HABIT: Human Action Benchmark for Interactive Traffic in CARLA

Mohan Ramesh* Mark Azer* Fabian B. Flohr
Intelligent Vehicles Lab, Munich University of Applied Sciences
{mohan.ramesh, mark.azer, fabian.flohr}@hm.edu

Figure 1. HABIT Benchmark: Integrating realistic pedestrian behavior into CARLA for high-fidelity autonomous driving evaluation. Left: retargeting real human motion into simulation. Center: interactive traffic scenarios with diverse agents. Right: testing perception, prediction, and planning systems under realistic pedestrian dynamics.

Abstract

Current autonomous driving (AD) simulations are critically limited by their inadequate representation of realistic and diverse human behavior, which is essential for ensuring safety and reliability. Existing benchmarks often simplify pedestrian interactions, failing to capture complex, dynamic intentions and varied responses critical for robust system deployment. To overcome this, we introduce HABIT (Human Action Benchmark for Interactive Traffic), a high-fidelity simulation benchmark. HABIT integrates real-world human motion, sourced from mocap and videos, into CARLA [1] (Car Learning to Act, a full autonomous driving simulator) via a modular, extensible, and physically consistent motion retargeting pipeline. From an initial pool of approximately 30,000 retargeted motions, we curate 4,730 traffic-compatible pedestrian motions, standardized in SMPL format for physically consistent trajectories. HABIT seamlessly integrates with CARLA's Leaderboard, enabling automated scenario generation and rigorous agent evaluation. Our safety metrics, including Abbreviated Injury Scale (AIS) and False Positive Braking Rate (FPBR), reveal critical failure modes in state-of-the-art AD agents missed by prior evaluations. Evaluating three state-of-the-art autonomous driving agents, InterFuser, TransFuser, and BEVDriver, demonstrates how HABIT exposes planner weaknesses that remain hidden in scripted simulations. Despite achieving close or equal to zero collisions per kilometer on the CARLA Leaderboard, the autonomous agents perform notably worse on HABIT, with up to 7.43 *collisions/km and a* 12.94% *AIS 3+ injury risk, and they brake unnecessarily in up to 33% of cases. All components are publicly released to support reproducible, pedestrian-aware AI research.*

1. Introduction

Autonomous driving agents face their greatest test in environments shaped by the unpredictable nature of human behavior. Pedestrian dynamics, in particular, introduce a uniquely complex challenge: their movements are rarely deterministic, instead being influenced by implicit social cues, environmental context, and their underlying inten-

*These authors contributed equally.

tions. Despite significant advancements in the visual realism of simulation platforms such as CARLA, LGSVL [2], AirSim [3], and DeepDrive [4], most current benchmarks fall short in capturing the full semantic and kinematic diversity of pedestrian behavior. While tools such as Omniverse [5] offer extensions for populating scenes with animated human actors and enabling basic scripted actions, they still lack support for generating diverse, context-aware trajectories that are truly grounded in real-world pedestrian dynamics. Instead, simulated agents are typically governed by simplistic, handcrafted state machines or deterministic scripts. This results in highly predictable, semantically shallow interactions that ultimately undermine the robustness and relevance of autonomous driving evaluation pipelines. This behavioral gap leads to models that overfit to synthetic priors, such as the assumption that pedestrians will always cross linearly or yield predictably. Consequently, these models exhibit poor generalization when confronted with ambiguous, rare, or socially nuanced behaviors in the real world. These limitations are especially critical in safety-sensitive scenarios, where accurate forecasting of human intent is essential. The current narrow focus on binary collision metrics in evaluations, which often overlooks the severity of impact, further exacerbates this issue. Without clinically meaningful safety measures, planners may overreact to low-risk encounters while inadvertently underestimating genuinely dangerous ones, thereby distorting our assessment of system competence.

To bridge the significant behavioral gap in autonomous driving simulations, we need to embed realistic human motion at scale, while ensuring both physical realism and contextual consistency. Our solution, HABIT (Human Action Benchmark for Interactive Traffic, see Figure 1), directly addresses this need. HABIT is based on a novel pipeline that integrates heterogeneous motion capture and video data into the CARLA simulator, generating highly realistic pedestrian trajectories. From over 30,000 motion clips, we meticulously curated 4,730 traffic-compatible sequences through a process of semantic filtering and visual validation. Unlike static motion clips, HABIT reconstructs temporally coherent trajectories by fusing SMPL [6] pose sequences with orientation-aligned global trajectory estimation. The resulting motions are smooth, physically grounded, and adaptable to diverse urban environments, faithfully capturing complex behaviors like jaywalking, non-yielding, and hesitant crossing. This rich behavioral diversity is crucial for developing robust autonomous driving systems capable of generalizing to real-world complexities. We validate HABIT's impact across critical perception, prediction, and planning tasks using state-of-the-art models like YOLOv11-Pose [7], Segment Anything Model (SAM) [8] and InterFuser [9]. Crucially, HABIT enables us to move beyond simple binary collision metrics by incorporating the Abbreviated Injury Scale (AIS) for a clinically grounded measure of injury severity, allowing for more meaningful safety evaluations.

Our key contributions are threefold. First, we introduce HABIT, the first CARLA-based benchmark that integrates thousands of semantically curated, real-world pedestrian motions with context-aware behavioral transitions. Our benchmark includes 4,730 high-fidelity traffic-compatible motions, further enhanced with randomized augmentations to increase behavioral diversity. Second, we develop a modular motion reconstruction pipeline that transforms heterogeneous motion sources, including motion capture and video data, into smooth, physically plausible SMPL-based trajectories. These are aligned globally for temporal coherence and integrated seamlessly into interactive driving simulations. Third, we propose novel evaluation metrics to assess model behavior beyond binary collision rates. We introduce a false-positive braking metric, revealing conservative failure modes in state-of-the-art planners, and a severity-aware collision assessment using the Abbreviated Injury Scale (AIS), enabling clinically meaningful evaluations of driving risk. Using these metrics, we evaluate InterFuser, TransFuser, and BEVDriver, showing that HABIT reveals planner-specific weaknesses and safety-conservatism trade-offs hidden in scripted benchmarks. We release all our 30,000 retargeted pedestrian motions, along with tools for motion processing, scenario generation, and evaluation to support reproducible and scalable research.

2. Related Works

Pedestrian Realism in Simulation Environments High-fidelity simulation is essential for autonomous driving development and evaluation, yet a persistent reality gap remains between simulated and real human behaviors. Here, we use the term "high-fidelity" to denote simulations that jointly optimize for appearance fidelity (photorealistic rendering), motion fidelity (biomechanically and physically consistent movements), and behavioral fidelity (context-aware, semantically diverse interactions). Platforms like CARLA [1], LGSVL [2], AirSim [3], and NVIDIA Omniverse (DRIVESim) [10] offer photorealistic environments and sensor fidelity. However, they still rely heavily on rule-based or deterministic pedestrian behaviors, resulting in semantically shallow and predictable interactions. Although they emphasize addressing both appearance and content gaps, particularly in actor behavior generation, the focus has predominantly been on vehicle traffic and sensor realism, not pedestrian-driven dynamics [11]. Furthermore, Omniverse supports data-driven human motions using USD-based animation blending, though its emphasis remains on high-fidelity rendering rather than pedestrian-interactive behavior. Prior work on pedestrian realism, such as hierarchical behavior models [12], remains sparse and rela-

Figure 2. Overview of the proposed motion data processing pipeline. The dotted box presents HABIT'S extensibility and scalability using our video based motion extraction.

tively immature. To our knowledge, no existing work integrates real-world pedestrian motion data into a large-scale, context-aware benchmark suitable for pedestrian-aware applications.

To enhance realism, generative and physics-based approaches are gaining traction. Wang *et al*. [13] introduced PACER+, a reinforcement-learned humanoid controller that follows a trajectory while reproducing motion styles and gestures in a physically plausible manner. This enables diverse behaviors, e.g., evasive movement, turning, or gesturing, grounded in both physics and scene context, however, the work does not scale to large scale interactive 3D simulation environment. Ramesh and Flohr [14] propose Walk-the-Talk a text to motion framework driven by large language models to generate semantically grounded pedestrian motion sequences such as jaywalking, intoxicated pedestrians for integration into a simulated environment. Other systems include the VRU Simulator [15], which is designed to evaluate pedestrian and cyclist interactions with vehicles, and GeoSim [16], which composites video snippets of real pedestrians into synthetic scenes. These approaches contribute behavioral diversity but often lack environmental grounding or integration into structured simulation and evaluation environments at scale.

While these efforts mark significant progress toward realistic human modeling, they remain limited in scalability, contextual diversity, or evaluation integration.

Pedestrian-Aware Prediction and Planning Classic models like the Social Force model [17] have been used to simulate pedestrian movements in traffic; while interpretable, these physics-based models require careful parameter tuning and struggle to capture the nuanced interactions of real pedestrians. Current state of research emphasizes integrating pedestrian intent and interaction into planning. Möller *et al*. [18] combine social force-based motion with risk-aware planning for navigating simulated crowded scenes. Probabilistic approaches further enhance prediction in simulation: Zhang *et al*. [19] model latent intent using Markov Decision Processes, while Dang *et al*. [20] apply game theory to predict pedestrian actions at unsignalized crossings. Occlusion-aware methods also improve safety realism. Tang *et al*. [21] use phantom agents based on map semantics to simulate hidden pedestrians, enabling cautious but efficient behavior. End-to-end approaches like InterFuser [9] have demonstrated strong performance and rank among the top performers on the CARLA Leaderboard.

Existing methods in this field rely on simplified simulation environments that fail to capture the complexity of real-world pedestrian behavior. HABIT addresses this gap by enabling realistic simulations in CARLA through semantically filtered motion-capture data, grounded trajectory reconstruction, and structured scenario generation.

3. Methodology

Figure 2 summarizes our pipeline. Motion data is processed through one of two independent pathways. The primary path uses the AMASS datasets [22], where sequences pass through the *Raw pose processing* and *Vector extraction* modules to obtain 3D joints, pose vectors, and velocities. These are integrated via the *Trajectory Reconstruction Module* to reconstruct global root trajectories aligned with local motion.

Additionally, pedestrian motions are extracted from open-source videos. These are processed through a *Video Encoder* to extract 2D keypoints, which are lifted to SMPL format. Regardless of the source, using *SMPLify* [23] and a temporal variant for videos, the keypoints are lifted to an

intermediate SMPL representation. The resulting SMPL sequences are sampled to match CARLA's target animation frequency and passed to the *Retargeting to CARLA* module for conversion into CARLA-compatible joint rotations. In the following sections, we explain our pipeline in detail.

3.1. Dataset filtering and processing

We begin by collecting raw human motion data from a variety of heterogeneous motion capture datasets from AMASS and open-source videos. These datasets capture a broad range of everyday activities, not originally tailored for traffic contexts, but offer diverse kinematics and pose information crucial for realistic pedestrian simulation. The initial motion pool contains approximately $30,000$ general human activities with associated metadata including natural language descriptions and temporal structure, sourced from the HumanML3D motion-language dataset [24], and the HumanSensing Lab [25] text-annotated dataset.

To adapt these general-purpose motions for autonomous driving scenarios, we apply a two-stage filtering pipeline designed to select behaviorally plausible pedestrian motions suitable for traffic environments. First, we process the textual motion annotations from HumanML3D using the spaCy *en_core_web_sm* model, an industrial-strength natural language processing library, performing keyword filtering for walking and movement-related terms. This step reduces the dataset from $30,000$ motions to $5,108$ candidate motions appropriate for traffic contexts. The resulting semantically filtered set then undergoes a structured visual validation process, serving both as a human verification step and as an additional filter to remove motions irrelevant to traffic scenarios. The final curated set consists of 4730 filtered motions in our current installation of the HABIT benchmark.

The resulting subset of filtered motions is then subjected to a standardized pre-processing pipeline. We begin by extracting joint positions from the AMASS pose data using a modified version of the raw pose processing technique introduced in HumanML3D. In addition to standard parameters, we compute linear and angular velocities of the root joint in a 6D rotation representation. These features are down-sampled to ensure temporal consistency with all the motions at 20Hz. Intermediate SMPL pose representations are then estimated using a pretrained SMPLify network. Finally, the computed velocities are leveraged to reconstruct global root trajectories and orientations (see Section Sec. 3.2), which are subsequently aligned with the coarse SMPL poses derived from the joint data to yield consistent full-body motion representations.

3.1.1. Motion from arbitrary videos

We designed a Video-to-CARLA pipeline inspired by recent SMPL lifting pipelines such as WHAM [26]. Our approach uses the ViTPose to detect 2D keypoints. The 2D keypoints are then uplifted into 3D SMPL poses using SMPLify. The global trajectory reconstruction algorithm described in Sec. 3.2, is complemented with DROID-SLAM[27] to use the estimated camera intrinsics to project trajectories into global reference frame.

We process a diverse set of static-camera videos sourced from YouTube [28], Pexels [29], and the HumanSensing Lab dataset [25], which provides frame-wise textual annotations. While the HABIT toolkit includes this processing framework, our goal here is to demonstrate HABIT's extensibility for creating novel benchmarks and cross-domain applications.

From the HumanSensing Lab dataset, we select a subset of 75 short, stationary pedestrian motions, however they are not a part of HABIT benchmark as they are positioned outside interaction range. These motions depict ambient sidewalk traffic in small groups and do not involve vehicle–pedestrian interactions. Importantly, these stationary motions differ from HABIT's *idle* motions: whose primary purpose is to augment interactive sequences by extending scenario duration and enriching context with naturalistic gestures, and not just sidewalk traffic.

3.2. Trajectory Reconstruction

We reconstruct physically plausible global root trajectories from local motion representations of the SMPL joints. Each motion sequence consists of a per-frame 6D root orientation vector $\mathbf{r}_t \in \mathbb{R}^6$ and a root-relative translational velocity vector $\mathbf{v}_t \in \mathbb{R}^3$.

The 6D root orientation vectors are converted into valid rotation matrices $\mathbf{R}_t \in \mathrm{SO}(3)$, the special orthogonal group of 3D rotations, through a continuous 6D-to-rotation matrix projection [30]. This step ensures smooth, gimbal-free orientation over time and forms the basis for global motion integration.

To recover the global root trajectory, we rotate each local velocity vector into the world frame:

$$\mathbf{v}_t^{\text{world}} = \mathbf{R}_t \cdot \mathbf{v}_t \tag{1}$$

and compute the global translation at time t via cumulative integration:

$$\mathbf{T}_t = \mathbf{T}_0 + \sum_{\tau=1}^{t} \mathbf{v}_\tau^{\text{world}} \tag{2}$$

where $\mathbf{T}_0 \in \mathbb{R}^3$ is the initial root position, typically set to the origin. This velocity-aligned trajectory reconstruction synchronizes global translation with local orientation at each timestep, enforcing temporal consistency by integrating motion along the pedestrian's forward direction rather than using frame-wise deltas (see Figure 3). It effectively reduces common SMPL artifacts such as abrupt heading changes, foot sliding, and inconsistent velocities, resulting in stable, realistic trajectories well-suited for simulation frameworks like CARLA.

Figure 3. Global root rotation over time: original (solid) vs. reconstructed (dashed).

3.3. SMPL-to-CARLA Motion Retargeting

SMPL provides human pose sequences as axis-angle representations $\boldsymbol{\theta} \in \mathbb{R}^{T\times J\times 3}$, defined per joint in a right-handed coordinate system. In contrast, CARLA, based on Unreal Engine [31], employs a left-handed convention and supports only local joint rotations as Euler angles relative to a predefined skeletal rest pose. Notably, CARLA's rest pose is neither a standard T-pose nor an A-pose, but rather a deformed body configuration, as illustrated in Figure 4. To reconcile these discrepancies, we design a motion retargeting pipeline that maps SMPL poses into CARLA-compatible Euler joint rotations $\boldsymbol{\phi} \in \mathbb{R}^{T\times J\times 3}$.

(a) SMPL zero pose (b) CARLA zero pose

Figure 4. Canonical poses of SMPL [32] and CARLA skeletons [33]. Note the difference in joint definitions, and limb angles

Each SMPL joint rotation vector $\boldsymbol{\omega}_{t,j} \in \mathbb{R}^3$ is first converted to a rotation matrix $R_{t,j} \in \mathrm{SO}(3)$ via the exponential map:

$$R_{t,j} = \exp\left(\hat{\boldsymbol{\omega}}_{t,j}\right), \tag{3}$$

where $\hat{\boldsymbol{\omega}}$ denotes the skew-symmetric matrix and the exponential map $\exp(\hat{\boldsymbol{\omega}})$ yields a valid 3D rotation.

To adapt SMPL's right-handed rotations into CARLA's left-handed frame, we apply a global coordinate transformation $C \in \mathrm{SO}(3)$:

$$R_{t,j}^{\text{CARLA}} = C R_{t,j} C^{-1}, \tag{4}$$

where C is constructed to flip the forward and up axes to align with CARLA's coordinate basis.

As CARLA defines a custom skeleton, we map joints based on most suitable closest match and certain SMPL joints must be merged to cater to the simplified representation in CARLA. For example, CARLA's `Spine1` is computed from the chained rotations of SMPL's `Spine2` and `Spine3`:

$$R_{t,\texttt{Spine1}}^{\text{CARLA}} = R_{t,\texttt{Spine2}}^{\text{CARLA}} \cdot R_{t,\texttt{Spine3}}^{\text{CARLA}}. \tag{5}$$

CARLA expects local joint rotations expressed relative to deformed zero pose, which we offset using a pre-defined standard T pose offsets for the CARLA skeleton, creating a new reference T pose for CARLA skeleton, $R_j^{\text{ref}} \in \mathrm{SO}(3)$. We compute the motion-relative joint rotation $\Delta R_{t,j}$ via:

$$\Delta R_{t,j} = \left(R_j^{\text{ref}}\right)^{-1} R_{t,j}^{\text{CARLA}} R_j^{\text{ref}}, \tag{6}$$

centering rotations around the reference pose. The final Euler joint angles are extracted using the intrinsic XYZ convention:

$$\boldsymbol{\phi}_{t,j} = \mathrm{Euler}_{\mathrm{XYZ}}(\Delta R_{t,j}), \tag{7}$$

yielding $\boldsymbol{\phi} \in \mathbb{R}^{T\times J\times 3}$ suitable for direct animation in CARLA.

4. HABIT Benchmark

The curated set of $4,730$ motion sequences obtained by our two-stage filtering pipeline is additionally annotated for structured scenario generation and benchmarking, retaining original natural language descriptions and unique identifiers. High-level behavior categories (e.g., walking, running, standing, falling) are inferred via NLP clustering, enabling reproducible behavior conditioning and semantic role assignment in simulation.

The current release of the HABIT benchmark comprises 110 routes under 12 distinct weather conditions, with each scenario populated by 30 vehicles, 20 behaviorally diverse pedestrians, and 10 ambient pedestrians serving as meaningful and background traffic.

Figure 5. Overview of the HABIT benchmark pipeline where retargeted motions are placed and controlled into CARLA scenes.

4.1. Simulation Scenario Generation

To evaluate autonomous agents under realistic human motion, we integrate the curated pedestrian dataset into interactive scenarios within the CARLA simulator (Figure 5). Each scenario features an ego vehicle navigating urban routes and encountering pedestrians exhibiting one of three behavior categories: not crossing, attempting to cross, or crossing. These categories are determined by forward displacement thresholds to preserve behavioral intent. The scenario generation pipeline defines agent placement, motion assignment, and interaction triggers for structured yet diverse evaluations within the CARLA Leaderboard.

4.1.1. Agent Configuration and Interaction Logic

Each simulation scenario consists of a single ego vehicle following a predefined route, controlled by CARLA's autopilot or external planners, and a fixed number of pedestrian agents. Background traffic behavior is governed by CARLA's ScenarioRunner module.

Pedestrians are pre-spawned at semantically appropriate locations such as sidewalks and curbs. Spawn points are randomly selected from CARLA's static sidewalk meshes within a bounded region around the ego route, ensuring pedestrian-viable placements through semantic tagging.

Initially, pedestrian agents exhibit idle, low-displacement motions (e.g., gesturing, scanning, subtle posture shifts), enhancing realism while ensuring scene stability. Behavior transitions to dynamic goal-oriented actions (e.g., crossing, attempting to cross) occur based on the ego vehicle's proximity and orientation or scenario-level triggers compatible with CARLA Leaderboard specifications. To maintain visual continuity, transitions employ short blending sequences interpolating between the initial idle pose and the dynamic motion start pose.

Pedestrian trajectories are continuously monitored for potential collisions with static objects (e.g., walls, parked vehicles) using bounding-box intersection checks. Predicted collisions prompt real-time trajectory rerouting to preserve locomotion and intent. Dynamic collisions with moving vehicles are permitted and logged to assess safety-critical scenarios. This selective collision logic maintains environmental realism and consistency for planner evaluation.

4.2. Evaluation Metrics

Our benchmark integrates all official evaluation metrics from the CARLA Leaderboard, along with two additional metrics designed to assess the behavioral and safety implications of pedestrian-agent interactions.

Injury Severity (AIS 3+): To provide a more informative assessment than binary collision counts, we incorporate a clinically grounded injury severity metric based on the Abbreviated Injury Scale (AIS) [34]. Specifically, we estimate the probability of a serious injury (MAIS 3+, i.e., AIS level 3 or higher) using a logistic regression model from Yanagisawa *et al*. [35], which relates injury risk to pedestrian impact speed v (in m/s):

$$P_{\text{MAIS 3+}}(v) = \frac{1}{1 + \exp(3.164 - 0.288v)} \tag{8}$$

False Positive Braking Rate (FPBR): We propose FPBR as a metric to quantify overcautious behavior in planners that incorporate pedestrian trajectory forecasting. FPBR is defined as the ratio of braking events that occur without an actual pedestrian crossing to the total number of braking events. This metric specifically targets systems like InterFuser [9] and PlanT [36], where the planner uses predicted pedestrian trajectories to assess collision risk and trigger braking. FPBR highlights unnecessary slowdowns caused by inaccurate or overly conservative forecasts.

Together, these metrics offer a multi-faceted evaluation, capturing both classical performance metrics and finer-grained behavioral sensitivity to complex pedestrian motion.

4.3. Dataset Statistics

Figure 6. Distribution of motion tags by behavior category.

The final benchmark includes 4,730 curated motion sequences, representing a diverse and balanced range of core pedestrian behaviors. To ensure comprehensive coverage, these behaviors are grouped into three broad classes, with NLP-derived behavioral tags proportionally distributed across them. Figure 6 visualizes this distribution, providing a clear overview of the dataset's behavioral diversity.

To characterize the dataset, we compute mean representative motions for each of the three main behavior categories. These averaged trajectories capture prototypical pedestrian dynamics and are visualized in Figure 7, highlighting distinct patterns and variability, especially among crossing behaviors.

5. Experiments

To demonstrate the realism and utility of the HABIT benchmark, we conduct comprehensive evaluations on driving

Figure 7. Trajectory statistics (mean, variance) by behavior class: Not Crossing, Attempting, Crossing.

models. Our goal is to analyze the impact of diverse, semantically grounded pedestrian behavior on planning robustness and prediction accuracy of autonomous agents. Prior to HABIT, such evaluations were either infeasible or uninformative on synthetic pedestrian data due to its limited behavioral diversity and visual domain gap. By scaling up realism in the CARLA simulator, HABIT bridges this gap, enabling meaningful benchmarking of pedestrian-aware autonomy under conditions that better reflect the complexities of the real world.

5.1. Prediction and Planning on HABIT

We evaluate three recent end-to-end autonomous driving agents on HABIT's semantically diverse pedestrian behaviors. Specifically, we include *InterFuser* [9], a transformer-based model that integrates perception, prediction, and planning; *TransFuser* [37], a transformer-based fusion architecture that jointly reasons over camera and LiDAR inputs; and *BEVDriver* [38], which leverages latent bird's-eye-view representations together with large language models (LLMs) for trajectory planning. For the latter, we adopt the route-to-language module introduced in the LangAuto benchmark [39], which encodes navigation waypoints into natural language form to condition planning. This evaluation enables a systematic comparison of how these agents generalize to richer and more ambiguous pedestrian motions than those found in scripted benchmark scenarios.

Model	Collisions/km ↓	pMAIS3+ (%) ↓	FPBR ↓	ADE ↓
InterFuser	5.24	10.96	0.33	1.64
TransFuser	7.43	12.94	0.12	1.87
BEVDriver	7.19	12.35	-	-

Table 1. Performance of three end-to-end planners on **HABIT Benchmark**. Metrics: safety (collisions/km, pMAIS3+), conservatism (FPBR), and pedestrian prediction (ADE). Lower is better; higher pMAIS3+ = more severe injuries.

Results in Table 1 highlight that HABIT reveals planner-specific weaknesses that remain invisible in scripted benchmarks. Most notably, InterFuser, despite achieving zero collisions/km on the CARLA Leaderboard, records **5.24 collisions/km** and a **10.96%** injury risk under HABIT. This sharp degradation underscores a critical limitation: its planner and controller, trained on rule-based pedestrians, adopt an overly conservative braking strategy (high FPBR of 0.33) yet still fails to anticipate diverse pedestrian motions, resulting in frequent collisions. In contrast, TransFuser displays the opposite tendency, with fewer unnecessary braking events (low FPBR of 0.12) but correspondingly higher collision frequency and prediction error, indicating underreactive planning. BEVDriver achieves safety scores comparable to TransFuser but lacks pedestrian intent forecasting, making FPBR and ADE evaluation inapplicable.

Taken together, these findings demonstrate that HABIT systematically stresses planning agents with realistic pedestrian behavior, surfacing trade-offs between safety and conservatism. In particular, the large performance drop of InterFuser emphasizes the need for benchmarks like HABIT to expose hidden biases and generalization failures in state-of-the-art driving agents.

5.2. Ablation studies

To isolate the effects of specific pedestrian behaviors, we evaluate InterFuser on two HABIT variants: (i) **HABIT-Idle**, where all pedestrians remain idle with no crossing intent, and (ii) **HABIT-Left**, where pedestrians exclusively originate from the left side and attempt rightward crossings.

Condition	Collisions/km ↓	pMAIS3+ (%) ↓	FPBR ↓	ADE ↓
HABIT-Idle	0.16	8.40	0.639	1.84
HABIT-Left	8.35	12.19	0.09	2.19

Table 2. InterFuser ablation: On **HABIT-Idle**, overly conservative behavior raises FPBR; on **HABIT-Left**, performance drops due to right-to-left bias from data imbalance.

Results in Table 2 show two complementary failure modes. Under HABIT-Idle, InterFuser records almost no collisions but a sharp increase in FPBR (0.639). This reflects an overcautious policy: the agent frequently brakes in response to idle pedestrians, misinterpreting their presence as crossing intent. In effect, the model treats standing humans as latent threats, revealing its reliance on detection cues without robust intent reasoning. Figure 8 illustrates this tendency, where the vehicle remains stopped for 12 seconds despite the pedestrian showing no crossing intent.

In contrast, HABIT-Left exposes a strong directional bias. As illustrated in Figure 9, InterFuser consistently forecasts right-to-left motion even when pedestrians cross from left to right. This misprediction, caused by data imbalance in its training distribution, leads to underreaction and late

Figure 8. **False Positive Braking over Time.** Front-view frames (T_0, T_6, T_{12}, , in seconds) show overcautious stopping: vehicle halts for a stationary pedestrian and resumes after 12 s.

braking. Quantitatively, this results in **8.35 collisions/km** and a pMAIS3+ rate of **12.19%**, alongside degraded prediction accuracy (ADE = 2.19). The low FPBR in this case indicates that the agent rarely brakes preemptively, further compounding safety risks.

Figure 9. **Directional prediction bias (qualitative).** Top: front camera input with ground truth pedestrian trajectory (green) moving right (T_0–T_2, in seconds). Bottom: predicted BEV maps over time; model incorrectly predicts left, showing directional bias.

Together, these ablations clarify why the performance degrades on the full HABIT benchmark (see Table 1): InterFuser oscillates between excessive caution in idle scenarios and unsafe underreaction when confronted with rare but realistic crossing behaviors. Our ablation results highlight the need for behaviorally diverse benchmarks to uncover planner biases that remain invisible in rule-based environments.

6. Discussion

Our experiments show that HABIT substantially advances the realism and diagnostic depth of pedestrian-aware benchmarks for autonomous driving. For prediction and planning, HABIT revealed critical shortcomings in state-of-the-art models. Interfuser, despite strong performance on the CARLA Leaderboard, struggled with HABIT's semantically rich scenarios, especially in anticipating crossing intents from uncommon spatial contexts. These failures, reflected in elevated collision rates, displacement errors, and MAIS 3+ injury risks, underscore the limitations of training on rule-based or scripted pedestrian behaviors. TransFuser faced similar difficulties, reacting less conservatively but with higher collision frequency and less accurate trajectory forecasts. BEVDriver performed comparably in safety terms but lacked explicit pedestrian forecasting, limiting its ability to capture fine-grained behavioral intent.

HABIT also enables fine-grained analysis of model weaknesses through controlled ablations. In idle-pedestrian scenarios, agents exhibited overcautious braking, while left-origin crossings exposed strong directional biases and systematic mispredictions. Together, these analyses show that HABIT surfaces subtle but safety-critical planner failures, highlighting trade-offs between conservatism and responsiveness. Its structured behavioral diversity ensures that planners are tested against realistic, unpredictable pedestrian interactions, rather than oversimplified patterns.

Despite this progress, challenges remain. Grounded motion extraction from arbitrary, moving-camera video sources is still difficult, requiring improved reconstruction techniques to maintain spatial accuracy. Additionally, traditional augmentation introduces redundant behaviors; future work could employ generative models to create richer intra-trajectory and pose variations [14]. Lastly, pedestrian spawning in simulation can be made more realistic by learning placement strategies from real-world data, enabling context-aware agent initialization. Addressing these areas will further enhance HABIT's scalability and realism for autonomous system evaluation.

7. Conclusion and future work

In this work, we introduced HABIT (Human Action Benchmark for Interactive Traffic), a high-fidelity simulation benchmark designed to address the critical limitations of current autonomous driving (AD) simulations in representing realistic human behavior. HABIT integrates real-world pedestrian motions into CARLA, offering a rigorous and realistic platform for evaluating pedestrian-aware autonomous driving systems. Our evaluations revealed that while modern end-to-end planning methods can achieve zero collisions per kilometer on the CARLA Leaderboard, they incur up to 7.43 collisions/km and a 12.94% AIS 3+ injury risk on HABIT, and unnecessarily brake in 33% of the cases. These findings highlight HABIT's ability to uncover critical failure modes and blind spots in pedestrian-aware autonomy that were missed by prior evaluations. Beyond AD, HABIT's pipeline offers a versatile solution for realistic human motion simulation, supporting embodied learning and enhancing human-agent interaction in robotics. To foster open research and reproducibility, the HABIT benchmark and its entire simulation pipeline are publicly released.

References

[1] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving

simulator. In *Proc. of the Conference on Robot Learning (CoRL)*, pages 1–16. PMLR, 2017. 1, 2

[2] Guodong Rong, Byung Hyun Shin, Hadi Tabatabaee, Qiang Lu, Steve Lemke, Mārtiņš Možeiko, Eric Boise, Geehoon Uhm, Mark Gerow, Shalin Mehta, et al. LGSVL simulator: A high fidelity simulator for autonomous driving. In *Proc. of the IEEE International conference on intelligent transportation systems (ITSC)*, pages 1–6. IEEE, 2020. 2, 11

[3] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics: Results of the 11th international conference*, pages 621–635. Springer, 2017. 2

[4] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Proc. of the Conference on robot learning (CoRL)*, pages 66–75. PMLR, 2020. 2

[5] NVIDIA. Omni.anim.people: Simulating human events in environments such as retail stores, warehouses, and traffic intersections. https://docs.isaacsim.omniverse.nvidia.com/4.2.0/features/warehouse_logistics/ext_omni_anim_people.html, 2023. [Accessed: 17-Jul-2025]. 2

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2

[7] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, Oct 2024. Preprint. 2

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Laura Rolland, Laura Gustafson, Chao Xiao, Spencer Whitehead, Adam Caine, Achal Patil, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 12

[9] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Proc. of the Conference on Robot Learning (CoRL)*, pages 726–737. PMLR, 2023. 2, 3, 6, 7

[10] NVIDIA Corporation. NVIDIA DRIVE Sim: A Scalable, Physically Accurate Simulator for Autonomous Vehicles. https://blogs.nvidia.com/blog/2021/04/12/nvidia-drive-sim-omniverse-early-access/, 2021. Accessed: 2025-07-19. 2

[11] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 892–901, 2021. 2

[12] Amir Rasouli and Iuliia Kotseruba. Intend-wait-cross: Towards modeling realistic pedestrian crossing behavior. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 83–90. IEEE, 2022. 2

[13] Jingbo Wang, Zhengyi Luo, Ye Yuan, Yixuan Li, and Bo Dai. Pacer+: On-demand pedestrian animation controller in driving scenarios. In *Proc. of. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[14] Mohan Ramesh and Fabian B Flohr. Walk-the-talk: Llm driven pedestrian motion generation. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 3057–3062. IEEE, 2024. 3, 8

[15] Martin Fischer, Gerald Temme, Kilian Gröne, Donaji Martinez Garcia, Gavin Grolms, and Johannes Rehm. A vru-simulator for the evaluation of pedestrian-and cyclist-vehicle interaction. design criteria and implementation. In *Driving Simulation and Virtual Reality Conference and Exhibition, DSC 2022 EUROPE*, pages 153–160. Driving Simulation Association, 2022. 3

[16] Dian Chen, Xinshuo Xu, Yi-Ting Chen, and Kris Kitani. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *CVPR*, 2021. 3

[17] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 3

[18] Korbinian Möller, Truls Nyberg, Jana Tumova, and Johannes Betz. Pedestrian-aware motion planning for autonomous driving in complex urban scenarios. *arXiv*, abs/2504.01409, 2025. URL https://arxiv.org/abs/2504.01409. Preprint. 3

[19] Chi Zhang, Florian Steinhauser, Gereon Hinz, and Alois Knoll. Improved occlusion scenario coverage with a pomdp-based behavior planner for autonomous urban driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 593–600. IEEE, 2021. 3

[20] Meiting Dang, Dezong Zhao, Yafei Wang, and Chongfeng Wei. Dynamic game-theoretical decision-making framework for vehicle-pedestrian interaction with human bounded rationality. *IEEE Transactions on Intelligent Transportation Systems (ITSC)*, 2025. 3

[21] Ming Tang, Tianrui Wang, and Dorsa Sadigh. Improved occlusion scenario coverage with a pomdp-based behavior planner for autonomous urban driving. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 3

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 3

[23] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. of European Conference on Computer Vision (ECCV)*, October 2016. 3

[24] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 4

[25] Stanislav Panev, Emily Kim, Sai Abhishek Si Namburu, Desislava Nikolova, Celso de Melo, Fernando De la Torre, and Jessica Hodgins. Exploring the impact of rendering method and motion quality on model performance when using multi-view synthetic data for action recognition. In *Proc. of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 4592–4602, 2024. 4

[26] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proc. of. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, June 2024. 4

[27] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:21114–21129, 2021. 4

[28] YouTube. Youtube: Video platform for user-generated content. https://www.youtube.com, 2025. Accessed: 2025-07-17. 4

[29] Pexels. Pexels: Free stock video platform. https://www.pexels.com/videos, 2025. Accessed: 2025-07-17. 4

[30] Yuxuan Zhou, Corey Barnes, Jimei Lu, and Qixing Yang. On the continuity of rotation representations in neural networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 4

[31] Epic Games. Unreal engine 5. https://www.unrealengine.com/, 2024. Computer software. Accessed on July 18, 2025. 5

[32] Pengpeng Hu, Edmond Shu-Lim Ho, and Adrian Munteanu. 3dbodynet: Fast reconstruction of 3d animatable human body shape from a single commodity depth camera. *IEEE transactions on multimedia*, 24:2139–2149, 2021. 5

[33] CARLA Simulator. Issue #7621: Is it possible to set pedestrian position manually in carla? https://github.com/carla-simulator/carla/issues/7621, 2023. Accessed: 2025-07-18. 5

[34] Xuemin Hu, Shen Li, Tingyu Huang, Bo Tang, Rouxing Huai, and Long Chen. How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. *IEEE Transactions on Intelligent Vehicles*, 9(1): 593–612, 2023. 6, 12

[35] Mikio Yanagisawa, Elizabeth D Swanson, Philip Azeredo, Wassim Najm, et al. Estimation of potential safety benefits for pedestrian crash avoidance/mitigation systems. Technical report, United States. Department of Transportation. National Highway Traffic Safety Administration, 2017. 6

[36] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A. Sophia Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. In *Proc. of the Conference on Robot Learning (CoRL)*, pages 459–470, 2022. 6

[37] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Fan Yu, Karl Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 108–118, 2022. 7

[38] Katharina Winter, Mark Azer, and Fabian B Flohr. Bevdriver: Leveraging bev maps in llms for robust closed-loop driving. *arXiv preprint arXiv:2503.03074*, 2025. 7

[39] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 7

[40] CARLA Team. Carla autonomous driving leaderboard. http://leaderboard.carla.org, 2024. 11

[41] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017. URL https://arxiv.org/abs/1705.05065. 11

[42] Zhanteng Xie and Philip Dames. Drl-vo: Learning to navigate through crowded dynamic scenes using velocity obstacles. *IEEE Transactions on Robotics*, 39(4):2700–2719, 2023. doi: 10.1109/TRO.2023.3257549. 11

[43] Maciej Wielgosz, Antonio M. López, and Muhammad Naveed Riaz. CARLA-BSP: a simulated dataset with pedestrians, May 2023. 12

[44] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, 2018. 12

[45] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017. 12

[46] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1465–1479, 2018. 12

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 12

[48] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 12

Appendix

A. HABIT Benchmark Illustrations and Details

The primary objective of the HABIT benchmark is to assess the driving competence of autonomous agents in complex and realistic traffic environments, with particular emphasis on interactions involving vulnerable road users such as pedestrians. A key limitation of existing benchmarks is their insufficient representation of pedestrians performing rare or unconventional gestures and behaviors. To address this gap, our initial evaluations focus on state-of-the-art agents from the CARLA leaderboard [40]. The current release of the HABIT benchmark comprises 110 routes under 12 distinct weather conditions, with each scenario populated by 30 vehicles, 20 behaviorally diverse pedestrians, and 10 ambient pedestrians serving as meaningful and background traffic. Representative examples from the benchmark are presented in Figure 10.

Figure 10. **HABIT Benchmark.** Examples illustrating diverse pedestrian behaviors, environmental conditions, and route configurations designed to bridge the reality gap.

B. Comparison of HABIT with AD Simulators

Figure 11 provides a comparative overview of three widely used simulation platforms—AirSim, Gazebo, and LGSVL—highlighting their respective limitations in representing pedestrians. Accurately modeling pedestrian behavior, especially rare or unconventional gestures, is critical for evaluating autonomous driving systems in realistic traffic scenarios.

AirSim [41] is an open-source simulator developed by Microsoft on top of Unreal Engine, primarily targeting autonomous vehicle and drone research. It offers high-fidelity visual and physical simulation, making it well-suited for environmental realism and sensor testing. However, AirSim lacks a native pedestrian asset class, limiting the ability to simulate complex pedestrian behaviors or interactions. Consequently, scenarios involving vulnerable road users cannot be fully represented, reducing the applicability of AirSim for comprehensive autonomous driving evaluation.

Airsim	**Gazebo**	**LGSVL**
No native pedestrian asset class	Social-force pedestrian model, no gestures	Limited assets, no skeleton-based control

Figure 11. Overview of existing simulators highlighting limitations in pedestrian representation.

Gazebo [42] is a versatile robotics simulator widely used in conjunction with the Robot Operating System (ROS). It incorporates a social-force model to simulate pedestrian movement and crowd dynamics, providing a basic framework for multi-agent interactions. Despite this, Gazebo does not support gestures or skeleton-based control, meaning pedestrians cannot perform nuanced behaviors such as hand signals, sudden evasive actions, or other uncommon motions. This restricts its utility in testing autonomous agents under diverse human behaviors.

LGSVL [2] offers realistic sensor modeling and urban traffic scenarios. While it includes pedestrian agents, these are limited in number and diversity, and the platform does not provide skeleton-based control for complex motion generation. Pedestrian movements are typically predefined along fixed paths, preventing the simulation of spontaneous or interactive behaviors that autonomous agents are likely to encounter in real-world conditions.

Collectively, these limitations illustrate a critical gap in existing simulation platforms: the inability to realistically model pedestrian behaviors, particularly rare or unconventional gestures. This gap motivates the design of the HABIT benchmark, which explicitly incorporates behaviorally diverse pedestrians capable of performing a wide range of motions and gestures. HABIT aims to rigorously evaluate autonomous agents under realistic and challenging traffic interactions. This ensures a more comprehensive assessment of agent competence in environments that closely mimic real-world complexities.

CARLA, by contrast, provides skeleton (bone) control of pedestrian models, enabling far finer-grained articula-

Figure 12. Examples from the ARCANE project (CARLA-BSP): scenes featuring a single pedestrian in crossing/non-crossing situations, following binary or simplified behavior labels.

tion and behavioral richness than many other simulators. However, widespread use of this capability remains unrealized owing to a variety of complications: retargeting motion capture or motion data to CARLA's coordinate systems, resolving mismatches in animation rigs and skeleton hierarchies, and implementing suitable collision, trajectory, and motion blending to avoid artifacts. One recent project, ARCANE [43], used CARLA's pedestrian catalogue for pedestrian intention estimation. Still, this work was constrained: pedestrians only followed pre-defined paths (without dynamic backtracking or deviations), demonstrated minimal interaction with the environment (e.g. no directional changes, no collision avoidance), and often suffered from simulation physics anomalies (e.g. pedestrians 'tunneling' through geometry or going underground) or lack of background traffic context. A few such failure modes are shown in Figure 12

Figure 13. Examples of rare pedestrian behaviours realized through the HABIT framework, demonstrating gestures, non-linear trajectories, interaction with environment, and reactive adaptation.

Our HABIT framework is engineered to overcome these limitations. We systematically retarget open motion sources to CARLA's pedestrian asset skeletons. By integrating velocity grounding, trajectory reconstruction, and collision avoidance, we convert general-purpose motion data into agents who behave as active traffic participants. In Figure 14 we present hand-crafted instances of rare or atypical pedestrian behaviour that are currently only possible with our approach.

C. Retargeting Fidelity

The motion transfer is fundamentally deterministic, with fidelity limited primarily by the mathematical constraints of Euler angle representation (gimbal lock) rather than algorithmic approximation. Extensive ground-truth comparison with SMPL data would not address the fundamental mathematical constraint we've isolated and solved.

Figure 14. Pose blend shape deformation due to Euler angle approximation of SMPL data.

Figure 15. **Retargeting fidelity** At extreme positions the character mesh deforms as a native limitation of Euler angle approximation, however the transfered pose remains skeletally same.

D. Perception realism of HABIT

Zero-shot perception tasks offer a robust method for evaluating visual realism and domain fidelity in simulated environments, particularly when using models trained solely on real-world data. Successful zero-shot performance in detection, segmentation, tracking, and pose estimation tasks strongly indicates alignment with real-world visual and contextual distributions [44–46]. Specifically, human-centric perception tasks implicitly validate realism through accurate modeling of human shape, articulation, and environmental interaction.

We perform zero-shot evaluation using *YOLOv11l-pose*, a keypoint-aware variant of YOLOv11 trained exclusively on real-world data [47, 48], on 12,000 randomly sampled HABIT frames (5 per motion sequence). Results in Table 3 confirm high performance, achieving 93.4% mAP@0.5 for detection and 95.7% PCK@0.5 for pose estimation, closely aligning with real-world results using various YOLOv11 models [44, 45]. These findings validate HABIT's visual and biomechanical fidelity and its utility for benchmarking pedestrian-aware systems [34].

To assess spatial consistency and trackability, we apply the Segment Anything Model (SAM) [8] to YOLO-

Figure 16. **Zero-shot pose estimation on simulated data.** Qualitative results of YOLOv11l-pose applied to HABIT.

Metric	HABIT (Synthetic)	Real-World (COCO)
PCK@0.2	0.692 ± 0.223	0.65–0.75
PCK@0.5	0.957 ± 0.144	> 0.95
MPJPE (px)	9.3 ± 9.6	8–12
OKS	0.418 ± 0.169	0.40–0.55
mAP@[.5:.95]	0.586 ± 0.269	0.55–0.65
mAP@0.5	0.934 ± 0.248	> 0.90
IoU (bbox overlap)	0.766 ± 0.164	0.75–0.80

Table 3. Comparison of pose estimation performance on the HABIT and real-world COCO datasets. Similar trends highlight the visual realism and fidelity of the HABIT benchmark.

pose detections. Using YOLO-derived keypoints for guidance, we generate segmentation masks and track identities across frames. Qualitative results (Figure 17) demonstrate accurate and temporally consistent segmentation, despite SAM being trained solely on real-world data, highlighting HABIT's visual continuity and its suitability for multi-object tracking without domain-specific adaptation."

Figure 17. **Multi-frame detection, segmentation, and tracking.** Consistent track IDs, even in edge-case motions, highlight pose consistency and temporal coherence of HABIT.

We further assess the realism of HABIT pedestrians through segmentation quality using the Segment Anything Model (SAM) trained solely on real-world data. This provides a zero-shot measure of how well simulated pedestrians align with real distributions. **We report two standard metrics:**

- Intersection over Union (IoU): overlap between predicted and reference masks.
- Dice Coefficient (DICE): similarity measure that is more sensitive to boundary alignment.

Table 4. SAM Evaluation Results on Pedestrians of CARLA

Metric	Mean ± Std	Median	90th Percentile
IoU Score	0.7364 ± 0.3333	0.8974	0.9392
DICE Score	0.7824 ± 0.3443	0.9459	–
Pipeline success rate: 95.9%			

Median IoU (0.90) and Dice (0.95) indicate high segmentation fidelity, with over 70% of cases achieving excellent quality. These results confirm that HABIT pedestrians exhibit strong visual and structural realism suitable for benchmarking perception models.