BEYOND FIXED BENCHMARKS: DATA SCALING OF AUTONOMOUS-DRIVING PLANNERS ON PHYSICALAI

Anonymous authors
Paper under double-blind review

ABSTRACT

Public autonomous-driving benchmarks typically compare planners at a single training-data size. Such comparisons do not show how their measured outcomes change as the training set grows. Using the NVIDIA Physical AI Autonomous Vehicles Dataset (PhysicalAI), we train five fixed planner configurations on one common chain of nested subsets ranging from 1 to 64 hours. We evaluate the selected checkpoints on the fixed, training-disjoint PhysicalAI reporting set using three-second average and final displacement errors (ADE3 and FDE3), and on NAVSIM v2 `navtest` using the closed-loop-relevant Extended Predictive Driver Model Score (EPDMS). On this subset chain, the observed ADE3 leader changes with training-data size, and the five planners show different data-doubling gains. PhysicalAI ADE3 reductions do not always coincide with post-adaptation EPDMS gains. Uni-World VLA has the highest post-adaptation EPDMS at every size, but changes by only 0.268 points from 1 to 64 hours after starting at 88.205; the high starting score limits interpretation of this small change. These curves describe fixed configurations under this data and evaluation protocol, rather than architecture-intrinsic scaling laws.

1 INTRODUCTION

A benchmark at one fixed training-data size answers a useful but narrow question: which planner has the best measured result at that size? It does not show how the measured results change along a sequence of larger training sets. A planner can start with the lowest error but improve slowly, while another can begin behind and later overtake it.

This distinction is familiar from scaling research. Neural-model performance can change systematically with model size, training-data size, and compute. Kaplan et al. reported power-law trends in language-model loss and different scaling behavior across model families (Kaplan et al., 2020). Tay et al. further showed that architecture rankings can change as training budgets increase (Tay et al., 2023). The corresponding question for autonomous driving is simple: does the top-ranked planner at one training-data size remain top-ranked after more driving data are added?

Autonomous-driving benchmarks such as nuScenes, the Waymo Open Dataset for End-to-End Driving (WOD-E2E), nuPlan, and NAVSIM v2 provide fixed splits for reproducible development (Caesar et al., 2020; Xu et al., 2026; Karnchanachari et al., 2024; Dauner et al., 2024). Studies using these benchmarks, however, typically train and evaluate each planner at one established training-data size (Bansal et al., 2019; Hu et al., 2023; Jiang et al., 2023; Sun et al., 2024; Liao et al., 2025; Guo et al., 2024). That result gives a ranking at one size. It does not show the observed changes after successive data additions.

Industrial studies demonstrate why scaling curves are useful, but their strongest results usually rely on private data, internal evaluation environments, and one production planner at a time (Baniodeh et al., 2025; Naumann et al., 2025; Zheng et al., 2024). Public studies such as BevAD and MOSAIC examine related architecture or data-selection questions in more controlled settings (Holtz et al., 2026; Dimlioglu et al., 2026). To our knowledge, prior public studies do not compare multiple planners across the same nested real-world training-data subsets under common evaluation procedures.

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.

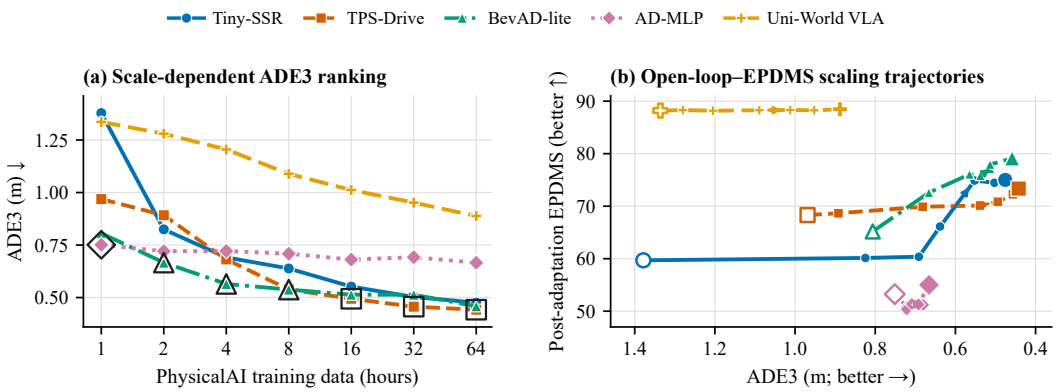


Figure 1: **Observed results across the nested PhysicalAI subsets.** (a) The ADE3 leader changes along this subset chain. (b) Movement toward the upper right means lower PhysicalAI ADE3 and higher post-adaptation NAVSIM v2 EPDMS. BevAD-lite improves on both metrics, whereas Uni-World VLA's EPDMS changes little despite lower ADE3. Its high starting EPDMS limits interpretation of the flat score. Open and filled markers denote the first and largest available training-data sizes.

We address this gap with the NVIDIA Physical AI Autonomous Vehicles Dataset (PhysicalAI). We train five complete planners at seven training-data sizes: a motion-only multilayer perceptron (AD-MLP), the bird's-eye-view (BEV) planner BevAD-lite, our compact sparse scene representation (SSR) planner Tiny-SSR, the vision-language-action (VLA) planner TPS-Drive, and the world-action planner Uni-World VLA (Holtz et al., 2026; Li et al., 2022; Li & Cui, 2025; Li et al., 2026; Liu et al., 2026). Within each planner, the model configuration and training recipe remain fixed while its training set expands along the nested chain. Each expansion also changes which scenes are present, so the curves do not isolate a pure quantity effect.

The PhysicalAI subsets contain 1, 2, 4, 8, 16, 32, and 64 hours of driving data (NVIDIA, 2026a;b). They are nested, so every larger subset retains the earlier samples and adds new ones. Every selected checkpoint is evaluated on the same training-disjoint PhysicalAI reporting set and on the full NAVSIM v2 `navtest` set. This design supports two descriptive comparisons: the observed planner ranking at each training-data size and the change after each data doubling.

Figure 1 previews the observed rank changes and unequal data-doubling gains. It also shows that lower PhysicalAI ADE3 need not coincide with higher post-adaptation EPDMS. This metric comparison must be read in light of each planner's starting EPDMS and the limits of NAVSIM v2 pseudo-simulation.

Repeating this comparison on other public datasets would test whether the rank changes and unequal gains persist beyond PhysicalAI.

Our contributions are:

- A public real-world data-scaling protocol. We create nested PhysicalAI training subsets, construct one fixed, training-disjoint 16-hour PhysicalAI reporting set, and use common ADE3/FDE3 and NAVSIM v2 EPDMS procedures at every training-data size.
- A scaling comparison of five complete planners. We train fixed configurations on the same seven nested training-data sizes and report the observed ranking and data-doubling gain at each size.
- A direct open-loop/EPDMS scaling analysis. We distinguish absolute metric values from changes across training-data sizes and report whether ADE3/FDE3 gains coincide with EPDMS gains.

2 Related Work

Scaling studies compare model families across data and compute budgets (Kaplan et al., 2020; Hoffmann et al., 2022; Tay et al., 2023). We use this experimental idea without assuming a power law for driving planners. A valid curve also requires controlled splits and evaluation: repeated scenes or training–evaluation overlap can make added data appear more useful than it is (Geirhos et al., 2020; Recht et al., 2019).

Driving studies have measured data or compute scaling for proprietary planners (Naumann et al., 2025; Baniodeh et al., 2025; Zheng et al., 2024). Public work has compared BevAD planning components across data sizes and used pilot scaling experiments to select training domains for one planner (Holtz et al., 2026; Dimlioglu et al., 2026). Our comparison instead holds the nested training subsets and evaluation procedures common across five complete planners. We report both trajectory error and NAVSIM v2 EPDMS because open-loop accuracy need not track planning quality (Zhai et al., 2023; Dauner et al., 2024; Cao et al., 2025).

Public driving datasets support reproducible comparison but are commonly used at fixed training sizes (Caesar et al., 2020; Xu et al., 2026; Karnchanachari et al., 2024). The released nuScenes scenes total about 5.6 hours, WOD-E2E's public training and validation segments about 14 hours, and nuPlan's raw camera and LiDAR data 128 hours; these durations have different definitions and are not directly comparable. PhysicalAI's larger release provides room for six doublings from 1 to 64 hours and a training-disjoint reporting set (NVIDIA, 2026a).

Our planners span motion-only, bird's-eye-view, vision-language-action, and world-action representations. Their different inputs and training objectives make it useful to compare complete systems across data sizes (Li et al., 2022; Li & Cui, 2025; Tian et al., 2024; Zhou et al., 2024; Li et al., 2026; 2025; Liu et al., 2026).

3 Task, Data, and Shared Scaling Protocol

Here, a *planner* is a complete planning pipeline; its *configuration* includes architecture, backbone, inputs, objectives, and optimization setup. We evaluate one fixed configuration per planner across training-data sizes.

We report absolute metric values at each size and *data-scaling behavior*, defined as the observed change along one nested subset chain. PhysicalAI ADE3/FDE3 measure open-loop trajectory accuracy, while NAVSIM v2 EPDMS is closed-loop-relevant. The curves remain conditional on the added scenes and fixed training recipes.

Figure 2 summarizes the experiment. Each larger PhysicalAI subset retains the earlier samples, and all planners use the same subsets and evaluation procedures. This permits descriptive comparisons of rankings, data-doubling gains, and whether ADE3/FDE3 gains coincide with EPDMS gains. Each planner retains its native inputs and optimization choices, so the comparison concerns complete configurations rather than isolated components or optimally retuned models.

3.1 Open-loop trajectory evaluation

All five planners predict the future ego trajectory while retaining the sensors, history length, route command, and auxiliary inputs of their original designs. Let the ground-truth position at future time t_k be $\mathbf{p}_k \in \mathbb{R}^2$ and the prediction be $\hat{\mathbf{p}}_k$. For a horizon with K sampled future points, we report average displacement error (ADE) and final displacement error (FDE):

$$\begin{aligned} \mathrm{ADE}_K &= \frac{1}{K}\sum_{k=1}^{K} \left\|\hat{\mathbf{p}}_k - \mathbf{p}_k\right\|_2 , \\ \mathrm{FDE}_K &= \left\|\hat{\mathbf{p}}_K - \mathbf{p}_K\right\|_2 . \end{aligned} \tag{1}$$

The number appended to each metric gives the horizon in seconds. ADE3 and FDE3 are therefore the common three-second open-loop metrics. One-, two-, four-, and five-second errors are retained as supplementary diagnostics and are not substituted for the shared ADE3/FDE3 comparison.

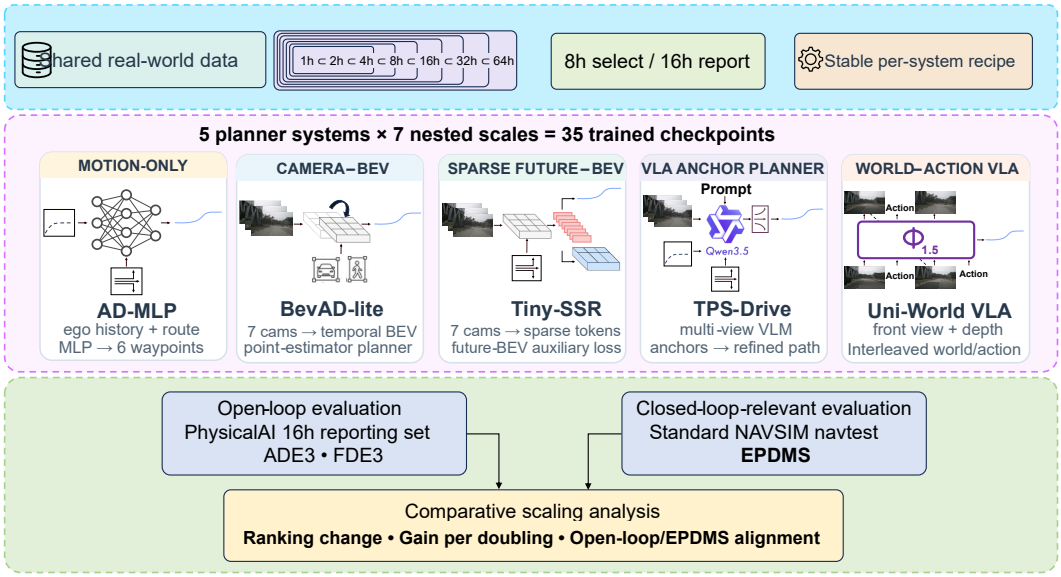


Figure 2: **Cross-framework data-scaling protocol.** Each of the five planners is trained at all seven nested PhysicalAI training-data sizes, producing 35 selected checkpoints. Within a planner, its configuration remains fixed as training-data size changes; configurations may differ among planners. Every selected checkpoint is evaluated with the same PhysicalAI ADE3/FDE3 procedure and the same NAVSIM v2 EPDMS procedure.

3.2 NAVSIM v2 EVALUATION

Trajectory error measures how closely a planner matches the logged future; it does not directly measure driving quality. We therefore also use the standard NAVSIM v2 `navtest` pseudo-simulation and its Extended Predictive Driver Model Score (EPDMS) (Cao et al., 2025). EPDMS combines four safety and rule-compliance terms—no at-fault collision (NC), drivable-area compliance (DAC), driving-direction compliance (DDC), and traffic-light compliance (TLC)—with ego progress (EP), time to collision within bound (TTC), lane keeping (LK), history comfort (HC), and extended comfort (EC). Higher EPDMS is better. We multiply the standard 0–1 score by 100 throughout. EPDMS is closed-loop-relevant because it evaluates safety, progress, and comfort in pseudo-simulation; it is not an on-road safety measurement.

3.3 PHYSICALAI NESTED TRAINING-DATA SUBSETS

We use seven training-data sizes: 1, 2, 4, 8, 16, 32, and 64 hours. The subsets form one fixed nested chain. We first construct the 64-hour pool, draw the 32-hour subset from it, and continue recursively through 16, 8, 4, 2, and 1 hours. This chain is built once before training and reused for every planner. When the training-data size increases from h to $2h$, all samples from the h-hour subset remain and new samples are added. Each interval compares two nested subsets; training-data size and scene composition change together.

An *anchor frame* is a camera frame selected after the original 30 Hz stream is downsampled to 8 Hz. It provides the temporal reference for one training or evaluation sample.

Table 1 reports the common source data at each training-data size. The number of training anchor frames grows from 23,808 to 1,430,464 across six doublings in nominal driving duration.

Every planner begins from the same anchor-frame list at a given training-data size. A planner may discard an anchor frame when its required history, future trajectory, cached depth, or target annotation is unavailable. The number of materialized examples can therefore differ across planners.

Table 1: The common nested PhysicalAI training-data subsets used by all five planners.

Training-data subset size	1h	2h	4h	8h	16h	32h	64h
Training anchor frames	23,808	45,632	93,620	187,612	365,552	721,680	1,430,464

3.4 Fixed, training-disjoint reporting set

The PhysicalAI reporting set is fixed and disjoint from training. We sample complete data chunks rather than individual clips, preventing clips from the same chunk from appearing on both sides of the split. Using seed 20260623, we select a 16-hour PhysicalAI reporting set outside the full 64-hour training pool. It contains 31 chunks, 2,933 clips, and 16.294 hours, with no chunk or clip overlap with training. The 8 Hz source timeline contains 363,692 frames; reported ADE3 and FDE3 values use the same 34,906 evaluation anchors sampled at 1 Hz. A change in the reported error therefore cannot be caused by a change in evaluation samples. The 8-hour validation subset is a prefix of this 16-hour reporting set.

3.5 Shared controls and planner-specific choices

At every training-data size, all planners begin with the same source anchor-frame list and are evaluated on the same final PhysicalAI samples. We also use the same three-second horizon, coordinate frame, ADE3/FDE3 implementation, checkpoint-selection score, and early-stopping patience. Each planner retains the inputs and optimization hyperparameters required by its design. These shared settings make the trajectory metrics directly comparable without forcing a VLA planner and a compact BEV planner to use the same training recipe.

3.6 Training-data-size-dependent ranking and data-doubling gain

At each training-data size, we first compare measured performance. Between adjacent sizes, we report the change along this subset chain. Let $E_s(h)$ be the error of planner s trained on h hours. For adjacent training-data sizes, we define the *ADE3 data-doubling gain* as the percentage reduction in error when the training data double:

$$G_s(h \to 2h) = \frac{E_s(h) - E_s(2h)}{E_s(h)} \times 100\%. \tag{2}$$

A positive value means that measured error falls between the two nested subsets, a value near zero means little change, and a negative value means measured error rises. This *data-doubling gain* is descriptive: it does not separate the effect of sample count from the composition of the added scenes. Several intervals are non-monotonic, so we do not fit a power law to the 1–64h range. FDE3 gain is computed in the same way. For pre- and post-adaptation EPDMS, where higher is better, the gain is $S_s(2h) - S_s(h)$ in EPDMS points.

4 Evaluated Planners

We select five planners that span different input and representation choices. AD-MLP is a motion-only baseline that predicts six waypoints from ego history and a route command. BevAD-lite is a seven-camera spatio-temporal BEV planner (Holtz et al., 2026; Li et al., 2022). Tiny-SSR is our compact implementation of the published SSR framework (Li & Cui, 2025). It uses an 18-layer residual network (ResNet-18) image encoder, compresses navigation-conditioned BEV features into 16 sparse tokens, and adds a future-BEV prediction objective. All Tiny-SSR results refer to this compact configuration.

The two pretrained multimodal planners use token representations. TPS-Drive scores and refines command-conditioned trajectory anchors using multi-view visual-language tokens (Li et al., 2026). Uni-World VLA interleaves future visual-token modeling and action generation from a front camera and cached depth (Liu et al., 2026). We treat all five implementations as complete planners, not as

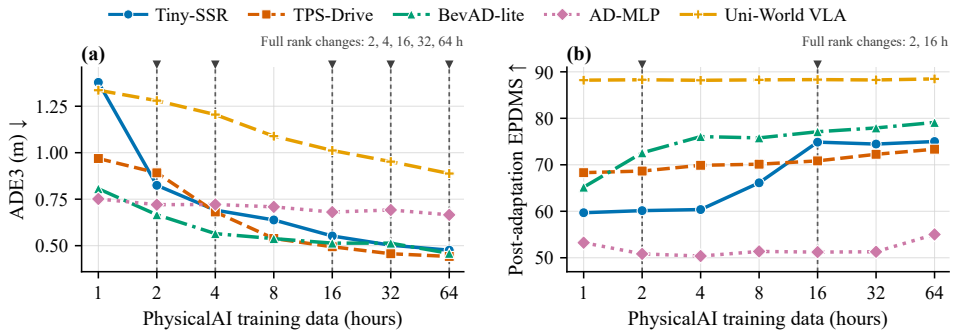


Figure 3: Measured outcomes across the nested PhysicalAI training subsets. Panel (a) reports ADE3 on the fixed, training-disjoint PhysicalAI reporting set. Panel (b) reports post-adaptation EPDMS after two epochs on NAVSIM v2 `navtrain` followed by evaluation on the full `navtest` set. Lower ADE3 and higher EPDMS are better. A dashed vertical marker at training-data size h indicates that the observed full planner ranking differs from the ranking at the preceding training-data size $h/2$. ADE3 ranking changes occur at 2, 4, 16, 32, and 64 hours; post-adaptation EPDMS ranking changes occur at 2 and 16 hours.

controlled ablations of one architecture. Their total/trainable parameter counts are 18M/18M (AD-MLP), 165M/67M (BevAD-lite), 17M/17M (Tiny-SSR), 4.5B/9.4M (TPS-Drive), and 1.8B/5.3M (Uni-World VLA). The billion-parameter totals for TPS-Drive and Uni-World VLA largely reflect frozen pretrained backbones; they do not imply billion-parameter adaptation. The supplementary material gives the detailed frozen and trainable components.

5 Experimental Setup

5.1 Training and checkpoint selection

Each planner keeps its original optimizer and batch size. We do not retune learning rate, weight decay, or batch size separately for each training-data size. Thus, the curves describe the fixed recipes, not the best performance each planner might attain after scale-specific tuning. During PhysicalAI training, checkpoints are selected on a fixed 8-hour validation subset using ADE3 plus 0.5 times FDE3 and a predefined early-stopping rule. We apply the same rule at all seven training-data sizes and recompute the final ADE3/FDE3 values on the full 16-hour reporting set. The validation subset is contained in this reporting set, so the PhysicalAI results are training-disjoint but partly used for checkpoint selection. The supplementary material reports the validation schedule, sampling rate, optimizer, batch size, stopping rule, selected checkpoint step, and training-compute estimates.

5.2 Common evaluation procedures

For each selected PhysicalAI checkpoint, *pre-adaptation EPDMS* is measured directly on the full NAVSIM v2 `navtest` set. For *post-adaptation EPDMS*, we fine-tune that checkpoint for two epochs on NAVSIM v2 `navtrain` and then evaluate it on the full `navtest` set. We measure pre-adaptation EPDMS before any NAVSIM v2 training. Within each planner, differences among the seven pre-adaptation scores arise from the different nested PhysicalAI training subsets under its fixed recipe. The EPDMS evaluation is separate from the PhysicalAI reporting set.

The supplementary NAVSIM v2 `navtest` analysis also reports trajectory ADE3 and FDE3 in metres before and after adaptation; lower is better.

5.3 Reported metrics

The main figures report PhysicalAI ADE3 and post-adaptation EPDMS. Table 2 adds PhysicalAI FDE3 and pre-adaptation EPDMS; complete rankings, data-doubling gains, checkpoint details, and compute estimates are provided in the supplementary material.

6 RESULTS

6.1 OBSERVED PLANNER RANKINGS VARY ACROSS TRAINING-DATA SIZES

The observed ADE3 leader depends on training-data size in this subset chain. Figure 3(a) shows AD-MLP at 1h, BevAD-lite at 2–8h, and TPS-Drive at 16–64h. The full ADE3 order changes after five of the six data doublings and contains nine pairwise rank reversals; only the 4h-to-8h interval preserves the complete order. These are descriptive ranks from one training run per planner and size. In particular, we cannot tell whether an individual reversal would persist with another training seed or subset chain.

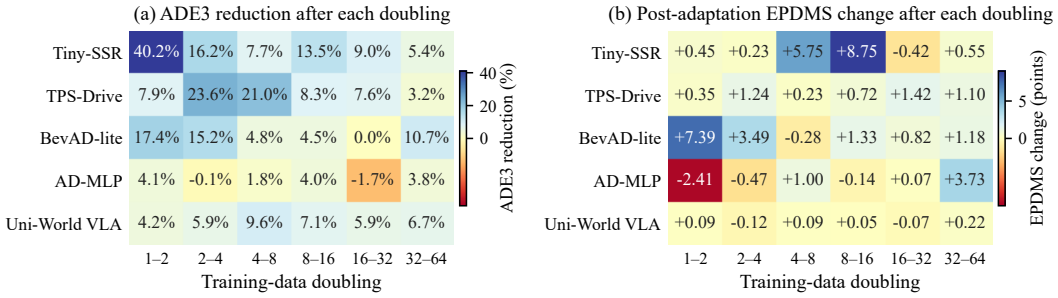


Figure 4: Observed data-doubling gains along one nested subset chain. Panel (a) reports the percentage reduction in ADE3; panel (b) reports the change in post-adaptation EPDMS. Positive values indicate improvement in the respective metric. Each interval also adds scenes with potentially different composition.

Observed rankings also depend on the metric. BevAD-lite has the lowest FDE3 from 2h through 16h, whereas Tiny-SSR has the lowest FDE3 at 32h and 64h. Uni-World VLA leads post-adaptation EPDMS at every training-data size, but the ordering beneath it still changes. Thus, a single size and metric do not describe the full ordering observed in this experiment. The supplementary material reports the complete rankings for all four metrics.

6.2 DATA-DOUBLING GAINS VARY ALONG THE SUBSET CHAIN

Figure 4 shows the observed data-doubling gain for each interval. BevAD-lite's ADE3 changes by only 0.02% from 16h to 32h, then falls by 10.70% from 32h to 64h. The late gain does not establish a smooth quantity-driven scaling trend: the 64h subset also adds scenes whose composition we have not analyzed.

Along this chain, all four vision-based planners have positive ADE3 data-doubling gains in all six intervals. Tiny-SSR has the largest total ADE3 reduction (65.49%), followed by TPS-Drive (54.37%), BevAD-lite (43.12%), and Uni-World VLA (33.53%). AD-MLP improves by only 11.37% and worsens after two doublings. Absolute metric values and measured changes across subsets are different: a planner with low error at one size need not show the largest reduction along this chain. The supplementary material reports the complete ADE3 and FDE3 gain after every doubling.

Table 2: Open-loop and EPDMS change from 1 to 64 hours. ADE3 and FDE3 columns report relative error reduction; EPDMS columns report point change. Positive values indicate improvement. The planners have different 1h baselines, which must be considered when comparing EPDMS point changes.

Planner	ADE3 reduction	FDE3 reduction	Pre-adaptation EPDMS change	Post-adaptation EPDMS change
Tiny-SSR	+65.49%	+56.69%	+19.736	+15.316
TPS-Drive	+54.37%	+45.96%	+12.842	+5.047
BevAD-lite	+43.12%	+36.51%	+22.943	+13.925
AD-MLP	+11.37%	+9.29%	+7.959	+1.776
Uni-World VLA	+33.53%	+29.44%	−0.961	+0.268

6.3 ADE3 AND EPDMS GAINS NEED NOT COINCIDE

PhysicalAI ADE3 and NAVSIM v2 EPDMS are measured on different datasets and capture different aspects of a selected checkpoint. Table 2 shows that Tiny-SSR has the largest 1–64h reductions in ADE3 and FDE3 and the largest post-adaptation EPDMS increase. BevAD-lite has the largest pre-adaptation EPDMS increase (22.943 points). The ordering of measured changes therefore depends on the metric and on whether NAVSIM v2 adaptation is applied.

Uni-World VLA illustrates an important interpretive limit. Its PhysicalAI ADE3 falls by 33.53%, while post-adaptation EPDMS starts at 88.205 and changes by only +0.268 points from 1h to 64h. A high starting score can compress measured point gains on the 0–100 EPDMS scale. Its pre-adaptation EPDMS starts at 67.438 and changes by -0.961 points; thus, a ceiling specific to the post-adaptation series would not explain the entire observed contrast. These measurements cannot determine whether metric compression, the fixed adaptation recipe, or the planner's frozen representations account for the small EPDMS changes. We do not interpret them as an intrinsic limit of Uni-World VLA or of VLA planners.

ADE3 and post-adaptation EPDMS improve in the same direction in all six intervals for TPS-Drive, two for AD-MLP, and four for Uni-World VLA; each Uni-World VLA EPDMS change is at most 0.220 points in magnitude. These counts describe the observed metric curves, without run-to-run uncertainty. The supported conclusion is that PhysicalAI ADE3 reductions do not consistently coincide with post-adaptation NAVSIM v2 EPDMS gains under this protocol.

7 CONCLUSION

For five fixed planner configurations trained on one nested PhysicalAI subset chain, the observed ADE3 ranking and data-doubling gains vary across training-data sizes. PhysicalAI ADE3 reductions do not consistently coincide with post-adaptation NAVSIM v2 EPDMS gains. Uni-World VLA's EPDMS changes by only +0.268 points from a high 1h baseline, but this does not establish poor intrinsic data scaling: metric headroom, frozen representations, and the fixed adaptation recipe remain possible explanations. These results support reporting multiple training-data sizes and both metrics, while treating the curves as conditional descriptions of the evaluated subsets and configurations rather than pure quantity effects or general laws for planner families.

AI USE STATEMENT

Generative AI tools were used for grammar checking, text polishing, LaTeX formatting. The authors take responsibility for the full paper, including its text, claims, citations, figures, and results.

REFERENCES

Mustafa Baniodeh, Kratarth Goel, Scott Ettinger, Carlos Fuertes, Ari Seff, Tim Shen, Cole Gulino, Chenjie Yang, Ghassen Jerfel, Dokook Choe, Rui Wang, Benjamin Charrow, Vinutha Kallem, Sergio Casas, Rami Al-Rfou, Benjamin Sapp, and Dragomir Anguelov. Scaling laws of motion forecasting and planning–technical report. *arXiv preprint arXiv:2506.08228*, 2025.

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *Proceedings of Robotics: Science and Systems*, 2019. doi: 10.15607/RSS.2019.XV.031.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning*, 2025.

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

Tolga Dimlioglu, Nadine Chang, Maying Shen, Rafid Mahmood, and Jose M. Alvarez. Scaling-aware data selection for end-to-end autonomous driving systems. *arXiv preprint arXiv:2604.08366*, 2026. CVPR 2026.

Robert Geirhos, Joern-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, and Liping Jing. End-to-end autonomous driving without costly modularization and 3d manual annotation. *arXiv preprint arXiv:2406.17680*, 2024. URL https://arxiv.org/abs/2406.17680.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

David Holtz, Niklas Hanselmann, Simon Doll, Marius Cordts, and Bernt Schiele. What matters for scalable and robust learning in end-to-end driving planners? *arXiv preprint arXiv:2603.15185*, 2026. CVPR Findings 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2212.10156.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. URL https://arxiv.org/abs/2303.12077.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *IEEE International Conference on Robotics and Automation*, pp. 629–636, 2024.

Jiaxiang Li, Yumao Liu, and Ke Ma. TPS-Drive: Task-guided representation purification for VLM-based autonomous driving. *arXiv preprint arXiv:2605.27038*, 2026. URL https://arxiv.org/abs/2605.27038.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *International Conference on Learning Representations*, 2025.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. DriveVLA-W0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2203.17270.

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. *arXiv preprint arXiv:2411.15139*, 2025. URL https://arxiv.org/abs/2411.15139.

Qiqi Liu, Huan Xu, Jingyu Li, Bin Sun, Zhihui Hao, Dangen She, Xiatian Zhu, and Li Zhang. Uni-World VLA: Interleaved world modeling and planning for autonomous driving. *arXiv preprint arXiv:2603.27287*, 2026.

Alexander Naumann, Xunjiang Gu, Tolga Dimlioglu, Mariusz Bojarski, Alperen Degirmenci, Alexander Popov, Devansh Bisla, Marco Pavone, Urs Muller, and Boris Ivanovic. Data scaling laws for end-to-end autonomous driving. *arXiv preprint arXiv:2504.04338*, 2025.

NVIDIA. NVIDIA PhysicalAI-Autonomous-Vehicles Dataset. Hugging Face dataset, 2026a. URL https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles. Authenticated metadata snapshot recorded in this repository on 2026-05-29.

NVIDIA. PhysicalAI AV Official Devkit. GitHub repository, 2026b. URL https://github.com/NVlabs/physical_ai_av.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. URL https://arxiv.org/abs/2405.19620.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs. model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://aclanthology.org/2023.findings-emnlp.825/.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Dragomir Anguelov. WOD-E2E: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3709–3718, 2026.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023.

Yupeng Zheng, Zhongpu Xia, Qichao Zhang, Teng Zhang, Ben Lu, Xiaochuang Huo, Chao Han, Yixian Li, Mengjie Yu, Bu Jin, Pengxuan Yang, Yuhang Zheng, Haifeng Yuan, Ke Jiang, Peng Jia, Xianpeng Lang, and Dongbin Zhao. Preliminary investigation into data scaling laws for imitation learning-based end-to-end autonomous driving. *arXiv preprint arXiv:2412.02689*, 2024.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bernhard L. Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.

A Supplementary Material: Scope and Notation

We use the NVIDIA Physical AI Autonomous Vehicles Dataset (PhysicalAI). Three-second average displacement error (ADE3) and three-second final displacement error (FDE3) measure trajectory error at the three-second horizon. Extended Predictive Driver Model Score (EPDMS) is NAVSIM v2's pseudo-simulation planning score, reported here after multiplying the standard 0–1 score by 100. Floating-point operations (FLOPs) measure training compute. We refer to the full NAVSIM v2 `navtest` split as NAVTEST.

This supplementary material reports the complete PhysicalAI ADE3 and FDE3 matrices, EPDMS before and after NAVSIM v2 adaptation, the EPDMS change produced by that adaptation, NAVTEST ADE3/FDE3, exact changes after every data doubling, selected-checkpoint training FLOPs, checkpoint steps, and planner implementation details. The primary comparison covers the seven nested PhysicalAI training-data sizes from 1 to 64 hours.

We report ADE3 and FDE3 on both the PhysicalAI reporting set and NAVTEST. We explicitly identify the evaluation dataset whenever these metrics are reported. Both trajectory errors are measured in metres, and lower values are better; higher EPDMS is better. *Pre-adaptation EPDMS* is obtained by evaluating the selected PhysicalAI checkpoint directly on the full NAVSIM v2 `navtest` set. *Post-adaptation EPDMS* is obtained after adapting the same parent checkpoint for two epochs on NAVSIM v2 `navtrain` and then evaluating it on the full `navtest` set. A dash denotes an unavailable measurement. We do not fill unavailable cells with interpolated or estimated performance.

For planner descriptions, multilayer perceptron (MLP), bird's-eye view (BEV), sparse scene representation (SSR), vision-language-action (VLA), low-rank adaptation (LoRA), exponential moving average (EMA), red-green-blue (RGB), and graphics processing unit (GPU) are written in abbreviated form after this point. The motion-only MLP baseline is denoted AD-MLP, and our compact SSR implementation is denoted Tiny-SSR. TPS-Drive, BevAD-lite, and Uni-World VLA are model names.

A.1 How the supplementary supports the main results

The main paper concentrates on three results: planner rankings that change with training-data size, the gain from each data doubling, and the relationship between open-loop and EPDMS improvement. This supplementary material provides five supporting analyses:

1. complete rank trajectories for ADE3, FDE3, pre-adaptation EPDMS, and post-adaptation EPDMS;
2. complete data-doubling gains for all four metrics;
3. the EPDMS change produced by the common NAVSIM v2 adaptation;
4. performance as a function of estimated PhysicalAI training FLOPs; and
5. additional NAVTEST ADE3/FDE3 results and the records needed to reproduce the comparison.

B Complete Rank Trajectories

Figure 5 reports the full planner order at each training-data size. Rank one denotes the best value for that metric and training-data size.

We count a pairwise rank reversal when two planners exchange their relative order between adjacent training-data sizes. A pair is counted only when both planners have results at both training-data sizes, so an unavailable result cannot create a rank reversal.

The ADE3 leader changes from AD-MLP at 1h to BevAD-lite at 2–8h and TPS-Drive at 16–64h. The full ADE3 ranking changes after five of the six data doublings; only the 4h-to-8h interval preserves the full ranking, and the other intervals contain nine pairwise rank reversals in total. The FDE3 leader changes from AD-MLP at 1h to BevAD-lite at 2–16h and Tiny-SSR at 32–64h. Uni-World VLA leads post-adaptation EPDMS at every training-data size, while BevAD-lite passes TPS-Drive at 2h and Tiny-SSR passes TPS-Drive at 16h. Uni-World VLA also leads pre-adaptation EPDMS at every

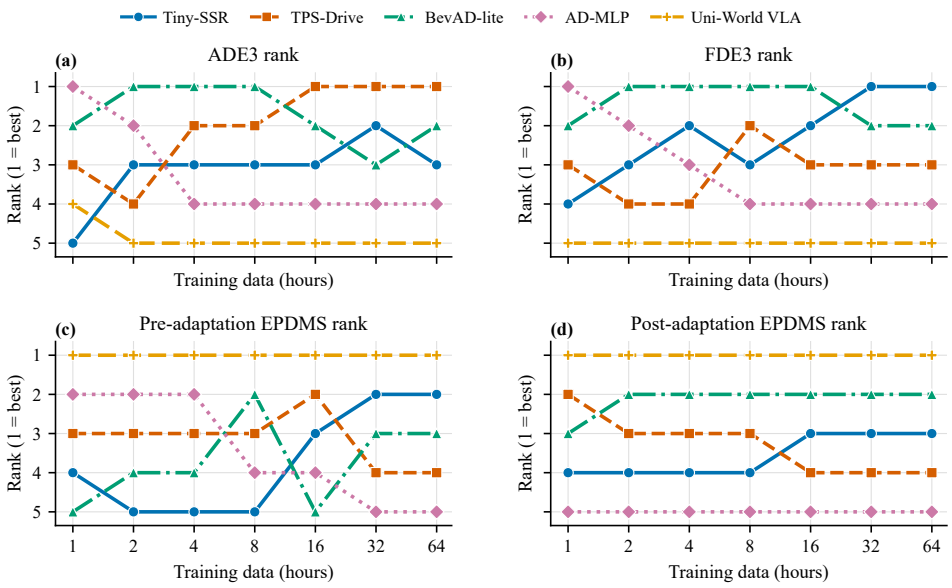


Figure 5: Planner ranks across PhysicalAI training-data sizes for (a) ADE3, (b) FDE3, (c) pre-adaptation EPDMS, and (d) post-adaptation EPDMS. Lower rank is better. Each rank is calculated from the results available at that training-data size.

training-data size. NAVSIM v2 adaptation can change the remaining order; for example, TPS-Drive is second in post-adaptation EPDMS at 1h, whereas AD-MLP is second in pre-adaptation EPDMS.

C Complete PhysicalAI Trajectory Results

Tables 3 and 4 report every available three-second trajectory result. Every value uses the same fixed, training-disjoint PhysicalAI reporting set, the same 1 Hz evaluation anchor-frame list, and the same metric implementation. The row label denotes the amount of PhysicalAI training data; the evaluation set does not change between columns.

Table 3: Complete PhysicalAI ADE3 results in metres. Lower is better.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	1.378333	0.824640	0.691223	0.638299	0.552132	0.502628	0.475720
TPS-Drive	0.968820	0.891807	0.681462	0.538609	0.494163	0.456595	0.442094
BevAD-lite	0.806231	0.666281	0.564751	0.537837	0.513648	0.513536	0.458610
AD-MLP	0.751200	0.720600	0.721500	0.708700	0.680700	0.692100	0.665800
Uni-World VLA	1.336000	1.280000	1.205000	1.089000	1.012000	0.952000	0.888000

Table 4: Complete PhysicalAI FDE3 results in metres. Lower is better.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	2.699978	1.833646	1.614946	1.529500	1.355750	1.227026	1.169230
TPS-Drive	2.296191	2.196872	1.813140	1.490224	1.384169	1.283863	1.240766
BevAD-lite	1.890384	1.672280	1.490135	1.421208	1.329182	1.264724	1.200251
AD-MLP	1.766500	1.706300	1.702700	1.683700	1.637800	1.652800	1.602400
Uni-World VLA	3.135000	3.010000	2.853000	2.642000	2.520000	2.306000	2.212000

The Tiny-SSR 64h row uses the selected EMA checkpoint `iter_1053000_ema`. Its ADE3 and FDE3 are 0.475720 m and 1.169230 m, respectively.

FDE3 provides a different ranking from ADE3 at larger training-data sizes. TPS-Drive has the lowest ADE3 from 16 to 64 hours, whereas BevAD-lite has the lowest FDE3 at 2–16 hours and Tiny-SSR has the lowest FDE3 at 32 and 64 hours. At 64 hours, Tiny-SSR reaches 1.169230 m FDE3, compared with 1.200251 m for BevAD-lite and 1.240766 m for TPS-Drive. The preferred planner can therefore change when the evaluation objective changes from average trajectory error to endpoint error.

D EPDMS Before and After NAVSIM v2 Adaptation

Tables 5 and 6 report all available EPDMS measurements. Pre-adaptation EPDMS evaluates the selected PhysicalAI checkpoint directly on NAVSIM v2 `navtest`. Post-adaptation EPDMS evaluates the same parent checkpoint after two epochs of NAVSIM v2 `navtrain` adaptation. Both measurements use the full `navtest` set.

Table 5: Pre-adaptation EPDMS. Higher is better.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	44.6078	39.2516	39.8947	40.8058	54.0769	62.2745	64.3439
TPS-Drive	47.4899	51.7701	49.7958	55.3833	57.2205	57.0607	60.3317
BevAD-lite	37.7328	39.4300	43.2768	59.8728	50.9477	58.3172	60.6759
AD-MLP	49.1719	53.0327	54.3937	53.2028	54.0549	55.1155	57.1311
Uni-World VLA	67.4380	70.7500	69.8300	70.5520	70.0480	65.4870	66.4770

Figure 6 makes two measured properties visible. Uni-World VLA has the highest EPDMS both before and after adaptation at every training-data size. The other planners show larger EPDMS changes along this nested PhysicalAI subset chain. From 1h to 64h, post-adaptation EPDMS increases by 15.316 points for Tiny-SSR, 13.925 points for BevAD-lite, and 5.047 points for TPS-Drive, but by only 0.268 points for Uni-World VLA. Its high starting post-adaptation EPDMS limits what this small point change can establish about data-scaling behavior.

Table 6: Post-adaptation EPDMS after two NAVSIM v2 `navtrain` epochs. Higher is better.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	59.6887	60.1388	60.3706	66.1227	74.8760	74.4588	75.0047
TPS-Drive	68.3077	68.6540	69.8920	70.1261	70.8419	72.2588	73.3548
BevAD-lite	65.1856	72.5726	76.0656	75.7872	77.1128	77.9306	79.1107
AD-MLP	53.2412	50.8353	50.3656	51.3608	51.2178	51.2918	55.0170
Uni-World VLA	88.2047	88.2971	88.1779	88.2724	88.3261	88.2526	88.4728

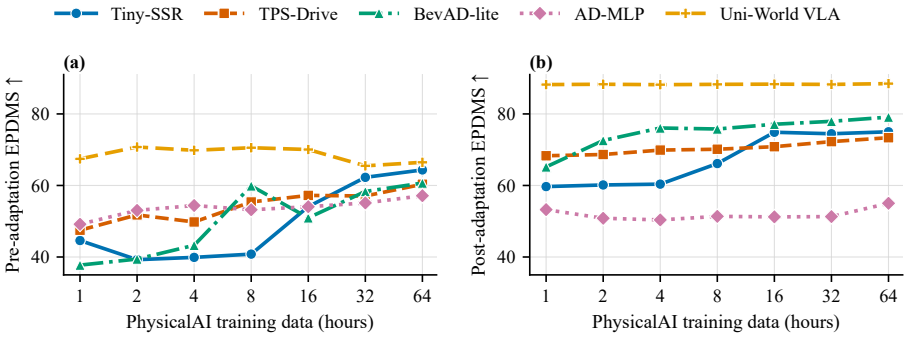


Figure 6: EPDMS scaling before and after the common NAVSIM v2 adaptation. (a) Pre-adaptation EPDMS from the selected PhysicalAI checkpoint. (b) Post-adaptation EPDMS after two `navtrain` epochs. The two scores are reported separately and are not combined into one curve.

D.1 Change produced by NAVSIM v2 adaptation

Table 7 subtracts pre-adaptation EPDMS from post-adaptation EPDMS at each training-data size. A positive value means that adaptation improves EPDMS; a negative value means that it reduces EPDMS.

Table 7: Post-adaptation EPDMS minus pre-adaptation EPDMS, in points.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	+15.081	+20.887	+20.476	+25.317	+20.799	+12.184	+10.661
TPS-Drive	+20.818	+16.884	+20.096	+14.743	+13.621	+15.198	+13.023
BevAD-lite	+27.453	+33.143	+32.789	+15.914	+26.165	+19.613	+18.435
AD-MLP	+4.069	-2.197	-4.028	-1.842	-2.837	-3.824	-2.114
Uni-World VLA	+20.767	+17.547	+18.348	+17.720	+18.278	+22.766	+21.996

The effect of the same adaptation procedure depends strongly on the planner. BevAD-lite gains between 15.914 and 33.143 EPDMS points at every training-data size. Uni-World VLA gains between 17.547 and 22.766 points and already reaches 88.2047 after adaptation of its 1h checkpoint. AD-MLP improves only at 1h; adaptation reduces its EPDMS at the other six training-data sizes. NAVSIM v2 adaptation is therefore not a neutral post-processing step: it interacts with each planner's representation and optimization recipe and can change both absolute performance and the measured scaling curve.

E NAVTEST Open-Loop ADE3/FDE3

This section reports ADE3 and FDE3 computed on the full NAVSIM v2 `navtest` set rather than on the PhysicalAI reporting set. Predictions and ground-truth trajectories are matched by NAVTEST scenario token and expressed in the current-ego local coordinate frame. Each complete trace aggregate contains 12,146 NAVTEST scenarios with zero failed scenarios. Complete pre- and post-adaptation traces are available for TPS-Drive, BevAD-lite, and AD-MLP. For Tiny-SSR, full 12,146-scenario traces with zero failures are available at all seven training-data sizes. The 32h NAVTEST ADE3/FDE3

trace uses `epoch_6_ema`; the EPDMS scaling row uses the selected `epoch_3_ema` checkpoint. NAVTEST ADE3/FDE3 was not collected for Uni-World VLA.

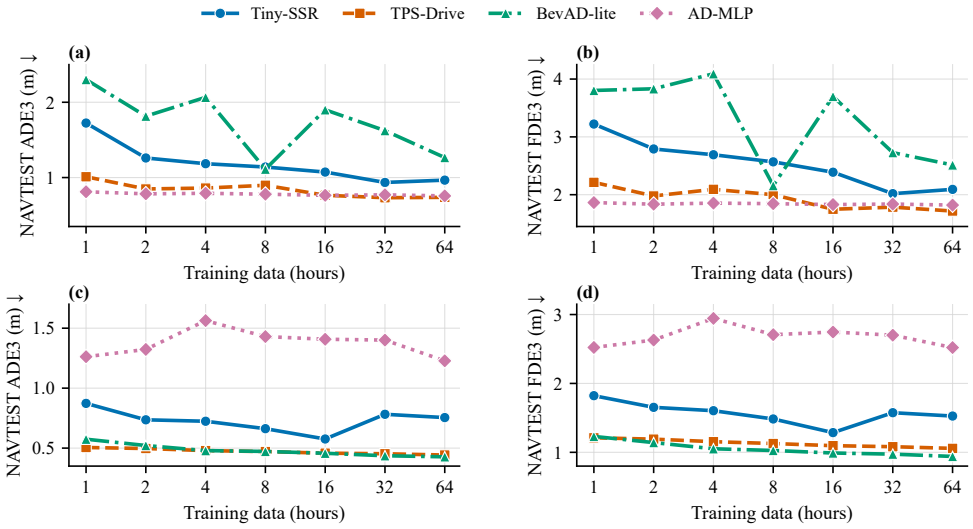


Figure 7: NAVTEST open-loop trajectory results. The top row reports pre-adaptation (a) ADE3 and (b) FDE3. The bottom row reports post-adaptation (c) ADE3 and (d) FDE3 after two NAVSIM v2 adaptation epochs. Lower is better.

Table 8: Pre-adaptation NAVTEST ADE3/FDE3 in metres. Lower is better.

Planner	Metric	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	ADE3	1.724	1.260	1.184	1.140	1.073	0.935	0.9657
	FDE3	3.222	2.791	2.691	2.568	2.389	2.018	2.0924
TPS-Drive	ADE3	1.0108	0.8486	0.8604	0.8975	0.7647	0.7322	0.7373
	FDE3	2.2136	1.9784	2.0927	2.0002	1.7482	1.7845	1.7174
BevAD-lite	ADE3	2.3013	1.8150	2.0657	1.1111	1.8999	1.6233	1.2657
	FDE3	3.8029	3.8317	4.0917	2.1576	3.6995	2.7276	2.5143
AD-MLP	ADE3	0.8112	0.7836	0.7910	0.7789	0.7643	0.7706	0.7563
	FDE3	1.8636	1.8359	1.8544	1.8447	1.8293	1.8383	1.8204

Table 9: Post-adaptation NAVTEST ADE3/FDE3 in metres after two NAVSIM v2 adaptation epochs. Lower is better.

Planner	Metric	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	ADE3	0.873	0.736	0.724	0.662	0.576	0.782	0.7546
	FDE3	1.822	1.653	1.604	1.484	1.286	1.575	1.5265
TPS-Drive	ADE3	0.5051	0.4966	0.4790	0.4711	0.4579	0.4530	0.4423
	FDE3	1.2107	1.1912	1.1543	1.1276	1.0964	1.0815	1.0572
BevAD-lite	ADE3	0.5742	0.5217	0.4790	0.4723	0.4571	0.4362	0.4274
	FDE3	1.2308	1.1413	1.0515	1.0267	0.9899	0.9734	0.9402
AD-MLP	ADE3	1.2620	1.3230	1.5632	1.4300	1.4078	1.4008	1.2278
	FDE3	2.5224	2.6297	2.9441	2.7076	2.7464	2.7004	2.5205

Table 10: Tiny-SSR NAVTEST trajectory errors in metres for the 64h `iter_1053000_ema` checkpoint. Post-adaptation values follow two NAVSIM v2 `navtrain` epochs on an A40 GPU. The 3s row supplies the NAVTEST ADE3/FDE3 values in Tables 8 and 9. Lower is better.

Horizon	Pre-adaptation ADE/FDE	Post-adaptation ADE/FDE
1s	0.2470 / 0.3557	0.2429 / 0.3325
2s	0.5462 / 1.0372	0.4631 / 0.8167
3s	0.9657 / 2.0924	0.7546 / 1.5265
4s	1.5061 / 3.4983	1.1287 / 2.5245

The Tiny-SSR 32h `epoch_6_ema` trace contains 12,146 valid scenarios and zero failures. Before adaptation, its EPDMS is 64.3317 and its NAVTEST ADE3/FDE3 is 0.935/2.018 m. After adaptation, these values are 75.4297 and 0.782/1.575 m. Tables 8 and 9 use these NAVTEST ADE3/FDE3 values. The EPDMS scaling tables retain the registered `epoch_3_ema` values.

At 64 hours, adaptation lowers NAVTEST ADE3/FDE3 by 0.2111/0.5659 m for Tiny-SSR, 0.2950/0.6602 m for TPS-Drive, and 0.8383/1.5741 m for BevAD-lite. For AD-MLP, adaptation instead increases NAVTEST ADE3/FDE3 by 0.4715/0.7001 m. This agrees with the EPDMS adaptation result: adaptation improves the registered Tiny-SSR, TPS-Drive, and BevAD-lite results but reduces the reported performance of AD-MLP.

From 1 to 64 hours, Tiny-SSR reduces pre-adaptation NAVTEST ADE3/FDE3 by 43.98%/35.06% and post-adaptation NAVTEST ADE3/FDE3 by 13.56%/16.22%. Thus, the Tiny-SSR NAVTEST traces provide the same qualitative scaling signal as its PhysicalAI ADE3/FDE3 curve: trajectory error at 64 hours is lower than at 1 hour both before and after NAVSIM v2 adaptation.

F Exact Change After Every Data Doubling

For ADE3 and FDE3, data-doubling gain is the relative error reduction

$$100\frac{E_h - E_{2h}}{E_h}.$$

For EPDMS, data-doubling gain is the point change

$$S_{2h} - S_h.$$

Positive values indicate improvement for all four tables below. A negative value means that performance became worse after that data doubling.

Table 11: Relative PhysicalAI ADE3 and FDE3 reduction after each training-data doubling, in percent.

Planner	Metric	1–2h	2–4h	4–8h	8–16h	16–32h	32–64h
Tiny-SSR	ADE3	+40.17	+16.18	+7.66	+13.50	+8.97	+5.35
	FDE3	+32.09	+11.93	+5.29	+11.36	+9.49	+4.71
TPS-Drive	ADE3	+7.95	+23.59	+20.96	+8.25	+7.60	+3.18
	FDE3	+4.33	+17.47	+17.81	+7.12	+7.25	+3.36
BevAD-lite	ADE3	+17.36	+15.24	+4.77	+4.50	+0.02	+10.70
	FDE3	+11.54	+10.89	+4.63	+6.48	+4.85	+5.10
AD-MLP	ADE3	+4.07	-0.12	+1.77	+3.95	-1.67	+3.80
	FDE3	+3.41	+0.21	+1.12	+2.73	-0.92	+3.05
Uni-World VLA	ADE3	+4.19	+5.86	+9.63	+7.07	+5.93	+6.72
	FDE3	+3.99	+5.22	+7.40	+4.62	+8.49	+4.08

The number of improving intervals provides a compact description of how consistently each metric responds to more data. It does not measure the size of the improvement and does not establish statistical significance.

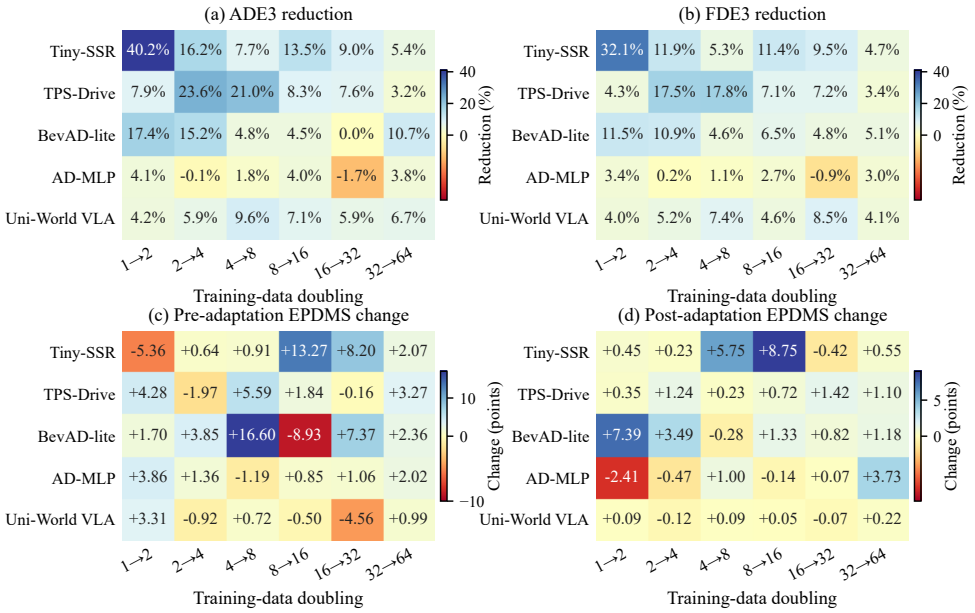


Figure 8: Complete data-doubling gains for (a) ADE3, (b) FDE3, (c) pre-adaptation EPDMS, and (d) post-adaptation EPDMS. Positive values indicate improvement. ADE3 and FDE3 use relative error reduction; EPDMS uses point change.

Table 12: Pre- and post-adaptation EPDMS change after each training-data doubling, in EPDMS points.

Planner	EPDMS measurement	1–2h	2–4h	4–8h	8–16h	16–32h	32–64h
Tiny-SSR	Pre-adaptation	-5.356	+0.643	+0.911	+13.271	+8.198	+2.069
	Post-adaptation	+0.450	+0.232	+5.752	+8.753	-0.417	+0.546
TPS-Drive	Pre-adaptation	+4.280	-1.974	+5.588	+1.837	-0.160	+3.271
	Post-adaptation	+0.346	+1.238	+0.234	+0.716	+1.417	+1.096
BevAD-lite	Pre-adaptation	+1.697	+3.847	+16.596	-8.925	+7.370	+2.359
	Post-adaptation	+7.387	+3.493	-0.278	+1.326	+0.818	+1.180
AD-MLP	Pre-adaptation	+3.861	+1.361	-1.191	+0.852	+1.061	+2.016
	Post-adaptation	-2.406	-0.470	+0.995	-0.143	+0.074	+3.725
Uni-World VLA	Pre-adaptation	+3.312	-0.920	+0.722	-0.504	-4.561	+0.990
	Post-adaptation	+0.092	-0.119	+0.095	+0.054	-0.073	+0.220

Table 13: Direction of improvement across adjacent training-data doublings. Panel (a) counts the intervals in which each metric improves. Panel (b) counts the intervals in which both listed metrics improve together. "Pre" and "post" refer to EPDMS before and after the common NAVSIM v2 adaptation. Each denominator is the number of intervals for which both required endpoint measurements are available. These counts describe directional consistency; Tables 11 and 12 report the corresponding magnitudes.

(a) Intervals in which each metric improves

Planner	ADE3	FDE3	Pre-adaptation EPDMS	Post-adaptation EPDMS
Tiny-SSR	6/6	6/6	5/6	5/6
TPS-Drive	6/6	6/6	4/6	6/6
BevAD-lite	6/6	6/6	5/6	5/6
AD-MLP	4/6	5/6	5/6	3/6
Uni-World VLA	6/6	6/6	3/6	4/6

(b) Intervals in which both listed metrics improve

Planner	ADE3 + pre EPDMS	ADE3 + post EPDMS	FDE3 + pre EPDMS	FDE3 + post EPDMS
Tiny-SSR	5/6	5/6	5/6	5/6
TPS-Drive	4/6	6/6	4/6	6/6
BevAD-lite	5/6	5/6	5/6	5/6
AD-MLP	3/6	2/6	4/6	2/6
Uni-World VLA	3/6	4/6	3/6	4/6

The interval-level results show where measured ADE3 decreases along this subset chain. Tiny-SSR obtains its largest ADE3 reduction, 40.17%, from 1h to 2h, whereas TPS-Drive obtains its largest reductions from 2h to 4h and from 4h to 8h. BevAD-lite changes by only 0.02% from 16h to 32h but then reduces ADE3 by 10.70% from 32h to 64h. The larger subset also adds new scenes, so this contrast does not identify a pure training-data-size effect or establish permanent saturation at 32h.

The additional open-loop/EPDMS comparisons show that directional agreement and total improvement are distinct. ADE3 and post-adaptation EPDMS improve together in 5/6 intervals for Tiny-SSR, 6/6 for TPS-Drive, 5/6 for BevAD-lite, 2/6 for AD-MLP, and 4/6 for Uni-World VLA. Uni-World VLA's post-adaptation EPDMS remains within 88.1779–88.4728, changes by only +0.268 points from 1h to 64h, and has the smallest 1-to-64-hour change among the five planners; its pre-adaptation EPDMS changes by -0.961 points, also the smallest change. Its post-adaptation EPDMS changes by at most 0.220 points in any interval even though its ADE3 decreases by 33.53%. The original Uni-World VLA paper reports an 89.4 Predictive Driver Model Score (PDMS) under its NAVSIM evaluation protocol (Liu et al., 2026). PDMS and EPDMS are not numerically comparable, but both results show strong absolute NAVSIM performance. TPS-Drive reduces ADE3 by 54.37%, compared with 43.12% for BevAD-lite, whereas BevAD-lite gains 22.943 and 13.925 EPDMS points before and after adaptation, compared with 12.842 and 5.047 points for TPS-Drive. Section D reports how the same NAVSIM v2 adaptation changes the five planners by different amounts. The high Uni-World VLA EPDMS baseline and lack of run-to-run uncertainty preclude an intrinsic planner-scaling conclusion from these point changes.

We do not attach confidence intervals or statistical significance tests to correlations across these curves. Each planner contributes only one training run per training-data size and at most six paired data-doubling intervals, so such intervals would not represent training-seed or subset-construction uncertainty. We instead report every paired direction and magnitude directly.

G Selected-Checkpoint Training Compute

Training-data size indexes the nested subsets; scene composition changes when new samples are added. Estimated PhysicalAI training FLOPs provide a hardware-independent description of the training cost used to reach each selected checkpoint. Let F_{train} denote training FLOPs per sample,

B_{global} the effective global batch size, and S the selected optimizer step. We calculate

$$\widehat{F}_{\text{PhysicalAI}} = F_{\text{train}} B_{\text{global}} S.$$

One multiply-add counts as two FLOPs. The accounting includes PhysicalAI training through the selected checkpoint and excludes historical pretraining, offline preprocessing, evaluation, checkpoint input/output, NAVSIM v2 adaptation, and NAVTEST evaluation.

TPS-Drive uses measured supported-operator forward, backward, and checkpoint-recomputation FLOPs. BevAD-lite uses measured supported operators with its registered recurrent-BEV accounting bound. Tiny-SSR uses its registered profiler approximation. AD-MLP uses three times its forward FLOPs with the registered checkpoint pairing. Uni-World VLA uses the registered decomposition of no-backward, frozen-weight operations that remain in autograd, and trainable operations.

Estimated training FLOPs are not required to increase monotonically with training-data size because the selected optimizer step can decrease. For example, the Uni-World VLA 4h checkpoint is selected at step 28,203, whereas the 8h and 16h checkpoints are selected at step 10,000. The Tiny-SSR 16h and 32h checkpoints likewise stop at similar steps. The compute curves therefore describe the realized cost of the selected checkpoints, not the nominal size of the training-data subset or intrinsic model complexity.

Figure 9 shows why training-data size and realized training compute should be reported separately. AD-MLP occupies a compute range several orders of magnitude below the vision-based planners, while TPS-Drive, BevAD-lite, and Uni-World VLA reach overlapping performance ranges with different amounts of training compute. The connected points are ordered by training-data size, but their horizontal coordinates need not increase monotonically because checkpoint selection can stop a larger-data run at an earlier optimizer step.

The 64h checkpoints make this distinction concrete. TPS-Drive uses an estimated 49.0970×10^{18} PhysicalAI training FLOPs, compared with 27.1908×10^{18} for BevAD-lite and 16.3232×10^{18} for Uni-World VLA. TPS-Drive therefore uses 1.81 times the estimated training FLOPs of BevAD-lite and 3.01 times those of Uni-World VLA at the same nominal training-data size. AD-MLP uses only 0.0019×10^{18} FLOPs at 64h, more than four orders of magnitude below the three vision-based planners. These differences explain why training-data size and training FLOPs provide complementary resource views rather than interchangeable horizontal axes.

The non-monotonic points also have identifiable causes. TPS-Drive selects 10,000 optimizer steps at 1h but 9,000 at 2h. Uni-World VLA selects 28,203 steps at 4h but 10,000 at both 8h and 16h. Tiny-SSR uses 91,388 steps at 16h and 90,210 at 32h, while AD-MLP selects 810 steps at 16h and 354 at 32h. A larger training subset can therefore require less realized compute when its selected checkpoint occurs earlier. The compute curves describe the cost of the reported checkpoints; they do not estimate the compute needed to train every planner for an equal number of epochs.

For post-adaptation EPDMS, the horizontal coordinate includes only the PhysicalAI training used to create the parent checkpoint. It excludes the subsequent two-epoch NAVSIM v2 adaptation. This convention isolates the cost of the scaling variable studied in the paper. It should not be interpreted as the total compute required to reproduce the complete EPDMS pipeline.

Table 14: Compute inputs reused across training-data sizes when the executed model graph and training input shape are unchanged. Training FLOPs per sample include backward computation and checkpoint recomputation according to each planner's registered calculation.

Planner	Forward FLOPs/sample	No-backward FLOPs/sample	Training FLOPs/sample	Global batch
Tiny-SSR	7.268×10^{11}	2.229×10^{11}	1.735×10^{12}	24
TPS-Drive	9.096×10^{12}	1.404×10^{12}	2.360×10^{13}	8
BevAD-lite	2.369×10^{12}	1.156×10^{10}	5.665×10^{12}	24
AD-MLP	3.601×10^{7}	0	1.080×10^{8}	4096
Uni-World VLA	2.140×10^{13}	1.752×10^{13}	2.915×10^{13}	16

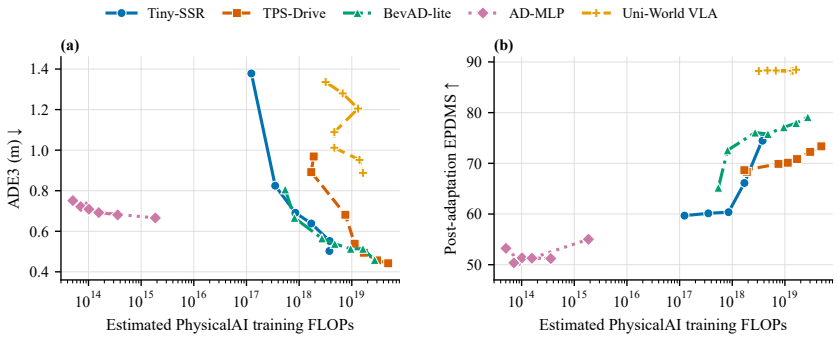


Figure 9: Complete-range compute comparison including AD-MLP. (a) PhysicalAI ADE3 and (b) post-adaptation EPDMS versus estimated PhysicalAI training FLOPs. Markers are selected checkpoints; lines connect checkpoints in increasing training-data-size order. The EPDMS coordinate uses the training FLOPs of its parent PhysicalAI checkpoint and excludes NAVSIM v2 adaptation.

Table 15: Estimated PhysicalAI training FLOPs to each selected checkpoint, reported in units of 10^{18} FLOPs.

Planner	1h	2h	4h	8h	16h	32h	64h
Tiny-SSR	0.1239	0.3510	0.8504	1.7009	3.8044	3.7554	–
TPS-Drive	1.8883	1.6995	7.5534	11.3301	16.9951	30.2135	49.0970
BevAD-lite	0.5438	0.8157	2.7191	4.7584	9.5168	16.3145	27.1908
AD-MLP	0.0001	0.0001	0.0001	0.0001	0.0004	0.0002	0.0019
Uni-World VLA	3.1881	6.7111	13.1532	4.6638	4.6638	13.9913	16.3232

Table 16: Optimizer, effective global batch size, and selected checkpoint step. Values are listed in ascending training-data-size order.

Planner	Optimizer	Global batch	Selected optimizer steps at 1/2/4/8/16/32/64h
AD-MLP	AdamW	4096	114 / 204 / 161 / 230 / 810 / 354 / 4,200
BevAD-lite	ScheduleFree AdamW	24	4,000 / 6,000 / 20,000 / 35,000 / 70,000 / 120,000 / 200,000
Tiny-SSR	AdamW	24	2,976 / 8,432 / 20,429 / 40,858 / 91,388 / 90,210 / –
TPS-Drive	AdamW	8	10,000 / 9,000 / 40,000 / 60,000 / 90,000 / 160,000 / 260,000
Uni-World VLA	AdamW	16	6,836 / 14,390 / 28,203 / 10,000 / 10,000 / 30,000 / 35,000

H Training, Checkpoint Selection, and Evaluation Details

H.1 Training anchor frames and evaluation anchor frames

We first construct the 64-hour PhysicalAI training pool. Using one fixed random draw, we select the 32-hour subset from that pool, then select the 16-hour subset from the saved 32-hour membership list, and recursively continue through 8, 4, 2, and 1 hours. The complete nested membership chain is fixed before training and reused for every planner. Thus, each larger subset adds data to the smaller subset rather than replacing its examples.

The original PhysicalAI camera stream is downsampled from 30 Hz to 8 Hz before anchor frames are selected. An anchor frame is a selected camera frame that serves as the temporal reference for one training or evaluation example. The seven nested training-data subsets contain 23,808; 45,632; 93,620; 187,612; 365,552; 721,680; and 1,430,464 training anchor frames. Every planner starts from the same training anchor-frame list at a given training-data size. A planner may discard an anchor frame when the history, future trajectory, cached depth, or target annotation required by that planner is unavailable. TPS-Drive consequently produces 17,430; 33,405; 68,584; 137,509; 267,949; 528,985; and 1,048,650 training examples.

The fixed PhysicalAI reporting set is selected at complete-chunk granularity outside the 64h training set. It contains 31 chunks, 2,933 clips, 16.294 hours, and 363,692 camera frames on the 8 Hz source timeline, with no chunk or clip shared with training. Every reported ADE3/FDE3 checkpoint is evaluated on the same 34,906 evaluation anchor frames sampled at 1 Hz. The 8h validation set used for checkpoint selection is a prefix of this 16h reporting set; the reported PhysicalAI scores are training-disjoint but partly used for model selection.

H.2 Checkpoint selection

During training, checkpoints are evaluated on a fixed 8h PhysicalAI validation set sampled at 1 Hz. The checkpoint-selection score is

$$\mathrm{ADE3} + 0.5\,\mathrm{FDE3}.$$

For smaller training-data sizes, checkpoints are usually evaluated approximately every 3,000 optimizer steps. For larger training-data sizes, the interval increases to approximately 20,000–30,000 steps. Training stops after three consecutive validation checks without a lower selection score, and the checkpoint with the minimum observed score is selected. This rule is defined before the final comparison among planners and is unchanged across the seven training-data sizes.

Preflight experiments evaluated selected checkpoints at 1, 2, and 8 Hz on the fixed PhysicalAI reporting set. ADE3 and FDE3 changed by only 0.001–0.002 m, and the planner rankings did not change. We therefore use one common 1 Hz evaluation anchor-frame list for all final PhysicalAI results. This sampling choice does not affect EPDMS. EPDMS uses two `navtrain` adaptation epochs followed by evaluation on the full `navtest` set.

H.3 One run per planner and training-data size

Each reported row comes from one training run followed by deterministic selected-checkpoint evaluation. We do not report seed variance, confidence intervals, or statistical significance. Differences smaller than possible training-run variation should therefore not be interpreted as statistically established. We also use one randomly constructed nested subset chain. The reported curves therefore do not quantify sensitivity to a different random subset construction or data ordering. Each doubling adds new scenes, so training-data size and scene composition change together. We do not analyze geography, weather, or scenario coverage in the added scenes. Within a planner, the optimization recipe is held fixed across sizes rather than retuned for each subset.

I Planner Implementation Details

Each planner keeps the same configuration and checkpoint-selection rule at all seven training-data sizes. The common PhysicalAI and NAVSIM v2 evaluation procedures are applied only after the selected checkpoint has been identified. Table 17 makes the principal differences among planners

explicit. These differences are part of each planner, so the experiments measure conditional planner-level scaling rather than an architecture-only effect.

Table 17: Planner configurations used in the scaling comparison. The planners intentionally retain their native inputs, pretraining, auxiliary supervision, and optimization recipes.

Planner	Inputs and temporal context	Pretraining and frozen components	PhysicalAI objectives	Optimizer
AD-MLP	No cameras; four 0.5s-spaced ego poses, velocity, acceleration, and route command; six future states over 3s	No pretrained component	Trajectory L1	AdamW, global batch 4096, learning rate 10^{-3}
BevAD-lite	Seven cameras, recurrent BEV state, ego state, route command, and current objects; 15 future steps over 3s	RADIO c-radio-v3-b image backbone with LoRA; base image weights frozen	Object detection and trajectory planning; map loss disabled	ScheduleFree AdamW, global batch 24, learning rate 2×10^{-4}
Tiny-SSR	Seven cameras, three-frame queue, ego state, and route command; six future steps over 3s	Torchvision ResNet-18 initialization; frozen image-backbone computation is retained in the executed graph	Trajectory regression and future-BEV reconstruction	AdamW, global batch 24, learning rate 5×10^{-5}
TPS-Drive	Multi-view cameras, ego-motion history, kinematics, route command, and command-specific trajectory anchors; common evaluation uses the first 3s	Qwen3.5-4B base frozen; LoRA in the last eight language layers and planner decoder trained	Anchor classification, trajectory regression, offset regularization, and acceleration smoothness	AdamW, global batch 8, base/decoder learning rates $3\times10^{-5}/2\times10^{-4}$
Uni-World VLA	Front-wide RGB, cached monocular depth, ego/status input, 2s history, and eight future actions over 4s; common evaluation uses the first 3s	Show-O/Phi-1.5 world-action model; visual tokenizer and depth encoder frozen	Visual-token modeling and trajectory regression	AdamW, global batch 16, learning rate 10^{-5}

Table 18: Parameter counts for the executed PhysicalAI training configurations. Counts include pretrained components that are loaded and used by the planner. Frozen is total minus trainable. The configuration is unchanged across all seven training-data sizes, so each planner has one parameter count.

Planner	Total	Trainable	Frozen	Trainable %
AD-MLP	18M	18M	0	100
BevAD-lite	165M	67M	98M	41
Tiny-SSR	17M	17M	0.17M	99
TPS-Drive	4.5B	9.4M	4.5B	0.21
Uni-World VLA	1.8B	5.3M	1.8B	0.29

AD-MLP trains all of its parameters, and Tiny-SSR is also almost fully trainable. BevAD-lite freezes most of its pretrained RADIO backbone. TPS-Drive and Uni-World VLA load billion-parameter pretrained models but train fewer than 10M parameters through adapters and planner-specific modules.

At 1, 2, 4, 8, 16, 32, and 64 hours, the common source-anchor counts are 23,808; 45,632; 93,620; 187,612; 365,552; 721,680; and 1,430,464. TPS-Drive retains 17,430; 33,405; 68,584; 137,509; 267,949; 528,985; and 1,048,650 examples at the corresponding sizes because it removes anchors without complete motion context. Separate retained-example ledgers are not available for the other four planners, so we report their common source anchors without claiming that every optimizer consumed every anchor.

AD-MLP. The motion-only baseline consumes four past ego poses, velocity, acceleration, and a three-way route command. The reported 18M-parameter MLP predicts six future waypoint states at 0.5-second intervals. Split-specific feature normalization and the same checkpoint-selection rule are used at every training-data size.

BevAD-lite. BevAD-lite consumes seven cameras, temporal ego state, a coarse route command, and current object supervision. Its PhysicalAI adaptation uses a RADIO image backbone with LoRA, a

256-dimensional neck, a spatio-temporal BEV encoder, and a query-based trajectory head. Detection and trajectory-planning losses are active; map supervision is disabled in this study. Training uses ScheduleFree AdamW with global batch 24 and base learning rate 2×10^{-4}.

Tiny-SSR. Tiny-SSR is our compact implementation of the published SSR framework described in the main paper. It uses seven cameras and a three-frame queue. A ResNet-18 image encoder and a BEVFormer-style camera-to-BEV module produce navigation-conditioned features. TokenLearner compresses the BEV representation into 16 sparse tokens. A latent transformer predicts a future BEV state and decodes command-conditioned trajectories. The active objectives are trajectory regression and future-BEV reconstruction. Training uses AdamW with global batch 24 and initial learning rate 5×10^{-5}.

TPS-Drive. TPS-Drive uses a Qwen3.5-4B vision-language backbone, multi-view imagery, ego history, kinematics, and command-conditioned trajectory anchors. The frozen backbone is adapted with LoRA. A transformer planner scores 64 anchors for each route command and predicts residual trajectory offsets. Training combines anchor classification, trajectory regression, offset regularization, and acceleration smoothness. TPS-Drive discards PhysicalAI anchor frames without the complete motion history and future trajectory required by the planner. Training uses AdamW with global batch 8, base learning rate 3×10^{-5}, and decoder learning rate 2×10^{-4}.

Uni-World VLA. Uni-World VLA interleaves future visual-token prediction and action generation. The PhysicalAI adaptation uses the front-wide camera, cached monocular depth, a frozen visual tokenizer and depth encoder, and a causal language backbone. It consumes two seconds of history and predicts eight action points over a four-second future. Training combines visual-token modeling and trajectory regression. AdamW uses global batch 16 and learning rate 10^{-5}.

J Uni-World VLA Multi-Horizon Measurements

The common comparison among planners uses ADE3/FDE3. Uni-World VLA also provides one-, two-, three-, and four-second PhysicalAI trajectory measurements. These additional horizons are reported in Table 19.

Table 19: Uni-World VLA PhysicalAI ADE/FDE in metres at one-, two-, three-, and four-second horizons.

Train h	Selected step	ADE1/FDE1	ADE2/FDE2	ADE3/FDE3	ADE4/FDE4
1	6,836	0.228/0.362	0.671/1.419	1.336/3.135	2.219/5.497
2	14,390	0.216/0.343	0.640/1.355	1.280/3.010	2.138/5.327
4	28,203	0.199/0.315	0.597/1.270	1.205/2.853	2.024/5.073
8	10,000	0.196/0.268	0.525/1.109	1.089/2.642	1.857/4.719
16	10,000	0.158/0.244	0.471/1.022	1.012/2.520	1.788/4.720
32	30,000	0.173/0.244	0.460/0.972	0.952/2.306	1.649/4.276
64	35,000	0.147/0.220	0.417/0.901	0.888/2.212	1.588/4.255

The 64h checkpoint has lower ADE and FDE than the 1h checkpoint at every reported horizon. The improvement is visible not only in the shared three-second comparison but also at one, two, and four seconds. The sequence is not strictly monotonic at every intermediate training-data size: for example, ADE1 increases from 0.158 m at 16h to 0.173 m at 32h before decreasing to 0.147 m at 64h. These measurements reinforce the main interpretation that a local regression does not determine the overall response across the full evaluated range.

K Result Availability and Remaining Measurements

Table 20 separates measured quantities from unavailable quantities. Counts use the seven training-data sizes from 1 to 64 hours.

Table 20: Number of available results among the seven 1–64h training-data sizes. PhysicalAI denotes the paired ADE3/FDE3 result. NAVTEST ADE3/FDE3 was not collected for Uni-World VLA.

Planner	PhysicalAI	EPDMS pre/post	NAVTEST ADE3/FDE3 pre/post
Tiny-SSR	7/7	7/7 / 7/7	7/7 / 7/7
TPS-Drive	7/7	7/7 / 7/7	7/7 / 7/7
BevAD-lite	7/7	7/7 / 7/7	7/7 / 7/7
AD-MLP	7/7	7/7 / 7/7	7/7 / 7/7
Uni-World VLA	7/7	7/7 / 7/7	0/7 / 0/7

The remaining unavailable measurements are:

- Tiny-SSR 64h estimated PhysicalAI training FLOPs;
- Uni-World VLA pre- and post-adaptation NAVTEST ADE3/FDE3, which was not part of that planner's registered evaluation series.

The performance tables now provide complete five-planner 64h rankings for PhysicalAI ADE3/FDE3 and post-adaptation EPDMS. The remaining gaps affect only the 64h training-FLOPs comparison and the supplementary NAVTEST ADE3/FDE3 analysis.

K.1 Evidence receipt status

For Tiny-SSR, TPS-Drive, and BevAD-lite, the PhysicalAI result receipts are available in the project records. TPS-Drive and BevAD-lite also have local NAVTEST EPDMS and ADE3/FDE3 receipts. Tiny-SSR EPDMS values and AD-MLP NAVTEST values were verified on their evaluation hosts; the corresponding trace aggregates are not copied into this anonymous paper directory. The Tiny-SSR 64h values use checkpoint `iter_1053000_ema`; its post-adaptation EPDMS and trajectory metrics were produced by the same two-epoch NAVSIM v2 adaptation on an A40 GPU. AD-MLP PhysicalAI values and Uni-World VLA PhysicalAI and EPDMS values come from the authors' registered result summaries. For AD-MLP, every pre- and post-adaptation NAVTEST ADE3/FDE3 row contains 12,146 aligned prediction/ground-truth pairs. Uni-World VLA NAVTEST ADE3/FDE3 was not collected.

L QUALITATIVE FUTURE-FRAME PREDICTIONS

Figure 10 shows five qualitative predictions from Uni-World VLA trained on the 8-hour PhysicalAI subset. In every strip, columns show prediction horizons from 0.5 to 4.0 seconds at 0.5-second intervals. The three rows show the ground-truth future frames, tokenizer reconstructions, and autoregressively predicted future frames, respectively. The model receives the observed history, current frame, ego status, and depth features; it does not receive future ground-truth frames.

Across the examples, short-horizon predictions generally preserve the dominant road layout, horizon, and large scene regions. Differences become more visible at longer horizons, where autoregressive uncertainty produces blur, color changes, or loss of small objects. These panels are included as qualitative diagnostics of the future-frame objective; they are not used to calculate ADE3, FDE3, or EPDMS.

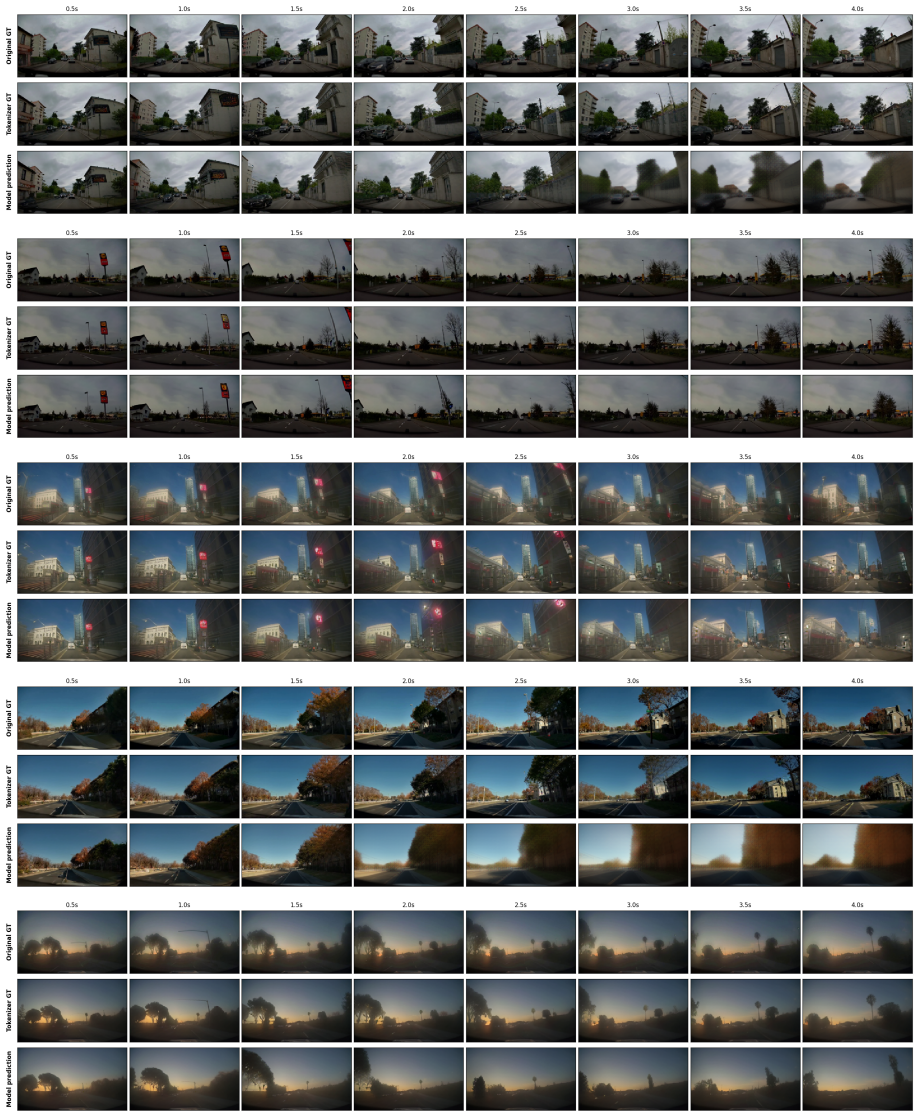


Figure 10: Qualitative future-frame predictions from Uni-World VLA trained on the 8-hour PhysicalAI subset (five selected samples).