BEYOND AVERAGE SUCCESS: OUTCOME REVERSALS AND DELAYED DIVERGENCE IN QUANTIZED VISION-AND-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization is widely used to compress models, reduce memory requirements, and potentially accelerate inference. When average performance is preserved, a natural expectation is that quantization introduces only local numerical perturbations while leaving model outputs and most individual outcomes largely unchanged. However, our experiments in streaming vision-and-language navigation (VLN) reveal substantial behavioral differences behind similar average scores. Changed trajectories can also alter subsequent observations and histories. Using instruction-paired R2R-CE episodes, we compare four-bit weight quantization of StreamVLN against its unquantized bfloat16 (BF16) model as the full-precision baseline. Success rate decreases by only 0.11 percentage points, yet 18.5% of outcomes reverse and 87.0% of action sequences change, as gains nearly offset losses. To examine this discrepancy, fixed-state probes separate numerical perturbations from immediate decisions and distinguish the effects of current computation and historical content. Further, controlled continuations show that history perturbations can induce later divergence even when the initial action queue is held fixed and subsequent computation returns to BF16. Together, these findings show that similar average performance does not imply behavioral fidelity, and immediate action agreement does not guarantee subsequent agreement in the tested states. Behavioral differences need not reduce task success. Quantized VLN deployment therefore calls for evaluating retained task outcomes and subsequent behavior alongside average quality, realized memory savings, and end-to-end latency.

1 INTRODUCTION

Vision-and-language navigation (VLN) requires an agent to follow language instructions through an environment, using visual observations to select actions and decide when to stop (Anderson et al., 2018b). Each action can change subsequent observations and the history available to later decisions. NaviLLM adapts pretrained language models to navigation, while StreamVLN maintains context over incoming video streams (Zheng et al., 2024; Wei et al., 2026). Model storage, repeated inference, and retained history create deployment costs, motivating compression that preserves navigation quality.

Quantization reduces numerical precision to save memory and potentially accelerate inference. NaVILA applies weight quantization to navigation and reports resource benefits while retaining performance in its evaluation (Cheng et al., 2025); in robotic manipulation, QVLA uses action sensitivity and temporal deviations to guide precision allocation (Xu et al., 2026). These developments make the behavioral consequences of quantization relevant to embodied deployment. Average task completion, however, does not reveal whether a compressed model solves the same instructions or follows the same trajectories.

When average scores remain close, a natural expectation is that quantization leaves most individual outcomes and behaviors unchanged. Our streaming VLN experiments reveal a different picture: gains and losses can cancel while trajectories change substantially. Those changes can in turn alter later observations and histories. We ask: *when average scores remain close, which outcomes and be-*

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.

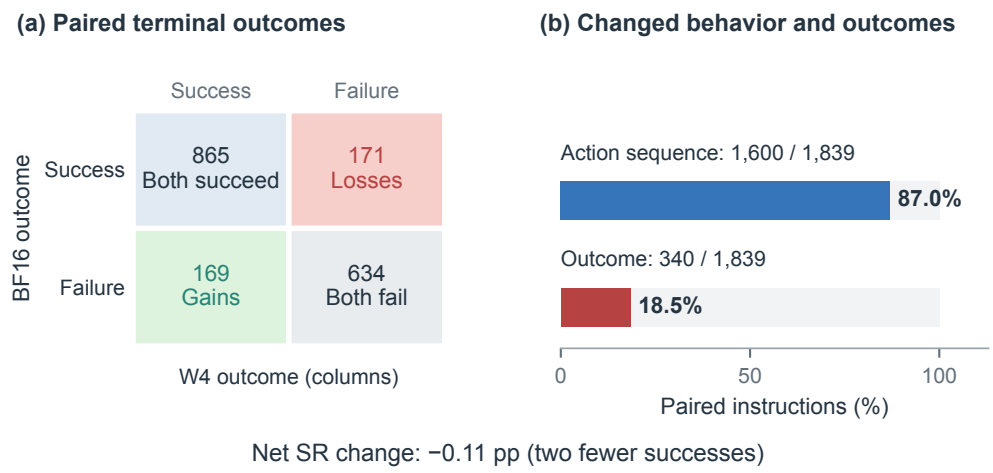

Figure 1: **Two fewer successes, but 340 reversed outcomes.** On 1,839 paired StreamVLN instructions, 169 gains nearly cancel 171 losses. The matrix compares terminal outcomes across models; 1,600 action sequences change. The net SR difference is -0.11 pp, computed before rounding.

haviors change, and can altered history change later navigation despite matched initial execution? We study this question through instruction-paired StreamVLN episodes, fixed-state probes, and controlled continuations, using the unquantized bfloat16 (BF16) model as the full-precision baseline. The experiments yield three findings.

Finding 1: Average success conceals individual reversals. On 1,839 paired R2R-CE instructions, four-bit weight quantization (W4) reduces success rate (SR) by only 0.11 percentage points, yet reverses 18.5% of success/failure outcomes and changes 87.0% of action sequences (Figure 1). Pairs that both succeed can also follow different trajectories. Controlled NaviLLM and static/dynamic eight-bit floating-point (FP8) comparisons show analogous near-cancellation within their own protocols, separating average quality from individual outcome and trajectory retention (Section 4).

Finding 2: Numerical changes need not change immediate decisions. With external inputs and teacher prefixes fixed, one probe population shows distribution changes without action changes. In a separate population, substituting either current computation or historical content alone can change actions. Within matched states, flips concentrate where competing actions have similar reference scores. These results distinguish numerical sensitivity from crossing an action decision boundary. The two probe populations are not state-matched, so their contrast cannot identify an effect of reference provenance (Section 5).

Finding 3: Matched initial execution can precede delayed divergence. After a history intervention, we enforce the reference initial action queue while retaining the branch's generated history and resuming BF16 computation. Key–value (KV) cache storage remains BF16 throughout. Despite matched initial execution, 16 of 128 continuations diverge at later model calls; an unperturbed queue control reproduces every reference continuation. This establishes a history-related pathway to delayed divergence in the tested states (Section 6). Together, the findings motivate evaluating retained outcomes and subsequent behavior alongside average quality and realized resource costs.

2 Related Work

2.1 Vision-and-Language Navigation with Retained History

VLN agents ground instructions in visual observations and retain context from earlier decisions. Speaker–Follower, EnvDrop, and PREVALENT strengthen supervision through instruction generation, environmental augmentation, and pretraining (Fried et al., 2018; Tan et al., 2019; Hao et al.,

2020). VLN-BERT, HAMT, and DUET develop recurrent state, hierarchical history encoding, and global topological planning, respectively (Hong et al., 2021; Chen et al., 2021; 2022). More recently, NaviLLM unifies navigation tasks through schema-based instructions (Zheng et al., 2024). StreamVLN combines slow-fast video context with bounded context and KV-cache reuse, while Efficient-VLN reduces history reprocessing through incremental cache reuse (Wei et al., 2026; Zheng et al., 2026). We examine which outcomes and trajectories survive lower precision through StreamVLN history interventions and a separate NaviLLM paired-outcome comparison.

2.2 Quantization and Behavioral Fidelity

FP8 supports low-precision floating-point computation (Micikevicius et al., 2022); GPTQ and AWQ enable accurate low-bit weight compression (Frantar et al., 2023; Lin et al., 2024). LLM.int8() isolates activation outliers in higher precision, while SmoothQuant rescales weights and activations for W8A8 inference (Dettmers et al., 2022; Xiao et al., 2023). KIVI and KVQuant extend quantization to KV caches for longer contexts (Liu et al., 2024; Hooper et al., 2024). In embodied deployment, NaVILA applies AWQ and reports memory and speed benefits while retaining navigation performance (Cheng et al., 2025). SQIL emphasizes mission-critical states during quantization-aware imitation learning (Park et al., 2025), while QVLA uses action-space sensitivity for precision allocation and analyzes temporal action deviations (Xu et al., 2026).

Similar average accuracy can coexist with conflicting individual predictions (Marx et al., 2020), and pruning can disproportionately affect particular examples (Hooker et al., 2019). Recent tool-agent results show that benchmark error budgets can mask increased process failures (Jang et al., 2026). We examine paired navigation gains and losses, then use fixed-state probes and controlled continuations to test whether altered history changes later behavior despite matched initial execution. Our history interventions retain BF16 KV storage, separating changes in historical content from cache compression.

3 Evaluation Protocol

Models and quantization. StreamVLN (Qwen2-7B) BF16/W4 pairs cover 1,839 R2R-CE val_unseen instructions on 613 routes in 11 scenes (Wei et al., 2026; Yang et al., 2024; Krantz et al., 2020). Controlled NaviLLM (Vicuna-7B) covers 2,349 R2R instructions on 783 routes, with stateless candidate ordering and no cross-step decoder cache (Zheng et al., 2024; Chiang et al., 2023). Our RxR evaluation (Ku et al., 2020) uses 1,024 frozen English routes and the R2R/RxR-trained StreamVLN checkpoint. StreamVLN W4 quantizes 196 decoder linears to group-64 INT4; other modules and KV storage remain BF16. Probes use unpacked W4 and deployment uses packed W4. Static E4M3 FP8 freezes activation scales, whereas dynamic FP8 recomputes them per forward pass. A4 uses activation-only INT4 quantize–dequantize operations rather than a deployed INT4 kernel. Each comparison retains its own simulator, checkpoint, and software branch (Appendices A and A.1).

Readouts and interventions. We measure SR, success weighted by path length (SPL), paired gains/losses, sequence-change rates, and normalized edit distance. Negative log-likelihood (NLL), direct Jensen–Shannon divergence (JS), and action/queue disagreement describe fixed-state responses; Table 6 gives complete definitions. Probes hold external inputs and teacher prefixes fixed while substituting current computation, history, or both. Single-call pulses then resume BF16; queue control additionally enforces the reference initial execution while retaining the branch's history (Section 6). Fixed-prefix horizons count model calls; continuation horizons count environment actions. Paired estimation retains all instructions from each sampled route to account for route dependence (Appendix B.1).

4 Outcome Reversals Behind Similar Average Scores

Do similar success rates preserve the same successful instructions and navigation behavior? Instruction-paired closed-loop episodes separate changes in terminal outcomes from changes in how agents navigate.

Table 1: **Similar average quality can conceal outcome and trajectory changes.** Each comparison retains its own protocol; arrows run from reference to comparison. (a) Absolute SR (ref./cmp., %), gain/loss counts, outcome flips (%), and executed-sequence changes (%, including length differences). (b) Normalized edit-distance mean/median (0–1), using actions for StreamVLN and emitted choices for NaviLLM. Stream/Navi denote StreamVLN/NaviLLM; FP8 S/D denote static/dynamic scales. Appendices B.1 and B.4 give uncertainty estimates and measurement details.

(a) Paired outcomes and sequence changes

Population / comparison	Pairs	SR ref./cmp.	Gain/loss	Flips	Seq. change
Stream R2R / BF16→W4	1,839	56.33/56.23	**169/171**	**18.5**	87.0
Stream R2R / FP8 S→D	1,839	56.28/57.04	116/102	11.9	63.5
Navi R2R / BF16→W4	2,349	65.77/65.69	121/123	10.4	37.9
Stream RxR / BF16→W4	1,024	55.18/51.86	101/135	23.0	92.6

(b) Sequence distance within outcome groups

R2R comparison	Both succeed	Gain	Loss	Both fail
Stream BF16→W4	.151/.139	.420/**.360**	.411/**.365**	.243/.192
Stream FP8 S→D	.093/.043	.406/.368	.363/.307	.157/.070
Navi BF16→W4	.062/.000	.432/.400	.450/.429	.147/.000

4.1 Paired Outcome Decomposition

For n paired episodes with G gains (reference failure to success) and L losses,

$$\text{outcome-flip rate} = \frac{G+L}{n}, \qquad \Delta\text{SR} = \frac{G-L}{n}.$$

While $G - L$ measures net success change, $G + L$ measures turnover in which episodes succeed. To relate average performance to the retention of individual outcomes, Table 1(a) reports absolute success rates alongside paired gains, losses, and executed-sequence changes.

On R2R-CE, StreamVLN's success rate changes from 56.33% under BF16 to 56.23% under W4, yet **169 gains nearly cancel 171 losses**, reversing outcomes for **18.5%** of paired episodes. The static/dynamic FP8 and NaviLLM comparisons also exhibit offsetting gains and losses within their respective protocols. On RxR, losses instead outnumber gains and sample success decreases, illustrating that the near-cancellation observed on R2R need not extend to another setting. Across all four comparisons, executed-sequence changes are more frequent than outcome reversals. We next examine how behavioral changes are distributed across outcome groups.

4.2 Behavioral Differences Across Outcome Groups

Outcome reversals capture only part of the behavioral change: even pairs that retain the same success label can follow different action sequences. Table 1(b) groups the R2R comparisons by both-success, gain, loss, and both-failure outcomes. StreamVLN distances compare executed actions; NaviLLM distances compare emitted viewpoint choices and termination. The nonzero median among StreamVLN BF16/W4 pairs that both succeed shows that outcome retention leaves room for behavioral change.

StreamVLN BF16/W4 gains and losses have nearly identical median edit distances, **.360 and .365**; the other comparisons also show substantial distances in both groups. Edit distance counts changed actions without measuring progress toward the goal. It therefore cannot distinguish useful deviations from lost successes, and these outcome-conditioned associations do not explain why an episode improves or fails.

Paired spatial measurements reveal a similar distinction: StreamVLN BF16/W4 has a mean path-length change of only $+0.028$ m but **2.406 m** mean endpoint separation. Signed length changes can cancel even when the agents finish at different locations; Appendix B.4 reports the full spatial comparison.

Table 2: **Numerical responses and immediate action changes.** (a) BF16-derived reference: numerical changes coexist with **0/1,823** first-action or queue changes; JS measures current substitution. (b) Separate W4-derived reference: action-disagreement rates average events within instructions, then instructions within routes. The two populations are not state-matched.

(a) BF16-derived reference: numerical response

Configuration	Current ΔNLL	History ΔNLL	JS (10^{-3} nats)
W4	.0436	.0162	1.268
A4	**.0889**	**.0385**	**2.661**
FP8 static	.0077	.0016	.270
FP8 dynamic	.0120	.0009	.280

(b) W4-derived reference: first-action disagreement (%)

Substitution	Current only	History only	Joint
Route-averaged rate	**2.44**	**2.49**	3.42

4.3 Cross-Setting Evidence and Scope

The static/dynamic FP8 pair holds quantized weights fixed, linking its contrast to activation-scale selection. Common-prefix audits establish comparability where checked for FP8 and controlled NaviLLM, before trajectories diverge (Appendices B.2 and B.3). These within-protocol results extend the aggregate–individual discrepancy without ranking model sensitivity.

On the 1,024 frozen English RxR routes, the -3.32 pp SR difference and lower path-quality point estimates retain their directions after removing any single scene (Appendix B.5). These estimates establish neither equivalence nor a population-wide quality decline; Appendix B.1 reports their uncertainty. The 26 unreachable-goal pairs remain failures in SR; SPL and navigation error use 998 finite pairs.

Takeaway. Similar average success can coexist with different successful instructions and navigation sequences. Paired outcome and behavior measurements are therefore needed to assess what quantization preserves beyond aggregate task completion.

5 From Numerical Perturbations to Immediate Decisions

When do numerical changes alter an action? Once trajectories diverge, their changing inputs obscure local sensitivity to quantization. We therefore hold external inputs and teacher prefixes fixed while perturbing current computation, historical content, or both. Probe outputs never enter later teacher inputs, isolating local responses before the closed-loop continuations.

5.1 Numerical and Action-Level Responses

A perturbed distribution can retain the same highest-scoring action. NLL and JS measure numerical responses, while action and queue agreement test whether those responses change a decoded decision.

A BF16-derived reference supplies two current-computation $\times$ history-provenance grids: BF16/W4/A4 and BF16/static FP8/dynamic FP8. They share the all-BF16 cell, giving 17 cells on 1,823 events over 613 routes and 11 scenes; 16 instructions naturally lack eligible history. Table 2a holds the other factor at BF16. NLL sums fixed teacher-action tokens, while direct JS compares first-action-token distributions in nats.

Measurable NLL/JS changes coexist with **zero first-action or queue mismatches on 1,823 events** (Table 2a). Even A4, which produces larger distribution shifts and relative activation error than FP8, preserves these decisions. Probability mass can move without changing the winning action, so numerical discrepancy alone does not determine local behavioral change. A4's separate closed-loop diagnostic lacks an eligible strictly paired behavior comparison (Appendices B.1 and C.3).

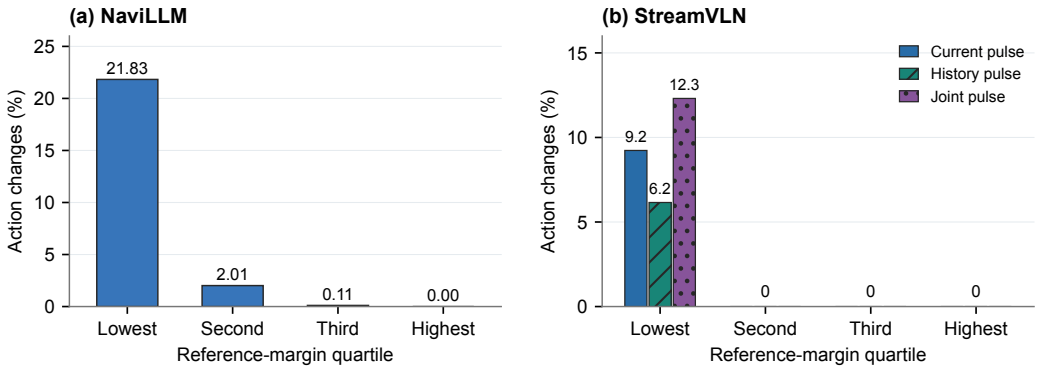

Figure 2: **Initial choice flips concentrate at small reference margins.** Bars show observed flip rates within each margin group. (a) NaviLLM matched BF16/W4 events. (b) StreamVLN current, history, and joint pulses on 256 routes each; all initial flips occur among the 65 lowest-margin routes. Model-specific scores do not define a shared threshold.

5.2 Current and History Perturbations

A separate W4-derived reference supplies 5,508 events from 1,836 instructions over 613 routes; three engineering failures remain missing. We substitute current computation, history, or both, with repeated-cell controls and the other factor held at BF16 in each single-factor contrast. History-only substitution changes actions at a route-averaged rate of **2.49%**, close to the 2.44% current-only rate (Table 2b). Thus BF16 current computation can still produce a different decision when its historical input changes. These rates establish sensitivity to both channels, without identifying dominance or additive effects.

Under prescribed future inputs, history-pulse effects attenuate in a fixed complete cohort, and rebuilding BF16 history removes measured differences in a separate reset cohort (Appendix C.4). These controls track historical state under shared inputs; they do not establish recovery after an altered environment trajectory.

The BF16-derived and W4-derived samples differ in selected events, observations, histories, and software branches. Their flip-rate contrast therefore cannot identify an effect of reference provenance. We instead examine decision margins within matched states.

5.3 Decision Margins and Action Sensitivity

The decision margin is the highest legal-action score minus the second highest. A small margin leaves less separation for a perturbation to overcome, although its direction also matters. We group 14,701 matched NaviLLM BF16/W4 events over 783 routes by their BF16 candidate margin, excluding 2,701 input- or candidate-mismatched events. StreamVLN uses first legal-action-token margins on 256 frozen continuation states. Appendix C.5 specifies the quartile boundaries, ties, and scoring interfaces.

Flips concentrate at small reference margins (Figure 2): NaviLLM's lowest-margin quartile has a **21.83%** flip rate, compared with 2.01% in the next quartile, and all StreamVLN initial flips occur in its lowest-margin group. This association is consistent with sensitivity at an action decision boundary; it does not validate a predictive threshold or explain the unmatched between-reference contrast. Moreover, a changed choice is distinct from execution: one of NaviLLM's 892 choice flips is intercepted by the step limit (Appendix C.5).

Takeaway. Under fixed inputs, numerical changes need not change the chosen action. Perturbing historical content can alter decisions, and flips concentrate at small reference margins. Immediate agreement leaves later behavioral effects unresolved.

Table 3: **Immediate changes and continuation outcomes.** Counts on 256 routes versus the unchanged reference. First action and queue concern the intervention call; sequence divergence and success gains/losses concern the remaining episode. Table 11 retains spatial and goal-distance readouts.

Intervention	First action differs	Queue differs	Sequence diverges	Gain/loss
Current pulse	6	31	53	2/2
History pulse	4	19	50	2/3
Joint pulse	8	38	**63**	5/2
Sustained W4	8	38	156	7/13

6 Delayed Divergence Despite Matched Initial Execution

Can a perturbation change later navigation despite immediate action agreement? We follow interventions to episode termination on 256 routes, then enforce the reference execution queue on a frozen 128-route subset. These experiments distinguish preserving the actions executed now from preserving the history used by future decisions.

6.1 Initial and Subsequent Divergence

From one frozen StreamVLN state on each of 256 routes in 11 scenes, we branch an unchanged BF16 reference, single-call current/history/joint pulses, and sustained W4. The reference exactly reproduces its source. A current pulse changes computation to W4, a history pulse substitutes the incoming KV prefix, and a joint pulse changes both. Each pulse then resumes BF16 without resetting altered history or pose; sustained W4 retains quantized computation. One call can emit several environment actions, so a matched first action can conceal a different initial queue, and a matched queue can conceal a different historical state.

The joint pulse changes **63/256 continuations**, including **25 with an identical initial queue but later divergence** (Table 3; classification in Appendix D.1). Returning to BF16 restores subsequent computation precision but does not reconstruct the reference history. Sustained W4 produces more divergent continuations despite identical initial action and queue change counts, showing that the continuation regime matters beyond the initial readout. This contrast combines further quantized computation with evolving histories and observations, so it does not isolate feedback amplification.

Divergence has no uniformly harmful direction: the pulses produce both success gains and losses (Table 3). Spatial reapproach, short-horizon separation, and goal-distance associations provide complementary readouts; their definitions and full results appear in Appendices D.1 and D.3.

6.2 History Interventions under Matched Execution

Naturally matching queues motivate a stronger intervention: enforce the reference queue while retaining the branch's own post-generation history. This fixes initial environment execution while preserving the historical difference whose later effect we test.

We select 128 routes by scene-proportional quotas and fixed route hashes, without selecting on outcomes, margins, or divergence. Branches share the frozen pose, instruction, visual features, KV, and memory buffers. A W4 shadow consumes the same BF16-generated prefix tokens and external inputs without executing W4 actions. Its KV prefix supplies the history pulse (H) with BF16 current computation, or the joint pulse (J) with W4 current computation; R is the unchanged reference.

For HQ/JQ, we replace the parsed execution queue with R's queue after normal generation, retaining the branch's generated tokens and post-generation history. Subsequent computation returns to BF16, and KV storage remains BF16 throughout. The no-op branch RQ reproduces all 128 reference input sequences, actions, and terminations; HQ/JQ match their archived H/J counterparts before queue replacement. Appendix D.2 gives the state transfers, parsing rules, and controls.

With initial execution fixed, HQ still changes **16/128 continuations (12.5%)**, with every first difference occurring one to five model calls later (Figure 3). Executing the reference queue preserves the initial physical evolution, but subsequent BF16 calls consume the branch's retained history. This

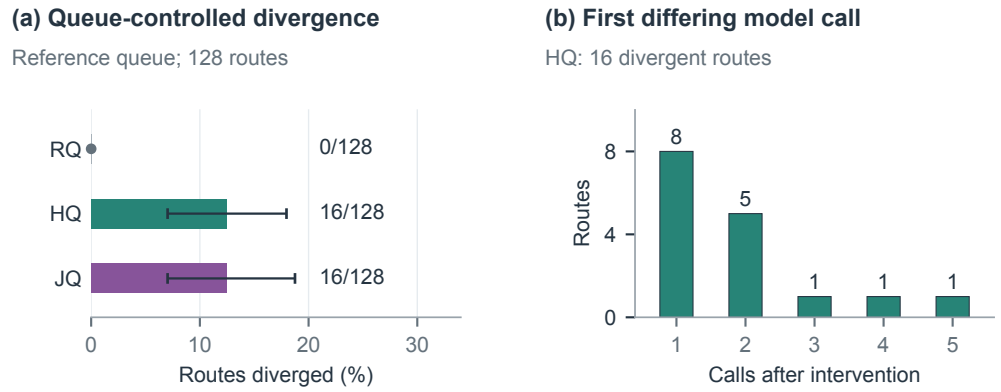

Figure 3: **Delayed divergence with the reference execution queue enforced.** (a) Full-continuation divergence on 128 routes, with descriptive 95% paired-route intervals. RQ is the no-op control; HQ/JQ are history/joint pulses with R's queue. (b) All 16 HQ first differences occur at later model calls, outside the controlled queue.

establishes a history-related pathway in the tested states without isolating the responsible history component or showing that its numerical magnitude grows over time. JQ has the same divergence count, which does not establish identical affected routes or additive channel effects.

Short-horizon summaries can miss these delayed changes: both controlled contrasts have zero median separation at eight actions despite full-continuation divergence. Terminal changes are sparse, with one HQ gain and two JQ gains and no losses; they do not establish a directional quality benefit (Appendix D.2). The subset comes from an inspected population, and contrast rates do not estimate causal pathway shares or the fraction of natural failures explained.

Takeaway. With the initial execution queue fixed, altered history still produces later divergence in the tested states. Preserving immediate actions is therefore insufficient to establish behavioral retention across a full episode.

7 Implications for Quantized Navigation

Real-workload evaluation. We measure BF16, packed W4, and static/dynamic E4M3 FP8 on an RTX PRO 5000 Blackwell, using 32 frozen natural routes and 64 captured inputs per model. Natural-episode timing includes each arm's actions, calls, and stopping behavior; fixed-call replay holds external inputs at a model-call boundary. Appendix E specifies calibration, repeated measurements, aggregation, and the quality and workload counts from these runs.

Realized benefits. Packed W4 slows both measured boundaries in both models (Table 4): natural episodes take 1.50× the BF16 time for StreamVLN and 3.28× for NaviLLM. Static FP8 reduces fixed-call time to 0.72× and 0.62×, respectively, yet StreamVLN episode time is 1.02×, with no resolved speed direction. These boundaries answer different questions because natural episodes also reflect changed workloads. The timing and workload contrasts do not identify the contribution of behavioral changes to elapsed time; full statistical estimates appear in Appendix E.

All compressed arms lower StreamVLN's natural allocated peak. NaviLLM W4 reduces allocated memory from 13,253 to 4,397 MiB while device-used memory rises from 14,327 to 15,085 MiB. PyTorch allocated memory counts tensors; NVML also includes runtime and driver residency. Quantized deployment therefore requires measuring both tensor and device memory, and checking whether faster fixed calls translate into faster episodes.

Table 4: **Realized costs of the tested implementations.** Time is the route-aggregated wall mean over two blocks; memory is the larger natural-episode peak. Allocated counts PyTorch tensors; NVIDIA Management Library (NVML) measures device-used memory. S/D denote static/dynamic FP8. Fixed-call memory appears in Table 13.

		Time (ms)		Natural memory (MiB)	
Model	Arm	Episode	Fixed call	Allocated	NVML
StreamVLN	BF16	11,088.8	149.6	16,743	21,794
	W4	16,604.2	239.1	8,143	13,126
	FP8-S	11,279.1	107.8	10,511	16,779
	FP8-D	14,662.3	184.2	10,511	16,779
NaviLLM	BF16	311.8	34.1	13,253	14,327
	W4	1,023.8	119.5	4,397	15,085
	FP8-S	200.1	21.0	7,075	14,143
	FP8-D	221.0	23.2	7,078	14,145

8 Conclusion

Quantized navigation can preserve average success while changing which tasks succeed. Paired episodes expose the gains and losses concealed by that average. Fixed-state probes distinguish numerical sensitivity from immediate action changes, while controlled continuations reveal delayed divergence: after a history intervention, 16/128 routes diverge despite the reference initial queue and subsequent BF16 computation. The no-op queue control reproduces every reference continuation. These findings support evaluating quantized replacements through retained task outcomes, behavior beyond immediate execution, and realized resource costs. They identify a history-related pathway in the tested states, without assigning a uniformly harmful direction to behavioral change.

Reproducibility Statement

Appendix A specifies checkpoints, protocol branches, populations, sampling units, missingness, software, and compute. The anonymous supplement's `README.md` identifies frozen route lists, per-route rows, intervention and analysis code, and independent audit receipts. Original model checkpoints and simulator assets are not redistributed.

AI Use Statement

AI assistants supported literature retrieval, research discussion, experiment implementation and orchestration, log analysis, statistical-code review, and manuscript drafting, revision, and interpretation. The authors take responsibility for the manuscript, claims, and evidence verification.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents, 2018a. arXiv:1807.06757.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018b.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 5834–5847, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16537–16547, 2022.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. Blog post, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13137–13146, 2020.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1643–1653, 2021.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget?, 2019. arXiv preprint arXiv:1911.05248.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. In *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2019.

Jiwon Jang, Kisu Yang, Heuiseok Lim, and Hyunwoo Park. Flat score, amplified failures: How the error budget masks damage in quantized llm agents, 2026. arXiv preprint arXiv:2607.27275.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision – ECCV 2020: 16th European Conference*, pp. 104–120, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems (MLSys)*, 2024.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Charles T. Marx, Flávio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 6765–6774, 2020.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. Fp8 formats for deep learning, 2022.

Seongmin Park, Hyungmin Kim, Sangwoo Kim, Wonseok Jeon, Juyoung Yang, Byeongwook Jeon, Yoonseon Oh, and Jungwook Choi. Saliency-aware quantized imitation learning for efficient robotic control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13140–13150, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621, 2019. doi: 10.18653/v1/N19-1268.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 38087–38099, 2023.

Yuhao Xu et al. Qvla: Not all channels are equal in vision-language-action model's quantization, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report, 2024.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13624–13634, 2024.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Efficient-VLN: A simple yet strong baseline for efficient vision-language navigation, 2026. arXiv:2512.10310v2.

A PROTOCOL DETAILS

Checkpoint and protocol identities. StreamVLN uses the released R2R/RxR/EnvDrop/ScaleVLN checkpoint with a Qwen2-7B language backbone. Its configuration has hidden size 3,584, intermediate size 18,944, 28 attention heads and four KV heads. NaviLLM uses its official `model_without_pretrain` checkpoint with Vicuna-7B-v1.1. The experiments preserve checkpoint, code, population, and result manifests; a shared instruction key is not treated as proof of identical observations across simulators.

Separate evidence families. StreamVLN's historical BF16/W4 closed loop uses Transformers 4.45.1, while the FP8 pair and older full mechanism branch use 4.51.0. A4 reconstructs the historical BF16/W4 protocol in an isolated run; it does not supply a new concurrent BF16/W4 control. The 17-cell matrix uses a BF16-derived reconstructed reference. The new StreamVLN continuation study branches from frozen common states. The controlled full-population NaviLLM BF16/W4 pair uses stateless candidate ordering; older randomized-order runs are not used for quantization attribution. Real-workload timing uses torch 2.8.0+cu128, torchao 0.10.0, and Transformers 4.51.0. Quality, mechanism, continuation, and timing estimands remain separate.

Coverage and missingness. Each StreamVLN closed-loop arm covers 1,839 instruction keys. The older full mechanism branch reuses 96 accepted instructions and completes 1,740 additional instructions: 1,836 usable units and 5,508 selected events. Three units fail after engineering recovery attempts and remain engineering missingness; they are not zero-filled or relabeled as natural termination. Its persistence coverage is 1,836/1,705/1,564/1,318 units at horizons 0/1/2/4, with a fixed complete cohort of 1,318 for the main curve. The newer matrix covers 1,823 eligible events per cell, in 114 batches, with 16 instructions naturally lacking an eligible history. Its final analysis uses the independently reviewed pooled-instruction estimator, not an earlier equal-scene estimator. The continuation study completes all five branches on 256 routes from 11 scenes and passes exact no-op restoration. NaviLLM covers 2,349 instruction keys per arm. Timing covers 32 frozen routes and 64 captured real inputs per model, with eight calibration routes excluded from the formal population.

Statistical units. The historical BF16/W4 quality report uses instruction-level paired outcomes and also reports a separate scene-level sensitivity analysis; neither is relabeled as a route-cluster analysis. Its historical nDTW-proxy difference is approximately -0.006. The older mechanism analysis averages selected events within instructions and then instructions within routes. The newer matrix reports pooled paired-instruction means. Audited supplementary analyses retain complete paired routes within scenes, rather than treating instructions from one route as independent clusters. Continuation target-distance readouts comprise four prespecified contrasts. Their references, aggregation schemes, and valid-pair populations remain distinct. McNemar values are instruction-level descriptive tests and do not address all dependence among instructions sharing a route. No equivalence margin was pre-specified.

Compute accounting. Recorded FP8 closed-loop work costs 12.69 GPU-hours, the older full W4 mechanism/timing branch 45.45, and the A4/reference-matrix workflow 74.27, including at-

Table 5: Evidence-family protocol map. Every comparison uses a within-row frozen population. The NaviLLM action interface is candidate scoring; StreamVLN produces action queues. Stream/Navi denote StreamVLN/NaviLLM; TF denotes Transformers. FP8 S/D denote static/dynamic scales. Checkpoint identities are specified above; route manifests accompany the reproducibility materials.

Evidence	Model	Population	Paired comparison	Software / selection
R2R quality	Stream	1,839 instr.; 613 routes	BF16/W4	TF 4.45.1; complete pairs
R2R FP8	Stream	1,839 instr.; 613 routes	static/dynamic	TF 4.51.0; complete pairs
R2R quality	Navi	2,349 instr.; 783 routes	BF16/W4	fixed candidate order
RxR-English	Stream	1,024 routes	BF16/W4	scene quota; route hash
Continuations	Stream	256 routes	frozen-state pulses	one event per route
Deployment	both	32 routes/model	BF16/W4/FP8 S/D	two process blocks

Table 6: **Readouts for paired episodes and fixed-state probes.**

Readout	Definition
Paired episodes	
SR / SPL	Success rate / success weighted by path length.
Gains / losses	Counts of failure→success / success→failure.
Outcome-flip rate	Gains plus losses, divided by the number of paired episodes.
Sequence-change rate	Fraction of paired sequences that differ.
Normalized edit distance	Levenshtein distance divided by the longer sequence length (minimum 1).
Fixed-state probes	
NLL / direct JS	Negative log-likelihood / Jensen–Shannon divergence.
First-action / queue changes	Disagreement in the first action / emitted action queue.

Sequences contain environment actions for StreamVLN and viewpoint choices plus termination for NaviLLM.

tempts. These totals are not added to their own component costs. The separate follow-up package totals 28.020 GPU-hours, including the 3.041-hour matched four-arm timing stage and discarded RustDesk-interrupted blocks. Failed attempts, component boundaries, host identity, and budget receipts remain in the source ledger.

Readout conventions. Quality readouts include success rate (SR), success weighted by path length (SPL), oracle success (OS), navigation error (NE), normalized dynamic time warping (nDTW), and success-weighted nDTW (SDTW) (Anderson et al., 2018a; Ilharco et al., 2019). Oracle success records whether the goal is reached at any point. The historical sparse-reference R2R nDTW proxy and dense-reference RxR nDTW are distinct estimands. Direct JS uses nats; model-specific margins subtract the second-highest legal-action score from the highest.

Audit provenance. Authoritative annotations identify 783 NaviLLM routes, 613 routes in each StreamVLN R2R pair, and 1,024 RxR routes, each spanning 11 scenes. The independently checked source audits preserve complete paired routes, instruction weights, and scene identities; their archived resampling recipes remain unchanged. Corrected supplementary route identities do not change per-instruction measurements or point estimates. RxR path-metric rows are aligned by scene and route key before weighting. These audits use unchanged model outputs without new inference and do not establish generalization to new scenes.

A.1 Precision and Execution

The 196 StreamVLN decoder targets comprise query, key, value, attention output, and feed-forward gate/up/down projections in each of 28 layers. Vision and other excluded modules remain unmodified. The numerical W4 reference includes unpacked weights; deployment uses packed torchao modules. StreamVLN uses greedy decoding with batch size one, seed 42, eight history frames, a 32-step cache window, and at most 200 generated tokens per call. NaviLLM instead ranks fixed candidate viewpoints. Checkpoint hashes, rather than directory-name substrings, identify the backbone. An additional repeated W4/W4 cell checks fixed-state repeatability. INT4 denotes four-bit integers; A4 denotes the activation-only four-bit diagnostic, implemented with quantize–dequantize (QDQ) operations. E4M3 is the FP8 format with four exponent bits and three fraction bits.

W4 keeps activations in BF16 and uses INT4 group-64 weights, unpacked for numerical probes or packed for deployment. Static FP8 freezes activation scales, whereas dynamic FP8 recomputes them per forward pass; both use E4M3 matrix multiplication. A4 uses BF16 weights and group-64 INT4 activation QDQ followed by BF16 matrix multiplication, not an INT4 activation kernel. KV cache storage remains BF16 throughout. For StreamVLN static FP8, the frozen *G* configuration calibrates each target Linear on the same 96 BF16 episodes (12,189 calls). It first freezes each layer's activation maximum a, then selects $r \in \{.25, .5, .75, 1, 1.25\}$ by the recorded activation QDQ error on those frozen inputs and uses $s = ar/448$ at inference. All 196 targets have a valid frozen scale; 43/77/76 select $r = .75/1/1.25$, respectively. The scale file and selection objectives are preserved

Table 7: **Quality changes and their scope.** Differences use unrounded comparison-minus-reference values. Panel (a) retains the original 98.33% scene-stratified route-cluster intervals. RxR SPL and NE use 998 finite pairs; its other metrics use 1,024. Panel (b) reports sample point estimates; nDTW/SDTW use a 0–100 scale.

(a) Paired changes: point [98.33% interval], pp

Comparison	ΔSR	ΔSPL
Stream R2R BF16→W4	$-.11\ [-2.58, 2.39]$	$-.34\ [-2.53, 1.88]$
Stream R2R FP8 static→dynamic	$+.76\ [-1.20, 2.66]$	$+.35\ [-1.24, 1.93]$
Navi R2R BF16→W4	$-.09\ [-1.68, 1.47]$	$+1.00\ [-.28, 2.22]$
Stream RxR BF16→W4	$\mathbf{-3.32}\ [-6.93, .29]$	$-2.88\ [-5.86, .04]$

(b) Additional RxR quality readouts

Metric	BF16	W4	Difference
NE (m)	5.52	6.24	+0.72
nDTW (points)	64.37	62.76	−1.61
SDTW (points)	46.74	43.97	−2.77

in the evidence package. G is a named calibration protocol, not a generic static-FP8 setting or a scale optimized on the deployment routes.

B Paired Results

The following analyses preserve within-protocol pairing. Differences between models, simulators, checkpoints, or software branches are not treated as matched effects.

B.1 Paired Estimation and Quality Scope

Table 1 reports absolute success rates, paired outcomes, and outcome-conditioned sequence distances. Table 7 provides the complete quality estimates and uncertainty, while Table 8 reports spatial readouts. Scene-stratified paired-route intervals use 10,000 resamples (seed 42), keeping all instructions from each selected route. Quality effects retain their original 98.33% comparison level; behavior-rate intervals are descriptive at the same level and do not redefine the comparison family. The source package preserves all frozen draws and secondary estimates. No equivalence margin was specified. For StreamVLN BF16/W4, the descriptive 98.33% intervals are $[84.80, 89.12]\%$ for behavior change and $[16.20, 20.83]\%$ for outcome flips.

The archived StreamVLN configurations span 0.82 pp in SR and 0.73 pp in SPL across distinct protocol branches, not a five-way equivalence test. A4's separately reconstructed protocol gives SR 56.39%, without a strictly paired behavior comparison. Its oracle success was not reported. The historical R2R nDTW proxy uses sparse references and is not interchangeable with the dense-reference RxR metric.

B.2 Static versus Dynamic FP8

This pair changes activation-scale selection while keeping quantized weights identical. Among alignable common-prefix input hashes, 21,571 match, none differ, and 169 are missing. The per-step analysis conditions on the shared prefix rather than treating later trajectory differences as independent perturbations. Tables 1 and 7 contain the paired outcomes and quality effects.

Why the full-population contrast uses static versus dynamic FP8. The historical BF16/W4 episodes use Transformers 4.45.1, whereas the 1,839-instruction FP8 episodes use 4.51.0. Joining these archives by instruction would mix precision changes with software-branch differences, so it would not isolate BF16-to-FP8 replacement. Matched BF16/static-FP8/dynamic-FP8 observations are available in the separate 32-route deployment study (Table 14) and in fixed-input probes (Table 2); neither supplies a matched full-population closed-loop comparison.

Table 8: **Physical geometry across complete paired paths.** Means over all 1,839 paired paths per StreamVLN comparison. Path-length difference is comparison minus reference; endpoint separation is between agents. FP8 S/D denote static/dynamic scales. NaviLLM lacks paired agent coordinates.

StreamVLN comparison	Path-length difference (m)	Endpoint separation (m)
BF16→W4	$+0.028$	**2.406**
FP8 S→D	-0.044	1.587

The dynamic-minus-static SR difference is $+1.69$ pp on 256 routes used in earlier exploration and $+0.09$ pp on 357 other routes. This descriptive exposure contrast limits extrapolation from the pilot; it neither tests exposure causally nor establishes a dynamic-scaling advantage.

B.3 Controlled NaviLLM

NaviLLM (Zheng et al., 2024) rebuilds a Vicuna-7B prompt for candidate-viewpoint scoring instead of carrying decoder KV across steps. Its official R2R checkpoint covers 2,349 val_unseen instructions, 783 routes, and 11 scenes in discrete MatterSim. Absolute scores are not matched to StreamVLN's continuous-environment evaluator.

The BF16/W4 pair uses eager attention, greedy selection, the same 224 decoder linears, and a stateless candidate permutation keyed by frozen event identity. All 14,701 retained common-prefix events pass external-input, candidate-order, inverse-mapping, and candidate-identity checks. Runtime audits find packed torchao INT4 group-64 modules without fallback. Tables 1 and 7 report the results; earlier randomized-order runs are excluded from quantization attribution because candidate order was uncontrolled. The 891 executed common-prefix differences comprise 67 BF16-STOP/W4-MOVE, 184 W4-STOP/BF16-MOVE, and 640 different movement candidates. STOP terminates an episode and MOVE denotes a movement choice; the step-cap exception is detailed in Appendix C.5.

B.4 Behavioral Measurement and Provenance

The sequence and spatial analyses in Tables 1(b) and 8 reuse three closed-loop archives: StreamVLN BF16/W4 and static/dynamic FP8 (1,839 instructions, 613 routes each), and controlled NaviLLM BF16/W4 (2,349 instructions, 783 routes). They are separate from the 256-route interventions. Source audits retain complete paired routes within scenes before forming outcome groups; corrected route identities leave all instruction-level distances, weights, and point estimates unchanged.

Levenshtein distance counts insertions, deletions, and substitutions and is divided by the maximum of both sequence lengths and one. Strict-prefix length differences are retained. StreamVLN sequences use environment-action symbols; NaviLLM uses emitted viewpoint identities and a termination symbol. These distances are not physical displacement. One NaviLLM choice flip is intercepted by the step limit, so executed-path changes count 891 pairs while emitted-choice distances retain all 892 choice flips. The archive separately flags 144 instructions with an unexecuted terminal movement choice; this alternative encoding does not replace the reported choice sequences.

All 1,839 StreamVLN pairs per comparison have paired agent paths. NaviLLM lacks paired agent-coordinate trajectories, precluding physical endpoint separation or path warping. The separately run A4 diagnostic has no eligible strictly paired trajectories. None of these definitions assigns a positive or negative task value to a changed action.

Across the three comparisons, 256, 294, and 385 outcome-stable pairs, respectively, lie in their comparison's overall upper edit-distance quartile. These outcome-conditioned summaries are associative. BF16 behavior is not assumed to be intrinsically optimal.

A near-zero mean path-length change accompanies substantial endpoint separation in both comparisons. Path length measures distance traveled, whereas endpoint separation records where the agents finish; signed length changes can cancel across episodes. Endpoint separation does not identify which agent is closer to the goal.

B.5 RxR Evaluation

Population and selection. The source val_unseen guide set contains 11,006 instruction episodes over 1,223 routes in 11 scenes; 3,669 instructions are English (1,223 US-English and 2,446 Indian-English locale entries). Before viewing quality, the protocol assigns scene quotas by largest remainders, hash-orders routes within scene, and freezes 1,024 routes. A global hash order assigns alternating English locales; where a route has two Indian-English instructions, a separate instruction hash selects one. Each route supplies one instruction, producing 512 entries per locale. A nested 512-route budget fallback was frozen but was not used. This is the specified subset, not the entire RxR benchmark or a random sample of all its English instructions.

Paired execution. Both arms use the same released R2R/RxR/EnvDrop/ScaleVLN checkpoint, manifest order, episode-hash prompt with seed 42, persistent episode environment, greedy decoding, and batch size one. The limits are 500 environment steps, 200 new tokens per call, model length 4,096, 32 frames, eight history frames, and four future steps. Pairing by canonical scene, episode identifier, and instruction text yields 1,024 pairs without duplicate or manifest-mismatched keys. W4 quantizes the 196 decoder linears with packed INT4 group size 64; other compression is disabled. This pairing does not make cross-benchmark outcomes a matched test of dataset difficulty.

Additional readouts. The frozen sample has 430 double successes, 358 double failures, and mean normalized edit distance .284. Oracle success is 64.45%/62.30% (BF16/W4; -2.15 pp). Primary outcomes, quality, and path readouts are in Tables 1 and 7.

Non-finite metrics explained. Both arms have non-finite SPL/NE on the same 26 pairs, all in scene `oLBMNvg9in8`; 16 use Indian-English and 10 US-English. Queries against the original navigation mesh show that their goals are unreachable from both the start and the final positions. NE is infinity, and the SPL formula contains an infinity-over-infinity ratio, yielding an undefined value. Source guide, compatibility-converted, and ground-truth coordinates agree; the audit finds no conversion-coordinate mismatch. The historical summary converted these non-finite values to null. They are not random missing data and are not replaced by zeros or by a zero difference between infinities.

Independent evaluation with the original metric classes in Habitat (Savva et al., 2019) reproduces all 2,048 arm records: finite NE/SPL values have zero discrepancy, and SR/OS agree. SR/OS retain the original 1,024-pair protocol, including 26 double failures; those unreachable tasks do not measure ability on valid reachable tasks. NE/SPL retain only 998 finite pairs. Capping NE at 10, 20, or 50 meters, or assigning zero SPL to failures, defines separate sensitivity estimands rather than recovering the missing finite primary metrics. Two external step logs have truncated tails (BF16 episode 10788: two steps; W4 episode 1169: one); complete embedded action records match all 2,048 stored paths. No missing steps are fabricated.

Scene-removal check. Removing the affected scene excludes all 96 routes, not only the 26 unreachable pairs. The remaining 928 finite pairs give differences of -3.02 pp SR, -2.77 pp SPL, and $+0.74$ m NE. Across all 11 scene-removal checks, SR differences range from -3.99 to -2.69 pp, SPL from -3.38 to -2.37 pp, and NE from $+0.59$ to $+0.82$ m. These post-hoc sample summaries do not establish new-scene generalization.

Resources and excluded claims. The supplementary workflow records 13.575 GPU-hours, including charged attempts, within its 48-hour limit. Margin, behavior, local-to-downstream, and subsequent metric audits reuse historical outputs without new forward passes. Runtime checks identify all 196 packed INT4 target modules with no silent fallback and 4,071,223,296 unique weight-storage bytes. This is storage identity, not total process memory. Matched own-process NVML sampling and a valid latency boundary are unavailable, so RxR adds no memory-peak or speed claim.

B.6 RxR Path Metrics

Definition and validation. We compute path metrics offline from the 2,048 stored arm paths using the official VLN-CE metric definitions (commit `729d141b2ee1`) and FastDTW 0.3.4's default pure-Python implementation. The calculation uses dense guide ground-truth positions,

three-dimensional Euclidean distance, and removes only consecutive exactly duplicate predicted positions. With predicted path P, dense reference path G, and $|G|$ its number of positions, $\mathrm{nDTW} = \exp[-\mathrm{FastDTW}(P, G)/(3|G|)]$, where 3 meters is the normalization distance. Success-weighted dynamic time warping is $\mathrm{SDTW} = s\,\mathrm{nDTW}$, where s is the validated original-protocol binary success label. Mean SDTW averages all routes, not just successful ones.

The independent check executes the pinned official nDTW/SDTW classes with unchanged arithmetic: all 2,048 final values match exactly. It additionally checks stepwise updates on 22 scene-hash-selected paths and synthetic and invalid-input cases. These checks validate this offline metric calculation; they do not make the action settings, compatibility conversion, or environment equivalent to an official leaderboard evaluation. The historical R2R sparse-reference proxy remains a separate metric and population.

Table 7b reports both path metrics. The 26 unreachable tasks retain validated success of zero; path metrics do not recover finite all-route NE/SPL. Route-key alignment preserves each stored path value and the accepted point estimates.

C Local Mechanisms

C.1 Fixed-State Controls

Section 5 specifies the 17-cell BF16-derived matrix, eligible population, and current/history substitutions. External inputs and teacher prefixes stay fixed; probe outputs do not enter later teacher inputs. The matrix is a fixed-state analysis, not a five-arm closed-loop evaluation.

C.2 Reference-condition audit

The supplied BF16-derived direct-JS event export contains 1,807 distinct instructions, a subset of the 1,823 eligible matrix events; the 16 absent rows are not imputed. It selects one call per instruction (index median 7, interquartile range 4–12), and its legal-action probability margin has median .896 (interquartile range .628–.987). The older W4-derived table contains three selected calls per 1,836 instructions (5,508 events; call-index median 7, interquartile range 3–12). Its per-instruction history-token count has median 2,332.5 (interquartile range 1,885–2,773); the supplied table does not retain a comparable reference margin. The BF16 export does not retain history-token count, and neither selected-event table retains the complete per-episode call denominator needed for a comparable normalized call position. The archives overlap in 1,804 scene/episode keys, but selected calls, external observations, and historical buffers have not been matched. The BF16 reconstruction follows the historical Transformers 4.45.1 branch and unpacked W4 diagnostic, whereas the older full-mechanism branch uses Transformers 4.51.0 and its own W4-derived selection. These descriptions cannot identify the effect of reference provenance on the different flip rates.

The numerical and action-level results appear together in Table 2. The zero-flip and nonzero-flip reference populations are not state-matched; overlapping instruction keys or similar call-index summaries do not identify a reference-provenance effect.

C.3 Distribution Readouts

Let p_{hc} be a model distribution with history provenance h and current precision c, where B denotes BF16 and W denotes W4, and let δ_y be a one-hot distribution for the natural reference action token. Direct perturbation measures include $\mathrm{JS}(p_{BB}, p_{BW})$ and $\mathrm{JS}(p_{BB}, p_{WB})$. The archived four-cell readout instead reports $J_{hc} = \mathrm{JS}(p_{hc}, \delta_y)$, averaged over action positions. Its values are 0.064598 (BB), 0.063560 (BW), 0.062935 (WB), and 0.061619 (WW). Thus its signed current-minus-history contrast is $(J_{BW} - J_{BB}) - (J_{WB} - J_{BB})$, about $+0.000626$ nats. This contrast is neither a distance between arms nor a comparison of absolute perturbation sizes. We retain the readout's definition but do not interpret that positive sign as current-computation dominance.

In the older reference, conditional first-action disagreement also depends on the fixed cell: current substitution with W4 history gives 2.37%, and history substitution with W4 current computation gives 2.27%. These differ from the BF16-conditioned 2.44% and 2.49% reported in the main text; they are not interchangeable summaries. The current-minus-history contrast for the BF16-

Table 9: **Perturbation channels and fixed-prefix persistence.** Panel (a) changes one or both factors on the W4-derived reference. Panel (b) follows a fixed complete cohort; horizons count model calls, not environment actions. These are different populations and contrasts.

(a) Immediate first-action disagreement (%)			
Substitution	Current only	History only	Joint
Route-averaged rate	**2.44**	**2.49**	3.42

(b) Persistence under prescribed future inputs				
Model-call horizon	0	1	2	4
Behavioral disagreement (%)	2.37	1.56	.836	**.334**
Action ΔNLL	.00617	.00327	.00235	.00255

conditioned rates is -0.054 pp before rounding; these rates identify neither a dominant channel nor additive effects.

The new matrix uses direct first-action JS and summed teacher-action NLL differences (Table 2). All additional W4/A4 and static/dynamic factorial comparisons have zero first-action and queue mismatches on 1,823 pairs. Their full numerical rows remain in the frozen source analysis; they do not add a closed-loop causal comparison.

Numerical diagnostic. On common BF16 layer inputs, relative activation MSE is 0.0257333 (A4), 0.000714703 (static FP8), and 0.000701646 (dynamic FP8). The isolated unpacked-W4 diagnostic changes action NLL by 0.044446 and direct first-action JS by 0.001268, with no action/queue changes at 1,823 events. This validates a numerical contrast, not a packed INT4 speed claim. The corresponding kernel diagnostic did not establish an all-CUDA, tensor-core execution path.

C.4 Persistence and Reset Protocol

Table 9 contains the immediate channel effects and fixed-prefix persistence curve. Current/history effects condition on BF16 for the other factor; their route-averaged rates are not the W4-conditioned readouts in Appendix C.3. The persistence curve uses only the 1,318 units over 598 routes observed at all four horizons, avoiding a changing cohort. Its initial contrast therefore need not equal the separate immediate history-only rate.

Reset readouts use 573 routes complete across reset measurements. The reference history is rebuilt from BF16 computation under prescribed future inputs; the measured zero differences at and after reset do not establish that a cache reset reverses a previously changed environment trajectory. The source analyses retain horizon-specific coverage and leave-one-scene-out checks; they do not measure repeated perturbations or free-running environmental feedback.

In the complete persistence cohort, direct JS falls from .001099 to .000170 nats, about 15% of its initial value. On the reset cohort, disagreement is 1.16% before reset and all measured differences are zero at and after rebuilding BF16 history. Growth is therefore not inevitable under prescribed future inputs.

C.5 Decision Margins

Within each model's analyzed population, the 25th, 50th, and 75th percentiles are computed by linear interpolation between adjacent sorted reference margins. Values equal to a cutoff enter the lower group, retaining ties rather than forcing equally sized groups.

For NaviLLM, the reference margin is the BF16 top-one minus top-two legal-candidate score, before comparing selected candidates under BF16 and W4. Among 17,402 common event keys, the identity filter retains 14,701 events with matching external-input hashes, canonical candidate sets, and candidate mappings. It excludes 2,701 identity-mismatched events, rather than treating post-divergence inputs as a controlled comparison. All retained events pass those checks.

Table 10: **Where the joint pulse first departs from the reference.** Exact route-level classification on 256 routes; "same"/"different" compare each readout with the unchanged reference. These four cells contain all routes; the remaining combinations are empty.

First action	Initial queue	Full continuation	Routes
Same	Same	Same	193
Same	Same	Different	**25**
Same	Different	Different	**30**
Different	Different	Different	8

Figure 2 reports NaviLLM's 892/14,701 choice flips. Quartile boundaries are 1.1875, 3.015625, and 5.84375, with boundary ties assigned to the lower group. Direct legal-action JS and NLL differences cannot be recovered from this archive.

For StreamVLN, all 256 initial margins reproduce saved first-action-token logits exactly. Ties yield quartile sizes 65/63/64/64; the current/history/joint pulses use the unchanged-reference margin. The sustained-versus-joint comparison instead uses the joint-pulse margin, with 70 routes in its lowest quartile, and has zero initial flips. These different scoring interfaces do not support a common numerical threshold or a causal margin interpretation.

Terminal-event reconciliation. The 892 choice flips, 891 executed differences, and 891 final path differences are compatible measurements. In the sole additional event, instruction `3484_2` at index 14, BF16 chooses STOP while W4 chooses MOVE, but the maximum-step limit intercepts the latter; neither arm moves. We retain that model-choice flip rather than force its count to equal the executed-path count. Post-termination best-stop graph-path appends are also distinguished from actual environment motion. The source audit uses the 783 annotated routes within 11 scenes as clusters, not the 2,349 instruction identifiers.

D Continuations

D.1 Continuation Readout Definitions

Tables 10 and 11 report the full 256-route joint classification and continuation outcomes. The joint pulse has 63 divergent continuations (24.6%, descriptive scene-stratified paired-route 95% interval 19.5–29.7%). Model calls can emit multiple environment actions. An action at zero-based offset k is counted after h executed actions when $k < h$. First-divergence counts include archived length-only differences. Reapproach uses two consecutive non-STOP matched steps within 0.25 m and 15 degrees after divergence; it denotes at least one return, not lasting recovery. Full-continuation observation lengths can differ across routes.

Spatial denominators. The joint pulse has 253/247/239/223 valid position pairs at 1/2/4/8 environment actions. Conditioning further on prior action divergence leaves 15/30/39/43 routes, whose separation medians are .020/.033/.250/.251 m. At eight actions the latter interquartile range is .185–.404 m. These conditional spatial summaries differ from the signed goal-distance contrasts and terminal outcomes. The archived process visualization carries terminal poses forward for its all-route position readout; that convention is not used for the actual-horizon queue-control pairs below. Sustained-W4 spatial rows compare with the joint pulse, whereas the three single-call pulses compare with the unchanged reference.

D.2 Execution-queue control

The 128-route subset retains one existing event per route. Scene quotas use largest remainders with at least two routes per scene: 18, 3, 9, 18, 18, 9, 11, 12, 2, 7, and 21 in sorted scene order. Fixed route hashes select within scenes without consulting outcomes, margins, or divergence. Checkpoint, decoding, history windows, and stopping limits match the original experiment; this inspected-population subset is not independent validation.

Table 11: **Divergence, spatial behavior, and task outcomes.** Panel (a) uses all 256 routes versus the unchanged reference; first action/queue concern the intervention call, and other columns the remaining episode. Reapproach means at least one return to proximity. Panel (b) retains the four prespecified goal-distance contrasts and 98.75% intervals; positive differences mean farther from the goal.

(a) Continuation counts

Intervention	First action differs	Initial queue differs	Sequence diverges	Spatial reapproach	Success Gain	Success Loss
Current pulse	6	31	53	23	2	2
History pulse	4	19	50	28	2	3
Joint pulse	8	38	**63**	**33**	5	2
Sustained W4	8	38	**156**	85	7	13

(b) Signed goal-distance changes at eight environment actions

Intervention	Comparator	Difference (m)	98.75% interval (m)
Current pulse	Reference	$+.0020$	$[-.0149, .0200]$
History pulse	Reference	$-.0012$	$[-.0146, .0135]$
Joint pulse	Reference	$-.0069$	$[-.0264, .0133]$
Sustained W4	Joint pulse	$-.0096$	$[-.0246, .0034]$

Each branch deep-copies its frozen state. Let $g(s)$ return generated tokens and updated model history, and let P parse a legal action queue with the existing empty-queue fallback. For RQ, JQ, and HQ, compute $(t, h') = g(s)$ normally, set $q_{\text{native}} = P(t)$, replace only $q_{\text{exec}} \leftarrow q_{\text{R}}$ immediately before the first environment step, and execute q_{exec} while retaining h'. RQ uses BF16 history/current computation, JQ W4 history/current, and HQ W4 history with BF16 current computation; all then resume BF16. R and J reuse the accepted no-control reference and joint pulse. This control never forces reference tokens during generation. The run logs t, q_{native}, and q_{exec} separately.

The history intervention is specifically the call's `past_key_values` argument. Before the event, a separate W4 shadow is teacher-forced with the actual BF16-generated tokens on the same external inputs, producing a W4 KV prefix without executing W4 actions. At the event, H/HQ supply that shadow KV to BF16 current computation, while J/JQ supply it to W4 current computation. If W4 current computation is used and future computation returns to BF16, the adapter deep-clones `cache`, `curr_t`, `_history_feature_cache`, `_dynamic_memory_cache`, `_last_memory_position_mask`, and the last video-compressor/voxel-pruning statistics from W4 to BF16 after generation. This semantic-state transfer is separate from the incoming-KV substitution. The implementation records incoming BF16/W4 KV lengths and hashes before generation (which may mutate caches in place), as well as selected-KV and output-KV hashes. Matching only an episode ID is insufficient to claim an identical intervention input.

Formal controls pass on all 128 routes: RQ reproduces full reference actions and terminations, and HQ/JQ match the archived H/J pre-intervention tokens, inputs, histories, and output KV. A separate script reconstructs contrasts from route rows. All branches reach natural termination or the existing step limit. Descriptive intervals use 10,000 scene-stratified paired-route resamples (seed 42).

HQ–R's first differing environment action has zero-based median offset 8.5 (interquartile range 6.5–11); its model-call counts appear in Figure 3. JQ–R's 16 first differences occur at later calls 1–5 in 10/1/3/1/1 routes. JQ–J differs on 21 routes within the controlled queue. Call offsets are reconstructed from each episode's observed call sequence, not run-global identifiers. Table 12 gives all terminal gain/loss counts; contrast rates do not decompose causal shares.

Spatial readouts require both branches to reach the actual 2/4/8-action horizon; valid paired denominators are 127/124/116 for HQ–R and JQ–R, without terminal-position imputation. At actions 2 and 4, neither controlled branch has a positive position separation from R. At action 8, HQ–R has 6/116 positive separations (mean over all valid pairs 0.010 m; median among positive pairs 0.250 m), and JQ–R has 9/116 (mean 0.021 m; positive-pair median 0.250 m). The all-pair medians remain zero. Mean signed goal-distance differences at action 8 are -0.0008 m for HQ–R and $+0.0159$ m for JQ–R, with the same 116 valid pairs; these short-horizon means do not replace full-episode out-

Table 12: **All queue-control contrasts.** Each contrast uses the same 128 routes. Divergence intervals are descriptive 95% scene-stratified paired-route intervals (10,000 draws). Gains/losses compare terminal success with the row comparator. RQ matches R exactly.

Contrast	Divergent / 128 (95% interval)	Gain	Loss
H vs. R	23 (11.7–25.0%)	1	0
HQ vs. R	**16 (7.0–18.0%)**	1	0
J vs. R	27 (14.8–28.1%)	4	0
JQ vs. R	16 (7.0–18.8%)	2	0
JQ vs. J	21 (10.2–22.7%)	0	2

Table 13: Fixed-call peak memory (MiB), taking the larger peak across the two real-input replay blocks. Natural-episode memory and both latency boundaries appear in Table 4. Allocated memory counts PyTorch tensors; NVML measures device-used memory.

Model	Memory scope	BF16	W4	FP8-S	FP8-D
StreamVLN	Allocated	16,552	7,950	10,318	10,318
	NVML	19,321	18,555	18,399	18,401
NaviLLM	Allocated	13,237	4,384	7,061	7,061
	NVML	14,289	15,055	14,115	14,115

Table 14: Quality and work from the same two 32-route natural-episode blocks as Table 4. Each block gives the same route-level results. Calls are model invocations; tokens count StreamVLN generated tokens. NaviLLM scores candidates without a comparable generated-token length.

Model	Arm	SR (%)	SPL (%)	Actions	Calls	Tokens
Stream	BF16	37.50	34.89	95.25	24.09	191.63
	W4	43.75	39.74	104.84	26.47	210.72
	FP8-S	40.63	37.21	104.19	26.34	209.56
	FP8-D	43.75	38.33	111.19	28.09	223.56
Navi	BF16	56.25	45.93	6.25	7.25	—
	W4	56.25	44.30	6.13	7.13	—
	FP8-S	59.38	49.51	6.22	7.22	—
	FP8-D	59.38	46.93	6.28	7.28	—

comes. Per-route rows, interval draws, control receipts, and the independent audits accompany the anonymous supplement.

D.3 Goal-Distance and Association Estimators

Table 11b retains the four prespecified signed goal-distance contrasts at eight actions and their accepted 98.75% intervals. Sustained W4 is compared with the joint pulse; the other three contrasts use the unchanged reference. These reference choices are not changed to match the terminal-outcome table.

The local-to-downstream join reuses 256 frozen routes without duplicate or missing route/contrast keys. It pairs initial direct legal-action JS with the absolute goal-distance difference at eight environment actions. Spearman intervals use 10,000 scene-stratified paired-route resamples (seed 42), retaining the original 98.75% four-contrast level. Spearman correlations for current, history, and joint pulses are .276, .143, and .259, with respective 98.75% intervals [.126,.410], [−.018,.294], and [.115,.390] (256 pairs each). These associations concern the magnitude of goal-distance change, not its harmful direction or mediation. The sustained-versus-joint initial distribution pair is unavailable, rather than zero. Small flip-conditioned groups do not identify channel dominance or a causal fraction of terminal failures.

E TIMING AND MEMORY

The StreamVLN episode-end STOP/500-step-cap counts are 30/2, 29/3, 30/2, and 29/3 in table order; all 32 NaviLLM routes stop explicitly in every arm. StreamVLN's separate model-generation logs label all call endings EOS; NaviLLM candidate scoring has no comparable generation-end field. Both block-level rows and reason counts are retained in the accompanying analysis, rather than treated as independent route samples. Natural paired-route intervals are 95% descriptive intervals, estimated after collapsing repeated timings within an input and keeping route identities within scenes.

Measurement protocol. Resource benefits provide the practical motivation for replacement and must be measured at the intended execution boundary. We time BF16, packed W4, and static/dynamic E4M3 FP8 on an RTX PRO 5000 Blackwell, using 32 frozen natural routes and 64 captured real inputs per model. Two independent process blocks per arm run in reverse order; fixed-call replay uses three warmups and five synchronized measurements per input. Repeats are aggregated within input before route pairing. The same natural runs provide the quality, action/call counts, generated lengths, and termination reasons in Table 14. The execution audits and fixed-call memory measurements are given below.

Paired timing contrasts. Packed W4 slows both measured boundaries in both models (Table 4); natural episode ratios relative to BF16 are $1.50\times$ (95% paired interval $[1.33, 1.68]$) for StreamVLN and $3.28\times$ $[3.05, 3.49]$ for NaviLLM. Static FP8 reduces fixed-call time to $0.72\times$ $[0.719, 0.722]$ and $0.62\times$ $[0.600, 0.634]$, respectively. Its StreamVLN episode ratio is $1.02\times$ $[0.89, 1.21]$, leaving the speed direction unresolved. Natural episode time includes the actions, calls, and stopping behavior of each arm; fixed replay holds external inputs at a model-call boundary. Neither this timing contrast nor the workload counts identify how much of the episode-time difference is caused by changed behavior.

Formal real-workload provenance. Table 4 uses independent BF16/W4/static-FP8/dynamic-FP8/dynamic-FP8/static-FP8/W4/BF16 blocks for each model and boundary. Frozen 32-route and 64-input manifests, repeated-call outputs, 100-ms NVML traces, exclusivity, and budget are checked against raw files; 16 latency means are independently reconstructed. Three RustDesk-interrupted blocks are excluded and fully rerun. PyTorch peaks reset after conversion. NVML records device-used memory (`nvmlDeviceGetMemoryInfo`), including context/driver residency, rather than allocated tensors or model-weight storage. StreamVLN FP8 targets 196 decoder Linear layers with frozen G scales; NaviLLM targets 224 W4-matched layers and derives each static E4M3 scale as $\max_c s_c(127/448)$ from `val_seen` INT8 calibration steps s_c. Dynamic FP8 uses $\max|x|/448$ per forward. Both execute E4M3 `torch._scaled_mm` with BF16 KV; other modules retain baseline behavior. Different calibration sources preclude a cross-model FP8 quality comparison.