F2-PLANNER: GENERATIVE REPLANNING FROM UNEXECUTED FUTURE TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative trajectory planners repeatedly refine future trajectories, although only a short prefix of each selected plan is executed before replanning. The unexecuted future offers a structured starting point, but vehicle motion and new observations can make it outdated. We propose F2-Planner, a future-to-future generative replanner that carries the unexecuted portion of the controller-selected trajectory across cycles as a revisable generative state. It first advances the previous plan and aligns the remaining waypoints with the current ego pose. A horizon-wise utility predictor then controls how much of this future is blended with current candidates and how broadly the resulting source is sampled. On Bench2Drive v0.0.4, F2-Planner achieves a mean Driving Score of 96.65 with two neural function evaluations across eight training seeds. It outperforms the matched four-evaluation cycle-local planner (96.01) while reducing planning-head latency by 27.77%. At the same two-evaluation budget and nearly matched planning-head latency, it improves over an execution-aligned warm-start by 0.42 Driving Score points. Separate event-response and corrupted-history tests show that the carried future can be revised when its intent becomes stale. The source update also improves a truncated-diffusion generator within the same planning stack. These results support reusing and revising an unexecuted future to reduce repeated refinement while retaining responsiveness in the evaluated scenarios.

1 INTRODUCTION

Trajectory planning is a core component of autonomous driving, where a planner predicts future motion from the current driving scene (Paden et al., 2016; Hu et al., 2023). Diffusion- and flow-matching-based planners have recently shown strong performance by modeling multiple plausible futures and refining them into executable trajectories (Zheng et al., 2025; Liao et al., 2025; Liu et al., 2026). These methods typically generate trajectories by iteratively transforming samples from a source distribution, requiring multiple neural function evaluations for each plan. In closed-loop driving, this refinement is repeated at every replanning cycle, making iterative generation a practical computational burden.

A natural way to reduce this cost is to improve the starting point of the generation process. Existing methods use structured trajectory anchors, learned priors, or previous planning outputs to move the initial distribution closer to plausible driving trajectories. More specifically, some methods construct scene-conditioned source distributions from trajectory anchors (Liao et al., 2025; Liu et al., 2026), while warm-start approaches reuse partially refined or noised trajectories from earlier planning cycles (Janner et al., 2022; Zhou et al., 2023; Guo et al., 2026). These strategies can reduce the correction required during inference. In closed-loop driving, a key question is how to account for actual execution and decide which parts of a previous plan should still shape the next source.

The trajectory most recently selected for vehicle control provides a concrete starting point. At each planning cycle, the planner predicts a full future trajectory, but only a short prefix is executed before the next observation arrives. Most of the selected trajectory therefore remains unexecuted. Unlike a generic trajectory anchor, this remaining future has just been generated and selected under a recent scene, and it contains structured information about the motion that the planner intends to execute. This raises a natural question: *can the unexecuted future from the previous cycle be used to construct the generative source for the next one?*

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.

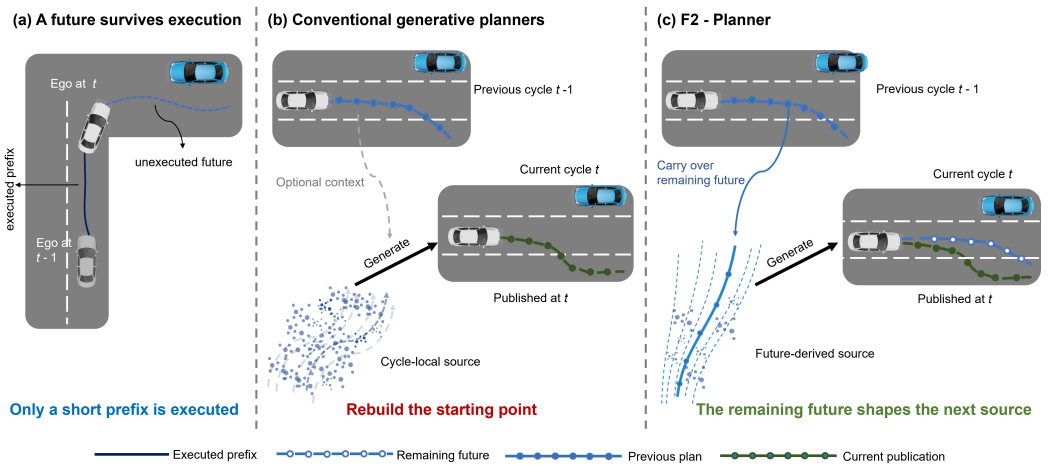


Figure 1: **From cycle-local generation to future-to-future replanning.** Only a short prefix of the selected trajectory is executed before replanning. F2-Planner advances and aligns the remaining path, then uses the current scene to adjust its contribution to the next generative source. The new trajectory selected for control is carried into the following cycle.

Simply reusing the previous trajectory is not sufficient. Between two planning cycles, the vehicle has already moved, so the remaining trajectory no longer lies in the correct ego frame. More importantly, new observations may invalidate part of the previous driving intent. A gap may close, another vehicle may cut in, or the realized ego motion may deviate from the previous plan. The previous future must therefore be reconciled with actual execution and revised according to the current scene. The key challenge is to preserve useful information from the previous decision without forcing the new plan to follow an intent that has become stale.

To address this challenge, we introduce F2-Planner, a future-to-future generative replanning framework that carries the unexecuted portion of the controller-selected trajectory across planning cycles as a revisable generative state. As illustrated in Figure 1, F2-Planner first reconciles the previous future with realized execution by advancing the trajectory and expressing the remaining waypoints in the current ego frame. It then predicts horizon-wise utility from the reconciled trajectory and the current scene. The utility determines how strongly each waypoint inherits the previous future and jointly controls the center and sampling spread of the next generative source. The resulting source is designed to preserve useful temporal structure while leaving room to revise outdated motion.

We evaluate F2-Planner in closed-loop driving on Bench2Drive v0.0.4. With two neural function evaluations, it achieves a mean Driving Score of 96.65, exceeding the matched four-evaluation cycle-local planner's 96.01 while reducing planning-head latency from 47.31 to 34.17 ms. At the same two-evaluation budget, it also improves on execution-aligned warm-start by 0.42 points at nearly equal planning-head latency. Source and component analyses examine these gains. Separate matched four-NFE tests measure response and recovery when recurrent intent becomes stale or corrupted. The source update also improves a second generative formulation within the shared planning stack.

Our contributions are threefold:

(1) We formulate *future-to-future generative replanning* around the unexecuted trajectory selected for vehicle control, allowing a recent driving future to serve as a recurrent generative state across planning cycles.

(2) We introduce an execution-reconciled, horizon-adaptive source update that jointly controls trajectory inheritance and sampling spread at each waypoint.

(3) We provide controlled closed-loop evidence that the proposed source construction improves the quality–compute trade-off under limited refinement. Separate matched four-NFE studies characterize temporal continuity, response, and recovery, and we evaluate the source update with two generative formulations in a shared planning stack.

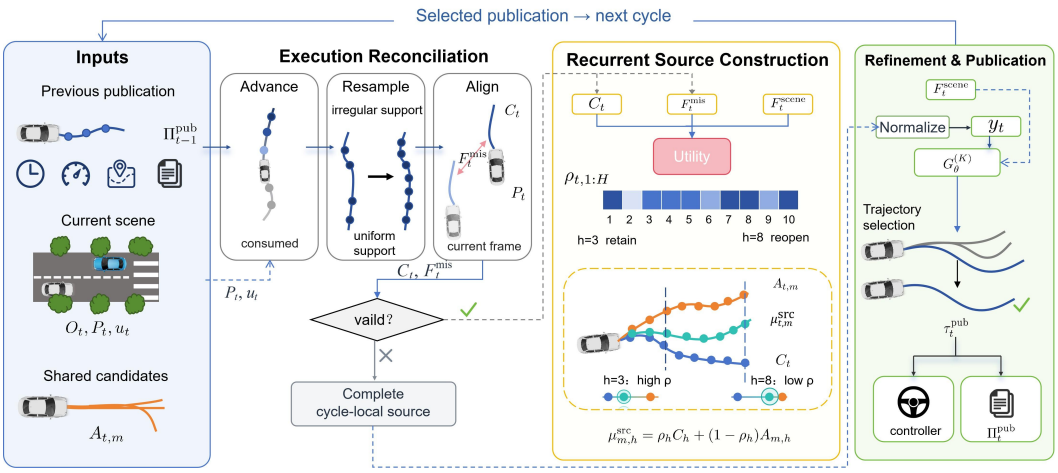


Figure 2: **Future-to-future generative replanning.** The trajectory selected in the previous cycle is advanced and aligned with realized execution. Current scene evidence then controls its contribution to the center and spread of the next generative source. Finite-step refinement produces a new controller-facing trajectory, which is carried into the following cycle. Π_t^{pub} denotes the cached publication record containing τ_t^{pub} and execution metadata; y_t is the normalized source.

2 Related Work

Generative trajectory planning and source design. Diffusion models learn iterative denoising (Ho et al., 2020), while flow matching learns vector fields between source and target distributions (Lipman et al., 2023). Diffuser and Diffusion Policy apply diffusion to trajectory and action generation (Janner et al., 2022; Chi et al., 2023). In driving, Diffusion Planner supports guided joint prediction and planning (Zheng et al., 2025); DiffusionDrive starts near multimodal anchors, DiffRefiner refines coarse proposals, and BridgeDrive learns an anchor-to-trajectory diffusion bridge (Liao et al., 2025; Yin et al., 2026; Liu et al., 2026). Our bridge implementation instead constructs the source from an execution-reconciled previous plan and current candidates.

Temporal reuse and history-informed generative planning. Diffuser warm-starts a previous plan, RDM uses plan likelihood to choose how to replan, and Temporal Diffusion Planner continues denoising across execution steps (Janner et al., 2022; Zhou et al., 2023; Guo et al., 2026). Real-time chunking overlaps generation with execution, while Streaming Flow Policy interleaves flow integration and action emission (Black et al., 2025; Jiang et al., 2025). BridgeAD reuses historical prediction and planning queries (Zhang et al., 2025); diffusion-forcing methods regulate temporal conditioning through independent noise levels and guidance (Chen et al., 2024; Zhang et al., 2026). A2A encodes executed-action history as a source for future actions (Jia et al., 2026). F2-Planner reconciles the unexecuted controller-selected trajectory with actual execution and adapts both source center and spread at each waypoint. Matched warm-start and component controls test these choices.

3 Future-to-Future Generative Replanning

F2-Planner carries the unexecuted part of the trajectory selected for control across planning cycles. It reconciles this future with realized execution, uses the current scene to construct the next source, and refines it into a new trajectory (Figure 2). The selected output is then stored for the next cycle. The generator architecture, current-cycle candidates, selector, and controller match those of the cycle-local planner.

3.1 Reconciling the Previous Future with Execution

Let $\tau_{t-1}^{\mathrm{pub}}$ be the previous controller-selected trajectory and γ_{t-1} its arc-length interpolant. Elapsed publication time and the cached speed profile give an estimate $\widehat{\ell}_t$ of consumed path length. We resample the remaining future on the planner's geometric support:

$$\bar{\tau}_{t,h} = \gamma_{t-1}\left(\widehat{\ell}_t + d_0 + (h-1)\delta_d\right), \qquad h = 1, \dots, H. \tag{1}$$

Here d_0 is the first waypoint offset and δ_d is the waypoint spacing. Let P_{t-1}^{pub} and P_t denote the publication and realized current poses, and let $T(P)$ map ego coordinates to world coordinates. After advancing progress, we express the remaining waypoints in the current ego frame:

$$\begin{bmatrix} C_{t,h} \\ 1 \end{bmatrix} = T(P_t)^{-1} T(P_{t-1}^{\mathrm{pub}}) \begin{bmatrix} \bar{\tau}_{t,h} \\ 1 \end{bmatrix}. \tag{2}$$

Progress estimation identifies the consumed portion of the path, while the pose transform expresses the remainder in the current frame. The transform alone does not correct errors in estimated along-path progress. We retain plan–execution mismatch features F_t^{mis} for source construction. Invalid temporal, pose, route, or geometric support triggers fallback to the complete cycle-local source. Progress estimation, support completion, and validity thresholds are detailed in Appendix B.

3.2 Constructing the Next Generative Source

Alignment locates the previous future in the current frame, but does not establish whether its maneuver remains appropriate. New observations may support part of the previous path while requiring a different motion farther ahead. Rather than accepting or discarding the whole trajectory, we predict how much to inherit at each waypoint. The current scene and reconciled future define a horizon-wise inheritance vector:

$$\boldsymbol{\rho}_t = \mathrm{sigmoid}\left(\mathcal{B}_\psi\left(C_t, F_t^{\mathrm{scene}}, F_t^{\mathrm{mis}}\right)\right), \qquad \boldsymbol{\rho}_t \in [0,1]^H, \tag{3}$$

where F_t^{scene} is the current scene representation. For the shared current-cycle candidates $A_t = \{A_{t,m}\}_{m=1}^{M}$, the same inheritance weight determines the source center:

$$\mu_{t,m,h}^{\mathrm{src}} = \rho_{t,h} C_{t,h} + (1-\rho_{t,h}) A_{t,m,h}. \tag{4}$$

It also sets the sampling scale:

$$\sigma_{t,h} = \sigma_{\min} + (1-\rho_{t,h})\left(\sigma_{\max} - \sigma_{\min}\right). \tag{5}$$

We draw the source as

$$X_{t,m,h}^{\mathrm{src}} = \mu_{t,m,h}^{\mathrm{src}} + \sigma_{t,h}\epsilon_{t,m,h}, \qquad \epsilon_{t,m,h} \sim \mathcal{N}(0, I_2). \tag{6}$$

A high $\rho_{t,h}$ keeps the source close to the reconciled future and concentrates samples around that decision. As inheritance decreases, the center moves toward current candidates and sampling broadens. The center therefore determines where refinement begins, while the spread determines how tightly the initial samples cluster there. This coupling is designed to retain useful geometry while leaving room to revise a stale maneuver. Its effect under limited refinement is tested by the center and spread ablations in Section 4.3.

The source centers are intermediate generative states, not controller outputs. Waypoint-wise mixing need not produce a smooth path before refinement; only the refined, selected trajectory is sent to the controller.

3.3 Learning for Finite-Step Refinement

We supervise inheritance by comparing the reconciled future with the highest-scoring current candidate, selected by the shared classifier:

$$\widehat{m}_t = \arg\max_m s_{t,m}. \tag{7}$$

Given the current target $\tau_t^\star$, the label at each valid waypoint is

$$q_{t,h} = \mathbf{1}\left[\left\|C_{t,h} - \tau_{t,h}^\star\right\|_2 < \left\|A_{t,\widehat{m}_t,h} - \tau_{t,h}^\star\right\|_2\right]. \tag{8}$$

Thus, $q_{t,h} = 1$ when the reconciled future is closer to the current target than the highest-scoring candidate at the same waypoint. This label measures relative geometric preference, not safety. The source constructor and generator are jointly trained with

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{src}}\mathcal{L}_{\text{src}}, \tag{9}$$

where $\mathcal{L}_{\text{gen}}$ is the original generator objective, $\mathcal{L}_{\text{rel}}$ is masked binary cross-entropy for inheritance, and $\mathcal{L}_{\text{src}}$ is an L_1 loss on the source centers. The source loss provides direct geometric supervision before iterative refinement. Appendix B gives the masks and reductions.

The primary diffusion-bridge generator holds the normalized source fixed through K neural function evaluations (NFEs), using the unchanged bridge schedule. Its selected output is sent to the controller and cached for the next cycle. The generator architecture is unchanged, but its parameters are trained jointly with the source constructor. Cached trajectories are model-generated and detached across cycles; ground-truth targets never enter history. The transfer study at the end of Section 4.4 evaluates the same source interface with truncated diffusion.

4 Experiments

We organize the evaluation around the quality–compute trade-off of future-to-future replanning. We first measure closed-loop scores and planning-head cost, then examine source geometry and design choices. Separate matched four-NFE studies assess continuity, event response, and recovery. Unless otherwise stated, all controlled comparisons use the same perception, candidate representation, selector, controller, and evaluation stack.

4.1 Experimental Setup and Main Results

Setup. We evaluate on the 220-route Bench2Drive v0.0.4 validation split (Jia et al., 2024; Bench2Drive Contributors, 2026) with CARLA 0.9.15 (Dosovitskiy et al., 2017). All controlled variants share the LEAD/TFv6 encoder, training data, 60-anchor bank, ten-waypoint support, selector, controller, and evaluator (Nguyen et al., 2026; Liu et al., 2026). Each configuration is trained with eight independent seeds and evaluated on all 220 routes per seed, yielding 1,760 route evaluations. Dev10 is used only for development. We report mean $\pm$ sample SD across training seeds, and paired intervals use matched seed-level results. We report Driving Score (DS) and Success Rate (SR). For a trained configuration, NFE varies only at inference, without retraining. In figures, F2 abbreviates F2-Planner.

All reported inference latencies refer to the planning head, measured at batch size 1 on a single RTX 4090. They include preprocessing, source construction, generation, selection, and cache update, but exclude perception, sensor transfer, simulator ticks, and downstream control. They are not end-to-end sensor-to-actuation latencies. Appendix A gives the evaluation and statistical procedures.

Table 1 summarizes the closed-loop results. With only two generator evaluations, F2-Planner reaches 96.65 ± 0.43 DS with a mean planning-head latency of 34.17 ms. This exceeds the 96.01 DS of Cycle-local-4 by 0.64 points in the paired eight-seed comparison (95% CI $[0.114, 1.166]$), while using half as many generator evaluations and reducing mean planning-head latency by 27.77%. The corresponding SR difference is 1.53 percentage points, with a paired interval of $[-1.153, 4.221]$ that includes zero.

At the same two-NFE budget and nearly matched planning-head latency, F2-Planner improves over aligned warm-start by 0.42 DS points (95% CI $[0.130, 0.710]$; 34.17 versus 33.88 ms). With four NFEs, F2-Planner reaches 97.07 ± 0.32 DS and $91.08 \pm 1.70\%$ SR. Relative to Cycle-local-4, the paired improvements are 1.06 DS points ($[0.739, 1.381]$) and 2.39 SR percentage points ($[1.278, 3.494]$).

We next examine the source geometry and design choices associated with these gains.

4.2 Understanding the Recurrent Source

A useful recurrent source should preserve geometry that the next planning cycle would otherwise need to reconstruct. Figure 3 measures average displacement error (ADE) of source centers before

Table 1: **Bench2Drive closed-loop overview.** Published rows use the official v0.0.4 results; their paper sources are listed in Appendix A.1. Controlled rows report mean ± sample SD across eight training seeds. Planning-head latency is measured at batch size 1 on a single RTX 4090.

Method	NFE	DS ↑	SR (%) ↑	Head (ms) ↓
Published Bench2Drive v0.0.4 context				
LEAD	–	95.59	87.28	–
SimLingo	–	86.55	70.45	–
Drive-π_0	–	69.71	45.91	–
DriveTransformer	–	60.20	35.91	–
Two-NFE matched comparison				
Aligned warm-start	2	96.23 ± 0.17	89.26 ± 1.87	33.88
F2-Planner	**2**	$\mathbf{96.65 \pm 0.43}$	$\mathbf{90.23 \pm 2.41}$	**34.17**
Four-NFE matched comparison				
Cycle-local	4	96.01 ± 0.36	88.69 ± 1.81	47.31
ExecAlign	4	96.73 ± 0.26	89.94 ± 1.74	–
Aligned warm-start	4	96.71 ± 0.32	90.06 ± 1.58	–
F2-Planner	**4**	$\mathbf{97.07 \pm 0.32}$	$\mathbf{91.08 \pm 1.70}$	50.44
F2-Planner-2 vs. Cycle-local-4: ΔDS $= +0.64$, 95% CI $[0.114, 1.166]$				

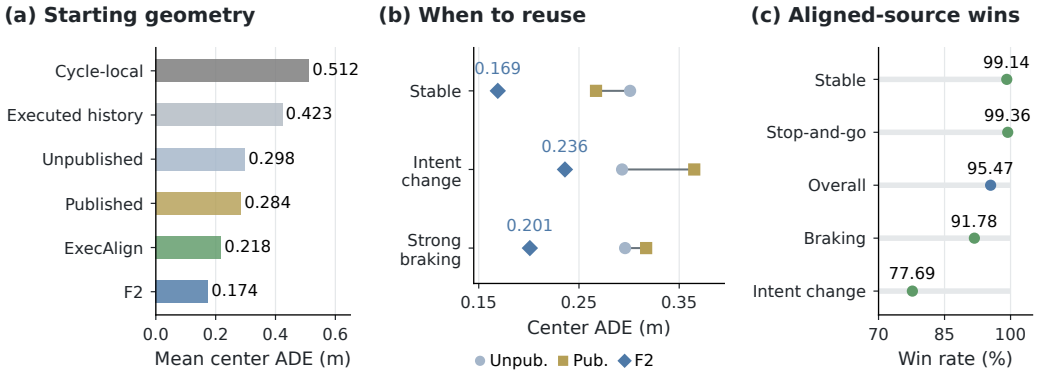


Figure 3: **The committed future provides useful but scene-dependent source geometry.** (a) Mean source-center ADE before stochastic sampling. (b) Published and pre-selection futures exchange order between stable motion and intent changes; diamonds show F2-Planner. The pre-selection future denotes the highest-scoring non-oracle trajectory available immediately before final controller-facing selection. (c) Execution-aligned source win rate against the selected cycle-local source. All comparisons use matched states and valid support.

stochastic sampling on 10,000 matched states. All source constructions are replayed on the same observations, ego states, references, candidate sets, and valid supports.

Mean ADE decreases from 0.512 m for the cycle-local source to 0.284 m for the previously selected future and to 0.218 m after execution reconciliation. The adaptive F2-Planner source further reduces the mean to 0.174 m. The execution-aligned source is closer to the current reference than the selected cycle-local source in 9,547 of the 10,000 matched comparisons.

The benefit of the previous future is not uniform across scenes. During stable motion, the published future is closer to the reference than its pre-selection counterpart (0.267 versus 0.301 m). During intent changes, the ordering reverses (0.365 versus 0.293 m), and the execution-aligned win rate falls to 77.69%. F2-Planner reaches 0.236 m mean source-center ADE in this regime. The result motivates selective reuse rather than unconditional inheritance: recently selected geometry can provide a strong starting point, but the source must remain able to move toward current-cycle evidence when the previous intent becomes stale.

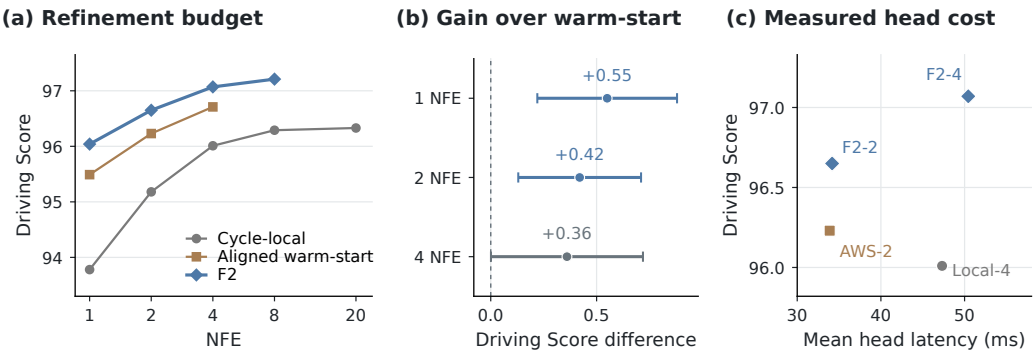


Figure 4: **Adaptive recurrent sources improve the measured quality–compute frontier.** (a) DS across evaluated NFE budgets, on a logarithmic NFE axis. (b) F2-Planner minus aligned-warm-start DS with pointwise paired 95% intervals; multiplicity-adjusted tests are reported in Appendix F. (c) DS versus measured mean planning-head latency. AWS denotes aligned warm-start; all markers are measured operating points.

Table 2: **Low-NFE source construction.** Mean ± sample SD across eight training seeds.

Method	1 NFE DS ↑	2 NFE DS ↑
Aligned warm-start	95.49 ± 0.28	96.23 ± 0.17
Fixed blend	95.66 ± 0.53	96.34 ± 0.49
Fusion only	95.79 ± 0.54	96.44 ± 0.37
Scalar utility	95.87 ± 0.53	96.50 ± 0.51
Fixed blend + adaptive σ	95.74 ± 0.51	96.40 ± 0.36
F2-Planner	$\mathbf{96.04 \pm 0.50}$	$\mathbf{96.65 \pm 0.43}$

A separate history-interface control helps distinguish this effect from history access alone. At four NFEs, capacity-matched execution-aligned conditioning and aligned warm-start have similar mean scores (96.69 versus 96.71 DS). The comparison therefore does not establish a generic advantage of supplying history through the source rather than through conditioning. The full adaptive F2-Planner source reaches 97.07 DS; the component analysis below examines the source-design choices associated with this gain. Appendix D reports the history-interface controls, while Appendix E gives regime-specific errors, provenance, and source diagnostics.

4.3 Planning under Limited Refinement

Figure 4 compares the source constructions as the refinement budget changes. The two-NFE cycle-local planner reaches 95.18 DS and 87.39% SR. Against aligned warm-start, F2-Planner gains 0.55, 0.42, and 0.36 DS points at one, two, and four NFEs, respectively, with paired 95% intervals $[0.220, 0.880]$, $[0.130, 0.710]$, and $[0.001, 0.719]$. The largest measured gain occurs at one NFE.

Table 2 separates the source-design choices. Fusion only learns waypoint-wise center mixing with fixed spread; fixed blend with adaptive σ keeps center mixing fixed and adapts the spread. Scalar utility replaces the waypoint-wise weights with one horizon-wide value. The full center–spread update has the highest mean DS at both low-NFE budgets, exceeding fusion only by 0.25 and 0.21 points. These comparisons support the combined update under the evaluated low-NFE settings. All ten one- and two-NFE F2-Planner-versus-control DS contrasts remain significant within the unified 11-test Holm family; Appendix F reports the complete paired statistics.

The measured planning-head speedup comes from fewer generator evaluations, not cheaper individual evaluations. At four NFEs, recurrence increases mean planning-head latency from 47.31 to 50.44 ms. At two NFEs, F2-Planner exceeds Cycle-local-4 in DS with 34.17 ms mean planning-head latency (Table 1). Appendix F.2 gives timing distributions and component costs.

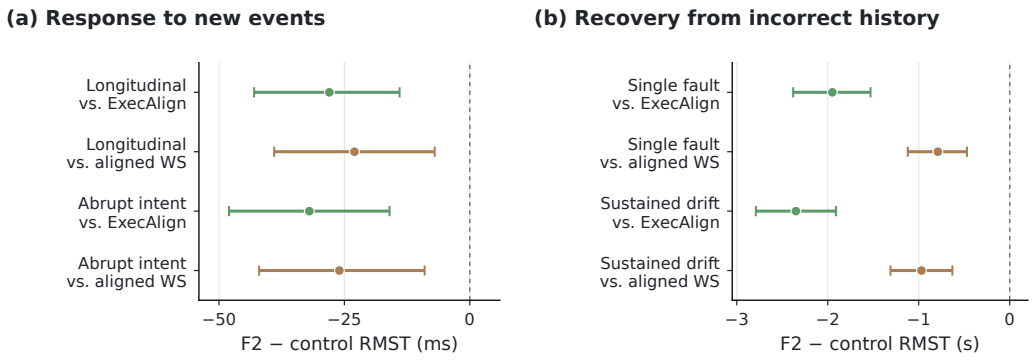


Figure 5: **Event response and recovery at four NFEs.** F2-Planner minus control RMST for event response (a) and recovery from corrupted recurrent state (b); negative values denote lower restricted mean time. All planners use four NFEs. Intervals are paired 95% bootstrap intervals. RMST windows are 2 s for response, 8 s for single-fault recovery, and 7 s for sustained-drift recovery.

A symmetric source–refiner crossover shows that the source effect depends on the refiner's training distribution (Appendix G). The frontier therefore describes matched source–refiner pairs, not a universally superior plug-in source.

4.4 Temporal Continuity and Revision at Four NFEs

All quantitative temporal comparisons use matched four-NFE configurations, with one denoiser call per NFE and no extra generator evaluations for F2-Planner. These studies are separate from the two-NFE benchmark results and do not establish response or recovery behavior at two NFEs. At four NFEs, F2-Planner reduces inter-cycle ADE from 0.441 m for Cycle-local to 0.221 m and lowers the switch rate from 18.7% to 9.8%. We then measure revision using shared, method-independent event suites (Appendix H).

We summarize time to response using restricted mean survival time (RMST), with nonresponses assigned the two-second window length. At four NFEs, F2-Planner reduces RMST relative to aligned warm-start from 0.486 to 0.463 s for longitudinal events and from 0.490 to 0.464 s for abrupt-intent events. The paired differences (F2-Planner minus aligned warm-start) are -23 ms (95% CI $[-39, -7]$) and -26 ms ($[-42, -9]$), respectively.

The learned inheritance weights change with the same events. Mean utility remains close to 0.79 around stable events. For stale-intent events, it decreases from 0.781 before the event to 0.612 at the event and 0.493 afterward, while the source spread increases. This four-NFE association is consistent with the intended source-control mechanism; utility is a geometric inheritance signal, not a safety estimate.

We also perturb a recurrent state directly while keeping the record inside the validity boundary, so the hard fallback mechanism is not activated. At the same four-NFE budget, F2-Planner reduces all-case recovery RMST relative to aligned warm-start from 2.33 to 1.54 s after a single fault and from 2.43 to 1.46 s under sustained drift. The paired recovery-rate intervals include zero; the supported difference concerns restricted recovery time.

Figure 5 summarizes the response and recovery effects. Appendices H and I give the event definitions, response coverage, corruption protocol, and paired inference.

Qualitative closed-loop behavior. Figure 6 shows selected closed-loop examples. The future evolves smoothly during vehicle entry and turning and is revised during the urban cut-in. The BridgeDrive cut-in rollout is annotated with a collision; these examples do not establish an aggregate safety advantage.

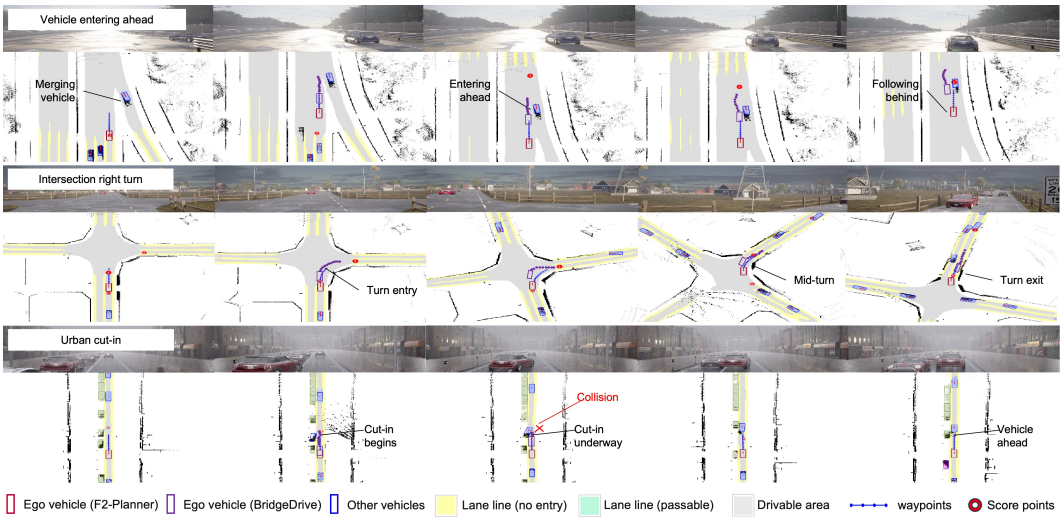


Figure 6: **Representative closed-loop behavior.** Five consecutive cycles for vehicle entry (top), an intersection right turn (middle), and an urban cut-in (bottom), comparing F2-Planner with BridgeDrive.

Transfer across generative formulations. We also evaluate the source update with a controlled DiffusionDrive-style truncated generator (Liao et al., 2025), changing only the source before forward noising within the shared planning stack. At four NFEs, F2-Planner improves the matched cycle-local generator by 1.03 DS points (95% CI $[0.697, 1.363]$) and 1.82 SR percentage points ($[0.471, 3.166]$). Appendix J reports the absolute scores and complete NFE frontier.

5 Conclusion

We presented F2-Planner, a future-to-future generative replanner that treats the unexecuted controller-selected trajectory as a revisable state across planning cycles. By reconciling this future with realized execution and adapting both the center and spread of the next generative source, F2-Planner reduces the amount of refinement required at each replanning step. On Bench2Drive, two NFEs achieve a higher Driving Score than the matched four-NFE cycle-local planner while reducing mean planning-head latency by 27.77%. Separate matched four-NFE experiments further show improved temporal continuity and shorter response and recovery RMSTs in the evaluated scenarios.

The source ablations and source–refiner crossover suggest that the benefit comes not simply from exposing the planner to history, but from how the previous future is transformed into the next generative source and how the refiner is trained with that source. The same source update also improves a second generative formulation within the shared planning stack, indicating that the idea is not specific to one refinement formulation.

Looking forward, an important direction is to study future-to-future replanning across different perception and control stacks, broader closed-loop benchmarks, and real vehicles. It is also worth examining longer recurrent horizons, where useful intent must be preserved without allowing accumulated errors to persist, and combining recurrent source construction with faster generative solvers. More broadly, our results suggest that efficient generative replanning depends not only on how trajectories are refined, but also on how each new generation process is initialized from the future that already exists.

REPRODUCIBILITY STATEMENT

Appendix A specifies evaluation units and statistical procedures. Appendix B reports cache conventions, source construction, auxiliary losses, and configuration values. The remaining appendices provide seed-level outcomes, history-interface controls, source geometry, quality–compute frontiers, source–refiner crossover, and event and recovery protocols. Closed-loop estimates, matched-state diagnostics, and event studies use the distinct replication units specified in Appendix A. The supplementary method module provides execution reconciliation, source construction, auxiliary losses, and a shared-generator interface; it does not include the complete benchmark evaluation stack.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, manuscript organization, literature identification, drafting and revising portions of the text, and improving the presentation of mathematical derivations and technical explanations. All methodological decisions, experimental settings, analyses, figures, quantitative results, references, and scientific claims were reviewed and verified by the authors. Generative AI outputs were not treated as authoritative sources. The authors take full responsibility for the final manuscript and its contents.

REFERENCES

Bench2Drive Contributors. Bench2Drive v0.0.4: Evaluation Update and Benchmark Results. Official repository documentation, 2026. URL `https://github.com/Thinklab-SJTU/Bench2Drive/blob/0.0.4/docs/v004_update.md`. Version 0.0.4. Accessed September 26, 2026.

Kevin Black, Manuel Galliker, and Sergey Levine. Real-Time Execution of Action Chunking Flow Policies. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1122.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0759.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 2017.

Jiaming Guo, Rui Zhang, Zerun Li, Yunkai Gao, Shaohui Peng, Siming Lan, Xing Hu, Zidong Du, Xishan Zhang, and Ling Li. Efficient Diffusion Planning with Temporal Diffusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21450–21458, 2026. doi: 10.1609/aaai.v40i26.39292.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-Oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with Diffusion for Flexible Behavior Synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022.

Jindou Jia, Gen Li, Xiangyu Chen, Tuo An, Yuxuan Hu, Jingliang Li, Xinying Guo, and Jianfei Yang. Action-to-Action Flow Matching. In *Proceedings of Robotics: Science and Systems*, 2026. doi: 10.15607/RSS.2026.XXII.209.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0025.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *International Conference on Learning Representations*, 2025.

Sunshine Jiang, Xiaolin Fang, Nicholas Roy, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Siddharth Ancha. Streaming Flow Policy: Simplifying diffusion/flow-matching policies by treating action trajectories as flow trajectories. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 238–257. PMLR, 2025.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12037–12047, 2025. doi: 10.1109/CVPR52734.2025.01124.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-Image Schrödinger Bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023a.

Shu Liu, Wenlin Chen, Weihao Li, Zheng Wang, Lijin Yang, Jianing Huang, Yipin Zhang, Zhongzhan Huang, Ze Cheng, and Hao Yang. BridgeDrive: Diffusion Bridge Policy for Closed-Loop Trajectory Planning in Autonomous Driving. In *International Conference on Learning Representations*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0418.

Long Nguyen, Micha Fauth, Bernhard Jaeger, Daniel Dauner, Maximilian Igl, Andreas Geiger, and Kashyap Chitta. LEAD: Minimizing Learner-Expert Asymmetry in End-to-End Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016. doi: 10.1109/TIV.2016.2578706.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11993–12003, 2025. doi: 10.1109/CVPR52734.2025.01120.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021.

Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. DriveMoE: Mixture-of-Experts for Vision-Language-Action Model in End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10678–10688, 2026.

Liuhan Yin, Runkun Ju, Guodong Guo, and Erkang Cheng. DiffRefiner: Coarse to Fine Trajectory Planning via Diffusion Refinement with Semantic Interaction for End to End Autonomous Driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):12009–12017, 2026. doi: 10.1609/aaai.v40i14.38189.

Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging Past and Future: End-to-End Autonomous Driving with Historical Prediction and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6854–6863, 2025.

Zehan Zhang, Neng Zhang, Yaoyi Li, Jia Cai, and Zhiling Wang. Diffusion Forcing Planner: History-Annealed Planning with Time-Dependent Guidance for Autonomous Driving. *arXiv preprint arXiv:2606.11019*, 2026. doi: 10.48550/arXiv.2606.11019.

Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Li, Xianyuan Zhan, and Jingjing Liu. Diffusion-Based Planning for Autonomous Driving with Flexible Guidance. In *International Conference on Learning Representations*, 2025.

Siyuan Zhou, Yilun Du, Shun Zhang, Mengdi Xu, Yikang Shen, Wei Xiao, Dit-Yan Yeung, and Chuang Gan. Adaptive Online Replanning with Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1909.

A Evaluation Design and Statistical Inference

A.1 Evaluation Design

Controlled closed-loop experiments use Bench2Drive v0.0.4 with CARLA 0.9.15. The validation split contains 220 routes across 44 scenarios (Jia et al., 2024; Bench2Drive Contributors, 2026). All controlled variants share the LEAD/TFv6 encoder, training data, candidate bank, selector, controller, route set, and evaluator. Dev10 is used only for development. We train eight independent seeds and evaluate every seed on all 220 routes, giving 1,760 route evaluations per configuration. Frozen-state diagnostics use matched observations, ego states, references, and valid supports.

We distinguish closed-loop replication from frozen-state replay. Closed-loop comparisons match seed and route identities, but later states can diverge because different planners take different actions. Frozen-state diagnostics instead replay the same stored state within each paired comparison. Diagnostic records retain route, seed, checkpoint, cycle, exogenous-state, regime, and provenance identifiers so that inference is performed at the appropriate sampling unit.

Published benchmark context. The published rows in Table 1 correspond to LEAD (Nguyen et al., 2026), SimLingo (Renz et al., 2025), Drive-π_0, and DriveTransformer (Jia et al., 2025). Drive-π_0 is the baseline introduced in the DriveMoE paper (Yang et al., 2026); the table does not report DriveMoE's own scores. All four rows use the official Bench2Drive v0.0.4 result document (Bench2Drive Contributors, 2026), rather than mixing scores from different benchmark versions. The shared LEAD/TFv6 encoder belongs to the TransFuser sensor-fusion family (Prakash et al., 2021; Nguyen et al., 2026).

Table A1: **Evaluation sizes and replication units.**

Study	Size
Training seeds	8
Routes per seed	220
Route evaluations per configuration	1,760
Source-geometry cycles	10,000
Source-swap cycles per seed	1,250
Evaluation instances / physical-event clusters	800 / 160
Instances per physical-event cluster	5
Recovery cases per fault / per seed	160 / 20
Timing sessions	5
Timing samples per configuration per session	3,000

Training uses eight RTX 4090 GPUs with global batch size 256. Planning-head latency uses one RTX 4090 with batch size 1 and excludes perception, sensor transfer, the simulator tick, and downstream control. All latency reductions refer to this planning-head boundary, not the complete sensor-to-actuation pipeline. Timing repetitions measure runtime variation rather than training variation.

A.2 Seed-Level Paired Estimates

For training seed i and route r, let $D_{ir}^{(a)}$ be the Driving Score of method a and $Z_{ir}^{(a)}$ its binary success. We compute

$$D_i^{(a)} = \frac{1}{220}\sum_{r=1}^{220} D_{ir}^{(a)}, \qquad S_i^{(a)} = \frac{100}{220}\sum_{r=1}^{220} Z_{ir}^{(a)}. \tag{10}$$

The reported mean and sample standard deviation use the eight seed-level values. Matched effects use $d_i = x_i^{(a)} - x_i^{(b)}$ and the interval

$$\bar{d} \;\pm\; t_{0.975,7}\frac{s_d}{\sqrt{8}}. \tag{11}$$

SR differences are reported in percentage points and are computed from the integer success counts before rounding. Benchmark intervals are pointwise paired intervals across the eight matched training seeds. Holm-adjusted values are reported separately for the unified 11-test source-construction DS family in Appendix F.

A.3 Event and Recovery Inference

All event-response, utility, single-fault, and sustained-drift comparisons use matched four-NFE configurations. Each NFE is one denoiser evaluation; F2-Planner receives no additional generator calls. These temporal results are separate from the two-NFE benchmark and planning-head latency results.

All-event RMST is the primary event-response estimand. Event eligibility, identities, onset times, directions, and reference signals are constructed once from the shared reference process before method comparison. The same retained physical-event IDs are then evaluated for every method, so all paired analyses operate on a common event set.

For event-response inference, each suite contains 800 evaluation instances grouped into 160 physical-event clusters, with five predefined neighboring instances per cluster. One cluster corresponds to one underlying physical scene transition. Its five instances are the five predefined evaluation instances associated with that event in the suite construction; they share route, scenario, physical-event identity, and event definition, and are not treated as independent physical events or independent replications. All methods are evaluated on the same five instances. They are therefore treated as dependent observations. We use 10,000 paired cluster-bootstrap replicates. In each replicate, 160 cluster IDs are sampled with replacement. Every evaluation instance belonging to a sampled cluster is included, and if a cluster is sampled k times, all five of its instances contribute k times. The identical sampled cluster multiset is used for F2-Planner and its comparator before computing the paired metric difference. Percentile 95% intervals are the empirical 0.025 and 0.975 quantiles of the 10,000 paired differences. The bootstrap random seed is 20250317.

Recovery uses 10,000 paired-case bootstrap replicates stratified by training seed and scenario. Within each observed seed–scenario stratum, the same number of case IDs as originally present in that stratum is sampled with replacement. The identical sampled case IDs are then applied to every compared method. Percentile 95% intervals again use the empirical 0.025 and 0.975 quantiles. The bootstrap random seed is 20250317.

Common-responder and nonresponse event effects are secondary estimands. Collision counts are descriptive four-NFE outcomes, not evidence of two-NFE safety. The source–refiner interaction is tested on eight seed-level contrasts rather than treating the 10,000 replayed cycles as independent training replications.

B Implementation and Training Details

B.1 Source-Construction Configuration

Table A2: **F2-Planner source-construction configuration.**

Configuration	Value
First waypoint distance	2.5 m
Minimum speed for time parameterization	0.5 m/s
Waypoint spacing	1.0 m
Maximum history age	0.6 s
Maximum tail extrapolation	2.0 m
Maximum execution translation mismatch	2.0 m
Maximum yaw mismatch	0.8 rad
$\sigma_{\min}$ / $\sigma_{\max}$	0.05 / 3.0 m
Utility hidden dimension	128
Utility attention heads	4
Candidates M / waypoints H	60 / 10
History dropout	0.15
λ_{rel} / λ_{src}	1.0 / 0.1
Global training batch size	256
Training hardware	8×RTX 4090
Base parameters	156,760,053
F2-Planner parameters	156,910,198
Added parameters	150,145 (0.0957%)

The reported percentage is the added parameter count divided by the total parameter count of F2-Planner.

B.2 Publication, Progress, and Coordinate Conventions

The committed future is the final controller-facing trajectory τ_t^{pub}. Its publication record Π_t^{pub} is written after final candidate selection and stores the controller-facing path together with publication pose and time, speed metadata, route state, and point-validity mask. Episode boundaries clear the record. The cache is detached across cycles.

During training, the recurrent input to cycle t is the detached publication selected at a preceding cycle. The current source is constructed from that record and the current observation, refined, and passed to the selector. The selected model output is stored for the following cycle through the same generate–select–commit path used at deployment. Ground-truth trajectories supervise only the current-cycle objectives and are never written into recurrent history. At sequence starts, episode boundaries, or history-dropout events, the planner uses the same cycle-local fallback as at deployment.

Execution mismatch is computed causally from the state implied by the stored publication and the realized current ego state, using only information available online. The representation contains pose/progress residual information and does not use the current target or future reference. Translation and yaw mismatch also define the hard validity boundary: values inside the configured limits remain available to learned source revision, whereas records outside those limits fall back to the cycle-local source.

For cached path points $p_0 = (0, 0), p_1, \ldots, p_H$, let

$$l_i = \|p_i - p_{i-1}\|_2, \qquad s_i = \sum_{j=1}^{i} l_j, \qquad S = s_H, \tag{12}$$

$$\bar{v}(s) = \max\left(v_0 + (v_1 - v_0)s/S,\ 0.5\right), \tag{13}$$

$$\delta u_i = \frac{2l_i}{\bar{v}(s_{i-1}) + \bar{v}(s_i)}, \qquad u_i = \sum_{j=1}^{i} \delta u_j. \tag{14}$$

With elapsed time $\Delta u = u_t - u_{t-1}^{\mathrm{pub}}$, the consumed-prefix estimate $\widehat{\ell}_t$ is obtained by interpolation on (u_i, s_i). The remaining path is resampled at $\widehat{\ell}_t + 2.5 + i$ m, $i = 0, \ldots, 9$. The implementation permits at most 2.0 m of history-tail extrapolation; remaining support is completed from the current route/anchor representation. Unusable support triggers fallback. The completed path is then transformed once from the publication frame through world coordinates into the realized current ego frame. The frame transform does not by itself correct error in the elapsed-time progress estimate.

B.3 Source Distribution and Auxiliary Losses

Suppressing batch indices, the source distribution is

$$p_t^{\mathrm{src}}(X) = \prod_{m=1}^{M} \prod_{h=1}^{H} \mathcal{N}\big(X_{m,h}; \mu_{t,m,h}^{\mathrm{src}}, \sigma_{t,h}^2 I_2\big)\,. \tag{15}$$

Sampling occurs in physical ego-frame coordinates, after which the shared base normalization is applied.

Let v_t^{hist} denote history validity and $w_{t,h}$ the common valid support. With $m_{t,h} = v_t^{\mathrm{hist}} w_{t,h}$ and clipped utility $\widetilde{\rho}_{t,h}$, the auxiliary losses are

$$\mathcal{L}_{\mathrm{rel},t} = \frac{\sum_h m_{t,h}\,\mathrm{BCE}(\widetilde{\rho}_{t,h}, q_{t,h})}{\sum_h m_{t,h} + \epsilon}, \tag{16}$$

$$\mathcal{L}_{\mathrm{src},t} = \frac{\sum_{m=1}^{M} \sum_h m_{t,h} \|\mu_{t,m,h}^{\mathrm{src}} - \tau_{t,h}^{\star}\|_1}{2M \sum_h m_{t,h} + \epsilon}. \tag{17}$$

The current target supervises the current prediction and utility target but is never written into recurrent history. Cached publications, discrete candidate selection, rigid transforms, and validity/fallback branches are stop-gradient. For each source-construction variant, all reported NFE operating points use the same trained checkpoint and differ only in the number of inference-time refinement evaluations. No NFE-specific retraining is used to construct the frontier.

B.4 Validity and Control Boundaries

Missing or non-finite metadata, episode resets, invalid temporal/geometric support, route-state invalidity, history age above 0.6 s, translation mismatch above 2.0 m, or yaw mismatch above 0.8 rad trigger hard fallback to the cycle-local source. For a valid record, decreasing ρ moves source centers toward the current candidates and increases the sampling scale. These checks screen the cached record; they do not guarantee collision avoidance.

HistoryCondition retains the cycle-local source and appends the encoded publication to conditioning memory. The capacity-matched history adapter adds 150,145 parameters, equal to the increase in F2-Planner. Aligned variants share the reconciled record and differ in its generative interface.

B.5 Source Geometry and Dispersion

For a fixed waypoint target y and standard two-dimensional Gaussian noise,

$$\mathbb{E}\|X - y\|_2^2 = \|\mu - y\|_2^2 + 2\sigma^2. \tag{18}$$

The identity separates center error from sampling spread in squared Euclidean distance. It does not define an additive decomposition of ADE. Because C_h and ρ_h are shared across candidates,

$$\mu_{m,h} - \mu_{n,h} = (1 - \rho_h)(A_{m,h} - A_{n,h}). \tag{19}$$

Hence the candidate-center variance satisfies

$$V_h^{\mu} = (1 - \rho_h)^2 V_h^{A}. \tag{20}$$

Lower inheritance restores more of the current candidate geometry. These identities describe the source distribution; the crossover in Appendix G measures the learned refiner response.

B.6 Source Construction and Generative Sampling

Generative efficiency can be improved at different levels. DDIM changes the reverse sampling process, and DPM-Solver uses dedicated high-order integration of diffusion ODEs (Song et al., 2021; Lu et al., 2022). Rectified Flow learns straighter transport paths between distributions (Liu et al., 2023b), while I^2SB learns conditional diffusion bridges between structurally related endpoints (Liu et al., 2023a). These works provide context for distinguishing sampling dynamics from the choice of source. Our controlled variants retain the generator family and schedule and change how the current source depends on the previous controller-selected trajectory. The reported generator formulations are the primary diffusion bridge and the truncated-diffusion variant in Appendix J.

C Complete Closed-Loop Results

Table A3: **Bridge controls.** Four NFEs unless stated otherwise. Mean ± sample SD across eight seeds.

Configuration	DS	SR (%)	Success / 1,760
Cycle-local	96.01 ± 0.36	88.69 ± 1.81	1561
HistoryCondition	96.25 ± 0.32	89.03 ± 1.49	1567
Capacity-matched condition	96.43 ± 0.32	89.38 ± 1.94	1573
Generic warm-start	96.27 ± 0.27	89.09 ± 1.68	1568
Executed-history extrapolation	96.17 ± 0.28	88.81 ± 1.87	1563
Published Future	96.41 ± 0.27	89.20 ± 1.64	1570
ExecAlign	96.73 ± 0.26	89.94 ± 1.74	1583
Aligned condition	96.55 ± 0.28	89.60 ± 1.71	1577
Aligned capacity condition	96.69 ± 0.36	89.94 ± 1.74	1583
Aligned warm-start	96.71 ± 0.32	90.06 ± 1.58	1585
Fixed blend	96.86 ± 0.32	90.34 ± 1.91	1590
Fusion only	96.98 ± 0.27	90.63 ± 1.88	1595
Fixed blend + adaptive σ	96.90 ± 0.28	90.40 ± 1.76	1591
Scalar utility	96.92 ± 0.27	90.51 ± 1.64	1593
No hard fallback	97.01 ± 0.26	90.74 ± 1.77	1597
No utility BCE	96.94 ± 0.28	90.57 ± 1.81	1594
No source proxy	96.91 ± 0.36	90.45 ± 1.75	1592
F2-Planner	97.07 ± 0.32	91.08 ± 1.70	1603
F2-Planner (2 NFE)	96.65 ± 0.43	90.23 ± 2.41	1588

Table A3 collects the four-NFE controls. Table A4 reports the corresponding core seed outcomes, and Table A5 reports pointwise paired effects. Positive mean differences do not all have intervals excluding zero. The low-NFE source-construction family is reported separately in Appendix F.

Table A4: **Seed-level core configurations.** Each cell is DS / integer success count on 220 routes.

Method	Seed 0	Seed 1	Seed 2	Seed 3
Cycle-local	95.79/189	96.41/201	96.07/197	96.37/195
ExecAlign	96.98/198	96.76/204	97.16/199	96.81/192
Aligned warm-start	96.21/195	96.29/201	97.04/197	96.64/199
Fusion only	96.82/195	97.35/201	97.19/205	96.80/197
Scalar utility	96.59/203	96.83/198	97.19/203	96.68/200
F2-Planner	96.89/194	97.00/203	96.83/205	96.95/202
F2-Planner (2 NFE)	96.33/197	96.24/199	96.91/195	97.20/207

Method	Seed 4	Seed 5	Seed 6	Seed 7
Cycle-local	96.43/195	95.74/190	95.77/198	95.50/196
ExecAlign	96.70/195	96.39/196	96.59/202	96.45/197
Aligned warm-start	96.83/203	96.73/192	96.85/200	97.09/198
Fusion only	97.08/205	96.58/199	97.24/199	96.78/194
Scalar utility	97.06/195	96.62/196	97.26/195	97.13/203
F2-Planner	97.70/198	96.95/198	97.43/199	96.81/204
F2-Planner (2 NFE)	96.15/191	97.11/202	96.32/194	96.94/203

Table A5: **Paired benchmark effects.** SR differences are in percentage points. These are unadjusted, pointwise comparisons.

Comparison	Metric	Mean Δ	95% CI	p
F2-Planner-4 – Cycle-local-4	DS	1.06	$[0.739, 1.381]$	0.0001
F2-Planner-4 – Cycle-local-4	SR	2.39	$[1.278, 3.494]$	0.0014
F2-Planner-2 – Cycle-local-4	DS	0.64	$[0.114, 1.166]$	0.0237
F2-Planner-2 – Cycle-local-4	SR	1.53	$[-1.153, 4.221]$	0.2190
F2-Planner-4 – ExecAlign-4	DS	0.34	$[-0.036, 0.716]$	0.0700
F2-Planner-4 – ExecAlign-4	SR	1.14	$[-0.758, 3.031]$	0.1991
F2-Planner-4 – Aligned warm-start-4	DS	0.36	$[0.001, 0.719]$	0.0495
F2-Planner-4 – Aligned warm-start-4	SR	1.02	$[-0.649, 2.695]$	0.1913
F2-Planner-4 – Fusion only-4	DS	0.09	$[-0.188, 0.368]$	0.4687
F2-Planner-4 – Fusion only-4	SR	0.45	$[-1.429, 2.338]$	0.5861
F2-Planner-4 – Scalar utility-4	DS	0.15	$[-0.131, 0.431]$	0.2474
F2-Planner-4 – Scalar utility-4	SR	0.57	$[-1.079, 2.215]$	0.4415
F2-Planner-4 – Fixed blend-4	DS	0.21	$[-0.201, 0.621]$	0.2661
F2-Planner-4 – Fixed blend-4	SR	0.74	$[-1.111, 2.588]$	0.3765
F2-Planner-4 – No hard fallback-4	DS	0.06	$[-0.183, 0.303]$	0.5783
F2-Planner-4 – No hard fallback-4	SR	0.34	$[-1.134, 1.816]$	0.6017
F2-Planner-4 – No utility BCE-4	DS	0.13	$[-0.177, 0.437]$	0.3497
F2-Planner-4 – No utility BCE-4	SR	0.51	$[-1.963, 2.986]$	0.6401
F2-Planner-4 – No source proxy-4	DS	0.16	$[-0.052, 0.372]$	0.1181
F2-Planner-4 – No source proxy-4	SR	0.63	$[-1.416, 2.666]$	0.4924
Truncated F2-Planner-4 – cycle-local-4	DS	1.03	$[0.697, 1.363]$	0.0002
Truncated F2-Planner-4 – cycle-local-4	SR	1.82	$[0.471, 3.166]$	0.0153
Truncated F2-Planner-4 – ExecAlign-4	DS	0.29	$[-0.015, 0.595]$	0.0595
Truncated F2-Planner-4 – ExecAlign-4	SR	0.57	$[-0.791, 1.927]$	0.3557

D History Interfaces and Source-Construction Controls

The history-interface controls separate two choices: what historical object is available, and where it enters the next generative process. HistoryCondition keeps the cycle-local source unchanged and supplies the encoded publication only as conditioning. Capacity-matched conditioning adds

a history adapter with the same additional parameter count as F2-Planner. Aligned controls first apply the same execution reconciliation and support handling used by F2-Planner, then differ in whether the aligned future is supplied as conditioning or injected into the source. Aligned warm-start uses the reconciled future as a conventional source without F2-Planner's learned adaptive source construction. Full F2-Planner uses the aligned future as a horizon-wise adaptive stochastic source.

Table A6: **History-interface controls.** The table isolates what historical object is available and where it enters the next generative process.

Setting	Historical object	Execution alignment	Interface
Cycle-local	none	–	cycle-local source
HistoryCondition	publication	no	condition
Capacity-matched condition	publication	no	condition
Aligned condition	publication	yes	condition
Aligned capacity condition	publication	yes	condition
Aligned warm-start	publication	yes	source / warm-start
F2-Planner	publication	yes	adaptive stochastic source

The complete four-NFE outcomes are reported in Table A3.

D.1 Operational Definitions of Source-Construction Controls

All controls share the same candidate bank, generator architecture, execution reconciliation, support completion, validity handling, training data, selector, and evaluator. The controlled difference is how the reconciled future enters the next generative source. *Aligned warm-start* uses the same execution-reconciled publication as a conventional warm-start distribution but does not use F2-Planner's learned horizon-wise source fusion. *Fixed blend* replaces the predicted inheritance vector by one fixed interpolation coefficient selected during development and held fixed for all reported evaluations, together with a non-adaptive source spread. *Fusion only* retains learned horizon-wise center fusion while keeping the source spread fixed. *Fixed blend + adaptive σ* keeps the same fixed center interpolation while allowing the spread to adapt. *Scalar utility* predicts one inheritance value shared across the full horizon. Full F2-Planner uses horizon-wise inheritance to control both center geometry and source dispersion. The "No utility BCE", "No source proxy", and "No hard fallback" controls remove only the named loss or validity branch while leaving the remaining source constructor unchanged.

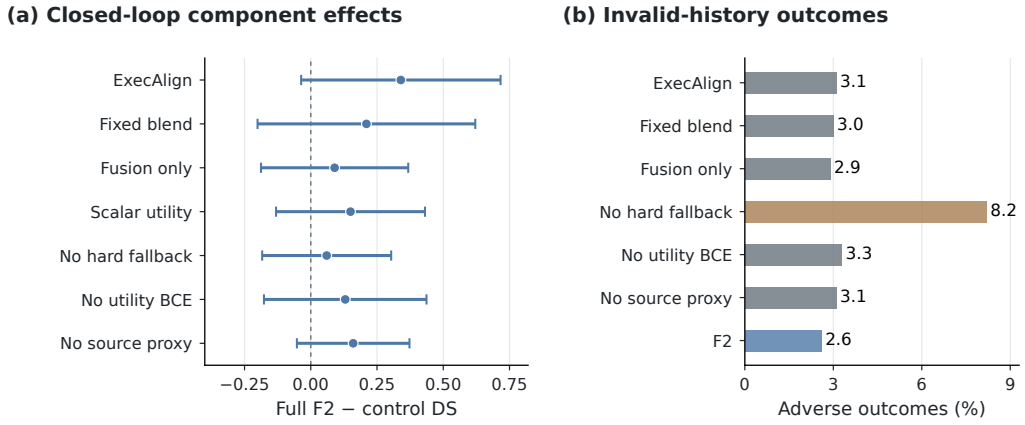


Figure A1: **Score increments and invalid-history handling.** (a) F2-Planner minus control DS with eight-seed pointwise paired 95% intervals. (b) Adverse-outcome percentages in the invalid-history stress test. The two panels measure different aspects of the constructor.

Table A7: **Generic warm-start controls at four NFEs.**

Setting	DS	SR (%)
No warm-start	96.01	88.69
Low forward noise	96.10	88.86
Medium forward noise	96.27	89.09
High forward noise	96.16	88.98
Truncated future only	96.13	88.81

Table A8: **Utility interventions at four NFEs.**

Strategy	DS	SR (%)
Fixed ρ	96.86	90.34
Shuffled utility	96.48	89.43
Learned utility	97.07	91.08
GT-label oracle	97.26	91.65

Table A9: **Adverse-outcome rate in the invalid-history stress suite.**

Configuration	Adverse rate
ExecAlign	3.1%
Fixed blend	3.0%
Fusion only	2.9%
No hard fallback	8.2%
No utility BCE	3.3%
No source proxy	3.1%
F2-Planner	**2.6%**

The GT-label oracle measures the headroom of utility prediction using target-informed labels. The invalid-history suite probes the hard validity boundary, while Appendix I studies valid records whose intent has been corrupted.

E SOURCE GEOMETRY AND PROVENANCE

The source-geometry analysis uses a shared frozen bank of 10,000 states. Every constructor is replayed on the same observation, ego state, previous publication, candidate bank, reference trajectory, and valid support. Source comparisons therefore isolate source construction rather than differences in visited-state distributions. Candidate-dependent summaries use the same highest-scoring deployment candidate selected without target information. The *unpublished future* is the highest-scoring trajectory available immediately before final controller-facing publication selection; no best-of-M oracle is used. Behavioral regimes are assigned once from the shared reference and exogenous state and reused unchanged for every source.

Average displacement error (ADE) averages Euclidean position error over valid waypoints; final displacement error (FDE) measures endpoint error on the same support. Both are reported in meters. Source-center error is measured before stochastic sampling. These are geometric diagnostics, not safety metrics or bounds on the required NFE. Table A11 reports the separate output error after one generator evaluation.

Table A10: **Complete source-center geometry over 10,000 matched cycles.** Geometric errors are in meters.

Source	Mean ADE	ADE P90	Mean FDE
Cycle-local	0.512	0.848	0.653
Executed-history extrap.	0.423	0.696	0.540
Unpublished future	0.298	0.496	0.382
Published Future	0.284	0.512	0.347
ExecAlign	0.218	0.404	0.289
F2-Planner	0.174	0.405	0.245

Table A11: **Residual error after one generator evaluation.** Geometric errors are in meters.

Source	ADE	FDE
Cycle-local	0.344	0.439
Executed-history extrap.	0.299	0.381
Published Future	0.248	0.317
ExecAlign	0.199	0.255
F2-Planner	0.162	0.209

Table A12: **Source ADE by behavioral regime.** Geometric errors are in meters.

Regime	Unpublished future	Published Future	ExecAlign	F2-Planner
Intent change	0.293	0.365	0.318	0.236
Stable motion	0.301	0.267	0.204	0.169
Strong braking	0.296	0.317	0.247	0.201

Table A13: **Execution-aligned source win rate against the selected cycle-local source.**

Population	Win rate
All matched cycles	95.47% (9,547 / 10,000)
Intent change	77.69%
Stable motion	99.14%
Stop-and-go	99.36%
Strong braking	91.78%

The execution-aligned source wins 9,547 of 10,000 matched comparisons. F2-Planner lowers mean center ADE to 0.174 m, while its P90 remains close to ExecAlign (0.405 versus 0.404 m). The improvement in the mean should therefore be distinguished from a reduction of the upper-tail statistic. The behavioral slices characterize how source value varies with the scene.

E.1 Valid-Point Provenance

Each valid waypoint is assigned one of four provenance classes: historical interpolation, historical extrapolation, route completion, or anchor completion.

Table A14: **Waypoint provenance and conditional errors.** Geometric errors are in meters.

Provenance	Point fraction	ExecAlign ADE	F2-Planner ADE
History interpolation	78.20%	0.211	0.176
History extrapolation	17.10%	0.276	0.234
Route completion	3.25%	0.286	0.241
Anchor completion	1.45%	0.291	0.246

Let $e_{i,h}$ be a valid waypoint error and H_i the number of valid points in cycle i. The headline uses cycle-weighted ADE,

$$\mathrm{ADE}_{\mathrm{cycle}} = \frac{1}{N}\sum_i \frac{1}{H_i}\sum_h e_{i,h}. \tag{21}$$

For provenance class c, its conditional waypoint-weighted error is

$$\mathrm{ADE}_c = \frac{\sum_{i,h}\mathbf{1}[z_{i,h}=c]e_{i,h}}{\sum_{i,h}\mathbf{1}[z_{i,h}=c]}. \tag{22}$$

Point fractions sum to 100%, and their weighted error average is the point-weighted overall ADE. It can differ from cycle-weighted ADE when valid support sizes differ across cycles. A cycle may contain multiple provenance types, so cycle-occurrence percentages need not sum to 100%.

F Refinement Budget and Planning-Head Runtime

F.1 Low-NFE Source-Construction Effects

This section measures how source construction affects finite-budget refinement. Aligned warm-start uses the same execution-reconciled publication, support completion, and validity handling as F2-Planner. The remaining controls change one aspect of the recurrent source while keeping the planning stack fixed.

All 11 paired DS contrasts reported here form one Holm family: F2-Planner versus aligned warm-start at 1, 2, and 4 NFEs, and F2-Planner versus four source-construction controls at 1 and 2 NFEs. We control family-wise error at $\alpha = 0.05$.

Table A15: **Low-NFE operating points.** Mean $\pm$ sample SD across eight training seeds.

Method	1-NFE DS	1-NFE SR (%)	2-NFE DS	2-NFE SR (%)
Aligned warm-start	95.49 ± 0.28	87.90 ± 1.52	96.23 ± 0.17	89.26 ± 1.87
Fixed blend	95.66 ± 0.53	88.18 ± 1.42	96.34 ± 0.49	89.49 ± 1.73
Fusion only	95.79 ± 0.54	88.41 ± 1.79	96.44 ± 0.37	89.66 ± 1.53
Scalar utility	95.87 ± 0.53	88.58 ± 1.43	96.50 ± 0.51	89.77 ± 1.65
Fixed blend + adaptive σ	95.74 ± 0.51	88.30 ± 1.43	96.40 ± 0.36	89.55 ± 1.50
F2-Planner	96.04 ± 0.50	88.92 ± 2.39	96.65 ± 0.43	90.23 ± 2.41

Table A16: **Paired DS effects for the low-NFE/source-construction analysis.** Each row reports F2-Planner minus the listed control. All 11 rows belong to one Holm family.

Control	NFE	Mean Δ	95% CI	Raw p	Holm p
Aligned warm-start	1	0.55	$[0.220, 0.880]$	0.00559	0.02237
Aligned warm-start	2	0.42	$[0.130, 0.710]$	0.01106	0.02237
Aligned warm-start	4	0.36	$[0.001, 0.719]$	0.04947	0.04947
Fixed blend	1	0.38	$[0.196, 0.564]$	0.00178	0.01961
Fusion only	1	0.25	$[0.116, 0.384]$	0.00308	0.02199
Scalar utility	1	0.17	$[0.078, 0.262]$	0.00328	0.02199
Fixed + adaptive σ	1	0.30	$[0.150, 0.450]$	0.00217	0.02173
Fixed blend	2	0.31	$[0.151, 0.469]$	0.00244	0.02199
Fusion only	2	0.21	$[0.101, 0.319]$	0.00258	0.02199
Scalar utility	2	0.15	$[0.058, 0.242]$	0.00625	0.02237
Fixed + adaptive σ	2	0.25	$[0.108, 0.392]$	0.00425	0.02199

The ten one- and two-NFE comparisons have Holm-adjusted values at most 0.02237. The four-NFE aligned-warm-start comparison has an adjusted value of 0.04947. These calculations reproduce from the displayed seed-level DS values, which are rounded to two decimals. The margin over aligned warm-start is largest at one NFE and decreases as more refinement is allowed. F2-Planner also exceeds fixed blend, fusion only, scalar utility, and fixed blend with adaptive σ at both low-NFE budgets. The observed trend associates the tested source update with higher DS under limited refinement; it is not a guarantee for other refinement budgets.

Table A17: **Seed-level low-NFE outcomes.** Each cell is DS / integer success count on 220 routes.

Method	Seed 0	Seed 1	Seed 2	Seed 3
Aligned warm-start (1 NFE)	95.38/191	95.62/196	95.41/192	96.00/199
Fixed blend (1 NFE)	95.61/192	95.35/195	96.17/194	96.37/200
Fusion only (1 NFE)	95.57/192	95.70/195	96.06/194	96.37/200
Scalar utility (1 NFE)	95.78/194	95.72/196	96.15/194	96.75/198
Fixed blend + adaptive σ (1 NFE)	95.55/195	95.33/196	95.76/192	96.59/198
F2-Planner (1 NFE)	95.90/192	95.95/200	96.22/192	96.75/204
Aligned warm-start (2 NFE)	96.28/195	96.22/194	96.17/194	96.47/204
Fixed blend (2 NFE)	96.11/194	96.08/195	96.69/196	97.18/204
Fusion only (2 NFE)	95.99/194	96.09/196	96.74/197	96.91/204
Scalar utility (2 NFE)	96.11/197	95.99/196	96.88/196	97.06/205
Fixed blend + adaptive σ (2 NFE)	96.04/195	96.26/198	96.55/196	96.98/204
F2-Planner (2 NFE)	96.33/197	96.24/199	96.91/195	97.20/207

Method	Seed 4	Seed 5	Seed 6	Seed 7
Aligned warm-start (1 NFE)	95.06/191	95.72/196	95.37/189	95.36/193
Fixed blend (1 NFE)	95.09/190	95.20/195	95.21/191	96.28/195
Fusion only (1 NFE)	94.71/187	95.88/198	95.61/194	96.42/196
Scalar utility (1 NFE)	95.11/189	95.60/199	95.47/193	96.38/196
Fixed blend + adaptive σ (1 NFE)	95.03/189	95.72/198	95.62/192	96.32/194
F2-Planner (1 NFE)	95.22/193	95.89/199	95.71/188	96.68/197
Aligned warm-start (2 NFE)	95.92/191	96.20/198	96.15/195	96.43/200
Fixed blend (2 NFE)	95.63/193	96.69/200	95.98/194	96.36/199
Fusion only (2 NFE)	96.07/194	96.82/199	96.32/195	96.58/199
Scalar utility (2 NFE)	95.99/194	96.98/200	96.02/194	96.97/198
Fixed blend + adaptive σ (2 NFE)	96.00/193	96.54/197	96.10/195	96.73/198
F2-Planner (2 NFE)	96.15/191	97.11/202	96.32/194	96.94/203

The F2-Planner–aligned-warm-start DS difference is positive for all eight paired seeds at both one and two NFEs. SR follows the same aggregate direction: +1.02 percentage points at one NFE and +0.97 points at two NFEs, with paired intervals $[0.217, 1.829]$ and $[0.194, 1.738]$.

The two-NFE planning-head latency comparison uses the nearest measured operating point within a 0.50 ms tolerance. F2-Planner-2 measures 34.17 ms and aligned-warm-start-2 measures 33.88 ms, a 0.29 ms gap. No interpolation is used. This comparison concerns DS and planning-head cost, not event response or recovery at two NFEs.

F.2 Quality–Compute Frontier and Runtime

All timing variants use the same profiling harness at batch size one on a single RTX 4090. Models run in evaluation/inference mode at deployment precision after a fixed warm-up. GPU work is synchronized at the start and end of each timed region, and every configuration uses the same timing API under an otherwise idle GPU. The reported planning-head time includes preprocessing, temporal shift, execution alignment, utility/source construction, generator evaluations, selection, cache commit, and dispatch jitter, but excludes perception, sensor transfer, simulator ticks, and downstream control. These measurements therefore do not represent end-to-end sensor-to-actuation latency.

Table A18: **Primary diffusion-bridge quality frontier.** Entries are DS / SR(%).

Source	1 NFE	2 NFE	4 NFE	8 NFE	20 NFE
Cycle-local	93.78/84.66	95.18/87.39	96.01/88.69	96.29/89.20	96.33/89.32
Published Future	94.82/86.48	95.84/88.13	96.41/89.20	96.55/89.43	–
ExecAlign	95.39/87.78	96.34/89.38	96.73/89.94	96.88/90.11	–
F2-Planner	96.04/88.92	96.65/90.23	97.07/91.08	97.21/91.36	–

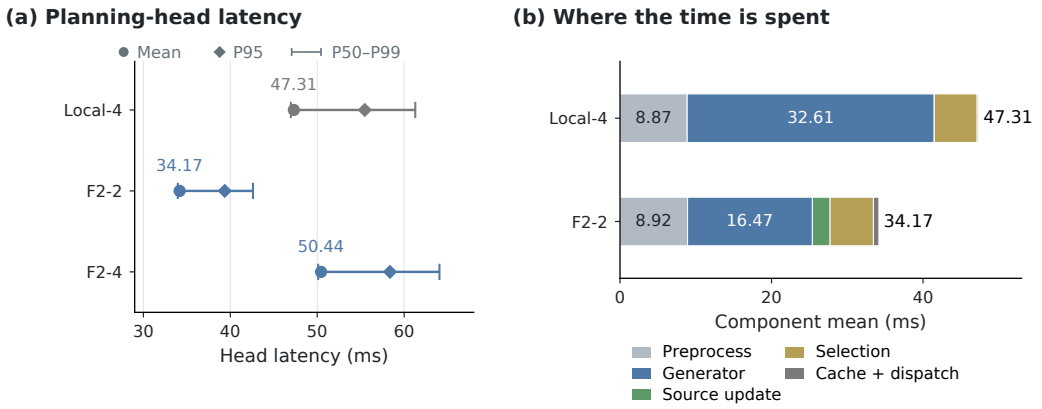


Figure A2: **Planning-head runtime distribution and component costs.** (a) Mean and quantiles of total planning-head latency. The P50–P99 segment is a runtime range, not a confidence interval. (b) Component means for Local-4 and F2-Planner-2. Source update combines shift, alignment, and utility construction. Perception and downstream control are excluded.

Table A19: **Repeated planning-head timing (ms).**

Configuration	Mean	P50	P95	P99
Cycle-local-4	47.31	46.97	55.47	61.28
F2-Planner-2	34.17	33.96	39.36	42.61
F2-Planner-4	50.44	50.11	58.37	64.06

Table A20: **Planning-head component means (ms).** The total head is reported separately from individual components.

Component	Local-4	F2-Planner-2
Preprocess	8.87	8.92
Temporal shift	0.00	0.49
Execution alignment	0.00	0.82
Utility/source	0.00	1.03
Generator	32.61	16.47
Selection	5.66	5.72
Cache commit	0.00	0.55
Dispatch jitter	0.17	0.17
Total head	47.31	34.17

Total-head summaries use five profiling sessions with 3,000 samples per configuration per session. P50, P95, and P99 are computed from total planning-head runtime samples. F2-Planner-2 lowers mean planning-head latency from 47.31 to 34.17 ms. F2-Planner-4 takes 50.44 ms. The decomposition attributes the measured planning-head reduction at F2-Planner-2 to halving generator evaluations; the recurrent source update adds the component costs listed in Table A20.

For F2-Planner-2, state-stratified means are 34.25 ms with normal history, 34.53 ms with completion, 31.74 ms under fallback, and 31.82 ms on the first cycle. The matched aligned-warm-start-2 configuration has a mean planning-head latency of 33.88 ms.

G SOURCE–REFINER CROSSOVER

The crossover tests whether source quality depends on the refiner's training distribution. It uses a symmetric 2×2 design. One refiner family is trained with the execution-aligned source and the other with the F2-Planner source. At evaluation time, each frozen refiner is paired with both source constructions. For each of eight paired training seeds, all four cells use the same 1,250 matched cycles, reference trajectories, and valid supports. Residual ADE is averaged within each seed before inference across the eight seed-level contrasts.

Table A21: **Symmetric source–refiner crossover.** Residual ADE (m), averaged over eight seeds and 1,250 matched cycles per seed.

	1 NFE		2 NFEs	
Refiner training source	ExecAlign src.	F2-Planner src.	ExecAlign src.	F2-Planner src.
ExecAlign	0.176	0.230	0.142	0.177
F2-Planner	0.191	**0.151**	0.155	**0.119**

Write $E^{(K)}_{i,g,s}$ for seed i's mean error, where $g \in \{E, F\}$ indexes refiners trained with the ExecAlign and F2-Planner sources, respectively, while $s \in \{E, F\}$ indexes the corresponding source supplied at evaluation. Define

$$d^{(K)}_{i,E} = E^{(K)}_{i,E,F} - E^{(K)}_{i,E,E}, \tag{23}$$

$$d^{(K)}_{i,F} = E^{(K)}_{i,F,F} - E^{(K)}_{i,F,E}, \tag{24}$$

$$I^{(K)}_{i} = d^{(K)}_{i,E} - d^{(K)}_{i,F}. \tag{25}$$

A positive interaction means that replacing ExecAlign by the F2-Planner source is less favorable under the ExecAlign-trained refiner than under the refiner trained with the F2-Planner source.

Table A22: **Symmetric source–refiner inference.** ADE effects are in meters; inference uses eight seed-level contrasts.

NFE	Contrast	Mean	95% CI	p
1	ExecAlign-trained source effect	0.054	$[0.049, 0.059]$	6.00e-08
1	F2-Planner-trained source effect	-0.040	$[-0.047, -0.033]$	2.65e-06
1	Source×refiner interaction	0.094	$[0.085, 0.103]$	5.00e-08
2	ExecAlign-trained source effect	0.035	$[0.026, 0.044]$	3.32e-05
2	F2-Planner-trained source effect	-0.036	$[-0.042, -0.030]$	2.41e-06
2	Source×refiner interaction	0.071	$[0.061, 0.081]$	8.90e-07

Table A23: **Source–refiner seed means at 1 NFE.** EE/EF denote ExecAlign-trained refiner with ExecAlign/F2-Planner source; FE/FF denote F2-Planner-trained refiner with ExecAlign/F2-Planner source.

Seed	EE	EF	FE	FF
0	0.174	0.228	0.196	0.153
1	0.193	0.244	0.210	0.172
2	0.180	0.225	0.202	0.154
3	0.152	0.210	0.170	0.119
4	0.165	0.216	0.173	0.145
5	0.182	0.234	0.183	0.154
6	0.192	0.246	0.200	0.156
7	0.170	0.237	0.194	0.155
Mean	0.176	0.230	0.191	0.151

Table A24: **Source–refiner seed means at 2 NFEs.** Notation follows Table A23.

Seed	EE	EF	FE	FF
0	0.116	0.168	0.142	0.095
1	0.144	0.173	0.146	0.114
2	0.141	0.168	0.147	0.101
3	0.147	0.166	0.161	0.121
4	0.130	0.165	0.139	0.109
5	0.151	0.195	0.171	0.142
6	0.157	0.189	0.168	0.134
7	0.150	0.192	0.166	0.136
Mean	0.142	0.177	0.155	0.119

The source effect changes sign across the two training distributions at both budgets. The interaction is 0.094 m after one NFE and 0.071 m after two NFEs. Both paired-seed intervals exclude zero. This crossover establishes a source–refiner interaction in the tested families: the F2-Planner source helps the F2-Planner-trained refiner but not the frozen ExecAlign-trained refiner. It does not by itself establish that joint optimization is the only possible training procedure.

H Temporal Continuity, Event Response, and Utility

All compared methods in the continuity, event-response, and utility studies use four NFEs. Each NFE counts one denoiser evaluation; F2-Planner receives no additional generator evaluations. The findings in this section apply to the evaluated four-NFE configurations and do not establish the corresponding behavior at two NFEs.

H.1 Cross-Cycle Continuity and Response Coverage

Inter-cycle ADE compares the current publication with the previous publication after the same progress advance and current-frame alignment used by the method, restricted to their common valid support. Switch rate applies one geometric threshold fixed before evaluation to this aligned cross-cycle change. The metric therefore measures trajectory-level switching rather than changes in a method-specific candidate identity.

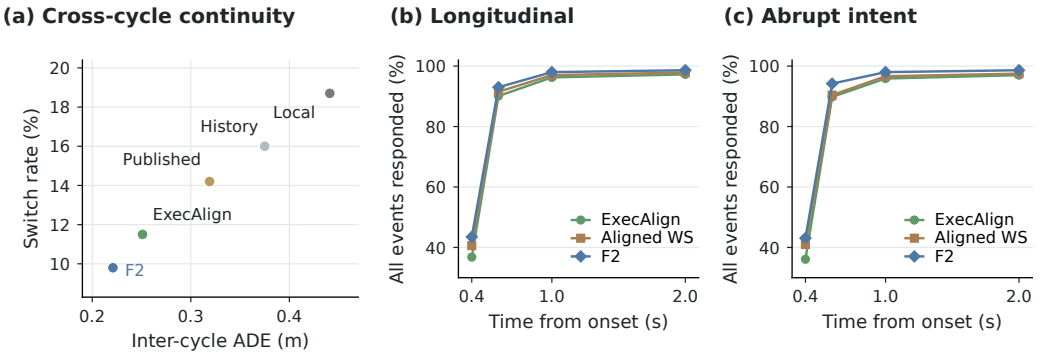


Figure A3: **Continuity and observed response coverage at four NFEs.** (a) Inter-cycle deviation versus switch rate. (b,c) Fractions of all evaluation instances responded by the reported time thresholds. All compared methods use four NFEs. Lines connect summaries at 0.4, 0.6, 1.0, and 2.0 s, not reconstructed event-level survival curves.

Table A25: **Cross-cycle continuity at four NFEs.**

Source	Inter-cycle ADE (m)	Switch rate (%)
Cycle-local	0.441	18.7
Executed-history extrap.	0.375	16.0
Published Future	0.319	14.2
ExecAlign	0.251	11.5
F2-Planner	**0.221**	**9.8**

Table A26: **Response-by-time summaries at four NFEs.** Percentages use all 800 evaluation instances.

Suite	Method	By 0.4 s (%)	By 0.6 s (%)	By 1 s (%)
Longitudinal	ExecAlign	36.8	90.1	96.25
Longitudinal	Aligned WS	40.6	91.5	97.00
Longitudinal	F2-Planner	43.5	93.0	98.00
Abrupt intent	ExecAlign	36.1	89.8	95.88
Abrupt intent	Aligned WS	41.0	90.4	96.63
Abrupt intent	F2-Planner	43.1	94.2	98.00

F2-Planner has the lowest inter-cycle ADE and switch rate in Table A25. Its response coverage is also highest at the reported early time thresholds. Responses between 1 and 2 s remain included in the all-instance RMST.

H.2 Event Definitions and Response Inference

The event bank is constructed once from the shared reference process before method comparison. Each of the 160 physical-event IDs contributes five predefined neighboring evaluation initializations

around the same underlying scene transition; these offsets are fixed before any method is evaluated. The five instances share route, scenario, physical-event identity, event definition, and exogenous scene realization, and remain a single cluster for paired bootstrap. Training-seed/checkpoint identities are retained so that all methods are evaluated on the same instances. Two-second trajectory quantities are obtained from the evaluation stack's speed-based parameterization of spatial trajectory support. No target-informed candidate selection is used.

Evaluation grid and event windows. All event-response quantities are evaluated on a 0.05 s grid at four NFEs. Response time is elapsed scenario time from event onset, not planning-head computation latency. The pre-event window is 1.00 s and the response window is 2.00 s. Event construction is method-independent: onset time t_0 and event direction are determined only from shared reference signals and are then reused unchanged for every method.

Longitudinal events. Let

$$a_{\mathrm{pre}} = \mathrm{median}\{a_{\mathrm{ref}}(t) : t \in [t_0 - 1.00\,\mathrm{s}, t_0)\}. \tag{26}$$

A longitudinal event begins at the first timestamp t_0 at which the shared reference longitudinal acceleration differs from a_{pre} by at least $1.50\,\mathrm{m/s^2}$. The sign of this change must remain unchanged through $t_0 + 0.25$ s on the 0.05 s evaluation grid. Equivalently, the qualifying direction must hold at every evaluation timestamp from t_0 through $t_0 + 0.25$ s, inclusive.

The response direction is

$$d_{\mathrm{long}} = \mathrm{sign}\left(a_{\mathrm{ref}}(t_0 + 0.25\,\mathrm{s}) - a_{\mathrm{pre}}\right), \tag{27}$$

and the method-specific response score is

$$s_{\mathrm{long}}(t) = d_{\mathrm{long}}\left(a_{\mathrm{target}}(t) - a_{\mathrm{pre}}\right). \tag{28}$$

A qualifying longitudinal response requires $s_{\mathrm{long}}(t) \geq 0.75\,\mathrm{m/s^2}$ for three consecutive evaluation steps.

Abrupt-intent events. Abrupt intent is defined from the shared reference trajectory, not from any method-specific publication. Let the shared pre-event reference trajectory be the reference available immediately before t_0. An abrupt-intent event begins at the first timestamp t_0 for which the shared reference at a 2.00 s horizon changes relative to that pre-event reference by at least 1.00 m laterally or 0.20 rad in heading. The new reference intent must remain in the same change direction through $t_0 + 0.25$ s on the 0.05 s evaluation grid. All retained abrupt-intent events also exhibit a lateral change of at least 1.00 m at the 2.00 s horizon; no heading-only event is included in the final 800-instance suite. All lateral and heading differences are evaluated in the common current ego-frame convention used by the trajectory metrics.

Let

$$d_{\mathrm{lat}} = \mathrm{sign}\left(y^{2\mathrm{s}}_{\mathrm{ref},\,t_0} - y^{2\mathrm{s}}_{\mathrm{ref},\,\mathrm{pre}}\right), \tag{29}$$

where $y^{2\mathrm{s}}$ denotes lateral position at the 2.00 s horizon. For an evaluated method, the lateral response score is

$$s_{\mathrm{lat}}(t) = d_{\mathrm{lat}}\left(y^{2\mathrm{s}}_{\mathrm{pred},\,t} - y^{2\mathrm{s}}_{\mathrm{pub},\,\mathrm{pre}}\right), \tag{30}$$

where $y^{2\mathrm{s}}_{\mathrm{pub},\,\mathrm{pre}}$ is that method's own pre-event published trajectory at the same horizon. A qualifying abrupt-intent response requires both

$$s_{\mathrm{lat}}(t) \geq 0.75\,\mathrm{m} \tag{31}$$

and a heading change of at least 0.10 rad toward the new shared reference intent, for three consecutive evaluation steps.

Event exclusions. Eligibility filtering is applied once during construction of the shared event bank. A candidate event is excluded if a collision occurs before t_0, if the previous publication is invalid at t_0, if less than 1.00 s of pre-event history is available, or if another qualifying event occurs within $[t_0 - 1.00\,\mathrm{s},\ t_0 + 2.00\,\mathrm{s}]$. The resulting retained physical-event IDs are fixed and reused unchanged for F2-Planner, aligned warm-start, and ExecAlign.

Response time and administrative censoring. A persistence run is counted only if all three qualifying evaluation timestamps lie within the 2.00 s response window. Thus, on the 0.05 s grid, the latest possible first timestamp of a qualifying three-step run is $t_0 + 1.90$ s. Response time is the timestamp of the first step in the first qualifying run minus t_0. If no fully contained qualifying run exists, the event is a nonresponse.

For evaluation instance i, define

$$T_i = \begin{cases} t_i^{\text{resp}}, & \text{if a qualifying response occurs,} \\ 2.00\,\text{s}, & \text{otherwise.} \end{cases} \tag{32}$$

The primary all-event restricted mean is

$$\text{RMST}(2) = \frac{1}{800} \sum_{i=1}^{800} T_i. \tag{33}$$

Response-by-t uses all 800 evaluation instances as the denominator:

$$R(t) = \frac{1}{800} \sum_{i} \mathbf{1}[T_i^{\text{resp}} \le t]. \tag{34}$$

Secondary event estimands. For a pairwise comparison, the common-responder set contains instances for which both methods respond within the 2.00 s window. The common-responder difference is

$$\frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left(t_{i,\text{F2-Planner}}^{\text{resp}} - t_{i,\text{comp}}^{\text{resp}} \right). \tag{35}$$

The nonresponse-rate difference is

$$100 \left(\frac{N_{\text{nonresp}}^{\text{F2-Planner}}}{800} - \frac{N_{\text{nonresp}}^{\text{comp}}}{800} \right) \tag{36}$$

in percentage points.

Physical-event clusters and paired bootstrap. The cluster identifier is `route_id + scenario_id + physical_event_id`. Each identifier corresponds to one underlying physical scene transition and contributes five predefined neighboring evaluation instances in the suite. The five instances share route, scenario, physical-event identity, and event definition; they are not separate physical events or independent repetitions. Every method is evaluated on the same five instances, so the five observations are kept together during resampling and treated as dependent. The paired bootstrap follows Appendix A.3: 160 physical-event cluster IDs are sampled with replacement for each of 10,000 replicates, repeated draws retain their multiplicity, and the same sampled cluster multiset is applied to both methods.

Table A27: **All-event response outcomes at four NFEs.** Each row has 800 evaluation instances; RMST is in seconds. Collisions are descriptive counts, not a two-NFE safety evaluation.

Suite	Method	Response	Nonresponse	RMST	Collisions
Longitudinal	ExecAlign	778	22	0.491	17
Longitudinal	Aligned WS	783	17	0.486	15
Longitudinal	F2-Planner	789	11	0.463	14
Abrupt intent	ExecAlign	776	24	0.496	18
Abrupt intent	Aligned WS	780	20	0.490	16
Abrupt intent	F2-Planner	789	11	0.464	12

Table A28: **Primary event effect at four NFEs: F2-Planner minus control RMST.** Effects are in milliseconds of response time, not computation latency; intervals are paired cluster-bootstrap percentile 95% intervals.

Suite	Comparator	Effect (ms)	95% CI
Longitudinal	ExecAlign	-28.0	$[-43.0, -14.0]$
Longitudinal	Aligned WS	-23.0	$[-39.0, -7.0]$
Abrupt intent	ExecAlign	-32.0	$[-48.0, -16.0]$
Abrupt intent	Aligned WS	-26.0	$[-42.0, -9.0]$

Table A29: **Secondary response contrasts at four NFEs.** Common-response contrasts use paired responder intersections; nonresponse uses all 800 evaluation instances. Intervals are pointwise.

Suite	Comparator	Metric	Effect	95% CI
Longitudinal	ExecAlign	Common response (ms)	-20.0	$[-32.0, -7.0]$
Longitudinal	ExecAlign	Nonresponse (pp)	-1.375	$[-2.500, -0.250]$
Longitudinal	Aligned WS	Common response (ms)	-13.0	$[-25.0, 1.0]$
Longitudinal	Aligned WS	Nonresponse (pp)	-0.750	$[-1.625, 0.125]$
Abrupt intent	ExecAlign	Common response (ms)	-21.0	$[-35.0, -6.0]$
Abrupt intent	ExecAlign	Nonresponse (pp)	-1.625	$[-2.875, -0.375]$
Abrupt intent	Aligned WS	Common response (ms)	-14.0	$[-28.0, 1.0]$
Abrupt intent	Aligned WS	Nonresponse (pp)	-1.125	$[-2.125, -0.125]$

Negative effects indicate shorter response times or fewer nonresponses at four NFEs. All-instance RMST is the primary summary because it retains both response time and nonresponse within a common physical-event population. These comparisons concern the specified response criteria, not a guarantee of collision avoidance.

H.3 Utility Dynamics and Source Revision

The pre/event/post summaries use fixed event-relative bins defined once for the shared event bank and applied identically to the stable and stale-intent groups. Phase assignment is independent of the evaluated method. The cycle-level utility score averages horizon weights over valid support:

$$r_t = \frac{\sum_h w_{t,h}\rho_{t,h}}{\sum_h w_{t,h}}. \tag{37}$$

Higher utility favors inheritance. We score strict geometric losses of the aligned source with $1 - r_t$ and compare continuous source advantage with r_t. Utility reaches 0.884 AUROC, 0.572 average precision, and 0.612 Spearman correlation. Strict losses occur in 453 of 10,000 cycles (4.53%). These four-NFE diagnostics measure geometric source usefulness, not the probability of safe execution.

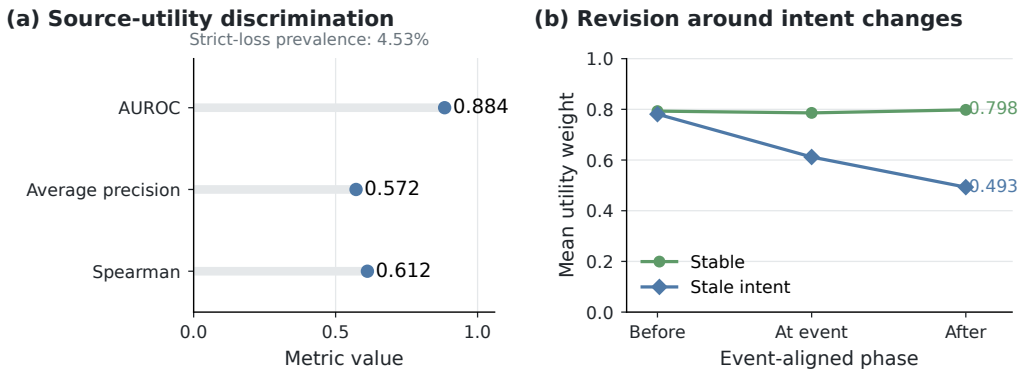


Figure A4: **Utility and revision at four NFEs.** (a) Source-utility discrimination and ranking summaries. (b) Mean horizon weights in the reported before/at/after event phases. Phase positions are categorical, not equally spaced physical times.

Table A30: **Event-aligned inheritance weights at four NFEs.**

Group	Pre-event ρ	Event ρ	Post-event ρ
Stable	0.793	0.786	0.798
Stale intent	0.781	0.612	0.493

The reported source scales are 0.681 m for stable events at onset, 1.195 m for stale-intent events at onset, and 1.546 m after stale intent. Scale is a deterministic function of utility under Equation 5. The diagram therefore plots utility rather than a second, redundant scale axis.

I Recovery from Incorrect Recurrent State

All methods in the single-fault and sustained-drift comparisons use matched four-NFE configurations, with one denoiser evaluation per NFE and no additional generator evaluations for F2-Planner. The results characterize these four-NFE stress tests, not recovery or safety at two NFEs.

Each corruption case is paired with an uncorrupted reference rollout that shares the same route, scenario, training seed/checkpoint, initial simulator state, and exogenous-agent realization. The recurrent publication corruption is the only intervention. Recovery is measured relative to this paired clean future. Returning to that reference is an operational recovery criterion, not proof that either rollout is safe.

Case counts and horizons. The single-fault and sustained-drift suites each contain 160 cases, with 20 cases per training seed. Recovery is evaluated on the same 0.05 s grid used for event response. The single-fault administrative horizon is 8.00 s. Sustained drift lasts 1.00 s; because the common rollout ends 8.00 s after initial fault injection, the post-drift recovery horizon is 7.00 s.

Wrong-publication injection. At injection time t_0, we construct a corrupted publication directly from the clean recurrent publication using a deterministic, validity-preserving intent perturbation. At the 2.00 s horizon, the perturbation applies a lateral offset of ± 1.25 m and a heading offset of ± 0.20 rad; the sign is fixed by case-ID parity. The original publication timestamp, support mask, route state, and speed metadata are retained. The constructed corruption is required to remain inside the recurrent-state validity boundary, including translation mismatch below 2.00 m and yaw mismatch below 0.80 rad. Thus the stress test changes the intended future while preserving a syntactically valid recurrent record.

Single-fault protocol. The wrong publication is injected for exactly one planning cycle at t_0. Normal recurrent updates resume immediately afterward. Recovery time is measured from t_0.

Sustained-drift protocol. During $[t_0, t_0 + 1.00\,\mathrm{s}]$, a linearly increasing perturbation is applied:

$$\Delta y(t) = q\,(1.25\,\mathrm{m})\frac{t - t_0}{1.00\,\mathrm{s}}, \tag{38}$$

$$\Delta\psi(t) = q\,(0.20\,\mathrm{rad})\frac{t - t_0}{1.00\,\mathrm{s}}, \tag{39}$$

where $q \in \{-1, +1\}$ is the case-parity sign. The perturbation terminates at $t_0 + 1.00$ s, and recovery time is measured from that termination instant.

Recovery criterion. At each evaluation step, both the evaluated publication and the clean reference publication are transformed into the current ego frame. Recovery path ADE is the mean Euclidean distance over the first ten geometrically corresponding waypoints. Recovery heading error is the absolute wrapped heading difference at the tenth waypoint, and recovery speed error is the absolute target-speed difference at that waypoint. A step is recovered if

$$\mathrm{ADE}_{\mathrm{rec}} \le 0.25\,\mathrm{m}, \tag{40}$$

$$|\Delta\psi_{\mathrm{rec}}| \le 0.10\,\mathrm{rad}, \tag{41}$$

$$|\Delta v_{\mathrm{rec}}| \le 0.50\,\mathrm{m/s}. \tag{42}$$

A recovery requires five consecutive recovered steps. The full five-step run must lie inside the administrative recovery horizon. Since the grid spacing is 0.05 s, the persistence run spans 0.20 s from its first to fifth timestamp.

For a single fault,

$$t_i^{\mathrm{rec}} = t_{i,\mathrm{first}}^{(5)} - t_0, \tag{43}$$

where $t_{i,\mathrm{first}}^{(5)}$ is the first timestamp of the first fully contained five-step recovered run. For sustained drift,

$$t_i^{\mathrm{rec}} = t_{i,\mathrm{first}}^{(5)} - (t_0 + 1.00\,\mathrm{s}). \tag{44}$$

Censoring, RMST, and recovery rate. For the single-fault suite,

$$T_i = \begin{cases} t_i^{\mathrm{rec}}, & t_i^{\mathrm{rec}} \le 8.00\,\mathrm{s}, \\ 8.00\,\mathrm{s}, & \text{otherwise}, \end{cases} \tag{45}$$

and for sustained drift,

$$T_i = \begin{cases} t_i^{\mathrm{rec}}, & t_i^{\mathrm{rec}} \le 7.00\,\mathrm{s}, \\ 7.00\,\mathrm{s}, & \text{otherwise}. \end{cases} \tag{46}$$

The all-case restricted mean is

$$\mathrm{RMST}(\tau) = \frac{1}{160}\sum_{i=1}^{160} T_i = \frac{N_{\mathrm{rec}}\bar{T}_{\mathrm{rec}} + N_{\mathrm{cens}}\tau}{160}, \tag{47}$$

with $\tau = 8$ s for single faults and $\tau = 7$ s for sustained drift. The recovery rate is the fraction of the 160 cases with a fully contained qualifying recovery run within the corresponding horizon.

Paired recovery bootstrap. Recovery intervals use 10,000 paired, seed–scenario-stratified bootstrap replicates. Within each observed training-seed $\times$ scenario stratum, case IDs are sampled with replacement using the original stratum size. The same sampled case multiset is applied to F2-Planner and the comparator before computing the metric difference. Repeated case draws retain their multiplicity. Percentile 95% intervals are the empirical 0.025 and 0.975 quantiles. The bootstrap random seed is 20250317.

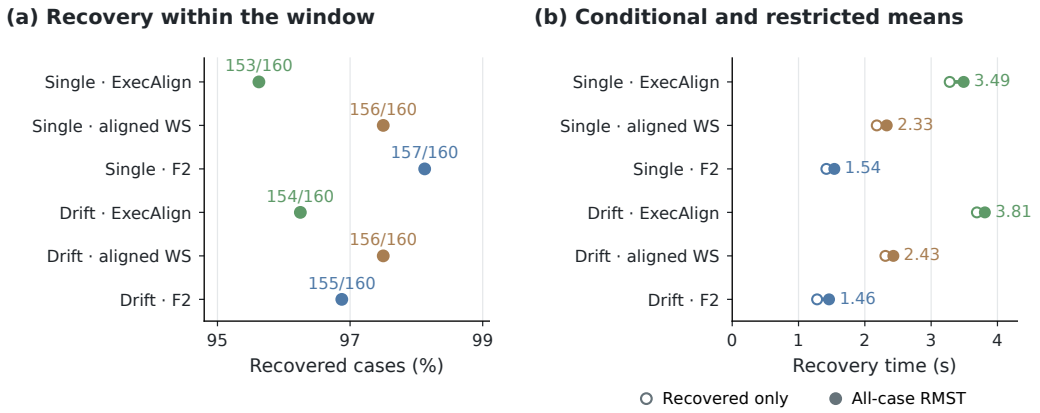


Figure A5: **Recovery probability and time at four NFEs.** (a) Recovered cases out of 160. (b) Conditional recovery means (open markers) and all-case RMST (filled markers). All methods use four NFEs. Single-fault and sustained-drift RMST use 8 s and 7 s horizons, respectively.

Table A31: **Recovery from valid but incorrect recurrent state at four NFEs.** Conditional means, P90, and RMST are in seconds.

Fault	Method	Rec.	Rate	Mean_{rec}	P90_{rec}	RMST
Single, $\tau = 8$s	ExecAlign	153/160	95.63%	3.28	4.10	3.49
	Aligned warm-start	156/160	97.50%	2.18	2.60	2.33
	F2-Planner	**157/160**	**98.13%**	**1.42**	**1.80**	**1.54**
Drift, $\tau = 7$s	ExecAlign	154/160	96.25%	3.69	4.40	3.81
	Aligned warm-start	**156/160**	**97.50%**	2.31	2.80	2.43
	F2-Planner	155/160	96.88%	**1.28**	**1.60**	**1.46**

Table A32: **Paired recovery effects at four NFEs.** RMST differences are seconds and recovery-rate differences are percentage points. Intervals use 10,000 paired, seed/scenario-stratified bootstrap replicates.

Fault	Comparator	ΔRMST	95% CI	Δrate	95% CI
Single	ExecAlign	-1.95	$[-2.38, -1.53]$	2.500	$[-0.625, 5.625]$
Single	Aligned WS	-0.79	$[-1.12, -0.47]$	0.625	$[-1.375, 2.750]$
Drift	ExecAlign	-2.35	$[-2.79, -1.91]$	0.625	$[-2.500, 3.750]$
Drift	Aligned WS	-0.97	$[-1.31, -0.63]$	-0.625	$[-3.125, 1.875]$

Mean final path errors are 0.089/0.071/0.049 m for single faults and 0.079/0.064/0.056 m under drift, in ExecAlign/aligned-warm-start/F2-Planner order. At four NFEs, F2-Planner has the lowest RMST under both fault types. The paired recovery-rate intervals include zero (Table A32). The supported comparison is a reduction in restricted recovery time under the tested faults, not a demonstrated increase in recovery probability.

J Transfer to Truncated Diffusion

The transfer study replaces the primary diffusion bridge with a controlled truncated-diffusion generator in the style of DiffusionDrive (Liao et al., 2025). F2-Planner changes only the trajectory source before forward noising; the remaining noise schedule and denoising updates are left unchanged. Perception, candidate generation, selection, and downstream control remain shared.

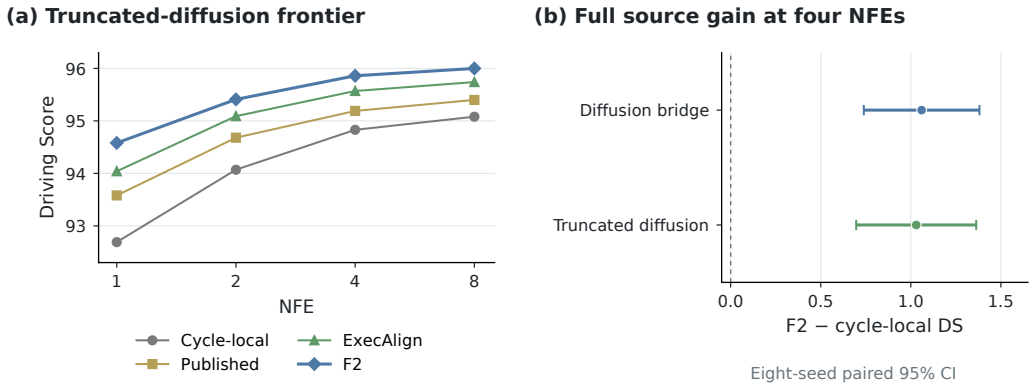


Figure A6: **Transfer within the shared planning stack.** (a) Truncated-diffusion DS versus NFE. (b) Eight-seed paired 95% intervals for F2-Planner minus cycle-local DS at four NFEs in each formulation.

Table A33: **Truncated-diffusion quality frontier.** Entries are DS / SR(%).

Source	1 NFE	2 NFE	4 NFE	8 NFE
Cycle-local	92.69/82.90	94.07/85.28	94.83/86.88	95.08/87.27
Published Future	93.58/84.20	94.68/86.36	95.19/87.44	95.40/87.73
ExecAlign	94.04/85.17	95.09/87.27	95.57/88.13	95.74/88.41
F2-Planner	94.58/86.02	95.41/87.90	95.86/88.69	96.00/89.03

Table A34: **Four-NFE truncated-diffusion seed-level outcomes.** Each cell is DS / integer success count on 220 routes.

Source	Seed 0	Seed 1	Seed 2	Seed 3
Cycle-local	95.15/193	94.67/189	95.09/196	94.90/191
Published Future	95.04/195	95.13/187	95.54/199	95.53/197
ExecAlign	95.72/196	95.08/195	95.83/201	95.82/192
F2-Planner	95.73/196	95.92/193	95.88/199	95.98/198

Source	Seed 4	Seed 5	Seed 6	Seed 7
Cycle-local	94.73/184	94.77/191	94.32/190	95.01/195
Published Future	95.08/190	95.26/193	94.68/190	95.26/188
ExecAlign	95.52/192	95.27/191	95.59/191	95.73/193
F2-Planner	95.28/194	95.95/189	96.09/192	96.05/200

At four NFEs, F2-Planner improves the matched truncated-diffusion planner from 94.83 ± 0.27 to 95.86 ± 0.26 DS and from $86.88 \pm 1.71\%$ to $88.69 \pm 1.73\%$ SR. The full source has the highest reported mean DS and SR at each measured NFE budget. These comparisons change the generator formulation within the shared stack; they do not evaluate transfer to a new perception or control architecture.