%**	28.65%	**32.99%**	28.04%	**31.97%**	29.27%	**33.20%**
	Far Left Turn	35.58%	**46.15%**	34.42%	**44.36%**	36.14%	**45.93%**	37.31%	**47.20%**
	Far Right Turn	38.95%	**45.39%**	40.21%	**45.37%**	40.12%	**45.66%**	41.54%	**47.42%**

Table 12: Intent accuracy across prediction horizons on Urban dataset (Pedestrian only): Trajectory-only (T) vs. Trajectory + 3D Pose (T+P)

Base Model	Intent Class	T=15 Frames		T=30 Frames		T=45 Frames		T=63 Frames	
		T	T+P	T	T+P	T	T+P	T	T+P
Social-LSTM	Left Turn	37.93%	**45.21%**	37.50%	**44.36%**	38.96%	**46.68%**	40.06%	**47.39%**
	Right Turn	36.32%	**43.05%**	37.10%	**43.67%**	37.96%	**44.54%**	39.11%	**46.02%**
	Far Left Turn	34.76%	**46.12%**	33.83%	**44.95%**	35.47%	**47.83%**	36.21%	**48.72%**
	Far Right Turn	35.20%	**46.85%**	34.87%	**45.64%**	36.05%	**47.98%**	37.22%	**49.31%**
Social-GAN	Left Turn	40.73%	**48.02%**	40.24%	**47.11%**	41.71%	**49.52%**	42.82%	**50.36%**
	Right Turn	38.86%	**45.48%**	39.65%	**46.71%**	40.61%	**47.33%**	41.75%	**48.92%**
	Far Left Turn	37.38%	**48.96%**	36.43%	**47.72%**	38.13%	**50.41%**	38.84%	**51.02%**
	Far Right Turn	37.84%	**49.31%**	37.50%	**48.27%**	38.62%	**50.06%**	39.83%	**51.63%**
AutoBots	Left Turn	46.31%	**53.79%**	45.74%	**52.41%**	47.26%	**55.18%**	48.38%	**56.09%**
	Right Turn	44.39%	**51.62%**	45.14%	**52.88%**	46.15%	**53.47%**	47.29%	**54.63%**
	Far Left Turn	42.85%	**55.36%**	41.89%	**53.74%**	43.58%	**56.21%**	44.34%	**57.68%**
	Far Right Turn	43.35%	**55.91%**	43.05%	**54.26%**	44.24%	**56.63%**	45.43%	**58.12%**
Social-Transmotion	Left Turn	49.16%	**56.94%**	48.56%	**55.81%**	50.11%	**58.02%**	51.22%	**59.46%**
	Right Turn	47.26%	**54.33%**	48.04%	**55.97%**	49.01%	**56.72%**	50.15%	**57.96%**
	Far Left Turn	45.76%	**58.42%**	44.74%	**56.98%**	46.49%	**59.87%**	47.23%	**60.92%**
	Far Right Turn	46.25%	**58.88%**	45.94%	**57.36%**	47.12%	**60.41%**	48.29%	**61.55%**
SingularTrajectory	Left Turn	53.80%	**61.24%**	53.23%	**60.31%**	54.77%	**62.05%**	55.95%	**63.18%**
	Right Turn	51.89%	**58.72%**	52.67%	**60.14%**	53.65%	**60.97%**	54.75%	**62.32%**
	Far Left Turn	50.33%	**61.02%**	49.37%	**59.54%**	51.10%	**61.28%**	51.88%	**60.73%**
	Far Right Turn	50.81%	**63.55%**	50.51%	**62.01%**	51.69%	**64.94%**	52.84%	**66.28%**
VLMTraj	Left Turn	55.74%	**63.41%**	55.14%	**62.28%**	56.65%	**64.72%**	57.83%	**65.91%**
	Right Turn	53.83%	**60.96%**	54.59%	**62.13%**	55.60%	**63.02%**	56.65%	**64.37%**
	Far Left Turn	52.33%	**62.01%**	51.33%	**61.44%**	53.03%	**62.27%**	53.78%	**62.58%**
	Far Right Turn	52.76%	**65.48%**	52.45%	**63.92%**	53.61%	**63.74%**	54.77%	**64.13%**