On the Metric Sensitivity of Heterogeneous Traffic Simulation: An Analysis of the HetroD Challenge via Scene-Coherent Perturbations

Yassine Achour[1] Mohamed Aziz Younes[1] Guillaume-Alexandre Bilodeau[1]
Nicolas Saunier[1] Wassim Bouachir[2]
[1]LITIV Lab, Polytechnique Montréal [2]Téluq University
{yassine.achour, mohamed-aziz-2.younes}@etud.polymtl.ca
{gabilodeau, nicolas.saunier}@polymtl.ca wassim.bouachir@teluq.ca

Abstract

A high simulation score need not imply that agents react to one another. We examine this distinction in the HetroD Challenge v1.0 through SRTG, a deterministic, training-free trajectory generator. Given eleven observed states, SRTG extrapolates paths from an anchored, validity-masked velocity fit and constructs 32 joint rollouts using shared speed profiles, gradual scene translations, and damped yaw modes that rotate the velocity vector. The translations preserve nominal pairwise distance traces, so proximity metrics are unchanged relative to the nominal path. Under the public histogram evaluator and type-balanced dataset aggregation, SRTG scores ***0.5607*** *on 100 held-out scenarios across five locations, against 0.5535 for the translation baseline without rotation (71 wins, 29 losses, paired* $t = 4.939$*,* $p = 3.18 \times 10^{-6}$*). A separate diagnostic supplied with ground-truth futures reaches 0.9010 under the same evaluator. On the development split, increasing the lateral sweep from 3 to 30 metres raises coverage from 0.0383 to 0.3588 but lowers the overall score from 0.9017 to 0.8712. The results motivate scoring reactive responses separately from proximity preservation.*

1. Introduction

SRTG scores 0.9010 when its nominal trajectories are taken from the future ground truth, yet its official dataset-aggregated score is 0.5607 when it extrapolates from observed history (0.5535 without the rotation modes). The comparison is diagnostic rather than a measure of interactive fidelity: the GT-assisted setup measures the effect of diversifying an already known future, and leaves the open-loop forecasting problem unresolved. Histogram evaluation confirms that preserving geometric proximity does not substitute for predicting actual traffic maneuvers.

HetroD [3] makes that distinction concrete. Each submission contains 32 joint futures for vehicles, pedestrians, and two-wheelers, spanning eight seconds after eleven observations at 10 Hz. The score combines kinematic likelihood, collisions and valid-region compliance, cross-type closest approach, and spatial coverage. Safety and cross-type interaction receive additive weights of 0.35 and 0.25. Coverage contributes at most 0.10, and that term is multiplied by kinematic realism and safety. A broader set of futures is therefore useful only when its occupancy gain compensates for changes in those other terms.

We study this trade-off with shared geometric perturbations. Translating every simulated agent by the same time-dependent vector leaves each pairwise distance trace unchanged. This is an exact property of translation, not evidence of yielding or negotiation. It also holds relative to the nominal trajectory, which may already predict the wrong encounter. SRTG combines this restricted form of diversification with shared speed modulation and damped yaw extrapolation. It is an inexpensive geometric baseline and a probe of what the score rewards. Our analysis separates the translation invariant from the properties that still require measurement: map compliance, timing under speed modulation, and realism when the future is hidden.

Scope of the report. We describe the generator, compare internal development variants, and report location-balanced validation results. The development study uses known futures to isolate perturbation effects; the history-only study evaluates extrapolation. We retain this distinction throughout. Neither a high score in the first setting nor exact distance preservation establishes reactive simulation. The practical contribution is a reproducible baseline whose restrictions make these differences visible.

2. Related Work

Rollout diversification in simulation. Closed-loop benchmarks such as the Waymo Open Sim Agents Challenge [5]

and nuPlan [1] score multiple sampled futures for realism and diversity. When a logged future is available as a reference path, structured geometric perturbations are a cheap baseline. Constant-velocity extrapolation remains a strong reference on pedestrian data [8]. We use that same kinematic core and add shared speed, translation, and damped-yaw modes so that the HetroD score can be probed without a learned policy.

Learning-based trajectory generators. Trajectron++ [7], Motion Transformer [9], and MotionDiffuser [4] predict diverse futures from history through trained models. SRTG has no learned parameters. It is not a substitute for those methods; it is a diagnostic baseline against which their interaction claims can be compared.

Social Force Models. The Social Force Model [2] and later tracking extensions [6] couple agents through repulsive potentials. In our development probes, SFM rollouts were not constrained to match the last observed state, and the kinematic score dropped accordingly. We treat that as a junction-initialization issue, not as a full appraisal of social-force dynamics.

3. Methodology

3.1. Task and Evaluation Settings

The input contains eleven observed states, indexed $t = 0, \dots, 10$, and the output contains 32 joint rollouts at $t = 11, \dots, 90$. The sampling interval is 0.1 s. Observations therefore span one second between their first and last timestamps, and the prediction horizon is eight seconds. No network is trained and no pretrained model is used.

We distinguish two experiments. In history-only generation, the nominal path is extrapolated from observations. In the GT-assisted diagnostic, a recorded future supplies the nominal path so that perturbation effects can be studied separately. Future targets are used for scoring history-only predictions, never for constructing them. Hyperparameter selection on validation data remains calibration even though it involves no model training.

3.2. Nominal Trajectory

For agent i, the unperturbed history-only rollout follows constant velocity:

$$\mathbf{p}_0^{(i)}(t) = \mathbf{p}^{(i)}(t_0) + \mathbf{v}^{(i)}(t_0) \cdot (t - t_0)\Delta t, \quad t \in \{11, \dots, 90\} \tag{1}$$

$$\psi_0^{(i)}(t) = \psi^{(i)}(t_0). \tag{2}$$

Here $t_0 = 10$ and $\Delta t = 0.1\,\text{s}$. Velocity is estimated by a recency-weighted linear fit on the last five observed positions, using validity masks and iterative Huber reweighting, anchored at the last observation. Stored velocity is used only when that window is unusable. Speeds below 0.1 m/s are set to zero. The same parser is used for holdout ground-truth files and ScenarioNet test inputs, so the reported validation numbers and the submitted rollouts share one initialization.

Static agents remain fixed in the nominal and speed-modulated modes. They can still move under the shared lateral translation described below.

3.3. The Observation-Prediction Boundary

The evaluator joins the observed and simulated states into $\mathbf{P} \in \mathbb{R}^{91\times 4}$. Its central-difference velocity at the first predicted frame is

$$\mathbf{v}(11) = \frac{\mathbf{p}_{\text{sim}}(12) - \mathbf{p}_{\text{GT}}(10)}{2\Delta t}. \tag{3}$$

Linear acceleration is the central difference of speed magnitude. A position error near the junction can therefore affect several derivative samples, with a scale proportional to $\delta/\Delta t^2$. Its exact contribution depends on the stencil and motion direction. The default evaluation configuration uses smoothed histograms, so a small jump does not necessarily receive zero probability.

Introducing the lateral offset. We apply the displacement gradually:

$$\rho(\tau) = \left(\frac{\tau}{T-1}\right)^{\alpha}, \quad \tau \in \{0, \dots, T-1\},\ T = 80. \tag{4}$$

The local index is $\tau = t - 11$, and $\alpha = 1.70$ was selected on the 30-scenario development split. The continuous extension satisfies

$$\rho(0) = 0, \quad \left.\frac{d\rho}{d\tau}\right|_{\tau=0} = 0. \tag{5}$$

These conditions prevent the perturbation from adding an instantaneous position or velocity jump at its onset. They do not repair a discontinuity in the nominal path. Nor is this ramp twice continuously differentiable: its second derivative is proportional to $\tau^{-0.30}$. The sampled derivatives are finite, but a universal acceleration bound cannot be inferred from $\mathcal{C}^1$ continuity alone.

3.4. Constructing the 32 Rollouts

SRTG constructs 32 joint rollouts in four families: one nominal path ($r = 0$), fifteen longitudinal speed variations ($r = 1, \dots, 15$), eight shared lateral translations ($r = 16, \dots, 23$), and eight damped yaw rotation modes ($r = 24, \dots, 31$). Each mode generates futures for all required agents.

Nominal mode ($r = 0$). This mode propagates the anchored estimated velocity $\mathbf{v}_0^{(i)}$ without perturbation and is the reference path for the translation family.

Speed modulation modes ($r = 1, \ldots, 15$)**.** A terminal scale factor modulates speed along the initial velocity direction:

$$v_r^{(i)}(\tau) = v_0^{(i)} \cdot [1 + (k_r - 1) \cdot S(\tau/(T-1))], \quad (6)$$

where $S(s) = 3s^2 - 2s^3$ satisfies $S'(0) = S'(1) = 0$. The fifteen k_r values are uniformly spaced over $[0, 2]$, from a stop ($k = 0$) to a doubling of speed ($k = 2$). Positions are updated by cumulative integration:

$$\mathbf{p}_r^{(i)}(\tau) = \mathbf{p}_0^{(i)}(0) + \hat{\mathbf{v}}_0^{(i)} \sum_{u=0}^{\tau} v_r^{(i)}(u)\Delta t. \quad (7)$$

Translation modes ($r = 16, \ldots, 23$)**.** All agents receive an identical time-ramped spatial translation along the perpendicular scene axis $\hat{\mathbf{n}}_\perp = (-\sin\bar{\psi}, \cos\bar{\psi})$:

$$\mathbf{p}_r^{(i)}(\tau) = \mathbf{p}_0^{(i)}(\tau) + \lambda_r \cdot \rho(\tau) \cdot \hat{\mathbf{n}}_\perp, \quad \forall i \in \mathcal{A}, \quad (8)$$

with eight amplitudes λ_r spanning $[-3, 3]$ m.

Damped yaw rotation modes ($r = 24, \ldots, 31$)**.** To follow observed curvature without a heading-only re-projection, we extrapolate yaw rate with exponential damping and rotate the initial velocity vector:

$$\Delta\psi_r(t) = \gamma_r \cdot c_i \cdot \hat{\omega}_i \cdot \tau_\omega \left(1 - e^{-t/\tau_\omega}\right), \quad (9)$$

$$\mathbf{v}_r^{(i)}(t) = \mathbf{R}(\Delta\psi_r(t))\mathbf{v}_0^{(i)}, \quad (10)$$

$$\mathbf{p}_r^{(i)}(\tau) = \mathbf{p}_0^{(i)}(0) + \sum_{u=0}^{\tau} \mathbf{v}_r^{(i)}(u)\Delta t, \quad \psi_r^{(i)}(\tau) = \psi_0^{(i)} + \Delta\psi_r(\tau), \quad (11)$$

where $\mathbf{R}(\theta)$ is the planar rotation matrix, $\tau_\omega = 2.0$ s, and $\gamma_r \in \{-0.5, -0.25, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$. Stationary agents ($\|v_0\| < 0.1$ m/s) keep $\omega_{\text{eff}} = 0$. Positions are integrated with the rotated velocity at interval midpoints, which keeps heading and displacement consistent after the last observation.

3.5. What Scene Coherence Preserves

For a common translation, the additive displacement cancels:

$$\|\mathbf{p}_r^{(i)}(\tau) - \mathbf{p}_r^{(j)}(\tau)\| = \|\mathbf{p}_0^{(i)}(\tau) - \mathbf{p}_0^{(j)}(\tau)\|. \quad (12)$$

Consequently, minimum distance and its time are unchanged relative to the nominal future. With unchanged box orientations and dimensions, translating both agents also preserves their box-overlap relation. These statements require both members of every evaluated pair to receive the same displacement.

If the nominal future is GT, these properties explain the cross-type score of 1.0000 in the translation diagnostic. If it is constant velocity, incorrect encounters remain incorrect. Translation also leaves road polygons fixed, so it cannot guarantee valid-region compliance. The near-perfect GT-assisted safety score is a measured result, not a general guarantee.

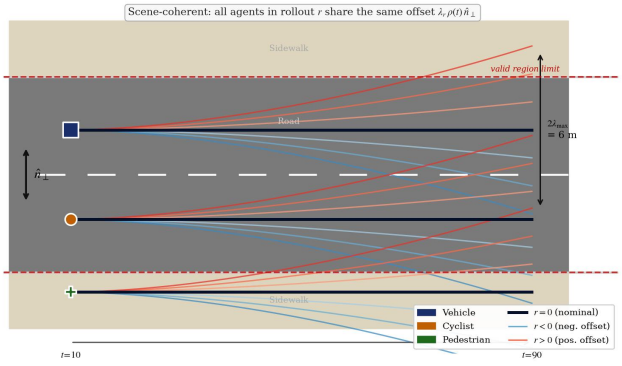

Figure 1. Shared translation of a nominal scene. At each time, the displacement is identical for all simulated agents. Distance preservation applies to this translation family relative to its nominal paths; it does not apply to the speed-modulated family.

Score trade-off. When all quality terms apply, the score is

$$S = \underbrace{0.30\, S_{\text{kin}} + 0.35\, S_{\text{saf}} + 0.25\, S_{\text{cross}}}_{\text{Quality Terms (Weight = 0.90)}} + \underbrace{0.10\, S_{\text{cov}} \cdot S_{\text{kin}} \cdot S_{\text{saf}}}_{\text{Coverage Bonus (Max = 0.10)}}. \quad (13)$$

The evaluator renormalizes the quality weights when cross-type interaction is inapplicable. Otherwise, at the reported holdout component means, a 0.10 increase in coverage contributes about 0.0023. A 0.02 decrease in safety costs about 0.0071. This local calculation motivates conservative diversification; it does not imply that every coverage/safety trade is unfavorable.

4. Experiments and Benchmark Analysis

4.1. Implementation and Evaluation Protocol

Team LITIV submits SRTG without learned weights, external data, or map-conditioned generation. The generator is analytical constant-velocity propagation with deterministic speed, translation, and damped-yaw modes; it has no trainable parameters. There is no training stage. Ramp and sweep amplitudes were selected on validation data, and no pretrained models were used. The author block lists all team members and affiliations. Each scene produces 32 arrays of 80 future states in global (x, y, z, yaw) coordinates. Scores below were computed locally with the public HetroD evaluator (`hetrod-0.8.0`) under the 2025 histogram textproto.

Table 1. Development progression. Baseline through per-agent calibrated: 5-scenario `loc3` probe. SRTG: GT-assisted translation diagnostic on the 30-scenario development set.

Method	Overall ↑	Kin. ↑	Saf. ↑	CT ↑
Official Baseline (GT+Noise)	0.8505	0.9818	0.8889	0.9777
SFM (Social Force)	0.6252	0.6784	0.6408	0.7427
SFM + Linear Ramp	0.6335	0.6661	0.6600	0.7861
SFM + Heading Align.	0.6520	0.6883	0.6901	0.7888
Ramped Fan-Out	0.8574	0.9931	0.8982	0.9729
Per-Agent Calibrated	0.8595	0.9946	0.9099	0.9629
SRTG (Ours)	**0.9017**	**0.9988**	**0.9951**	**1.0000**

They are validation records, not hidden-test leaderboard numbers. Generation is CPU-only.

The development split contains 30 scenarios from `loc3`. It was used to choose the ramp and sweep amplitude with GT futures available as nominal paths. The history-only split contains 100 scenarios, twenty each from `loc0`, `loc1`, `loc3`, `loc4`, and `loc5`, with no scenario overlap with development. Location `loc2` is test-only. Scenario disjointness does not by itself establish independence between nearby windows from the same recording.

The official evaluator balances samples and agent types within each location before taking its location macro-average. Where we also report a paired t-test, that test is computed on per-scenario scores and is not interchangeable with the official aggregate.

4.2. Benchmark Progression on Development Split

Table 1 summarizes internal development variants. Early rows were measured on a 5-scenario `loc3` probe; the SRTG row is the GT-assisted translation diagnostic on the 30-scenario development split.

Interpreting the SFM comparison. On the 5-scenario probe, social-force variants score from 0.6252 to 0.6520, compared with 0.8574 for ramped fan-out. Their kinematic scores are also lower, but these experiments change more than boundary initialization. They support checking the junction and the resulting speed distributions; they do not isolate the fraction of the loss caused by either effect. A boundary-only ablation would be needed for that attribution.

The effect of shared translation. On the 30-scenario development split, the GT-assisted translation diagnostic reaches 0.9017, with safety 0.9951 and cross-type interaction 1.0000. The earlier per-agent calibrated variant scored 0.8595 on the 5-scenario probe and is not a matched ablation. The translation invariant explains why cross-type interaction can remain 1.0000 when nominal futures are supplied by GT. It does not show that the translated rollouts contain more realistic responses between agents.

Table 2. Ablation on maximum lateral sweep $\lambda_{\max}$ (Development set, $N = 30$). GT-assisted translation diagnostic: cross-type interaction remains 1.0000 for all tested offsets.

$\lambda_{\max}$	Overall ↑	Safety ↑	Coverage ↑	CrossType ↑
3.0 m	**0.9017**	**0.9951**	0.0383	**1.0000**
5.0 m	0.9014	0.9885	0.0667	**1.0000**
8.0 m	0.8999	0.9791	0.1087	**1.0000**
12.0 m	0.8962	0.9659	0.1636	**1.0000**
20.0 m	0.8855	0.9379	0.2634	**1.0000**
30.0 m	0.8712	0.9083	**0.3588**	**1.0000**

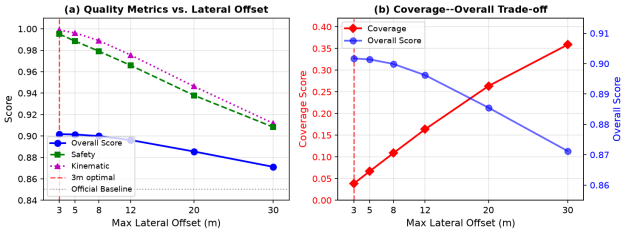

Figure 2. Ablation of maximum lateral sweep $\lambda_{\max}$ on the development set. **Left**: Quality metrics decline as $\lambda_{\max}$ increases. **Right**: Coverage versus overall score; $\lambda_{\max} = 3.0$ m is the best tested setting under Eq. 13.

4.3. Ablation: The Safety vs. Coverage Trade-off

Table 2 and Figure 2 detail the ablation of the maximum lateral sweep $\lambda_{\max}$ across six magnitudes on the development set.

Increasing the sweep from 3 to 30 m raises coverage from 0.0383 to 0.3588 while reducing safety from 0.9951 to 0.9083. The overall score falls by 0.0305. Map violations are consistent with this pattern, but the safety total alone does not identify which agent types or map boundaries account for the loss. The change in safety cannot explain the total score decrease without the corresponding kinematic and coverage contributions. The 3 m setting is the best tested amplitude on this development split, not a universal optimum.

4.4. Held-out Validation Evaluation (Masked-History Protocol)

Table 3 reports the validation results under the official textproto histogram configuration and type-balanced dataset aggregation. The generator uses only observed history ($t = 0, \ldots, 10$).

Known futures versus extrapolated futures. The GT-assisted diagnostic reaches 0.9010 under dataset aggregation (kinematic realism 0.9983, safety 0.9942, cross-type 1.0000). When extrapolating from history only, the official score is 0.5607. Geometric translation preserves proximity metrics when a valid future is already given; open-loop extrapolation

Table 3. History-only evaluation on the held-out validation set ($N = 100$) using the public histogram configuration and type-balanced dataset aggregation. Per-location σ is the population standard deviation of the 20 scenario scores. The dataset row is the official location-macro aggregate.

Location ($N = 20$)	**Overall** ($S \pm \sigma$)	**Kin.** ↑	**Saf.** ↑	**CT** ↑	**Cov.** ↑
`loc0`	$\mathbf{0.6719 \pm 0.0798}$	**0.4464**	**0.8977**	**0.8764**	0.1174
`loc1`	0.5006 ± 0.1114	0.2347	0.6837	0.7550	**0.1301**
`loc3` (holdout)	0.5669 ± 0.0882	0.2995	0.7912	0.7905	0.1048
`loc4`	0.5620 ± 0.1238	0.2629	0.8245	0.7672	0.1279
`loc5`	0.5020 ± 0.0930	0.2387	0.7087	0.7213	0.1161
Dataset Aggregated ($N = 100$)	**0.5607**	**0.2965**	**0.7812**	**0.7821**	**0.1193**

remains limited by unobserved intentions and road curvature.

Paired ablation: effect of damped yaw modulation. On the same 100 holdout scenarios, the translation baseline without rotation scores 0.5535. Adding the damped rotation modes raises the official aggregate to 0.5607. Per-scenario scores favor SRTG in 71 cases and the baseline in 29 (mean difference $+0.0182 \pm 0.0366$, paired $t = 4.939$, $p = 3.18 \times 10^{-6}$). The official-component gains are $+0.0241$ in kinematic realism (0.2965 vs. 0.2724) and $+0.0132$ in safety (0.7812 vs. 0.7680), against a cross-type drop of 0.0207 (0.7821 vs. 0.8028). Coverage rises from 0.1013 to 0.1193. The paired mean and the official aggregate differ because the latter is a type-balanced location macro-average, not a mean of scenario scores.

5. Discussion and Benchmark Recommendations

5.1. Limits of the Generator

Shared translation preserves a nominal encounter but cannot decide whether an agent should yield. Shared speed modulation adds temporal variation, yet it imposes the same terminal factor on agents with different intentions. An initially stationary agent remains stationary in the speed modes, while lateral modes can move it sideways. These are restrictions of the generator, not properties of heterogeneous traffic.

The development sweep exposes a second limit. A tenfold increase in coverage accompanies a lower score, and the best amplitude is selected from one location. A single scene axis also becomes ambiguous when vehicles travel in opposite directions. Extending this construction to intersections would require conditioning motion on local routes and separating nearby interacting groups from unrelated agents. Such extensions would sacrifice the global invariant and would need fresh evaluation.

5.2. What the Evaluation Should Make Visible

Report contributions, not only a total. At the history-only component means, the gated coverage term contributes 0.0028. Reporting that contribution alongside the raw coverage score would make its influence explicit. The development change from 3 to 30 m is a useful diagnostic case: coverage rises by 0.3205 while the total falls by 0.0305. Alternative weightings should be tested on such cases before choosing a replacement objective.

Separate proximity from response. Cross-type interaction reaches 1.0000 in the GT-assisted translation experiment despite the absence of behavioral feedback. A complementary intervention test could brake a leading vehicle and measure the timing and magnitude of the follower's response. This would ask a different question from whether minimum distance and closest-approach time resemble the recorded encounter.

Expose junction errors without hiding them. Report kinematic diagnostics near the first predicted frames separately from the rest of the horizon. This would help determine whether a low likelihood arises from initialization or sustained motion. Automatically smoothing submitted trajectories would change the object being evaluated; a documented initialization convention and an explicit junction diagnostic would be easier to interpret.

6. Conclusion

SRTG provides a training-free baseline for studying how HetroD scores joint futures. Shared translations preserve nominal pairwise distance traces, whereas shared speed profiles change their timing, and damped yaw rotation modes follow observed curvature without splitting heading from displacement. The distinction between problem settings matters: the GT-assisted diagnostic reaches 0.9010 under dataset aggregation, whereas official history-only evaluation reaches 0.5607. Our sweep ablation also shows that substantially greater coverage can accompany a lower total score. The supported claim concerns geometric invariance and continuous kinematic extrapolation, not reactive traffic simulation. The next step is to test whether map-conditioned paths and explicit yielding improve history-only performance while retaining sensible responses to interventions that were absent from the observed scene.

References

[1] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. In *CVPR Workshops*, 2022. 2

[2] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995. 2

[3] HetroD Challenge Organizers. HetroD Challenge v1.0: Heterogeneous driving agent simulation. https://github.com/bob020416/HetroD-Challenge-v1.0-Evaluation, 2026. ECCV 2026 DriveX Workshop. 1

[4] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, and Dragomir Anguelov. MotionDiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, pages 9644–9653, 2023. 2

[5] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, Brandyn White, and Dragomir Anguelov. The waymo open sim agents challenge. In *NeurIPS*, 2023. 1

[6] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009. 2

[7] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, pages 683–700, 2020. 2

[8] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020. 2

[9] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. In *NeurIPS*, 2022. 2