MORE CAMERA FRAMES, BETTER DRIVING? WITHIN-SCENE INFORMATION SCALING IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous driving requires models to learn complex interactions among the ego vehicle, surrounding agents, and dynamic environments from continuously evolving driving processes. Existing work focuses on scaling model capacity and scaling training information exposure, largely treating driving scenes as the fundamental unit of training information. However, how the information exposed within each scene should scale with model capacity remains underexplored. We study *within-scene information scaling*, using temporal sampling frequency as a controlled probe to vary camera-frame density within fixed driving scenes. We formulate a minimal information–capacity model that captures the interaction among bounded driving-relevant information, increasing driving-irrelevant exposure, and finite model capacity. The model characterizes the capacity-dependent scaling behavior of within-scene information and predicts that the preferred temporal sampling frequency increases with model capacity under its assumptions. We conduct controlled frequency sweeps across compact end-to-end models, large vision-language-action (VLA) models, and world-action models on Waymo, nuScenes, and PAVE and in NAVSIM-based settings, with additional iteration-matched comparisons. AutoVLA and DriveVLA-W0 attain their lowest reported nuScenes trajectory errors at the highest tested frequency, yet further reductions above 4 Hz are much smaller for DriveVLA-W0. Drive-JEPA achieves strong driving performance on NAVTEST already at 2 Hz. These findings distinguish how much a model improves with denser temporal exposure from how well it already performs under sparse temporal exposure. They motivate evaluating driving performance alongside frequency-response curves and jointly developing model architectures, learning objectives, and within-scene data construction to utilize driving-relevant information that affects the driving decision while limiting the influence of driving-irrelevant exposure.

1 INTRODUCTION

Autonomous driving requires models to learn complex interactions among the ego vehicle, surrounding agents, and dynamic environments from continuously evolving driving processes (Chen et al., 2024; Hu et al., 2023). This requirement poses the core challenge for models to effectively extract and utilize information from large-scale driving scenes. Researchers identified two important research directions to address this challenge: scaling model capacity and scaling training information exposure (Zhou et al., 2025; Li et al., 2025b; Dimlioglu et al., 2026).

Existing works have explored these two scaling directions from complementary perspectives. On the model side, advances in spatiotemporal representation learning and unified driving architectures have improved how models represent and reason about driving scenes. BEVFormer develops spatiotemporal scene representations that aggregate observations across time (Li et al., 2022), while UniAD integrates perception, prediction, and planning within a unified framework (Hu et al., 2023). Recent vision-language-action and world-action models further extend this direction with larger-scale models and richer supervision (Zhou et al., 2025; Li et al., 2025b). On the information side, existing approaches mainly focus on expanding driving data coverage and improving scene-level data composition. For example, MOSAIC selects informative driving scene clips based on their estimated

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.

marginal gains (Dimlioglu et al., 2026). However, existing scaling strategies largely treat driving scenes as the fundamental unit of training information: they study which scenes should be included, how scenes should be represented, and how models should learn from these scenes, while leaving unexplored how the information contained within each individual scene should scale with model capacity.

In this work, we study a complementary scaling dimension: *within-scene information scaling*. Beyond scaling the number of driving scenes, we ask: *how should the information exposed within each scene scale with the model?* A driving scene is a continuous process. During data collection, this continuous process is discretized into camera frames with a certain temporal sampling frequency. Within-scene information scaling therefore corresponds to increasing the density of camera frames by increasing the temporal sampling frequency. These camera frames provide driving-relevant information, such as vehicle motion, agent interactions, and scene evolution, but also contain driving-irrelevant exposure (Feichtenhofer et al., 2019; Hu et al., 2023), including redundant content and visual variations unrelated to the driving decision (Selvam et al., 2025).

We hypothesize that increasing temporal sampling frequency changes the composition of information exposed to the model. At low temporal sampling frequency, sparse sampling can miss driving-relevant information such as temporal changes and interactions. As temporal sampling frequency increases, driving-relevant information increases but eventually approaches a finite limit, whereas driving-irrelevant exposure can continue increasing through redundant content and visual variations unrelated to the driving decision. Under finite model capacity, increasing driving-irrelevant exposure competes for capacity, so different models may utilize different amounts of within-scene information. We formulate a minimal information–capacity model that captures the interaction among driving-relevant information, driving-irrelevant exposure, and finite model capacity. The model characterizes a capacity-dependent scaling behavior of within-scene information. A concrete implication of this behavior is that the preferred temporal sampling frequency depends on model capacity. Figure 1 illustrates this capacity-dependent scaling behavior by relating temporal sampling frequency to the changing composition of within-scene information and the resulting frequency-response curves across model capacities.

To empirically study this capacity-dependent scaling behavior, we use temporal sampling frequency as a controlled probe of within-scene information scaling. Within each frequency sweep, we fix the scene amount in the training set, model architecture, input construction, prediction target, and evaluation protocol, varying only temporal sampling frequency. We evaluate frequency-response curves across compact end-to-end models, large VLA models, and world-action models on Waymo (Xu et al., 2025), nuScenes (Caesar et al., 2020), PAVE (Li et al., 2025a), and in NAVSIM-based open-loop and closed-loop settings. Iteration-matched comparisons further assess the effect of the number of optimization updates on these responses. AutoVLA and DriveVLA-W0 attain their lowest reported nuScenes ADE at the highest tested frequency, yet the further ADE reduction above 4 Hz is much smaller for DriveVLA-W0. Drive-JEPA already achieves low trajectory ADE/FDE and high NAVTEST EPDMS at 2 Hz. These findings distinguish how much a model improves with denser temporal exposure from how well it already performs under sparse temporal exposure. They motivate evaluating both driving performance and frequency-response curves, and jointly developing model architectures, learning objectives, and within-scene data construction to utilize driving-relevant information that affects the driving decision.

Our contributions are summarized as follows:

- A within-scene information scaling perspective for autonomous driving models. We study how the information exposed within each driving scene should scale with model capacity, identifying a scaling dimension complementary to model capacity scaling and scene-level data scaling.
- A minimal information–capacity model for within-scene information scaling. We characterize the capacity-dependent scaling behavior of within-scene information through the interaction among driving-relevant information, driving-irrelevant exposure, and finite model capacity. Under the model assumptions, the preferred temporal sampling frequency increases with model capacity.
- Controlled frequency-response evaluation across models and driving settings. Using temporal sampling frequency as a controlled probe, we distinguish driving performance under sparse temporal exposure from improvement under denser temporal exposure. These findings motivate using frequency-response curves to guide the joint development of model architectures, learning objectives, and within-scene data construction.

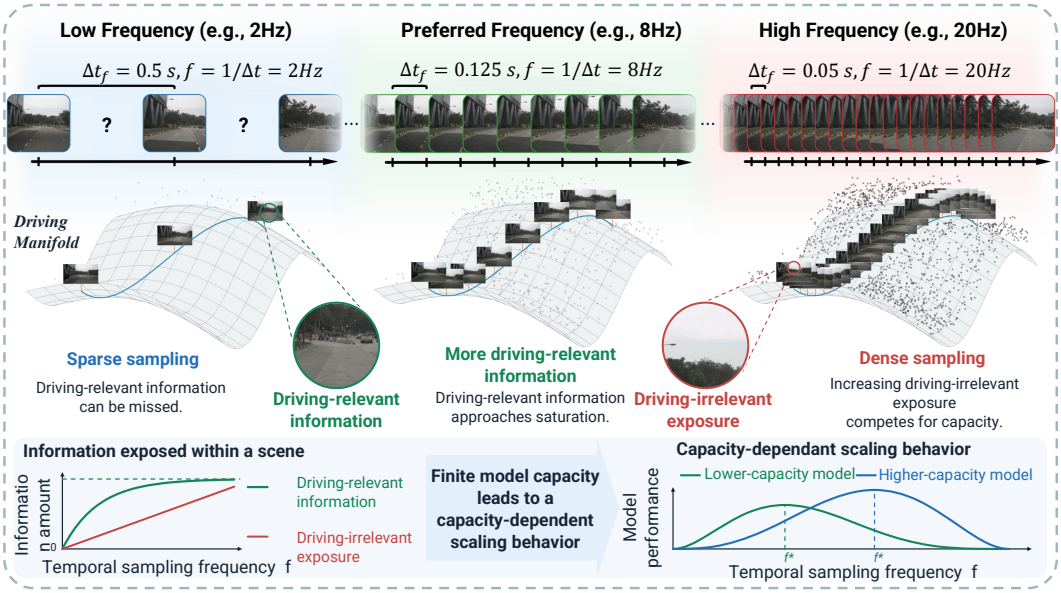

Figure 1: **Within-scene information scaling under finite model capacity.** Sparse temporal sampling can miss driving-relevant information. As temporal sampling frequency increases, driving-relevant information increases toward a finite limit, while driving-irrelevant exposure can continue increasing. Their interaction with finite model capacity shapes frequency-response curves, illustrating the capacity-dependent scaling behavior of within-scene information. Under the model assumptions, the preferred temporal sampling frequency increases with model capacity.

2 Related Work

Model capacity scaling and driving representations, architectures, and supervision. On the model-capacity-scaling side, existing works improve how driving models represent and reason about driving scenes through advances in spatiotemporal representations, unified driving architectures, and richer supervision. BEVFormer aggregates spatial and temporal observations into a bird's-eye-view representation (Li et al., 2022). UniAD combines perception, prediction, and planning in a joint system (Hu et al., 2023), while VAD and SparseDrive explore vectorized or sparse scene representations (Jiang et al., 2023; Sun et al., 2024). AutoVLA extends this direction through a vision-language-action formulation (Zhou et al., 2025). DriveVLA-W0 adds future-image prediction to action supervision and studies how world modeling affects data scaling (Li et al., 2025b).

Driving data coverage and scene-level data composition. Data scaling also depends on which driving scenes are included in training. MOSAIC estimates how clips from different driving domains contribute to evaluation metrics and allocates a collection budget accordingly (Dimlioglu et al., 2026). Such scene-level selection is complementary to within-scene information scaling: scene-level scaling changes which driving scenes are included, whereas within-scene scaling changes the density of camera frames within fixed driving scenes.

Temporal information and redundancy. Camera frames provide driving-relevant information, such as vehicle motion, agent interactions, and scene evolution (Hu et al., 2023; Feichtenhofer et al., 2019). SlowFast uses pathways with different temporal rates for video recognition (Feichtenhofer et al., 2019). EVE avoids redundant computation across point-cloud frames (Peng et al., 2021), and SimCache reuses VLM computations for similar scene regions (Selvam et al., 2025). These methods primarily study temporal representation or computational reuse rather than how the composition of information exposed within a fixed driving scene changes with temporal sampling frequency.

Within-scene information scaling and finite model capacity. Scaling-law studies relate performance to model size, data volume, and training compute (Kaplan et al., 2020; Hoffmann et al., 2022). The Back to Basics study provides a different perspective by examining how learning data

structure and high-dimensional noise place different demands on a model (Li & He, 2025). Motivated by these perspectives, our formulation models within-scene information scaling as an interaction among driving-relevant information, driving-irrelevant exposure, and finite model capacity, yielding a capacity-dependent preferred temporal sampling frequency.

Driving evaluation. Trajectory displacement errors measure agreement with recorded futures, but they do not capture every aspect of driving performance. NAVSIM closed-loop evaluations score predicted plans with simulation-based metrics that capture aspects of driving beyond displacement errors (Dauner et al., 2024; Cao et al., 2025). We therefore report trajectory ADE/FDE together with NAVSIM EPDMS in the corresponding evaluation settings.

3 Within-Scene Information Scaling under Finite Model Capacity

As temporal sampling frequency $f > 0$ increases, driving-relevant information $\mathcal{I}(f)$ approaches a finite limit, whereas driving-irrelevant exposure $\mathcal{B}(f)$ can continue increasing. The interval between sampled camera frames within each driving scene is $\Delta t = 1/f$. Within each frequency sweep, the scene amount in the training set, model architecture, input construction, prediction target, and evaluation protocol remain fixed. We model this distinction as

$$\mathcal{I}(f) = \Omega \frac{(f/f_0)^p}{1 + (f/f_0)^p}, \qquad \mathcal{B}(f) \approx \nu f, \tag{1}$$

where $\Omega, f_0, \nu > 0$ and $p > 1$. Here, Ω is the finite limit of driving-relevant information, half of which is exposed at f_0; p controls the increase under sparse temporal sampling. Driving-irrelevant exposure includes redundant content and visual variations unrelated to the driving decision, with ν setting its linear growth rate. We use this linear approximation for the analysis. Both quantities are defined on a common normalized scale, not measured as Shannon information.

Let $\mathcal{R}(f) := \mathcal{I}(f) + \mathcal{B}(f)$ denote the combined modeled exposure and $C > 0$ denote finite model capacity on the same normalized scale. We model the driving-relevant information effectively utilized through the normalized measure

$$\alpha_C(f) := \widetilde{\mathcal{I}}_C(f) = \frac{\mathcal{I}(f)}{\mathcal{R}(f)} \frac{C}{C + \mathcal{R}(f)} = \frac{C\,\mathcal{I}(f)}{\mathcal{R}(f)\big(C + \mathcal{R}(f)\big)}. \tag{2}$$

The first factor measures the driving-relevant share of the combined modeled exposure. The second models the fraction that can be processed: it decreases with exposure and increases with model capacity. At fixed C, we assume that driving performance $\mathcal{P}_C(f)$, oriented so that higher is better, is a strictly increasing function of $\alpha_C(f)$. The preferred temporal sampling frequency therefore satisfies

$$f_C^\star \in \arg\max_{f>0} \mathcal{P}_C(f) = \arg\max_{f>0} \alpha_C(f). \tag{3}$$

For $p > 1$, $\alpha_C(f)$ approaches zero at both frequency extremes: sparse temporal sampling misses driving-relevant information, while very dense temporal sampling adds driving-irrelevant exposure after driving-relevant information has nearly saturated. The model therefore has a finite interior preferred temporal sampling frequency. Holding the fixed driving scenes and the parameters Ω, f_0, p, ν unchanged, a nondegenerate interior optimum that varies smoothly with C satisfies

$$\frac{\mathrm{d}f_C^\star}{\mathrm{d}C} > 0. \tag{4}$$

Thus, greater model capacity allows denser temporal exposure to be processed, shifting the preferred temporal sampling frequency upward under the model assumptions. On a finite tested frequency set, this frequency can remain unchanged or increase. This shift is one consequence of the capacity-dependent scaling behavior of within-scene information. Appendix B provides the derivation.

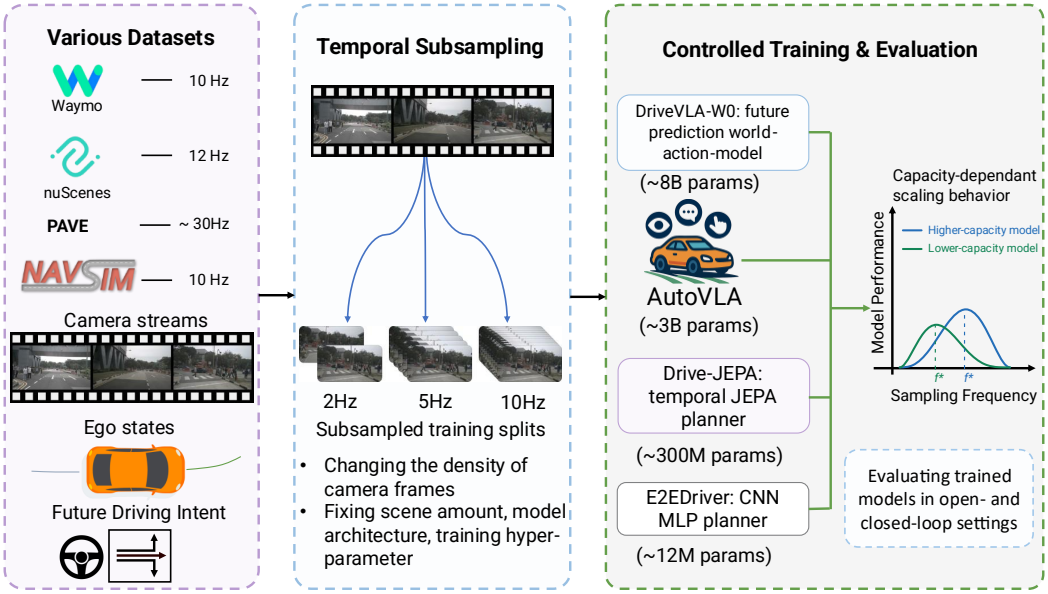

Figure 2: **Temporal sampling frequency as a controlled probe of within-scene information scaling.** Training sets contain different densities of sampled camera frames from the same driving scenes, while model architecture, input construction, prediction target, and evaluation protocol remain fixed within each frequency sweep. The resulting frequency-response curves characterize how different models utilize increasingly dense temporal exposure.

4 Experiments

We use temporal sampling frequency as a controlled probe of within-scene information scaling across compact end-to-end models, large VLA models, and world-action models. We analyze frequency-response curves to examine both driving performance under sparse temporal exposure and its changes as temporal sampling becomes denser. Iteration-matched comparisons assess the effect of the number of optimization updates on these responses. Figure 2 illustrates the training-set construction for the open-loop and NAVSIM experiments. Dataset-handling details are provided in Appendix D.

Task and metrics. We study command-conditioned ego-trajectory prediction (Codevilla et al., 2018). At timestamp t, each model receives model-specific observations and predicts a future ego trajectory. The open-loop sweeps report 3-second ADE and FDE in meters (Ettinger et al., 2021), with 5-second results where available. The NAVSIM closed-loop evaluations report EPDMS as a percentage (Dauner et al., 2024; Cao et al., 2025) alongside 3-second ADE/FDE. On NAVTEST, ADE averages positional error at six targets from 0.5 to 3.0 seconds, and FDE uses the 3.0-second target. EPDMS includes all 12,146 scenarios; ADE/FDE use valid trajectories only. Higher EPDMS and lower ADE/FDE indicate better driving performance under their respective evaluation protocols.

Datasets. The open-loop sweeps use Waymo (Xu et al., 2025), nuScenes (Caesar et al., 2020), and PAVE (Li et al., 2025a), with highest evaluated temporal sampling frequencies of 10 Hz, 12 Hz, and 20 Hz, respectively. The PAVE sweep stops below its approximately 30 Hz native camera frequency. Table 1 summarizes the frequency ranges, driving domains, and training-set statistics. NAVSIM-10Hz uses native camera frames from the corresponding nuPlan recordings to construct a 10 Hz version of the official 2 Hz NAVSIM training clips within the same driving scenes (Caesar et al., 2021; Dauner et al., 2024).

Models and training. The open-loop sweeps evaluate E2EDriver, BEV-E2EDriver, Tiny-SSR, and AutoVLA. The three compact end-to-end models have approximately 10–18M parameters and use CNN-based multi-view image encoding, a bird's-eye-view representation, and a compact sparse-scene representation, respectively. AutoVLA uses a roughly 3B-parameter vision-language-

Table 1: Dataset summary and training-set motion statistics. Mean ego speed and displacement per sampled camera frame are computed at each dataset's highest evaluated temporal sampling frequency. NAVSIM-10Hz uses native nuPlan camera frames within the NAVTRAIN driving scenes, not upsampled OpenScene images.

	Waymo	**nuScenes**	**PAVE**	**NAVSIM-10Hz**
Camera-frame frequency	10 Hz	12 Hz	∼30 Hz	10 Hz
Highest evaluated f	10 Hz	12 Hz	20 Hz	10 Hz
Cameras	8	6	8	8
Dominant scenario	Urban/suburban	Urban	Highway/expressway	Urban/suburban
Training samples	∼365k	∼116k	∼167k	∼425k
Mean ego speed (m/s)	5.80	5.05	13.75	4.67
Displacement/frame (m)	0.58	0.42	0.69	0.46
Source	Public	Public	Public	Public

action backbone to predict trajectory action tokens (Zhou et al., 2025). DriveVLA-W0 (Li et al., 2025b), which uses world-model supervision, is additionally evaluated on nuScenes. The NAVTEST evaluations compare Drive-JEPA (Wang et al., 2026), a temporal JEPA planner, with E2EDriver, AutoVLA, and DriveVLA-W0. Within each frequency sweep, model architecture, input construction (including the command representation), and prediction target remain fixed. Detailed model and loss descriptions for the open-loop sweeps are provided in Appendix C. The open-loop sweeps keep the epoch count fixed within each model–dataset pair. Increasing temporal sampling frequency therefore increases the number of sampled camera frames per epoch and the number of optimization updates, motivating the iteration-matched comparisons below.

Open-loop frequency-response curves depend on model and dataset. Table 2 reports the 3-second open-loop sweeps, and Figure 3 summarizes the corresponding ADE frequency-response curves. Compact end-to-end models often show intermediate minima or near-plateau regions, whereas AutoVLA reaches its lowest reported ADE at the highest tested temporal sampling frequency on each dataset. This contrast shows that increasing the density of sampled camera frames does not produce uniform improvements across models. Importantly, the improvement from denser temporal exposure and the level of driving performance are distinct: Tiny-SSR at 2 Hz already has lower ADE than AutoVLA at its highest tested frequency on all three datasets. A higher preferred temporal sampling frequency within the tested range is therefore not, by itself, evidence of better driving performance. The repeated runs also distinguish different frequency-response shapes: E2EDriver has a pronounced intermediate ADE minimum on nuScenes but a near-plateau region on Waymo. Near such a plateau, the range of temporal sampling frequencies with similar performance is more informative than the frequency with the lowest reported mean. These observations motivate studying within-scene information scaling jointly with model architecture and the driving scenes, using frequency-response curves to assess performance under sparse temporal exposure, changes under denser temporal exposure, and the reported run-to-run variation.

Iteration-matched controls. We assess whether the frequency-response curves can be explained by the number of optimization updates alone. We pair lower temporal sampling frequencies with more epochs and higher frequencies with fewer epochs, using frequency $\times$ epochs to approximately match the number of optimization updates. The comparisons in Appendix A and Table 5 show that lower-frequency configurations retain comparable or lower ADE after this adjustment. For E2EDriver, 6 Hz for 10 epochs has 12.05% lower ADE than 12 Hz for 5 epochs on nuScenes, whereas the analogous difference on Waymo is only 0.12%. Increasing epochs revisits sampled camera frames, whereas increasing temporal sampling frequency changes within-scene information exposure. The results support treating within-scene information scaling and the number of optimization updates as distinct training choices: denser temporal exposure does not necessarily improve driving performance even when the update counts are approximately matched.

5-second results. The 5-second evaluations in Appendix E retain the model- and dataset-dependent frequency-response patterns, while the preferred temporal sampling frequency can change with the

Table 2: **Three-second open-loop trajectory prediction across temporal sampling frequencies.** ADE/FDE are in meters. ADE is reported as mean ± standard deviation over three runs where available. Bold marks the lowest mean ADE for each model–dataset pair and the FDE at that setting. Some high-frequency results are grouped into ranges.

		Waymo (10 Hz)		**nuScenes** (12 Hz)		**PAVE** (20 Hz eval.)	
Model	**Freq.**	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓
E2EDriver	2 Hz	0.564±.002	1.584	1.188±.004	3.082	1.594±.216	4.123
	4 Hz	0.538±.001	1.527	1.206±.011	3.150	1.384±.093	3.716
	6 Hz	0.526±.002	1.500	**1.099**±.006	**2.856**	1.145±.016	2.523
	8 Hz	**0.517**±.002	**1.468**	1.200±.005	3.134	1.109±.031	2.478
	10 Hz	0.518±.003	1.469	1.195±.002	3.114	**1.049**±.007	**2.397**
	12 Hz	—	—	1.183±.007	3.093	1.223±.208	2.666
	15 Hz	—	—	—	—	1.118±.039	2.437
	18–20 Hz	—	—	—	—	1.083–1.105	2.374–2.478
BEV-E2EDriver	2 Hz	0.531±.003	1.485	1.214±.012	3.137	1.527±.185	3.952
	4 Hz	0.524±.003	1.479	1.208±.012	3.121	1.378±.147	3.629
	6 Hz	0.523±.003	1.491	**1.117**±.016	**2.875**	1.175±.010	2.588
	8 Hz	**0.518**±.001	**1.464**	1.213±.006	3.126	1.098±.003	2.491
	10 Hz	0.519±.003	1.464	1.191±.031	3.105	**1.075**±.014	**2.422**
	12 Hz	—	—	1.194±.004	3.094	1.094±.036	2.478
	15–20 Hz	—	—	—	—	1.081–1.097	2.423–2.435
Tiny-SSR	2 Hz	0.581±.008	1.443	0.628±.001	1.465	0.682±.011	1.587
	4 Hz	0.543±.003	1.348	0.621±.002	1.459	0.575±.005	1.412
	6 Hz	0.542±.001	1.347	0.618±.002	1.455	0.551±.004	1.159
	8 Hz	**0.538**±.002	**1.334**	**0.608**±.001	**1.440**	0.520±.005	1.049
	10 Hz	0.540±.002	1.340	0.611±.003	1.447	0.518±.003	1.049
	12 Hz	—	—	0.611±.003	1.448	0.516±.005	1.052
	15 Hz	—	—	—	—	**0.515**±.003	**1.047**
	18–20 Hz	—	—	—	—	0.518±.002	1.062
AutoVLA	2 Hz	1.282±.120	2.881	1.503±.227	3.331	1.157±.040	2.450
	4 Hz	1.201±.020	2.755	1.024±.016	2.398	1.062±.022	2.200
	6 Hz	1.178±.018	2.684	0.924±.007	2.184	0.961±.015	2.016
	8 Hz	1.162±.019	2.718	0.930±.034	2.176	0.949±.005	1.973
	10 Hz	**1.146**±.019	**2.635**	0.868±.021	2.040	0.942±.002	1.930
	12 Hz	—	—	**0.855**±.008	**2.029**	0.932±.025	1.907
	15 Hz	—	—	—	—	0.899±.004	1.844
	18 Hz	—	—	—	—	0.944±.019	1.941
	20 Hz	—	—	—	—	**0.875**±.004	**1.823**

prediction target. This motivates evaluating within-scene information scaling for the prediction target rather than assigning each model a fixed preferred temporal sampling frequency.

DriveVLA-W0 frequency response on nuScenes. Table 3 shows that most of DriveVLA-W0's ADE reduction occurs between 2 Hz and 4 Hz. Increasing temporal sampling frequency from 4 Hz to 12 Hz further reduces its mean ADE by approximately 1.6%, compared with 16.5% for AutoVLA on nuScenes (Table 2). Although both models reach their lowest reported ADE at 12 Hz, the extent of improvement from denser temporal exposure differs substantially. The frequency-response curve is therefore more informative for within-scene data construction than the best tested frequency alone. These results motivate studying how temporal sampling can expose additional driving-relevant information that affects the driving decision, rather than assuming that VLA and world-action models all require uniformly dense temporal exposure.

NAVTEST frequency-response curves and adaptive temporal sampling. Table 4 reports NAVSIM closed-loop evaluations on NAVTEST over three random seeds per setting. Drive-JEPA and

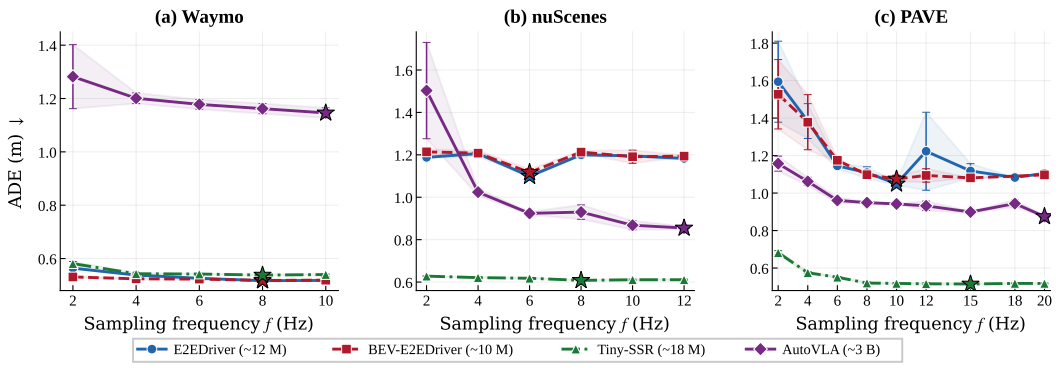

Figure 3: **Open-loop frequency-response curves across models and datasets.** Compact end-to-end models show intermediate ADE minima or near-plateau regions, whereas AutoVLA reaches its lowest reported ADE at the highest tested temporal sampling frequency on each dataset.

Table 3: **DriveVLA-W0 frequency response on nuScenes.** ADE/FDE are in meters; lower is better. Bold marks the lowest reported mean for each metric.

Training frequency	ADE (m) ↓	FDE (m) ↓
2 Hz	1.321±0.002	2.964±0.003
4 Hz	1.256±0.001	2.915±0.005
6 Hz	1.265±0.000	2.919±0.002
8 Hz	1.258±0.002	2.922±0.004
10 Hz	1.242±0.003	2.872±0.006
12 Hz	**1.236**±0.004	**2.856**±0.006

E2EDriver use uniform temporal sampling at 2, 4, 6, 8, and 10 Hz; AutoVLA and DriveVLA-W0 are evaluated with uniform 2 Hz, motion-based within-scene adaptive 4 Hz, and uniform 10 Hz training. The adaptive setting uses motion analysis to vary the sampling frequency over time within each scene, sampling more densely during highly dynamic periods and more sparsely when motion changes slowly. We distinguish the uniform frequency comparisons from adaptive temporal sampling, which also changes how camera frames are sampled. Drive-JEPA at 2 Hz already achieves higher mean EPDMS and lower mean ADE/FDE than AutoVLA and DriveVLA-W0 at uniform 10 Hz. It maintains high EPDMS throughout its sweep, while its lowest mean trajectory errors occur at 2 Hz. This contrast distinguishes achieving strong driving performance under sparse temporal exposure from improving performance through denser temporal exposure. It motivates evaluating model architectures and learning objectives by both the performance attained at lower temporal sampling frequencies and their response to additional exposure, rather than treating dependence on high-frequency training as an indicator of model capability.

The adaptive comparisons show why within-scene information scaling should not be assessed by temporal sampling frequency or trajectory errors alone. For DriveVLA-W0, adaptive 4 Hz yields higher mean EPDMS than uniform 10 Hz, whereas uniform 10 Hz yields lower ADE/FDE. For AutoVLA, the two settings have similar trajectory errors, yet uniform 10 Hz yields higher mean EPDMS. These contrasting results motivate studying which camera frames are sampled alongside how densely each driving scene is sampled, with the choice evaluated for each model and driving-performance metric. The objective is to expose driving-relevant information that affects the driving decision, rather than simply increasing camera-frame density. Frequency-response curves should therefore be assessed using both trajectory errors and NAVSIM metrics, with the preferred temporal sampling frequency interpreted for the chosen metric and evaluation protocol.

Discussion. The experiments motivate treating within-scene information scaling as a joint problem of model architecture, learning objectives, and data construction. For VLA and world-action models, the need for denser temporal exposure should be evaluated through frequency-response curves rather than inferred from model capacity alone. Drive-JEPA's strong performance under sparse temporal

Table 4: **NAVTEST evaluation with uniform and adaptive temporal sampling.** Every metric is reported as mean ± sample standard deviation over three random seeds, rounded to three decimal places. EPDMS is in percent and includes all 12,146 NAVTEST scenarios; 3-second ADE/FDE are in meters and use valid trajectories. A4 denotes adaptive 4 Hz training; all other settings use uniform temporal sampling. Bold marks the best reported mean within each model and metric.

Model	Training frequency	EPDMS (%) ↑	3-s ADE (m) ↓	3-s FDE (m) ↓
Drive-JEPA	2 Hz	89.554±0.085	**0.350**±0.002	**0.811**±0.010
	4 Hz	**89.958**±0.068	0.389±0.003	0.878±0.005
	6 Hz	89.548±0.062	0.370±0.005	0.850±0.004
	8 Hz	89.608±0.016	0.378±0.001	0.881±0.004
	10 Hz	89.597±0.119	0.386±0.002	0.883±0.006
DriveVLA-W0	2 Hz	85.673±0.078	0.621±0.001	1.236±0.003
	A4	**86.663**±0.127	0.467±0.001	1.016±0.002
	10 Hz	86.409±0.132	**0.422**±0.007	**0.947**±0.007
AutoVLA	2 Hz	78.793±0.062	0.506±0.000	1.150±0.001
	A4	80.319±0.022	**0.484**±0.000	**1.102**±0.001
	10 Hz	**81.135**±0.085	0.485±0.000	1.103±0.001
E2EDriver	2 Hz	76.888±0.298	0.395±0.014	0.919±0.013
	4 Hz	79.138±0.308	0.382±0.006	0.893±0.005
	6 Hz	80.079±0.570	0.391±0.024	0.898±0.030
	8 Hz	**81.558**±0.078	**0.371**±0.001	**0.876**±0.006
	10 Hz	81.343±0.059	0.373±0.008	0.879±0.007

exposure further motivates developing representations and learning objectives that effectively utilize the driving-relevant information already exposed at lower temporal sampling frequencies. When increasing temporal sampling frequency produces little improvement, the next design question is whether to improve the selection of sampled camera frames, the model's ability to utilize driving-relevant information, or both. The minimal information–capacity model provides a framework for studying the capacity-dependent scaling behavior of within-scene information through the interaction among driving-relevant information, driving-irrelevant exposure, and finite model capacity. Adaptive temporal sampling offers a complementary direction for data construction: studying which camera frames to expose, not only how densely to sample, with choices evaluated for each model and driving-performance metric. The objective is to achieve strong driving performance when sparse temporal exposure contains sufficient driving-relevant information, exploit additional driving-relevant information when denser temporal sampling reveals it, and limit the influence of driving-irrelevant exposure. This objective does not require flat frequency-response curves or a preference for low or high temporal sampling frequencies. Instead, frequency-response curves should guide model and data development toward learning the driving-relevant information that affects the driving decision.

5 Conclusion

We study within-scene information scaling in autonomous driving using temporal sampling frequency as a controlled probe. Our minimal information–capacity model characterizes the capacity-dependent scaling behavior of within-scene information through the interaction among driving-relevant information, driving-irrelevant exposure, and finite model capacity. The VLA and world-action model comparisons show that the best tested frequency alone does not capture the improvement from denser temporal exposure, while Drive-JEPA achieves strong driving performance under sparse temporal exposure. These findings distinguish how much a model improves with denser temporal exposure from how well it already performs under sparse temporal exposure. Model evaluation should therefore consider both driving performance and frequency-response curves, rather than only the preferred temporal sampling frequency. The adaptive temporal sampling results further motivate studying which camera frames are exposed, not only their density. Together, these findings motivate joint development of model architectures, learning objectives, and within-scene data construction. The objective is to utilize driving-relevant information that affects the driving decision while limiting the influence of driving-irrelevant exposure, rather than favoring low or high temporal sampling frequencies.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used to assist with grammar checking and text polishing, LaTeX formatting. All AI-assisted changes were reviewed by the authors for technical accuracy and consistency with the reported work. The authors take responsibility for the final content of the manuscript, including text and artifacts produced with AI assistance.

REFERENCES

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 *CVPR*, pp. 11621–11631, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

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. URL https://arxiv.org/abs/2506.04218.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024. doi: 10.1109/TPAMI.2024.3435937.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 4693–4700. IEEE, 2018.

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, 2024. URL https://arxiv.org/abs/2406.15349.

Tolga Dimlioglu, Nadine Chang, Maying Shen, Rafid Mahmood, and Jose M. Alvarez. Scaling-aware data selection for end-to-end autonomous driving systems, 2026. URL https://arxiv.org/abs/2604.08366.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, pp. 9710–9719, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ettinger_Large_Scale_Interactive_Motion_Forecasting_for_Autonomous_Driving_The_Waymo_ICCV_2021_paper.html.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556.

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 *CVPR*, 2023.

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. In *ICCV*, 2023.

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. doi: 10.48550/arXiv.2001.08361.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *ICLR*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Xiangyu Li, Chen Wang, Yumao Liu, Dengbo He, Jiahao Zhang, and Ke Ma. Pave: An end-to-end dataset for production autonomous vehicle evaluation. *arXiv preprint arXiv:2511.14185*, 2025a.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025b. URL https://arxiv.org/abs/2510.12796.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html.

Adeel Mufti. e2edriver. https://github.com/AdeelMufti/e2edriver, 2026. GitHub repository. Accessed: 2026-03-04.

Gao Peng, Bo Pang, and Cewu Lu. Efficient 3d video engine using frame redundancy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3792–3802, 2021.

Surya Selvam, Ravi K. Rajendran, Murugan Sankaradas, Anand Raghunathan, and Srimat T. Chakradhar. Simcache: Similarity caching for efficient VLM-based scene understanding. In *CVPRW*, 2025.

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.

Linhan Wang, Zichong Yang, Chen Bai, Guoxiang Zhang, Xiaotong Liu, Xiaoyin Zheng, Xiao-Xiao Long, Chang-Tien Lu, and Cheng Lu. Drive-jepa: Video jepa meets multimodal trajectory distillation for end-to-end driving. *arXiv preprint arXiv:2601.22032*, 2026. URL https://arxiv.org/abs/2601.22032.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina 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. *arXiv preprint arXiv:2510.26125*, 2025. URL https://arxiv.org/abs/2510.26125.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025.

A ITERATION-MATCHED CONTROLS

Table 5: **Lower-frequency configurations retain comparable or lower ADE after approximate training-volume matching.** Three-second ADE/FDE are in meters. Each configuration lists sampling frequency × epochs, followed by ADE/FDE. Training volume is approximated by frequency × epochs; Δ is the ADE change relative to the higher-frequency setting, computed before rounding.

Model	Dataset	Lower config; ADE/FDE	Higher config; ADE/FDE	Δ ADE
E2EDriver	Waymo	6 × 8; **0.518 / 1.473**	10 × 5; 0.519 / 1.474	−0.12%
E2EDriver	nuScenes	6 × 10; **1.042 / 2.709**	12 × 5; 1.185 / 3.096	−12.05%
E2EDriver	PAVE	10 × 10; **1.022 / 2.252**	20 × 5; 1.083 / 2.374	−5.65%
BEV-E2EDriver	Waymo	6 × 8; **0.517 / 1.465**	10 × 5; 0.519 / 1.467	−0.39%
BEV-E2EDriver	nuScenes	6 × 10; **1.074 / 2.759**	12 × 5; 1.189 / 3.098	−9.67%
BEV-E2EDriver	PAVE	10 × 10; **1.052 / 2.416**	20 × 5; 1.081 / 2.423	−2.71%
Tiny-SSR	Waymo	8 × 10; **0.538 / 1.338**	10 × 8; 0.559 / 1.399	−3.76%
Tiny-SSR	nuScenes	10 × 12; **0.665 / 1.604**	12 × 10; 0.726 / 1.706	−8.40%
Tiny-SSR	PAVE	15 × 16; **0.528 / 1.106**	20 × 12; 0.539 / 1.144	−2.04%

B DERIVATION OF THE CAPACITY-DEPENDENT PREFERRED TEMPORAL SAMPLING FREQUENCY

This section restates the information–capacity model and gives the derivation of the result stated in Section 3.

Temporal sampling exposes bounded driving-relevant information and increasing driving-irrelevant exposure. As described in Section 1, temporal sampling frequency $f > 0$ denotes the rate at which camera frames are sampled from each driving scene. The interval between sampled camera frames is $\Delta t = 1/f$. Within each frequency sweep, the scene amount in training set, model architecture, input construction, prediction target, and evaluation protocol remain fixed, while only the temporal sampling frequency varies. This provides a controlled setting for studying how different models utilize increasingly dense temporal exposure.

Let $\mathcal{I}(f)$ denote the driving-relevant information exposed at frequency f. Sparsely sampled camera frames may miss ego motion, agent interactions, and scene evolution, whereas densely sampled camera frames increasingly revisit information already contained within the same driving scenes. Let $\mathcal{B}(f)$ denote the driving-irrelevant exposure at frequency f. For $\Omega > 0$, $f_0 > 0$, $p > 1$, and $\nu > 0$, we model $\mathcal{I}(f)$ and $\mathcal{B}(f)$ as follows:

$$\mathcal{I}(f) = \Omega \frac{(f/f_0)^p}{1 + (f/f_0)^p} \qquad \mathcal{B}(f) \approx \nu f. \tag{5}$$

Here, Ω is the finite limit of driving-relevant information within the fixed driving scenes for the prediction target, f_0 is the characteristic temporal sampling frequency governing saturation, and p controls how rapidly $\mathcal{I}(f)$ changes under sparse temporal sampling. Thus, $\mathcal{I}(f) \to 0$ as $f \to 0^+$ and $\mathcal{I}(f) \to \Omega$ as $f \to \infty$. As temporal sampling becomes denser, sampled camera frames increasingly repeat driving-relevant information already exposed by nearby frames and introduce visual variations unrelated to the driving decision. Following the terminology in Section 1, we group this redundant content and these unrelated visual variations under *driving-irrelevant exposure*, with $\mathcal{B}(f)$ growing approximately linearly over the tested range. Thus, driving-relevant information is bounded within the fixed driving scenes, whereas driving-irrelevant exposure can continue to increase with denser temporal sampling. We treat $\mathcal{I}(f)$ and $\mathcal{B}(f)$ as phenomenological quantities on a common normalized scale, rather than direct measurements of Shannon information, so that they are comparable and can be combined within the model.

Finite model capacity determines how much driving-relevant information can be utilized. Let $\mathcal{R}(f) := \mathcal{I}(f) + \mathcal{B}(f)$ denote the combined modeled exposure at frequency f. The fraction of this exposure that is driving-relevant is $\mathcal{I}(f)/\mathcal{R}(f)$. We represent finite model capacity by a positive

normalized parameter C that is commensurate with $\mathcal{R}(f)$ within the phenomenological model, and model the fraction of the combined exposure that can be effectively processed as $C/[C + \mathcal{R}(f)]$. Their product defines a normalized measure of the driving-relevant information effectively utilized by the model:

$$\alpha_C(f) := \widetilde{\mathcal{I}}_C(f) = \frac{\mathcal{I}(f)}{\mathcal{R}(f)} \frac{C}{C + \mathcal{R}(f)} = \frac{C\,\mathcal{I}(f)}{\mathcal{R}(f)\big(C + \mathcal{R}(f)\big)}. \tag{6}$$

The first factor measures the driving-relevant share of the combined modeled exposure. The second factor specifies how finite model capacity limits the fraction of that exposure that can be effectively processed. Their product is therefore a dimensionless normalized measure of how much of the exposed driving-relevant information can be utilized by a model of capacity C in the presence of driving-irrelevant exposure.

The endpoint limits also ensure that a finite preferred temporal sampling frequency exists. As $f \to 0^+$, $\mathcal{I}(f) \sim \Omega f^p / f_0^p$ and $\mathcal{R}(f) \sim \nu f$, so

$$\alpha_C(f) \sim \frac{\Omega}{\nu f_0^p} f^{p-1} \longrightarrow 0. \tag{7}$$

As $f \to \infty$, $\mathcal{I}(f) \to \Omega$ and $\mathcal{R}(f) \sim \nu f$, giving

$$\alpha_C(f) \sim \frac{C\Omega}{\nu^2 f^2} \longrightarrow 0. \tag{8}$$

Since $\alpha_C(f) > 0$ and is continuous for finite $f > 0$, it attains a maximum at a finite positive frequency. For the stationary calculation, the information derivative is

$$\mathcal{I}'(f) = \frac{\Omega p f_0^p f^{p-1}}{(f_0^p + f^p)^2} > 0. \tag{9}$$

Driving performance increases as more driving-relevant information is effectively utilized. Let $\mathcal{P}_C(f)$ denote a driving-performance metric at frequency f. It may be an open-loop trajectory score or a closed-loop driving metric. We orient $\mathcal{P}_C(f)$ so that larger values mean better performance; an error metric can be negated or transformed monotonically. We assume that, for fixed model capacity, driving performance is monotonically increasing in the normalized measure $\widetilde{\mathcal{I}}_C(f)$:

$$\frac{\partial \mathcal{P}_C(f)}{\partial \widetilde{\mathcal{I}}_C(f)} > 0. \tag{10}$$

Under this monotonicity assumption, maximizing driving performance over temporal sampling frequency is equivalent to maximizing $\widetilde{\mathcal{I}}_C(f)$:

$$f_C^\star \in \arg\max_{f>0} \mathcal{P}_C(f) = \arg\max_{f>0} \widetilde{\mathcal{I}}_C(f). \tag{11}$$

A capacity-dependent preferred temporal sampling frequency. For the derivative, write $I(f) := \mathcal{I}(f)$ and $R(f) := I(f) + \nu f$, so $R'(f) = I'(f) + \nu$. Applying the quotient rule to Eq. equation 6 gives

$$\frac{\mathrm{d}\widetilde{\mathcal{I}}_C(f)}{\mathrm{d}f} = C\,\frac{I'(f)R(f)\big(C + R(f)\big) - I(f)R'(f)\big(C + 2R(f)\big)}{\big[R(f)\big(C + R(f)\big)\big]^2}. \tag{12}$$

At an interior preferred temporal sampling frequency, the derivative vanishes. Hence,

$$
\begin{aligned}
0 = \left.\frac{\mathrm{d}\widetilde{\mathcal{I}}_C(f)}{\mathrm{d}f}\right|_{f=f_C^\star} \\
\iff \quad & I'(f_C^\star)R(f_C^\star)\big(C + R(f_C^\star)\big) \\
& = I(f_C^\star)R'(f_C^\star)\big(C + 2R(f_C^\star)\big).
\end{aligned} \tag{13}
$$

Dividing by the positive factors yields

$$
\frac{I'(f_C^\star)}{I(f_C^\star)} = R'(f_C^\star)\left[\frac{1}{R(f_C^\star)} + \frac{1}{C + R(f_C^\star)}\right]. \tag{14}
$$

At a stationary point, the relative increase in driving-relevant information balances the combined effect of increasing modeled exposure in the two denominator factors associated with exposure composition and finite model capacity.

To determine how model capacity changes the preferred temporal sampling frequency, define $G(f, C) := \frac{\partial}{\partial f}\log\widetilde{\mathcal{I}}_C(f)$. Because $R'(f) = I'(f) + \nu > 0$, the capacity derivative of G is

$$
\frac{\partial G(f,C)}{\partial C} = \frac{R'(f)}{\big(C + R(f)\big)^2} > 0. \tag{15}
$$

Suppose that $f_C^\star$ is a nondegenerate interior maximum and changes smoothly with C. Then $G(f_C^\star, C) = 0$ and $\partial G/\partial f < 0$. Implicit differentiation gives

$$
\frac{\mathrm{d}f_C^\star}{\mathrm{d}C} = -\frac{\partial G/\partial C}{\partial G/\partial f} > 0. \tag{16}
$$

Holding the fixed driving scenes and the parameters governing driving-relevant information and driving-irrelevant exposure unchanged, increasing model capacity raises the capacity-dependent preferred temporal sampling frequency under these assumptions. On a finite tested frequency set, the preferred temporal sampling frequency can therefore remain unchanged or shift to a higher tested frequency as model capacity increases. This ordering provides a testable signature of the modeled information–capacity interaction.

For completeness, the discrete result does not require differentiating an optimum. For two tested frequencies $f_2 > f_1$, the modeled exposure satisfies $R(f_2) > R(f_1)$. Differentiating their log-effective-information difference gives

$$
\frac{\mathrm{d}}{\mathrm{d}C}\Big[\log\widetilde{\mathcal{I}}_C(f_2) - \log\widetilde{\mathcal{I}}_C(f_1)\Big] = \frac{1}{C + R(f_1)} - \frac{1}{C + R(f_2)} > 0. \tag{17}
$$

Thus, increasing C favors the higher member of every frequency pair. When $\mathcal{I}$ and $\mathcal{B}$ remain fixed across the capacity comparison, a maximizer on a finite tested set cannot move to a lower frequency as capacity increases.

C Model and Loss Details

E2EDriver ($\sim$ 12M parameters). E2EDriver (Mufti, 2026) uses the current multi-view camera-frame at anchor t, a 4-second ego-state history, and a high-level navigation command. A per-view CNN extracts image features, token-level and cross-camera fusion modules form a global visual embedding, and an MLP branch encodes the ego-state history and command. The fused representation is passed to deterministic trajectory, Gaussian-mixture, and MNLL-style probabilistic heads. Across temporal sampling frequencies, the current-frame visual input, ego-state history, command encoding, and future ego-trajectory target are unchanged.

BEV-E2EDriver ($\sim$ 10M parameters). BEV-E2EDriver keeps the ego-state branch, command input, output heads, loss formulation, and future ego-trajectory target of E2EDriver, but replaces the image-space encoder with a BEVFormer-style bird's-eye-view module (Li et al., 2022). Multi-view image features from the current anchor are aggregated with BEV queries and spatial cross-attention to produce a global BEV embedding. This model provides a controlled comparison between image-space and BEV-space perception under the same temporal-frequency intervention.

Tiny-SSR ($\sim$ 18M parameters). Tiny-SSR is a compact variant of the Sparse Scene Representation framework (Li & Cui, 2025). It replaces the baseline ResNet-50 backbone with a ResNet-18 (He et al., 2016) plus FPN (Lin et al., 2017), uses a 256-dimensional embedding, reduces the BEV grid from 100×100 to 50×50, and uses one encoder/decoder layer. Each sample uses the current multi-view frame, one previous visual/BEV context frame at a fixed 0.5-second offset, ego-state information, and a navigation command. The model predicts a command-conditioned sequence of future ego waypoints over a 3-second horizon at 0.5-second spacing. Across temporal sampling frequencies, this current-plus-previous-frame input pattern and target horizon are unchanged.

AutoVLA ($\sim$ 3B parameters). AutoVLA is instantiated as a Qwen2.5-VL-3B-based multimodal autoregressive trajectory-prediction model. At each anchor t, an AutoVLA sample follows a fixed multimodal template: four temporal image frames ending at t, three forward-facing camera views per temporal frame, a fixed system prompt, a fixed multimodal/user prompt, and a fixed ego-state conditioning block. The future ego-trajectory is represented as action tokens. A learned action codebook, tokenizer, and rollout decoder convert action tokens back into continuous ego waypoints. Across the frequency study, the pretrained VLM backbone, visual interface, prompt contract, ego-state pathway, action vocabulary, trajectory decoder, and prediction target remain fixed. Dataset-specific preprocessing is used only to construct the corresponding frequency-sweep training sets under the original AutoVLA codebase interface.

Drive-JEPA. Drive-JEPA (Wang et al., 2026) combines a video joint-embedding predictive architecture (V-JEPA) encoder with a proposal-centric trajectory planner. Its published framework pretrains a vision transformer on driving videos, refines multiple waypoint proposals with multimodal trajectory distillation, and selects a trajectory using a momentum-aware score. Within the NAVTEST frequency sweep, the model configuration and per-sample input and prediction target remain fixed.

DriveVLA-W0. DriveVLA-W0 (Li et al., 2025b) augments vision-language-action learning with a world-model objective that predicts future visual content. Its published framework combines a VLA backbone with an action expert for trajectory generation and includes autoregressive visual-token and diffusion-latent world-model variants. Within each frequency comparison, the selected model configuration and trajectory target remain fixed.

Training losses. For E2EDriver and BEV-E2EDriver, the deterministic trajectory head is supervised with mean-squared error (MSE), the Gaussian-mixture head is supervised with a Gaussian-mixture negative log-likelihood (GMM-NLL), and the modal negative log-likelihood (MNLL) head assigns supervision to the best-matching candidate trajectory mode while learning its mode probability. When multiple heads are enabled, the objective is

$$\mathcal{L}_{\text{E2E}} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{gmm}}\mathcal{L}_{\text{gmm}} + \lambda_{\text{mnll}}\mathcal{L}_{\text{mnll}}. \tag{18}$$

Here, $\mathcal{L}_{\text{E2E}}$ denotes the total training loss for E2EDriver or BEV-E2EDriver, $\mathcal{L}_{\text{mse}}$ denotes the deterministic trajectory MSE loss, $\mathcal{L}_{\text{gmm}}$ denotes the Gaussian-mixture negative log-likelihood loss, and $\mathcal{L}_{\text{mnll}}$ denotes the modal negative log-likelihood loss. The coefficients λ_{mse}, λ_{gmm}, and λ_{mnll} are nonnegative loss weights controlling the contribution of the corresponding heads.

Tiny-SSR follows the SSR objective, which combines imitation learning for command-conditioned waypoint prediction with BEV feature reconstruction. Let $\mathcal{L}_{\text{imi}}$ denote the imitation-learning loss for command-conditioned waypoint prediction. Let $\hat{\mathbf{B}}_{t+1}$ denote the predicted future BEV feature at the next prediction step, and let $\mathbf{B}_{t+1}$ denote the corresponding target BEV feature:

$$\mathcal{L}_{\text{bev}} = \left\| \hat{\mathbf{B}}_{t+1} - \mathbf{B}_{t+1} \right\|_2, \qquad \mathcal{L}_{\text{SSR}} = \mathcal{L}_{\text{imi}} + \mathcal{L}_{\text{bev}}. \tag{19}$$

Here, $\mathcal{L}_{\text{bev}}$ denotes the BEV feature reconstruction loss, and $\mathcal{L}_{\text{SSR}}$ denotes the total Tiny-SSR training loss.

AutoVLA is trained with supervised fine-tuning (SFT) of the multimodal autoregressive model (Zhou et al., 2025). Given camera inputs $\mathbf{C}$, prompt tokens $\mathbf{I}$, and ego-state input $\mathbf{S}$, the model predicts the assistant response containing future-trajectory action tokens. Here, $\mathbf{C}$ denotes the input camera observations, $\mathbf{I}$ denotes the tokenized language prompt, and $\mathbf{S}$ denotes the ego-state input used by the model. The frequency-response study configuration uses an answer-only target and omits Chain-of-Thought (CoT) reasoning and reinforcement fine-tuning (RFT) from the original AutoVLA methodology:

$$\mathcal{L}_i^{\text{SFT}} = \mathcal{L}_{\text{LM},i} + \lambda_a \mathcal{L}_{\text{action},i}. \tag{20}$$

Here, i indexes a training example, $\mathcal{L}_i^{\text{SFT}}$ denotes the SFT loss for example i, $\mathcal{L}_{\text{LM},i}$ denotes the language-modeling loss for the assistant response tokens, and $\mathcal{L}_{\text{action},i}$ denotes the action-token loss. The coefficient λ_a is a nonnegative loss weight for the action-token loss. Predicted action tokens are decoded into continuous future ego waypoints and evaluated with average displacement error (ADE) and final displacement error (FDE).

In Drive-JEPA's published planner, the training objective combines trajectory-proposal loss with proposal-score, mapping, and collision-prediction losses (Wang et al., 2026). Its video encoder is pretrained with the V-JEPA latent-prediction objective. The frequency sweep retains the chosen model and training objectives across settings.

DriveVLA-W0's published world-model variants combine action prediction loss with a visual prediction loss (Li et al., 2025b). The autoregressive variant predicts visual tokens, while the diffusion variant denoises a future-image latent. The chosen objective remains fixed within a frequency comparison.

D Timestamp-Anchor Construction and Dataset Handling

For each scene timeline, native camera timestamps are sorted in temporal order. For a requested temporal sampling frequency f, the consecutive camera frame timestamp gap is $\Delta t_f = 1/f$. The frequency-induced anchor sequence is obtained by retaining anchors at this interval along the scene timeline, subject to model-specific validity constraints. The validity operator $\text{Valid}_m(\mathcal{S}_i, t)$ excludes anchors for which the required model input, ego-state history, camera views, previous context frame, or future ego-trajectory supervision is unavailable. Thus, changing f changes which anchors are retained, but not the semantics of a retained sample.

For irregular timestamps, missing frames, or synchronization failures, the intended behavior is to retain only anchors satisfying the same model-specific input and future-target requirements used at other frequencies. Exact reconstruction of each frequency-induced training set requires dataset-specific scene identifiers, timestamp manifests, and the validity rules used for every model.

The new NAVSIM comparison starts from the official 2 Hz training clips (Dauner et al., 2024) and uses corresponding nuPlan recordings (Caesar et al., 2021) to construct a denser 10 Hz training version. The reported 2 Hz and 10 Hz results compare models trained with these two anchor densities and evaluated with NAVSIM metrics. This construction makes additional within-scene timestamps available for the frequency intervention.

E 5-Second trajectory-prediction

The main comparison uses 3-second trajectory-prediction. We also report 5-second results where measured, because longer horizons introduce additional trajectory uncertainty and may change the preferred temporal sampling frequency. The 5-second results are broadly consistent with the 3-second frequency-response, although the preferred frequency can shift with prediction horizon. On Waymo, E2EDriver and BEV-E2EDriver again perform best at 8 Hz rather than 10 Hz. On nuScenes, E2EDriver and BEV-E2EDriver also remain below the highest evaluated frequency, with the best ADE at 6 Hz and 8 Hz, respectively. On PAVE, their best frequencies shift upward, with E2EDriver best at 12 Hz and BEV-E2EDriver best at 15 Hz. AutoVLA performs best at 10 Hz on Waymo, 12 Hz on nuScenes, and approximately 18 Hz on PAVE, where 18 Hz and 20 Hz are very close. Thus, longer-horizon evaluation does not remove the main pattern: the preferred temporal sampling frequency remains model- and dataset-dependent.

Table 6: 5-second future ego-trajectory prediction (ADE / FDE in meters, single run where available). **Bold** marks the best ADE setting for each model–dataset pair.

		Waymo (10 Hz)		**nuScenes** (12 Hz)		**PAVE** (20 Hz eval.)	
Model	**Freq.**	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓
E2EDriver	2 Hz	1.731	4.605	2.492	7.715	—	—
	4 Hz	1.526	4.435	2.536	7.881	—	—
	6 Hz	1.488	4.323	**2.327**	**7.142**	2.494	6.015
	8 Hz	**1.471**	**4.245**	2.521	7.843	2.180	5.711
	10 Hz	1.516	4.419	2.510	7.796	2.300	5.478
	12 Hz	—	—	2.489	7.731	**1.942**	**4.916**
	15 Hz	—	—	—	—	1.982	4.942
	18–20 Hz	—	—	—	—	2.28–2.48	5.44–5.98
BEV-E2EDriver	2 Hz	1.591	4.561	2.559	7.843	—	—
	4 Hz	1.565	4.516	2.541	7.807	—	—
	6 Hz	1.482	4.320	2.557	7.821	2.574	7.963
	8 Hz	**1.464**	**4.218**	**2.351**	**7.197**	2.311	5.758
	10 Hz	1.486	4.320	2.506	7.762	2.330	5.759
	12 Hz	—	—	2.514	7.748	2.254	5.700
	15 Hz	—	—	—	—	**2.054**	**5.174**
	18–20 Hz	—	—	—	—	2.27–2.48	5.42–5.69
AutoVLA	2 Hz	2.831	6.975	3.236	8.621	2.482	5.665
	4 Hz	2.795	6.883	2.111	5.988	1.970	4.582
	6 Hz	2.766	6.820	1.910	5.474	1.953	**4.353**
	8 Hz	2.780	6.914	1.827	5.204	1.955	4.519
	10 Hz	**2.758**	**6.833**	1.811	5.173	1.929	4.436
	12 Hz	—	—	**1.741**	**4.994**	1.905	4.421
	15 Hz	—	—	—	—	1.878	4.403
	18 Hz	—	—	—	—	**1.878**	4.377
	20 Hz	—	—	—	—	1.880	4.485