

Supplementary Material: HABIT: Human Action Benchmark for Interactive Traffic in CARLA

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 [1]. 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 1.



Figure 1. **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 2. Overview of existing simulators highlighting limitations in pedestrian representation.

Figure 2 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 [2] 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.

Gazebo [3] 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 [4] 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.



Figure 3. Examples from the ARCAN project (CARLA-BSP): scenes featuring a single pedestrian in crossing/non-crossing situations, following binary or simplified behavior labels.

CARLA, by contrast, provides skeleton (bone) control of pedestrian models, enabling far finer-grained articulation 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, ARCAN [5], 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 3

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 5



Figure 4. Examples of rare pedestrian behaviours realized through the HABIT framework, demonstrating gestures, non-linear trajectories, interaction with environment, and reactive adaptation.

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.

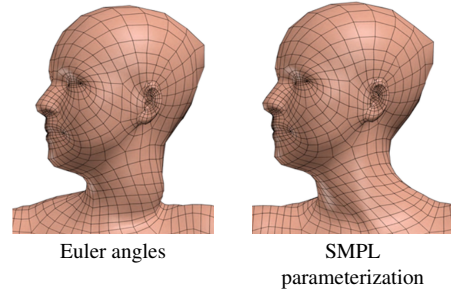


Figure 5. Pose blend shape deformation due to Euler angle approximation of SMPL data.



Figure 6. **Retargeting fidelity** At extreme positions the character mesh deforms as a native limitation of Euler angle approximation, however the transferred 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 [6–8]. Specifically, human-centric perception tasks implicitly validate realism through accurate modeling of human shape, articulation, and environmental interaction.



Figure 7. **Zero-shot pose estimation on simulated data.** Qualitative results of YOLOv11-pose applied to HABIT.

We perform zero-shot evaluation using *YOLOv11-pose*, a keypoint-aware variant of YOLOv11 trained exclusively on real-world data [9, 10], on 12,000 randomly sampled HABIT frames (5 per motion sequence). Results in Table 1 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 [6, 7]. These findings validate HABIT’s visual and biomechanical fidelity and its utility for benchmarking pedestrian-aware systems [11].

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 1. 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.

To assess spatial consistency and trackability, we apply the Segment Anything Model (SAM) [12] to YOLO-pose detections. Using YOLO-derived keypoints for guidance, we generate segmentation masks and track identities across frames. Qualitative results (Figure 8) 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 8. **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 2. SAM Evaluation Results on Pedestrians of CARLA

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

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.

References

- [1] CARLA Team. Carla autonomous driving leaderboard. <http://leaderboard.carla.org>, 2024. 1
- [2] 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>. 1
- [3] 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. 1
- [4] 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. 1
- [5] Maciej Wielgosz, Antonio M. López, and Muhammad Naveed Riaz. CARLA-BSP: a simulated dataset with pedestrians, May 2023. 2

- [6] 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. [3](#)
- [7] 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. [3](#)
- [8] 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. [3](#)
- [9] 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. [3](#)
- [10] 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. [3](#)
- [11] 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. [3](#)
- [12] 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. [3](#)