

Peer Review

Review of: "Don't Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper starts from a problem that anyone who has watched a multi-modal planner frame by frame will recognise: the top-scoring trajectory can jump between modes from one timestep to the next, so the car's intended path "shakes" even when nothing in the scene has changed. The authors propose MomAD, built on SparseDrive, with two planning components. Topological Trajectory Matching (TTM, §3.1, Eq. 1–3) transforms the current candidates into the ego frame of the previous step (Eq. 1, "from moment t to $t-1$ ") and picks, among the $K = 6 \times 3$ candidates, the one with the smallest Hausdorff distance to it. The Momentum Planning Interactor (MPI, §3.2, Figure 3) then lets the selected planning query cross-attend to the previous frame's planning queries, mixed through an LSTM, and regenerates the trajectories. A third piece, instance denoising via Gaussian perturbation during training (§3.3), targets the perception features the planner depends on.

To measure what they care about, the authors define a Trajectory Prediction Consistency metric (TPC, Eq. 6) and build a Turning-nuScenes validation subset (17 scenes, 680 samples, Appendix A.2). The main results are open-loop on nuScenes (Table 1), on Turning-nuScenes (Table 2, Table A1), at 4–6 s horizons (Table 3, Table A2), under corruptions from nuScenes-C (Table A3), and closed-loop on Bench2Drive (Table 4). The abstract reports a 26% collision reduction and a 0.97 m (33.45%) TPC improvement at 6 s on Turning-nuScenes, and up to 16.3% higher success rate on Bench2Drive. Code is linked.

2. Whether the conclusion holds on the data

What I value most here is that the paper names a real failure mode and then builds the instruments to see it. The Turning-nuScenes subset is a sensible response to the known dominance of straight driving in

nuScenes, and the long-horizon comparison in Table 3 is where the method's advantage is clearest: on Turning-nuScenes at 6 s, L2 falls from 3.36 to 2.51 (Table 3, row T-nuScenes, L2, column 6s, SparseDrive vs MomAD), and TPC from 2.90 to 1.93. I also appreciate that Table 4 includes a "MomAD (Euclidean)" row, which is exactly the control one wants for the Hausdorff choice, and that the ablations in Tables 6–8 are internally consistent with the headline tables. I checked: the full-model row of Table 6 ($\sigma = 0.1$, both modules) reproduces Table 5's MomAD perception numbers (mAP 0.423, NDS 0.531, AMOTA 0.391) and Table 1's VAD-protocol planning averages (L2 0.60, Col. 0.09, TPC 0.54). The $t = 2$ row of Table 7 matches Table 2's MomAD row exactly in the columns they share (2s, 3s, Avg.). That kind of consistency is not a given.

Where I think the text runs ahead of the numbers is in a handful of specific places.

Full nuScenes, VAD protocol. §4.2.1 says MomAD "shows SOTA performance in planning results". In Table 1 (lower block, VAD protocol), SparseDrive has L2 Avg. 0.61 and MomAD 0.60, but SparseDrive's collision rate is lower: Col. Rate Avg. 0.08 against 0.09, and at 3s 0.18 against 0.22. L2 at 1s is also 0.29 for SparseDrive and 0.31 for MomAD. So on the full set the honest reading is: comparable accuracy, slightly higher collision rate, clearly better consistency (TPC Avg. 0.57 to 0.54; 3s 0.85 to 0.78). I would say exactly that. The gain is in TPC and in turning scenes, and the paper does not need the broader claim.

The "26%" collision figure. The abstract says MomAD "reduces the collision rate by 26%" at 6 s on Turning-nuScenes. Table 3 (row T-nuScenes, Col. Rate (%), column 6s) gives 2.57 for SparseDrive and 2.31 for MomAD. The difference is 0.26 percentage points. As a relative reduction it is $0.26 / 2.57 = 10.1\%$. I believe the abstract has turned an absolute difference in a column already expressed in percent into a relative percentage. It should read "by 0.26 percentage points (about 10%)".

The TPC percentage. The abstract gives the 6 s TPC gain as 0.97 m (33.45%); §4.2.1, "Long Trajectory Prediction" paragraph, gives the same 0.97 m as 32.45%. From Table 3 (row T-nuScenes, TPC, column 6s): $0.97 / 2.90 = 33.45\%$. The abstract is right and the body has a typo.

One delta in Table 3. In the nuScenes block, L2 at 4s is 1.75 for SparseDrive and 1.67 for MomAD, a difference of 0.08, but the delta row shows -0.09 and the text says "0.09m (5.14%)". 5.14% is $0.09/1.75$, so the error propagated. With 0.08 it is 4.57%. The same mismatch appears in Table A2. Either a cell or the delta is wrong.

Bench2Drive percentages. The text under Table 4 says MomAD improves success rate "by 16.3% and 8.4% over the VAD multi-modal variant and SparseDrive". I recomputed from Table 4. Against

SparseDrive*, SR 16.71 to 18.11 is $(18.11 - 16.71)/16.71 = 8.4\%$, which matches. But 16.3% is the gain over plain VAD (15.00 to 17.44: $2.44/15.00 = 16.3\%$), not over VADmmt, where it is $1.53/15.91 = 9.6\%$. The Comfort figure has the same issue: 7.2% is against VAD (46.01 to 49.34), and against VADmmt (47.22) it is 4.5%. I would also report absolute differences. On 220 routes, 1.40 SR points is roughly three routes ($0.014 \times 220 = 3.1$). CARLA closed-loop results move by that much between seeds, so without repeated runs this row supports "no worse, probably somewhat better", not a firm 8.4% gain. I noticed the SR values (e.g. 17.44, 18.11) are not multiples of $1/220$; if they are averages over several runs, saying so and adding the spread would strengthen the table considerably.

nuScenes-C "on average". Appendix A.5 says MomAD beats SparseDrive on nuScenes-C "by 22.9% (detection), 27.1% (tracking), 25.1% (mapping), 24.2% (motion), and 40.0% (planning) on average". I checked each against Table A3, and every one of these figures is the *Snow* row alone: mAP 0.140 to 0.172 is 22.9%; AMOTA 0.133 to 0.169 is 27.1%; mapping mAP 22.3 to 27.9 is 25.1%; mADE 0.95 to 0.72 is 24.2%; Col. 0.30 to 0.18 is 40.0%. Averaged over Snow, Rain and Fog, detection mAP improves by $(22.9 + 16.4 + 18.4)/3 = 19.2\%$. The direction holds, but the phrase "on average" should either be dropped or the averages recomputed.

The metric and the method pull in the same direction. TTM selects the candidate closest (in Hausdorff distance) to the previous plan, and TPC (Eq. 6) rewards small distance between the current and previous plan. Part of the TPC improvement is therefore expected by construction. This does not make the result wrong, and the authors are right to report TPC next to L2 and collision. I would still want to see two things. First, a TTM-only row (selection without MPI) on Turning-nuScenes, so the reader can separate "we picked the consistent mode" from "we planned better". Second, a trivial reference: re-using the previous plan shifted by one step. That baseline would have excellent TPC and poor L2, and would show the reader where on the consistency/accuracy trade-off MomAD sits.

Sample size and variance. Turning-nuScenes is 17 scenes and 680 samples (Appendix A.2). Several differences in Tables 2, 7 and 8 are 0.01–0.04 in collision rate (e.g. Table 2, Col. Rate, column 1s: 0.04 vs 0.03). At that scale a single collision event moves the number. No table reports seeds, repeated training runs, or confidence intervals. I would at least report the mean and spread over three training seeds for Table 2 and Table 3, or a per-scene bootstrap interval, which costs nothing extra to compute.

Motion prediction. §4.2.2 says MomAD "outperforms SparseDrive and UniAD" in motion prediction. In Table 5 it does on mADE (0.61 vs 0.62), mFDE (0.98 vs 0.99) and EPA (0.499 vs 0.482), but MR is 0.137 for MomAD against 0.136 for SparseDrive. A small point; "better on three of four motion metrics" is accurate.

3. Reproducibility

This part is better than average. The code repository is linked in the abstract. Appendix A.3 gives the backbone (ResNet50), input size (256×704), perception ranges, the two-stage schedule, stage-2 epochs (10), total batch size (48), learning rate (3×10^{-4}), the loss weights (appendix Eq. 5–7), and the hardware (8 $\times$ RTX 4090 24 GB). Bench2Drive is pinned to V0.0.3 with the base training set and the official 220 routes (Table 4 caption and §4.1). Baselines marked * were reproduced from official checkpoints (Table 1 caption).

What I could not find, or could not read, is the following.

- Random seeds and the number of training runs per configuration. Nothing in the text says whether any number is a mean.
- Stage-1 epochs and learning-rate schedule. Only stage 2 is specified.
- The exact list of Turning-nuScenes sample tokens. The selection rule is given (Appendix A.2, formula 2), but see the next point; releasing the token list would remove all doubt.
- The unit of the threshold $\varepsilon = 25$ in formula 2, and the frame it is applied in. The appendix says the rule assumes "the ego vehicle's driving direction aligns with the y-axis of the world coordinate system". In nuScenes, the global frame is fixed per map, and the ego heading varies, so a vehicle driving straight along the global x-axis would show a large x-change. I could not tell from the text whether the future trajectory is taken in the ego frame (where the rule is sensible) or in the world frame (where it is not). The main text (§4.1) describes the same rule as a threshold "between 3s and 0.5s of 'gt_ego_fut_traj's' to 25", which suggests the ego frame; stating this explicitly would settle it.
- Which evaluation protocol Tables 2 and 3 use. Their captions say VAD; the captions of Tables A1 and A2, which contain the same MomAD and SparseDrive numbers, say ST-P3.
- As rendered on Qeios, the column headers of Tables 6, 7, and 8 (the ED, MP, σ , t, and Add/LQM/TP indicators) are reduced to tick marks and are partly not legible. I reconstructed Table 7 from the text, but in Table 8, I could not tell with confidence which modules the second row enables.
- Inference latency is given (FPS 7.8, Table 1), but on RTX 4090 for MomAD and on A100 for UniAD, so the FPS column compares different hardware. A note in the caption would help.

4. What specifically to fix

1. Abstract, collision sentence: replace "reduces the collision rate by 26%" with the absolute and relative values from Table 3, row T-nuScenes, Col. Rate, column 6s (2.57 to 2.31, i.e. 0.26 points, about 10%).
2. §4.2.1, "Long Trajectory Prediction": 32.45% should be 33.45% (0.97/2.90).
3. Table 3 and Table A2, nuScenes block, L2 at 4s: reconcile $1.75 - 1.67 = 0.08$ with the reported -0.09 and the 5.14% in the text.
4. Text under Table 4. The 16.3% SR gain and the 7.2% Comfort gain are relative to VAD, not VADmmt; against VADmmt they are 9.6% and 4.5%. I would state which base each MomAD row is built on (the table leaves the reader to infer it from grouping) and give absolute SR differences next to the relative ones.
5. Appendix A.5: the "on average" percentages are the Snow row of Table A3. Recompute them over Snow, Rain and Fog, or say "under snow".
6. §4.2.1, nuScenes paragraph: soften "SOTA performance in planning results". Table 1 (VAD protocol) shows a higher collision rate than SparseDrive (0.09 vs 0.08 Avg.; 0.22 vs 0.18 at 3s). The consistency result stands on its own.
7. Add a TTM-only ablation and a "previous plan, shifted" reference row to Table 7 or Table 8, so that the TPC gain can be separated from the selection criterion that directly targets it.
8. Report seeds and variance, at minimum for Tables 2, 3 and 4. For Bench2Drive, three evaluation runs with mean and range would already change how much weight a 1.4-point SR difference can carry.
9. Appendix A.2, formula 2: state the unit of ε and the coordinate frame, and release the Turning-nuScenes token list alongside the code.
10. Captions of Tables 2/3 vs A1/A2: use one protocol name. In A.4 "Tables A2 and A3 in the main" should presumably read "Tables 2 and 3".
11. §4.2.2: "outperforms" in motion prediction should acknowledge MR (Table 5, 0.137 vs 0.136).
12. Table 6: if the first row (no ED, no MP, $\sigma = 0.0$) is meant to be the SparseDrive reproduction, its Col. 0.14 differs from SparseDrive's 0.08 in Table 1 (VAD protocol). A sentence explaining what that row is would prevent the question.
13. Smaller things: "Collition Rate" (§4.1), "Turing-nuScenes" (§4.2.1), "molels" (§4.2.1), "perception files" (abstract, Figure 3 caption; presumably "fields"), "Planing" (§4.3 heading).

5. Recommendation

Minor revisions. The idea is simple, well motivated and cheap to add to an existing planner, and the evidence for better temporal consistency, especially at long horizons and in turns (Table 3), is convincing in direction. The fixes I ask for are mostly about stating the numbers precisely: one abstract figure is an absolute difference presented as a relative one, a few percentages are computed against the wrong baseline or a single row, and there is no variance anywhere. None of that requires new modelling. The two extra ablation rows and seed runs would take some compute, but they would turn a good paper into one whose consistency claim cannot be read as a by-product of the selection rule.

Declarations

Potential competing interests: No potential competing interests to declare.