StyleFinger: Vehicle-Level Interaction Tendencies for Controlled Interaction Behavior in Autonomous-Driving Tests

Anonymous authors
Paper under double-blind review

Abstract

Interactive autonomous-driving tests need surrounding vehicles that can respond differently to the same interaction demand while remaining compatible with observed traffic behavior. Replay preserves realism but fixes one response, whereas an unstructured latent sample does not tell a tester which response factor changed. We introduce StyleFinger, which extracts a low-dimensional vehicle-level interaction tendency from a history-only segment and exposes it as a condition separate from the current scenario. Under a fixed target history and local interaction scenario, with logged partner replay, decoder, and eight-candidate budget, we replace only the tendency code and re-evaluate the same interaction-intention adapter. On a vehicle-disjoint, non-overlapping NGSIM two-car protocol, the code gives a retrieval AUC of 0.754 (and 0.631 with a 250-frame gap), produces response dispersion of 0.637 ± 0.013, and reaches higher observed-future matching and coverage than matched repetition (0.661 ± 0.032/0.893 ± 0.039 versus 0.528 ± 0.019/0.742 ± 0.039). A training-time tendency shuffle retains dispersion but lowers observed-future matching to 0.554 ± 0.048, separating structured variation from arbitrary spread. These results support an auditable fixed-scenario response condition for interaction-focused testing; they do not establish a semantic aggressiveness axis, full behavior-space coverage, causal control, or closed-loop safety improvement.

1 Introduction

Interactive autonomous-driving tests need surrounding vehicles that can respond differently to the same interaction demand. A merge, following episode, or conflict-resolution event can admit several plausible responses even when road geometry and local states are unchanged. A test suite that exposes only one replayed response can therefore miss failures that depend on how another vehicle reacts. The goal is to vary the response while preserving the interaction demand being tested.

Existing generation interfaces leave a gap between these requirements. Replay preserves empirical realism but fixes one response. Multimodal predictors can sample several futures, yet a latent sample does not tell a tester which response factor changed. Scene perturbations alter the situation itself, making it difficult to separate a new interaction demand from a new response. Driving-style or behavior embeddings provide a possible handle, but they are often evaluated for recognition or prediction rather than as a matched, vehicle-level response interface. The missing piece is an auditable condition that can be inferred from history, substituted for one target vehicle, and evaluated while the local interaction demand and partner behavior remain fixed.

We address this gap with StyleFinger. Figure 1 summarizes the three stages: a non-overlapping source interaction window provides historical evidence, the local interaction scenario and partner replay are represented together with the target history, and a fixed-scenario intervention swaps the tendency code z while the scenario and replay remain unchanged. A history-only source interaction window, containing the target and available neighbor histories but no future coordinates, is mapped to a continuous vehicle-level interaction tendency, while the current target history and local interaction scenario provide the interaction demand. The tendency is operational: it summarizes response-relevant regularities in historical interaction trajectories. It is not a psychological personality label, a vehicle-

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

identity claim, or a universal aggressive–conservative coordinate. Given target history H_t, scenario S_t, and a non-overlapping source window $(H^{\mathrm{src}}, S^{\mathrm{src}})$, the model computes $z = P(F(H^{\mathrm{src}}, S^{\mathrm{src}}))$ and generates $p_\theta(\tau \mid H_t, S_t, z)$. The generator is a validation carrier for this interface. In the central intervention, target history, scenario, logged partner replay, decoder, residual latent policy, and candidate budget remain fixed; the supplied tendency is replaced and the same tendency-conditioned interaction-intention adapter is re-evaluated.

The design follows a representation-to-response path. The first representation network encodes action-grounded motion and neighbor relations. The second stage uses interaction-intention supervision and non-overlapping same-vehicle groups to refine that signal into a compact condition whose meaning is tied to how a vehicle responds under an interaction demand. This distinction matters: the tendency is a response condition, not a complete description of the current interaction demand, and the scenario does not determine which vehicle-level response is rendered on its own. The resulting code is useful only if it changes the generated response and remains aligned with observed interaction behavior.

We evaluate the interface with four linked questions. First, do separated history segments contain a measurable vehicle-level signal, and does intention-aware refinement add information beyond the action-only control? Second, does replacing the tendency change the target response under a matched local scenario? Third, does the variation expand observed behavior bins while preserving future matching, ordering, and open-loop intent agreement? Fourth, do temporal-gap, no-intention, repetition, and training-time tendency-shuffle audits delimit the mechanism and its failure modes? All claims are evaluated on a vehicle-disjoint NGSIM two-car protocol with logged partner replay; the study does not claim planner-in-the-loop failure discovery or closed-loop safety improvement.

The contributions are:

- We formulate a vehicle-level interaction tendency as a replaceable response condition learned from history-only interaction segments, and implement an action-grounded to intention-aware representation path that separates the current interaction demand from the vehicle's response regularities.
- We provide a matched tendency-conditioned interaction interface: the target history, local scenario, logged partner replay, decoder, and eight-candidate budget are fixed, while the tendency code is replaced and the same interaction-intention adapter is re-evaluated. This makes response variation measurable at the model level.
- We establish an evidence chain for structured variation. Vehicle-disjoint retrieval supports the representation claim, while role probes test the division between scenario information and the tendency condition; equal-budget intervention yields non-zero response dispersion; observed-future matching, observed-bin coverage, rank agreement, and a tendency-shuffle audit test whether the diversity remains aligned with logged interactions.

The scope is intentionally narrow. The experiments use logged partner replay and compare generated candidates with observed two-car interactions. They do not measure a semantic aggressiveness axis, full behavior-space coverage, cross-dataset transfer, causal effects in traffic, planner-in-the-loop failure discovery, or closed-loop safety improvement. These are downstream uses of the interface rather than claims established by the present study.

2 Related Work

2.1 Scenario Generation and Testing for Autonomous Driving

Scenario-generation and testing systems vary road layouts, initial states, traffic participants, or perturbations to expose an autonomous system to rare and interactive situations. Scenario Diffusion and SimGen study controllable scene synthesis, while NAVSIM and HAD-Gen provide evaluation settings for planning and human-like driving behavior (Pronovost et al., 2023; Zhou et al., 2024a; Dauner et al., 2024; Wang et al., 2025). Safety-oriented systems also use adaptive or adversarial agents to search for failures (Feng et al., 2023; Chang et al., 2024). These approaches motivate our setting: a test suite may need several responses inside one interaction scenario, in addition to changing the scene itself. Our focus is the narrower response interface that can be evaluated before a planner-in-the-loop study.

2.2 Vehicle-Level Behavior and Driving-Style Representation

Multi-agent predictors and traffic simulators model plausible futures from heterogeneous histories, interactions, and multimodal latent variables. Trajectron++ and AgentFormer are representative trajectory models, while Waymax and the Waymo Open Sim Agents Challenge provide simulation and evaluation substrates (Salzmann et al., 2020; Yuan et al., 2021; Gulino et al., 2023; Montali et al., 2023). A separate line learns driver-character variables, behavior embeddings, or editing operations for recognition, prediction, or policy conditioning (Chang et al., 2023; Li et al., 2025; Pronovost et al., 2023; Niedoba et al., 2023). Style2Risk illustrates why style-conditioned agents are relevant to downstream autonomous-vehicle testing (Lai et al., 2026). These works show that behavior-related representations can be useful, but they do not by themselves establish a matched vehicle-level condition for changing one response while keeping the local interaction demand fixed.

2.3 Interaction-Aware and Intention-Conditioned Behavior Generation

Interaction-aware generation introduces neighboring agents, intent variables, or structured conditions to make future behavior more coherent (Wu et al., 2024; Zhou et al., 2024b; Chang et al., 2024; 2025; Xia et al., 2024). The distinction relevant to this paper is between *what the current scenario demands* and *how a vehicle tends to respond*. A scene-level condition or a sampled latent can change the whole future without identifying that distinction; an intent label alone does not provide a vehicle-level history condition. StyleFinger connects the two through a history-derived tendency, an intention-aware representation path, and a fixed-scenario substitution audit. The generator is used to expose this interface and measure its scope, rather than to claim a new state-of-the-art simulator or a completed closed-loop testing system.

3 Methodology

3.1 Problem Formulation and Overall Framework

We study controlled variation of one target vehicle's response inside a recorded two-vehicle interaction. For target window j, let H_j be the target history, S_j the current local scenario represented by neighboring states and their relative relations, and τ_j the future target response. Together, H_j and S_j define the current target interaction input. A source interaction window $X^{\mathrm{src}}_{r(j)} = (H^{\mathrm{src}}_{r(j)}, S^{\mathrm{src}}_{r(j)})$ is sampled from the same vehicle at a temporally non-overlapping interval and is used only to infer a response condition. The training tuple is

$$\mathcal{W}_j = (H_j, S_j, i_j, \tau_j, v_j), \qquad h^{\mathrm{src}}_{r(j)} = \mathrm{Pool}_t\Big(F_\phi(X^{\mathrm{src}}_{r(j)})\Big), z_j = P_\psi\Big(h^{\mathrm{src}}_{r(j)}\Big), \tag{1}$$

where i_j is the lateral–longitudinal interaction-demand label and v_j is the vehicle identifier. The source encoder receives only the target and neighbor histories in $X^{\mathrm{src}}_{r(j)}$; the future coordinates are used only as training labels for the interaction-demand heads. The central evaluation therefore replaces z_j while holding H_j, S_j, partner replay, and the decoder fixed.

The framework separates three roles. The current scenario describes what the vehicle is facing; the tendency summarizes how the vehicle has responded in historical interactions; and the conditional decoder renders a future trajectory that makes the tendency substitution observable. We call z a *vehicle-level interaction tendency* because this definition is tied to a measurable response interface. It is not a psychological personality label, a vehicle-identity classifier, or a universal aggressive–conservative axis.

3.2 Vehicle-Level Interaction Tendency Representation Learning

The representation is learned along an action-to-intention path. In the implementation, *v0* names the action-grounded control and *v1* names the intention-aware two-stage variant; these are training aliases, not separate deployed modules. The control starts from motion and neighbor relations, whereas the two-stage variant adds supervision for the interaction demand before aggregating separated windows from the same vehicle. Thus, intention is introduced to lift the representation from what the vehicle did toward how it tends to respond within an interaction task.

3.2.1 ACTION-GROUNDED TENDENCY REPRESENTATION

For each history window, a target branch and a neighbor branch encode the first 25 history steps. The implementation uses an LSTM for the target stream and an LSTM for each neighbor stream, followed by linear projections and masked attention over valid neighbors. The encoder returns a 25-step fused sequence; the representation used by the heads, projector, and generator is its temporal mean,

$$h_j = \mathrm{Pool}_t(F_\phi(H_j,\ S_j)) \in \mathbb{R}^{128}. \tag{2}$$

The v0 control disables the interaction-intention loss and uses this action-grounded representation for the tendency projector. This control asks whether motion and neighbor relations alone are sufficient, without treating an intent label as a substitute for the tendency. A vehicle-level projector maps the fused representation to a low-dimensional code,

$$z_j = P_\psi(h_j) \in \mathbb{R}^{d_z}, \qquad d_z = 5. \tag{3}$$

3.2.2 INTENTION-AWARE TENDENCY REFINEMENT

The v1 encoder retains the action representation but jointly predicts lateral and longitudinal interaction factors. With class-weighted cross-entropy on the natural training split and an auxiliary vehicle-identity loss, the Stage-1 objective is

$$\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{lat}} + \mathcal{L}_{\mathrm{lon}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}. \tag{4}$$

The intent heads supply an interaction-aware training signal; they are not assumed to define a complete ontology of driving intention. After Stage 1 is frozen, Stage 2 groups non-overlapping windows from the same vehicle and trains the projector with a supervised-contrastive term, a vehicle-classification term, and an within-group variance penalty:

$$\mathcal{L}_{\mathrm{stage2}} = \mathcal{L}_{\mathrm{supcon}}(z, v) + 0.5\,\mathcal{L}_{\mathrm{ID}}(\bar{z}, v) + 0.1\,\mathrm{Var}_{r\in v}(z_r). \tag{5}$$

The non-overlap rule prevents adjacent copies of one event from acting as independent evidence. The two-stage design therefore moves from action-level motion to how this vehicle responds within an interaction demand while retaining a history-only source interaction window at inference. The no-intention retrieval control and the temporal-gap audit in Section 5 quantify how much of this refinement is supported.

3.3 TENDENCY-CONDITIONED INTERACTION BEHAVIOR GENERATION

The generator validates whether a tendency code can be used as a response condition. The frozen Stage-1 encoder produces a fused local-interaction representation h_j from the target history and available neighbor histories. An interaction-intention adapter maps this representation and the tendency code to a six-way interaction-demand distribution:

$$q_j(z) = \mathrm{softmax}(A_\omega([h_j, z]))\,. \tag{6}$$

A recurrent pair-CVAE then conditions on target history, partner history, the local-interaction representation h_j, $q_j(z)$, and z:

$$p_\theta(\tau_j \mid H_j, S_j, z) = p_\theta\big(\tau_j \mid H_j, H_j^{\mathrm{partner}}, h_j, q_j(z), z\big)\,. \tag{7}$$

The implemented decoder uses GRU encoders for the target and partner histories and a GRUCell rollout for the target future. Its training objective is a reconstruction term plus a small CVAE KL term,

$$\mathcal{L}_{\mathrm{gen}} = \|\hat{\tau}_j - \tau_j\|_2^2 + 0.01\,D_{\mathrm{KL}}(q_\theta(u \mid \tau_j, \cdot)\,\|\,\mathcal{N}(0, I))\,. \tag{8}$$

For the claim-bearing intervention, the residual latent u is fixed to its prior mean. When z is replaced, the adapter is re-evaluated as $q_j(z)$; the reported effect is therefore a joint tendency-conditioned response intervention, rather than a fixed-intent or causal manipulation of a real vehicle.

3.4 TENDENCY-CONDITIONED RESPONSE CONSTRUCTION

For each held-out target window, we construct a finite candidate set by selecting $K = 8$ tendency prototypes from training vehicles with a pre-specified farthest-first rule in the learned code space. The

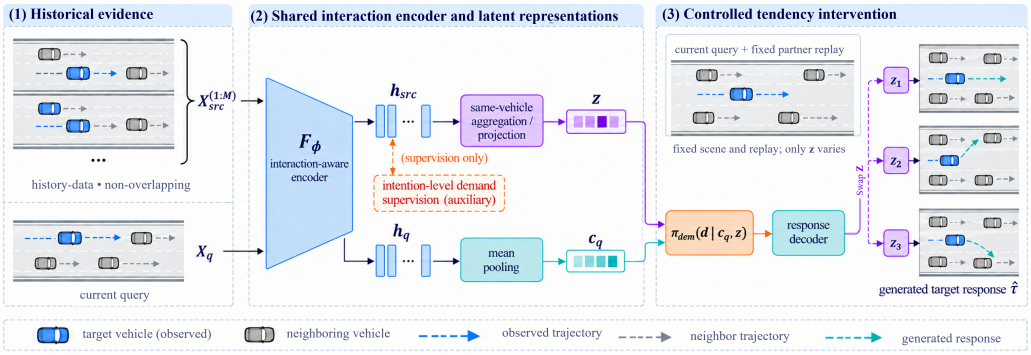

Figure 1: Overview of StyleFinger. (1) A non-overlapping source interaction window provides the historical evidence from which the vehicle-level tendency is learned. (2) The target history is represented together with the local interaction scenario and partner replay. (3) During the fixed-scenario intervention, the local interaction scenario and partner replay remain fixed while the tendency code z is swapped, producing alternative target responses.

target history, current local interaction scenario, logged partner replay, decoder parameters, residual-latent policy, and candidate budget are shared across all eight candidates. Only the supplied tendency prototype changes, and each replacement is passed through the same adapter and decoder. The resulting target trajectories form a controlled response family within one local interaction scenario: the external interaction demand is fixed while the target response is varied.

The comparison conditions are defined at the level of this construction. *Repeat* places the same deterministic response in all candidate slots and measures the no-variation floor. *StyleShuffle* permutes scenario–tendency assignments during training while preserving the architecture, data volume, optimizer, and evaluation budget; it tests whether observed-data alignment depends on the learned pairing rather than on dispersion alone. Partner behavior is replayed from logged data, so the construction is an open-loop validation layer. It supports an auditable behavior-level interface for testing, but it does not establish reciprocal online simulation, causal control, full behavior-space coverage, or planner-in-the-loop safety gains.

4 Experimental Setup

4.1 Dataset, Interaction Construction, and Data Splits

We use 30,259 two-car windows from 732 vehicles in four roadway scenes extracted from NGSIM US-101 (Colyar & Halkias, 2006). The vehicle-disjoint split contains 504/111/117 train/validation/test vehicles. Each target input uses the first 25 history steps; the future is never supplied to the tendency encoder. Source and target windows are selected under the declared temporal non-overlap rule. A separate 250-frame audit increases the minimum start-after-previous-end gap to test sensitivity to short temporal overlap. The generation evaluation contains 4,123 held-out target windows.

The current local interaction input contains the target history and neighbor state; we refer to this fixed pre-response condition as the local interaction scenario, while the partner vehicle follows its logged trajectory. Only the target response is generated. This partner-replay protocol isolates the response interface, but all generation results are open-loop and do not model reciprocal partner reactions.

4.2 Compared Models, Evaluation Protocol, and Metrics

For each held-out target window, eight tendency prototypes are selected from training vehicles using a pre-specified farthest-first rule. The same target history, local interaction scenario, partner replay, decoder, residual-latent policy, random-seed policy, and candidate budget are used for all conditions. At inference the residual latent is fixed to its prior mean; replacing the tendency code and recomputing the adapter is the intended model-level intervention. All claim-bearing generation values are means

Table 1: Primary evaluation protocol and leakage controls. Counts and budgets are fixed before evaluation; generation metrics are aggregated per held-out target window over three seeds.

Item	Primary setting
Data	30,259 two-car windows; 732 vehicles; four scenes
Split	Vehicle-disjoint 504/111/117 train/validation/test vehicles
Source/target	History-only inputs; declared temporal non-overlap rule
Target windows	4,123
Candidates	$K = 8$; farthest-first prototypes from training vehicles
Retrieval	111 held-out vehicles; one gallery centroid per vehicle; remaining windows as queries
Partner	Logged partner replay; target response generated
Seeds / unit	Three seeds; per-window generation aggregates
Evaluation	Open-loop; observed-bin and geometric proxy metrics

over three seeds, with the target window as the unit for generation metrics and the query–candidate pair as the unit for retrieval.

The representation audit reports vehicle-pair ROC-AUC, MRR, and the same-minus-cross similarity margin. The role audit reports macro-F1 and macro-recall over six joint interaction-demand labels. The intervention reports response dispersion D_{behavior}, Unique@0.25, intervention trajectory distance, observed-future matching, observed-bin coverage, and tendency–behavior rank agreement. ADE/FDE, intent agreement, style re-encoding, and partner-gap quantities are open-loop diagnostics. The matched *Repeat* condition measures the no-variation floor, while *StyleShuffle* permutes training-time scenario–tendency assignments without changing the model or budget. The complete protocol and leakage controls are summarized in Table 1.

5 Results

5.1 Vehicle-Level Tendency Representation

The learned code captures a vehicle-associated response signal across separated history segments, and the intention-aware refinement is complementary to the action-only representation. Under the vehicle-disjoint, history-only retrieval audit, the action-only representation reaches ROC-AUC 0.698 ± 0.035, MRR 0.098 ± 0.016, and a same-minus-cross similarity margin of 0.218 ± 0.107. The intention-aware representation reaches 0.754, 0.123, and 0.334, respectively. The comparison is consistent with complementary interaction supervision in this retrieval audit. Retrieval separability alone does not establish improved behavioral recovery, a universal tendency coordinate, or immutable vehicle behavior. The implementation aliases used in the appendix are *v0* for the action-only control and *v1* for the intention-aware two-stage representation; they identify training variants rather than separate deployed modules.

The current scenario contributes information that is complementary to the tendency condition. A tendency-only probe reaches macro-F1 0.220, whereas the joint scenario–tendency condition reaches macro-F1 0.659 and macro-recall 0.664 over six joint labels. Because a scenario-only control is not available in this audit, the comparison does not prove strict necessity or full disentanglement. The separate local-interaction representation-plus-style probe reaches approximately 0.678 macro-F1 under its original context-plus-style aggregation (Appendix Table 7). These results support the role assignment used by the method: a source history supplies a vehicle-level response condition, while the current scenario supplies the interaction demand. With a 250-frame minimum temporal gap, retrieval decreases to AUC 0.631 and margin 0.211; the signal weakens with separation but remains above chance. The gap audit does not establish permanent identity or cross-dataset invariance.

5.2 Tendency Substitution Changes the Target Response

Replacing the tendency changes the target response under a matched local scenario. The central intervention keeps target history, local interaction scenario, logged partner replay, decoder, residual latent, random-seed policy, and the eight-candidate budget fixed. The supplied tendency prototype is replaced and the same adapter is re-evaluated for each prototype. StyleFinger obtains $D_{\text{behavior}} =$

Table 2: Vehicle-level tendency representation and role audits. Retrieval compares non-overlapping history segments under the vehicle-disjoint split; role metrics are macro averages over six joint interaction-demand labels. Descriptive condition names are used here; the implementation aliases *v0*/*v1* are defined in the method and appendix.

Representation condition	ROC-AUC	MRR	Same–cross margin
Action-only representation	0.698 ± 0.035	0.098 ± 0.016	0.218 ± 0.107
Intention-aware representation	0.754	0.123	0.334
Intention-aware, 250-frame gap	0.631	–	0.211

Role condition	Macro-F1	Macro-recall
Tendency only	0.220	–
Scenario + tendency	0.659	0.664

Retrieval AUC is the probability that a randomly selected same-vehicle pair receives a higher similarity score than a cross-vehicle pair; it is not vehicle-identification accuracy. The base audit uses the held-out vehicle split and history-only source/target segments. The 250-frame result is a sensitivity check against short temporal overlap, not a test of permanent identity. Exact query counts, class support, and the separate local-interaction representation-plus-style probe are reported in the appendix under its original context-plus-style label.

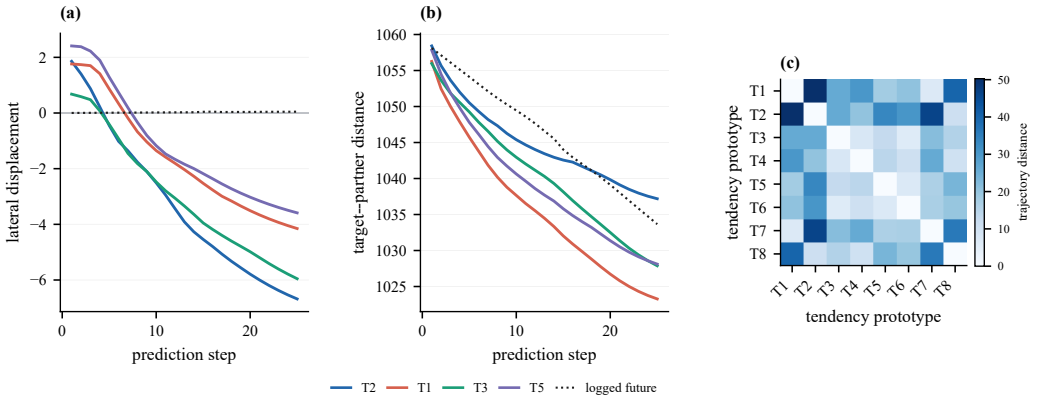

Figure 2: Fixed-scenario tendency substitution. The target history, local interaction scenario, logged partner replay, decoder, residual-latent policy, and eight-candidate budget are fixed while only the tendency code changes. (a) Future lateral-response profiles for four representative prototypes; the logged future is the dotted reference. (b) Target–partner geometric-distance profiles for the same prototypes under the fixed replay. (c) The complete pairwise trajectory-distance matrix for all eight candidates. The prototype indices identify conditions and do not encode an aggressive–conservative direction; the geometric distance is a diagnostic, not a safety metric.

0.637 ± 0.013, Unique@0.25 $= 0.839 \pm 0.008$, and an intervention trajectory distance of 1.542 ± 0.242. The Repeat control has zero dispersion by construction. These values show a measurable model-level response change; they do not identify a causal effect in traffic.

The smoothed StyleFinger generator audit provides a complementary open-loop check (Appendix Table 8). It reports ADE 3.916 ± 0.181, FDE 7.659 ± 0.138, style re-encoding cosine similarity 0.664 (vehicle-bootstrap 95% interval $[0.636, 0.690]$), and intention agreement of approximately 0.959 with the predicted interaction demand and 0.966 with the logged demand. These diagnostics indicate encoder-consistency of the substituted responses while retaining prediction error and protocol limitations. They are not a safety or closed-loop metric.

Figure 2 makes the intervention visible on a held-out interaction instance through three synchronized views. Panel (a) plots lateral displacement after the common history, panel (b) plots the target–partner geometric distance under the same logged partner replay, and panel (c) summarizes the complete pairwise distance matrix for all eight candidate conditions. Thus the local interaction scenario, partner replay, decoder, and candidate budget stay fixed while only the supplied tendency changes. The geometric distance is a response diagnostic rather than a collision or safety metric. Across the three held-out interaction instances in the corrected qualitative export, Figure 3(a) shows the corresponding response envelopes; the bands are min–max ranges over the eight candidates, not confidence intervals.

Table 3: Main fixed-scenario evidence on 4,123 held-out target windows with an equal eight-candidate budget. Values are three-seed means ± seed standard deviation. Panel A measures response change; Panel B measures agreement with observed interaction outcomes. D_{behavior} and intervention distance are descriptive variation measures, not quality scores.

A. Controlled response variation

Condition	D_{behavior}	Unique@0.25	Intervention dist.
StyleFinger	0.637 ± 0.013	0.839 ± 0.008	1.542 ± 0.242
Repeat	0.000	N/A	N/A
StyleShuffle	0.675 ± 0.054	0.843 ± 0.011	0.315 ± 0.005

B. Agreement with observed interactions

Condition	Observed-future match	Coverage	Rank agreement
StyleFinger	0.661 ± 0.032	0.893 ± 0.039	0.886 ± 0.039
Repeat	0.528 ± 0.019	0.742 ± 0.039	N/A
StyleShuffle	0.554 ± 0.048	0.830 ± 0.019	0.729 ± 0.099

Coverage is the fraction of observed behavior bins reached by at least one candidate; it is not full behavior-space coverage. Rank agreement is computed only for the tendency–behavior ordering defined by the evaluation pipeline. StyleShuffle uses the same architecture, data volume, optimizer, and candidate budget but permutes scenario–tendency assignments during training.

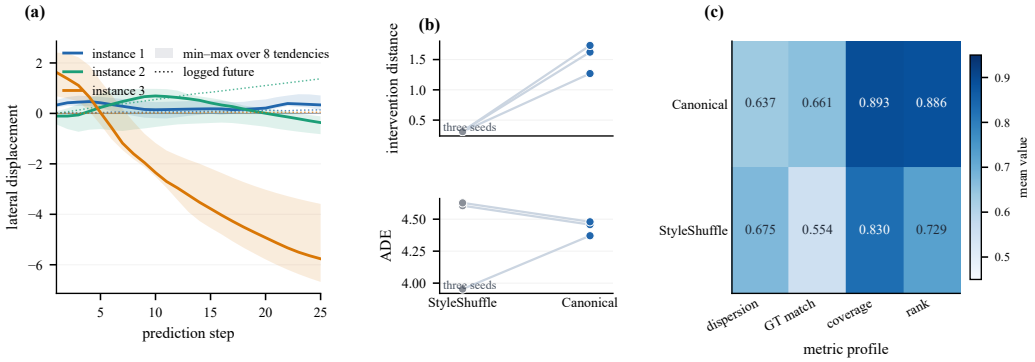

Figure 3: Three complementary views of the response evidence. (a) Across three held-out interaction instances, each band is the min–max lateral response over eight tendency candidates, the solid line is the candidate median, and the dotted line is the logged future. (b) Paired three-seed M9 audits: canonical conditioning increases intervention distance while ADE remains comparable to the shuffled control. (c) Three-seed M7/M10 metric profile. Canonical pairing preserves response dispersion while improving observed-future matching, observed-bin coverage, and rank agreement. Exact means and seed deviations are reported in Table 3; the envelopes in (a) are qualitative ranges, not uncertainty intervals.

5.3 Canonical Pairing Preserves Observed Interaction Structure

The generated response family is useful only if variation remains aligned with observed interactions. Under the same eight-candidate budget, StyleFinger reaches observed-future matching 0.661 ± 0.032, observed-bin coverage 0.893 ± 0.039, and tendency–behavior rank agreement 0.886 ± 0.039. Repeat reaches 0.528 ± 0.019 matching and 0.742 ± 0.039 coverage because it supplies no alternative response.

StyleShuffle separates structured variation from raw spread. It retains comparable dispersion ($D_{\text{behavior}} = 0.675 \pm 0.054$ and Unique@0.25 $= 0.843 \pm 0.011$), but lowers intervention distance to 0.315 ± 0.005, observed-future matching to 0.554 ± 0.048, coverage to 0.830 ± 0.019, and rank agreement to 0.729 ± 0.099. The shuffled model therefore still produces diverse trajectories; what degrades is their alignment with observed interactions. Figure 3 makes this distinction visible in three complementary forms: response envelopes, paired intervention/fidelity slopes, and a canonical-versus-shuffled metric profile. The exact aggregate values remain in Table 3.

5.4 Ablations and Robustness Checks

The auxiliary audits narrow the interpretation of the main results. The physical-statistics retrieval proxy reaches AUC 0.662 with margin 0.198, while the strict IDM+MOBIL history proxy reaches AUC 0.581 with margin 0.108; these references show that raw kinematics and a bounded hand-crafted history model do not reproduce the full learned signal. An end-to-end ablation reaches a higher retrieval AUC (0.772) but a lower same-minus-cross margin (0.153), so the two-stage design should not be described as uniformly superior on every metric (Appendix Table 5). The natural test split contains a zero-support `lat3_lon2` joint class; this support boundary limits interpretation of the macro metrics.

Together, the paired representation audit, temporal-gap check, Repeat floor, and StyleShuffle mechanism audit establish the evidence boundary for the paper. They support a vehicle-level, intention-aware response condition and structured fixed-scenario variation under logged partner replay. They do not support claims of causal control, full behavior-space coverage, cross-scenario transfer, improved ADS failure discovery, or closed-loop testing efficiency.

6 Discussion and Limitations

The intervention is a model-level conditioning effect. With target history, local interaction scenario, partner replay, decoder, and candidate budget fixed, replacing the tendency changes the generated response. Because the same interaction-intention adapter is re-evaluated for each supplied code, this is a joint tendency-conditioned response intervention rather than a causal manipulation of a real vehicle. The role audit gives the corresponding interpretation: history supplies a vehicle-associated response condition, while the joint scenario–tendency probe carries more interaction-demand information than the tendency-only condition. Since no scenario-only probe is available, the result supports complementarity rather than strict necessity or disentanglement.

StyleShuffle preserves raw response dispersion but lowers agreement with observed interactions and weakens response ordering, showing why diversity alone is insufficient. The evidence is therefore limited to an offline validation layer with logged partner replay, finite data-defined coverage, and geometric gap proxies. It establishes an auditable response-generation interface for the tested vehicle-disjoint NGSIM two-car protocol, not causal control, full behavior-space coverage, cross-dataset transfer, planner-in-the-loop failure discovery, or closed-loop safety improvement.

7 Conclusion

StyleFinger turns a history-only interaction segment into a vehicle-level tendency that can be substituted while the current interaction scenario is held fixed. The method separates the response condition from the demand being tested: the scenario and logged partner replay define the local situation, the tendency-conditioned adapter interprets a supplied vehicle history signal, and the recurrent decoder renders a target response for inspection.

The evidence supports this interface at three levels. The action-grounded to intention-aware representation path produces a measurable vehicle-associated signal under vehicle-disjoint retrieval, while the role probes show that adding the current scenario carries more interaction-demand information than the tendency-only condition. Because no scenario-only control is reported, this supports complementarity rather than strict necessity. Equal-budget tendency substitution then produces non-zero response variation under a matched local interaction scenario. Finally, observed-future matching, observed-bin coverage, rank agreement, and the tendency-shuffle audit show that the canonical pairing is more aligned with logged interactions than raw dispersion alone.

The contribution is deliberately bounded to an offline, open-loop NGSIM two-car protocol with logged partner replay. It does not establish a semantic aggressiveness ordering, full behavior-space coverage, cross-scenario or cross-dataset transfer, causal control, planner-in-the-loop failure discovery, or closed-loop safety improvement. The next step is to place the same auditable interface inside a reactive simulator and measure whether tendency-controlled responses expose failures that replay and unstructured sampling miss.

AI USE STATEMENT

The authors used AI-assisted tools for literature organization, experiment documentation, language editing, and draft review. The authors remain responsible for the scientific claims, experimental results, implementation, citations, and final submitted manuscript.

REFERENCES

Wei-Jer Chang, Chen Tang, Chenran Li, Yeping Hu, Masayoshi Tomizuka, and Wei Zhan. Editing driver character: Socially-controllable behavior generation for interactive traffic simulation. *IEEE Robotics and Automation Letters*, 8(9):5432–5439, 2023. doi: 10.1109/LRA.2023.3291897.

Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *Computer Vision – ECCV 2024*, pp. 242–258. Springer, 2024. doi: 10.1007/978-3-031-72664-4_14.

Wei-Jer Chang, Wei Zhan, Masayoshi Tomizuka, Manmohan Chandraker, and Francesco Pittaluga. Langtraj: Diffusion model and dataset for language-conditioned trajectory simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26622–26631, 2025. URL https://langtraj.github.io/.

James Colyar and John Halkias. Us highway 101 dataset. Technical Report FHWA-HRT-07-030, Federal Highway Administration, 2006.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0902.

Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, Henry Liang, Yiheng Sun, et al. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615:620–627, 2023. doi: 10.1038/s41586-023-05732-2.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0339.

Jintao Lai, Meng Wang, Zhen Zhang, Yiming Guo, Shixingyue Hu, Chengyuan Ma, Sa Gao, and Sijin Liu. Style2risk: A closed-loop testing system for safety evaluation of automated vehicles under style-conditioned adversarial interactions. *Systems*, 14(9):1144, 2026. doi: 10.3390/systems14091144. URL https://doi.org/10.3390/systems14091144.

Chengzhang Li, Sijin Liu, Meng Wang, Zhen Zhang, Jintao Lai, and Xiaoguang Yang. Condensed representation learning for interactive driving styles recognition. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, 2025. doi: 10.1145/3704413.3764458. URL https://escholarship.org/uc/item/7n1744js.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2582.

Matthew Niedoba, Jonathan Lavington, Yunpeng Liu, Vasileios Lioutas, Justice Sefas, Xiaoxuan Liang, Dylan Green, Setareh Dabiri, Berend Zwartsenberg, Adam Scibior, and Frank Wood. A diffusion-model of joint interactive navigation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2442.

Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nicholas Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3014.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2020.

Table 4: Descriptive dataset audit used by the two-car protocol. These counts document the support of the retained NGSIM windows; they are not an additional performance metric.

Quantity	Value
Trajectory windows	30,259
Encoded / original vehicles	732 / 732
Roadway scenes	4 (IDs 0, 1, 4, 5)
Vehicles observed in at least two scenes	594
Windows per vehicle (min / median / mean / max)	2 / 22 / 41.34 / 556
Held-out target windows	4,123

The vehicle-disjoint train/validation/test split and the history-only, non-overlap rule are defined in the main experimental protocol. The counts above are descriptive support checks and should not be read as evidence of cross-scene transfer.

Cheng Wang, Lingxin Kong, Massimiliano Tamborski, and Stefano V. Albrecht. Had-gen: Human-like and diverse driving behavior modeling for controllable scenario generation. *Accident Analysis & Prevention*, 223: 108270, 2025. doi: 10.1016/j.aap.2025.108270.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3622.

Junkai Xia, Chenxin Xu, Qingyao Xu, Yanfeng Wang, and Siheng Chen. Language-driven interactive traffic trajectory generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2474.

Ye Yuan, Xinyu Weng, Yang Ou, and Kris M. Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9813–9823, 2021.

Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. Simgen: Simulator-conditioned driving scene generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1548.

Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2527.

A Protocol, data, and metric definitions

This appendix follows the evidence order used in the main paper: first the data and protocol boundary, then representation and role audits, then generation and mechanism checks, and finally qualitative diagnostics and exclusions. Every reported quantity below is either a frozen aggregate already used by the main text or an auxiliary diagnostic with its protocol named explicitly.

Dataset and split. We use the 30,259 two-car windows from 732 vehicles and four roadway scenes described in the main protocol. The vehicle-disjoint train/validation/test split contains 504/111/117 vehicles. Generation metrics use 4,123 held-out target windows. Table 4 records the descriptive support audit; it does not substitute for the declared split or non-overlap rule.

Fixed-scenario intervention. For each held-out interaction window, the local two-vehicle history and partner replay are held fixed. We sample $K = 8$ outputs while changing only the history-derived vehicle-level interaction tendency code. We use the same decoder, residual-latent policy, number of candidates, and post-processing for all compared conditions.

Retrieval and role metrics. The retrieval evaluation compares representations from source and target segments under the declared temporal non-overlap rule, with cross-vehicle negatives. ROC-AUC is a same-vehicle-versus-cross-vehicle ranking probability, not vehicle-identification accuracy. The role audit reports macro-F1 and macro-recall over six joint interaction-demand labels; support limitations are stated with the corresponding table.

Table 5: Supplementary representation and robustness audits. These rows diagnose the action-to-intention refinement, temporal separation, and history-only reference signals; they are not a leaderboard.

Condition	ROC-AUC	same–cross margin	Interpretation
v0 action-only / no intention	$0.698 \pm .035$	$0.218 \pm .107$	Action-grounded control
v1 intention-aware two-stage	0.754	0.334	Main representation
End-to-end ablation	0.772	0.153	Higher AUC but lower separation margin
Physical-statistics proxy	0.662	0.198	Hand-crafted history reference
IDM+MOBIL proxy	0.581	0.108	Strict bounded history reference
v1, 250-frame gap	0.631	0.211	Temporal-separation sensitivity

The v0 and v1 rows use the vehicle-disjoint retrieval protocol. The end-to-end, physical, and IDM+MOBIL rows are diagnostic references with their own aggregation artifacts. The natural test split has zero support for the `lat3_lon2` joint label; this support boundary is retained here to prevent macro metrics from being read as complete six-class coverage.

Table 6: Supplementary support for the action-to-intention refinement. The v0 row is the action-grounded/no-intention control; the final row reports absolute v1–v0 differences for the same retrieval audit. These differences are descriptive contrasts, not a separate significance test.

Condition	ROC-AUC	MRR	Same–cross margin
v0 action-only / no intention	$0.698 \pm .035$	$0.098 \pm .016$	$0.218 \pm .107$
v1 intention-aware two-stage	0.754	0.123	0.334
v1, 250-frame minimum gap	0.631	–	0.211
Absolute v1 $-$ v0	$+0.056$	$+0.025$	$+0.116$

All rows use the vehicle-disjoint, history-only retrieval audit. The v0 row shows that motion and neighbor relations alone produce a measurable finite signal. The v1 row is the intention-aware refinement, while the 250-frame row tests sensitivity to temporal separation and is not a second v0 model. AUC is a same-vehicle-versus-cross-vehicle ranking probability, not vehicle-identification accuracy.

Behavior metrics. We discretize the behavior descriptors defined by the existing evaluation pipeline. Coverage is the fraction of bins observed in the recorded futures that are reached by at least one generated sample under the finite sampling budget; it is not an estimate of complete behavior-space volume. The recorded future is one realization of the interaction, so future-match is an empirical similarity proxy rather than a binary correctness label. Gap, acceleration, jerk, and collision-proxy quantities are reported in dataset coordinates.

B Representation, role, and robustness audits

The main text reports the v0/v1 comparison needed to interpret the action-to-intention design. The supplementary audit makes the control explicit, adds the end-to-end and history-proxy references, and records the temporal-gap sensitivity. The claim-facing v1 retrieval export uses the non-overlap supervised-contrastive Stage-2 projector: Stage-1 features are cached under no gradient, and the projector combines a supervised-contrastive term, vehicle-classification term, and within-group variance penalty. The compatibility Stage-2 objective used in an earlier sanity script is not the source of the claim-bearing SupCon numbers. The end-to-end row has a higher retrieval AUC but a lower same-minus-cross margin than v1, so the two-stage design is not described as uniformly superior.

The v0 control answers whether motion and neighbor relations alone contain a finite vehicle-associated response signal. Table 6 gives the matched retrieval scale, the intention-aware refinement, and the temporal-gap sensitivity in one place. The absolute v1–v0 differences are descriptive contrasts from the frozen aggregate export; they do not establish that intention supervision improves every downstream behavior metric.

C Generation fidelity and mechanism audits

This section separates open-loop fidelity diagnostics from the claim-bearing fixed-scenario and StyleShuffle aggregates. The adapted public predictors are retained as reproducibility references for ADE/FDE only; they are not a same-architecture leaderboard.

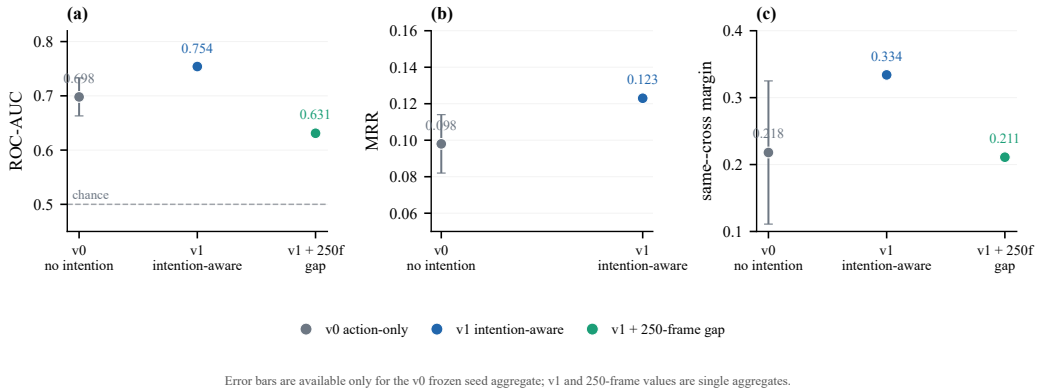

Figure 4: Visual summary of the frozen v0/v1 representation audit. (a) ROC-AUC stays above chance for the action-only v0 control, increases for v1, and weakens under the 250-frame separation check. (b) MRR compares v0 and v1 on the same retrieval protocol. (c) The same–cross margin follows the same refinement and temporal-sensitivity pattern. Error bars are shown only where the frozen export provides the v0 seed spread; v1 and 250-frame values are single aggregates.

Table 7: Scenario–tendency role audit. The table tests whether adding the current interaction scenario to the tendency condition carries more interaction-demand information than the tendency-only probe; it does not provide a scenario-only baseline for a strict necessity claim. The local-interaction representation-plus-style row retains the original context-plus-style aggregation and is a separate diagnostic probe.

Condition	Macro-F1	Macro-recall	Interpretation
Tendency only	0.220	–	Response condition without current demand
Scenario + tendency	0.659	0.664	Joint six-label interaction-demand probe
Local representation + style*	$\approx$ 0.678	–	Auxiliary original-label diagnostic

* This row was originally recorded as "context + style"; the natural test split contains zero support for the `lat3_lon2` joint label. Macro averages therefore describe the supported classes and do not establish complete six-class intent coverage or causal disentanglement.

D FIXED-SCENARIO QUALITATIVE DIAGNOSTICS

The corrected qualitative export makes the fixed-scenario intervention inspectable without turning the main paper into a trajectory atlas. Each row in Figure 6 keeps the target history and logged partner replay fixed while all eight tendency prototypes are evaluated. These are descriptive views of the same audited outputs and do not introduce a new training objective or headline claim.

E EVIDENCE BOUNDARIES AND EXCLUDED DIAGNOSTICS

Scenario role and intention distributions. The retained generator export contains aggregate intention-agreement audits, but not a frozen per-window probability vector, independent intention label, and class-support manifest for every compared condition. We therefore do not use an intention-distribution plot as main evidence. If such an export is released, the admissible analysis is a matched v0/v1/no-intention distribution on the same split, with class support and calibration reported alongside it; a posterior probe without independent labels would remain descriptive only.

Reader consistency. The earlier atlas and behavior-attribute table are intentionally withheld because their profile exports were produced before the corrected time-major partner reader was audited. The main-text aggregates and retained qualitative export use the audited canonical reader; legacy PDF/TeX assets remain in the repository for provenance only.

Incomplete or superseded controls. Context-only (the implementation label for a local-interaction encoding only), random-code, and stochastic equal-budget controls are not included in the current claim-bearing snapshot because their frozen provenance is incomplete. Historical dimension sweeps

Table 8: Mechanism and open-loop generation audits. Panel A compares canonical and training-time shuffled tendency pairing under the same architecture, data volume, optimizer, and candidate budget. Panel B reports the smoothed StyleFinger generator audit and protocol-matched trajectory-prediction references; the references are not a general SOTA comparison.

A. Pairing mechanism

Training	Intervention dist.	D_{behavior}	Observed match	Coverage	Rank agreement
Canonical	$\mathbf{1.542 \pm .242}$	$0.637 \pm .013$	$\mathbf{0.661 \pm .032}$	$\mathbf{0.893 \pm .039}$	$\mathbf{0.886 \pm .039}$
StyleShuffle	$0.315 \pm .005$	$\mathbf{0.675 \pm .054}$	$0.554 \pm .048$	$0.830 \pm .019$	$0.729 \pm .099$

B. Open-loop generation audit

Condition	ADE ↓	FDE ↓	Intent agr. ↑	GT intent agr. ↑	Style re-encode ↑
StyleFinger generator (smooth)	$3.916 \pm .181$	$7.659 \pm .138$	0.959	0.966	0.664
Trajectron++ (protocol matched)	$5.148 \pm .128$	$12.952 \pm .440$	–	–	–
AgentFormer (protocol matched)	27.665 ± 7.721	37.870 ± 11.818	–	–	–

Values in Panel A are three-seed mean ± seed standard deviation. The smooth-audit style re-encoding cosine has a vehicle-bootstrap 95% interval of $[0.636, 0.690]$. Panel B uses the adapted two-car, target-only, partner-replay protocol; the public references are included for scale and are not claimed as same-dataset leaderboard baselines. Coverage is observed-bin coverage, not full behavior-space coverage.

Table 9: Reference trajectory predictors under the two-car evaluation adapter. Values are three-seed means ± seed standard deviation over the reported 4,123 test windows where available. These adapted public models are reference points for trajectory error only, not a same-architecture SOTA comparison for the tendency interface.

Method	ADE ↓	FDE ↓	Role
StyleFinger (ours)	4.436 ± 0.058	8.778 ± 0.257	tendency-conditioned generator
Trajectron++	5.148 ± 0.128	12.952 ± 0.440	adapted public predictor
AgentFormer	27.665 ± 7.721	37.870 ± 11.818	adapted public predictor

All rows use the 25-history/25-future two-car adapter, target-vehicle prediction, logged partner replay, and the same held-out target-window list when the public adapter produced a finite output. The table is included to document the completed reference runs; it is not used to claim trajectory-prediction SOTA because the public models require different adapters and preprocessing.

and the earlier closed-set retrieval files use different readers, splits, or aggregation conventions and are retained for provenance rather than promoted as current evidence. A naive leave-one-scene-out analysis is also not used as a transfer result because scene IDs and interaction labels are confounded.

F Reproducibility Statement

The paper reports the data split, history and prediction horizons, non-overlap rule, sampling budget, random seeds, model configuration, and metric definitions in the main paper and supplementary material. The released code should include the exact configuration files used for the reported checkpoint, the result ledger, and scripts that regenerate each table and figure. The remote manifest hash and exact checkpoint/config paths remain release-time provenance items.

G Ethics Statement

This work uses an existing driving trajectory dataset for offline modeling and does not deploy a controller on public roads or interact with human subjects. Generated trajectories are research artifacts for simulation and testing; they must not be interpreted as safety guarantees or used directly for vehicle control without independent validation.

Table 10: Auxiliary open-loop diagnostics recorded in the journal audit. The derivative-smoothed checkpoint reduces kinematic tails while preserving the reported style re-encoding value; these quantities are geometric or kinematic proxies, not safety guarantees.

Metric	Original audit	Derivative-smoothed audit	Reading
ADE	3.935	$3.916 \pm .181$	essentially unchanged
p95 $\lvert a \rvert$	39.91	14.93	lower tail magnitude
p95 $\lvert j \rvert$	550.79	191.70	lower tail magnitude
True style re-encode cosine	0.664	0.664	unchanged
FDE	–	$7.659 \pm .138$	smoothed aggregate only
Fixed-scenario code-swap distance	–	$1.066 \pm .254$	measurable response change
Intent flip rate under code swap	–	0.0085	mostly continuous change

The smoothed generator reports predicted/re-encoded intent agreement of $0.959/0.966$ and a vehicle-bootstrap 95% interval $[0.636, 0.690]$ for the true style cosine. The fixed-scenario code-swap distance and flip rate use the smooth-audit protocol and are distinct from the three-seed M7/M9 intervention aggregates in the main text.

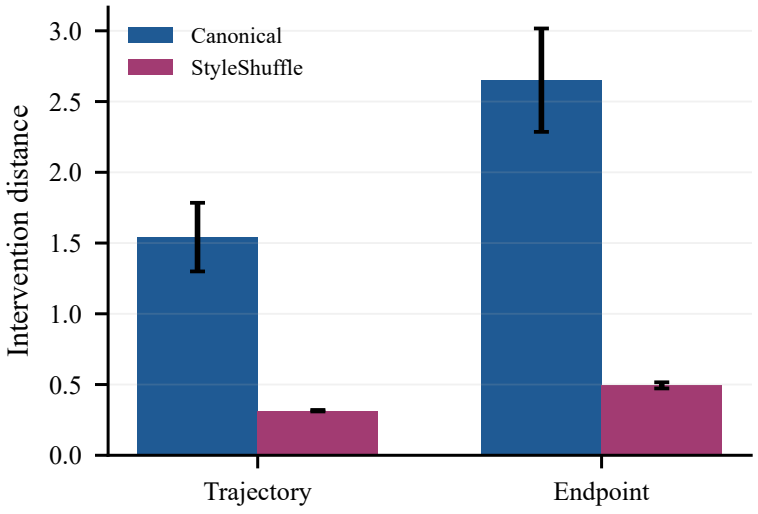

Figure 5: Mechanism-level code-swap audit. Canonical scenario–tendency pairing produces a larger response change than training-time tendency shuffling under the same architecture and training budget. The corresponding aggregate values and limitations are reported in Table 8.

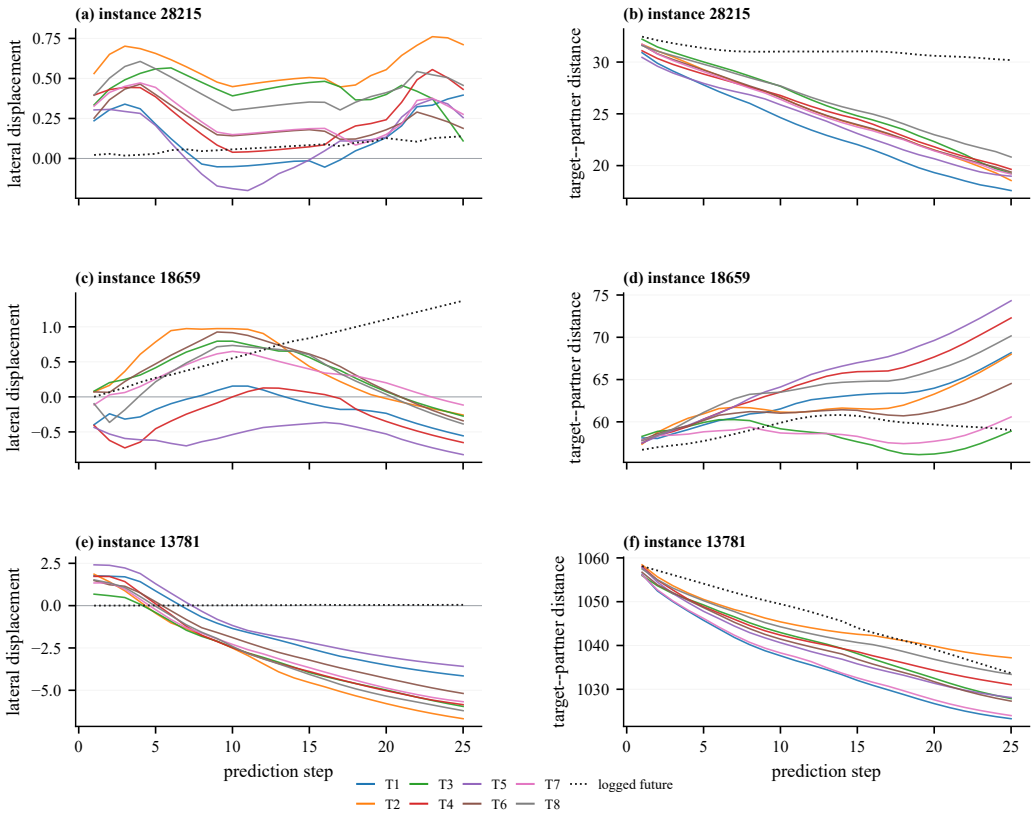

Figure 6: Complete corrected qualitative export across the three held-out interaction instances. Each row keeps the target history and logged partner replay fixed while all eight tendency prototypes are evaluated. The left column shows lateral displacement and the right column shows target–partner geometric distance; the dotted gray curve is the logged future. This appendix diagnostic exposes the full candidate family behind the representative main-text view and is not a new aggregate metric or a safety evaluation.