SCALING ACTION-FREE VIDEO LATENT PREDICTION FOR TRANSFERABLE REPRESENTATIONS IN END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving policies require effective visual representations that capture both spatial scene structure and temporal dynamics. Video latent prediction has emerged as an effective paradigm for learning spatiotemporal visual representations, yet it remains unclear whether scaling this objective with action-free driving videos can yield representations that benefit end-to-end planning. We study action-free video latent prediction as a scalable pretraining objective and introduce ScaleDrive, a driving-oriented framework pretrained exclusively on large-scale internet open-world driving videos, without action annotations or benchmark-specific data. We keep the pretrained video encoder frozen during downstream end-to-end planning training to isolate the transferability of representations learned from internet driving videos. Ego-motion and depth probes show that scaling action-free video pretraining progressively strengthens the temporal and geometric information encoded in the frozen representation. We systematically study the visual representation transfer across different end-to-end driving policies and closed-loop benchmarks, where ScaleDrive consistently improves downstream performance and demonstrates that scaling action-free video latent prediction yields representations that generalize across both planning policy architectures and evaluation scenarios and protocols.

1 INTRODUCTION

Learning generalizable visual representations is a central challenge for end-to-end autonomous driving. A strong representation should capture not only scene appearance, but also spatial structure, interactions among traffic participants, and temporal evolution. Existing driving representations are often learned through task-specific objectives, including object detection (Huang et al., 2021; Li et al., 2024b; Liu et al., 2022), online mapping (Liao et al., 2022; Liu et al., 2024; Liao et al., 2025b), occupancy prediction (Tian et al., 2023; Huang et al., 2023; Zhang et al., 2023), and vision-language-action modeling (Wang et al., 2025; Zhou et al., 2026a;b). Although effective within their respective task formulations, these objectives primarily encourage representations to capture information required by predefined semantic, geometric, or planning targets, rather than broadly modeling the general structure and dynamics of the driving world.

Self-supervised video learning provides a complementary paradigm by learning directly from how visual scenes are structured and evolve over time, without requiring task-specific supervision (Tong et al., 2022; Wang et al., 2024b; Assran et al., 2025). In particular, video latent prediction learns spatiotemporal representations by predicting masked content in representation space (Bardes et al., 2024; Assran et al., 2025). This paradigm is particularly attractive for autonomous driving: raw driving videos naturally capture road geometry, traffic participants, interactions, and scene dynamics without requiring additional detection, mapping, occupancy, reasoning, or action annotations. They can therefore support large-scale representation pretraining without additional task-specific labeling. Although prior works have explored self-supervised pretraining from driving videos (Shi et al., 2025; Bartoccioni et al., 2025; Wang et al., 2026b; Strong et al., 2026), an important question remains underexplored: *does driving representation quality continue to improve as raw driving video pretraining data scales? More importantly, do these representation gains translate into better downstream planning and transfer across different planning policies and driving scenarios?*

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.

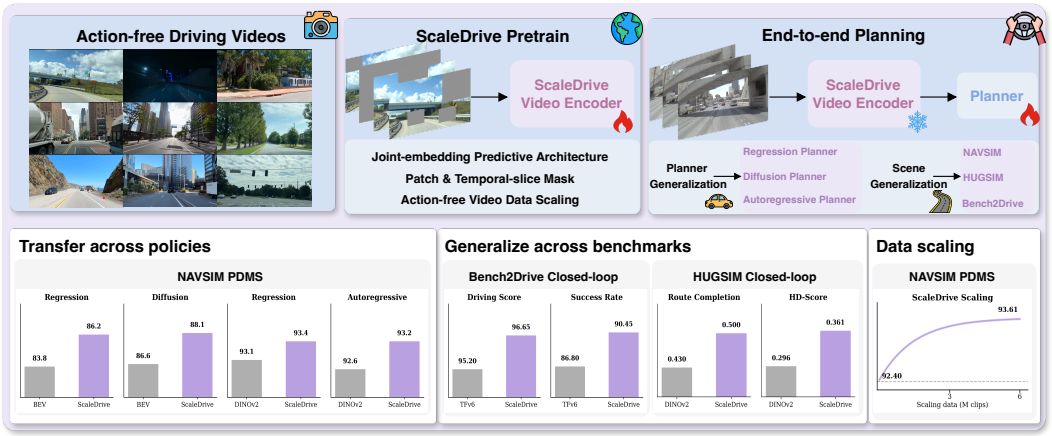

Figure 1: ScaleDrive learns transferable driving representations through self-supervised predictive pretraining on large-scale raw driving videos. The pretrained encoder is kept frozen and transferred across heterogeneous end-to-end planners and closed-loop driving environments, consistently improving downstream planning performance.

To study this question, we introduce **ScaleDrive**, a scalable self-supervised video representation learning framework for autonomous driving, as illustrated in Fig. 1. ScaleDrive learns solely from raw driving videos, without action or task-specific supervision. It follows a joint-embedding predictive architecture (JEPA) (Bardes et al., 2024; Assran et al., 2025), where visible video regions are encoded by a context encoder, and the latent representations of masked regions are predicted under the supervision of targets produced by the momentum target encoder. To strengthen temporal modeling, we augment conventional patch-level masking with complete temporal-slice masking, encouraging the model to capture both spatial scene structure and temporal dynamics. For downstream evaluation, we freeze the pretrained video encoder and train only the planning module, helping isolate the contribution of the learned representation from downstream representation adaptation.

Our experiments reveal a consistent representation scaling trend. As pretraining data increases, downstream planning performance improves while the visual encoder remains frozen. Ego-motion and monocular depth probing further show that temporal-motion and spatial-geometric information becomes increasingly accessible from the learned representation. Together, these results indicate that scaling video latent prediction pretraining produces increasingly useful driving representations.

We further evaluate whether these gains transfer beyond a single downstream planning configuration. The same frozen ScaleDrive representation consistently improves regression, diffusion, and autoregressive planners, and remains effective across distinct closed-loop benchmarks including HUGSIM (Zhou et al., 2025) and Bench2Drive (Jia et al., 2024). These results show that scaling self-supervised predictive pretraining on raw driving videos can progressively strengthen transferable driving representations and improve downstream planning across heterogeneous policy architectures and driving environments.

The main contributions are summarized as follows:

- We introduce **ScaleDrive**, a scalable self-supervised pretraining framework that learns transferable driving representations solely from large-scale raw driving videos, without action annotations or task-specific supervision.
- We develop a **driving-oriented predictive pretraining strategy** that combines patch-level masking with temporal-slice masking to capture both spatial scene structure and temporal dynamics.
- We conduct a **systematic study of representation scaling**. As video pretraining data increases, downstream planning performance improves, while ego-motion and depth probing reveal progressively stronger temporal-motion and spatial-geometric information in the frozen representation.
- ScaleDrive achieves **competitive end-to-end planning performance with a frozen visual encoder** and consistently improves regression, diffusion, and autoregressive planners and generalizes across distinct benchmarks including NAVSIM, HUGSIM and Bench2Drive.

2 RELATED WORKS

Existing studies on visual representation learning for autonomous driving span task-specific intermediate representations such as BEV and occupancy (Li et al., 2024b; Huang et al., 2021; Tian et al., 2023; Zhang et al., 2023), generic foundation-model features from large-scale pretrained vision and vision-language models (Oquab et al., 2023; Yang et al., 2025; Wang et al., 2024a; Kirby et al., 2026; Zhang et al., 2025; Gui et al., 2026; Zhou et al., 2026b; Jiang et al., 2026), and video-based joint-embedding predictive learning (Assran et al., 2025; Wang et al., 2026b; Yang et al., 2026). In parallel, scaling studies show that driving policies benefit from more action-supervised data (Naumann et al., 2025; Zheng et al., 2024; Tian et al., 2026; Li et al., 2026c), and representation learning benefits from larger-scale data and models (Kong et al., 2025; Li et al., 2026b). However, whether *action-free* driving representations can be systematically scaled and transferred to end-to-end planning remains underexplored. The full version of related works can be found in Appendix B.

3 SCALEDRIVE

ScaleDrive learns transferable driving representations from action-free videos through self-supervised latent prediction. It builds upon the joint-embedding predictive architecture, using raw driving videos without trajectory or perception annotations for pretraining. After pretraining, the predictor is discarded and the video encoder is frozen for downstream end-to-end driving policies, allowing downstream performance to more directly reflect the quality of the learned representation. A key design challenge is that conventional patch-level prediction may exploit same-frame spatial redundancy without sufficiently modeling temporal dynamics. We therefore introduce complete temporal-slice masking to complement patch-level masking, encouraging prediction from temporally distributed visual context. We next provide the theoretical motivation for this design and describe the resulting spatiotemporal latent prediction framework.

3.1 DYNAMICS-GUIDED TEMPORAL PREDICTION

We first analyze latent prediction for driving videos from a dynamic-system perspective. The purpose of this analysis is not to model the full complexity of real driving dynamics, but to characterize a sufficient condition under which patch-level prediction admits a same-frame shortcut and to explain why complete temporal masking provides a stronger temporal learning signal.

Driving dynamic system. A driving video can be viewed as a sequence of partial observations generated by an underlying dynamic scene state. Let $\mathbf{s}_t \in \mathbb{R}^{d_s}$ denote the latent physical state of the driving scene, including ego motion, surrounding-agent states, and local scene geometry. Its temporal evolution and visual observation can be abstracted as

$$\mathbf{s}_{t+1} = F(\mathbf{s}_t) + \mathbf{w}_t, \qquad \mathbf{y}_t = G(\mathbf{s}_t) + \mathbf{v}_t, \tag{1}$$

where $\mathbf{w}_t$ and $\mathbf{v}_t$ denote process and observation noise, respectively, and $\mathbf{y}_t$ denotes a temporally indexed latent observation corresponding to the predictive-learning target. Here, $\mathbf{s}_t$ is an abstract physical latent introduced only for analysis and we do not assume that a neural encoder explicitly recovers this state.

Local linearization. Over a sufficiently short video clip, we analyze the local behavior of the nonlinear system through a first-order linearization around the observed trajectory:

$$\delta\mathbf{s}_{t+1} = \mathbf{A}_t\delta\mathbf{s}_t + \mathbf{w}_t, \qquad \delta\mathbf{y}_t = \mathbf{H}_t\delta\mathbf{s}_t + \mathbf{v}_t. \tag{2}$$

For analytical clarity, we consider a short local window in which the Jacobians vary slowly and can be approximated as $\mathbf{A}_t \approx \mathbf{A}$ and $\mathbf{H}_t \approx \mathbf{H}$. This local time-invariant approximation is used only for the following analysis. The complete nonlinear-to-linear formulation and assumptions are provided in Appendix A.1.

Proposition 1 (Existence of a same-frame shortcut under spatial redundancy)**.** *Let the latent observation at time t be partitioned into visible and masked components,*

$$\delta\mathbf{y}_t^V = \mathbf{H}_V\delta\mathbf{s}_t + \mathbf{v}_t^V, \qquad \delta\mathbf{y}_t^M = \mathbf{H}_M\delta\mathbf{s}_t + \mathbf{v}_t^M. \tag{3}$$

If there exists a mapping $\mathbf{C}$ *such that*

$$\|\mathbf{H}_M - \mathbf{C}\mathbf{H}_V\|_2 \le \epsilon_s, \tag{4}$$

then the same-frame predictor $\widehat{\delta\mathbf{y}_t^M} = \mathbf{C}\delta\mathbf{y}_t^V$ *achieves*

$$\left\|\delta\mathbf{y}_t^M - \mathbf{C}\delta\mathbf{y}_t^V\right\| \le \epsilon_s \left\|\delta\mathbf{s}_t\right\| + \left\|\mathbf{v}_t^M - \mathbf{C}\mathbf{v}_t^V\right\|. \tag{5}$$

Proposition 1 provides a sufficient condition under which masked latent regions can be predicted from visible regions at the same temporal instant. Importantly, neither the predictor nor its error bound depends on the temporal transition operator $\mathbf{A}$. Therefore, when strong spatial redundancy exists, low patch-prediction error can be achieved without explicitly exploiting temporal transition dynamics. We refer to this possible solution as a *same-frame spatial shortcut*. The proposition does not imply that every patch-level prediction problem admits such a shortcut; rather, it characterizes when such a solution can exist. The proof is provided in Appendix A.2.

Proposition 2 (Temporal recoverability under complete temporal masking)**.** *Suppose the complete latent observation at time* t *is unavailable. For the preceding* K *observations, define*

$$\mathbf{Y}_K = \begin{bmatrix} \delta\mathbf{y}_{t-K} \\ \vdots \\ \delta\mathbf{y}_{t-1} \end{bmatrix}, \qquad \mathbf{O}_K = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}\mathbf{A} \\ \vdots \\ \mathbf{H}\mathbf{A}^{K-1} \end{bmatrix}, \qquad \mathbf{Y}_K = \mathbf{O}_K\delta\mathbf{s}_{t-K} + \boldsymbol{\eta}, \tag{6}$$

where $\boldsymbol{\eta}$ *collects process, observation, and local-linearization residuals. If the local system is observable,* $\mathrm{rank}(\mathbf{O}_K) = d_s$*, and*

$$\delta\mathbf{y}_t = \mathbf{H}\mathbf{A}^K\delta\mathbf{s}_{t-K} + \mathbf{r}_t, \qquad \|\mathbf{r}_t\| \le \epsilon_{\mathrm{dyn}}, \tag{7}$$

then the predictor $\widehat{\delta\mathbf{y}_t} = \mathbf{H}\mathbf{A}^K\mathbf{O}_K^{\dagger}\mathbf{Y}_K$ *satisfies*

$$\left\|\widehat{\delta\mathbf{y}_t} - \delta\mathbf{y}_t\right\| \le \frac{\|\mathbf{H}\mathbf{A}^K\|_2}{\sigma_{\min}(\mathbf{O}_K)}\|\boldsymbol{\eta}\| + \epsilon_{\mathrm{dyn}}. \tag{8}$$

The past-only formulation in Proposition 2 is used for analytical clarity and the actual predictor can exploit context from both temporal directions. Complete temporal masking removes the direct same-frame shortcut characterized in Proposition 1. Under local observability, however, the masked observation remains recoverable from temporally distributed context through the local transition dynamics. This is a recoverability result rather than a guarantee that a finite-capacity neural network explicitly reconstructs the physical state. The proof is provided in Appendix A.3.

Design implication. The two propositions motivate complementary masking strategies. Patch-level masking provides dense spatial supervision but may admit same-frame shortcuts under strong spatial redundancy. Temporal-slice masking removes all observations at selected temporal indices, eliminating this direct same-frame evidence and increasing reliance on temporally distributed context. ScaleDrive therefore combines patch-level and temporal-slice masking to jointly strengthen spatial and temporal representation learning.

3.2 Spatial-Temporal Latent Prediction

Guided by the above analysis, we instantiate ScaleDrive using a joint-embedding predictive architecture on action-free driving videos.

Input and latent prediction. Given an action-free driving video $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^T \in \mathbb{R}^{T\times H\times W\times 3}$, we partition it into non-overlapping spatiotemporal patches and project them into visual tokens $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N\times D}$, where N and D denote the number and dimension of tokens, respectively. The temporal and spatial coordinates are encoded using 3D RoPE Su et al. (2024).

Rather than reconstructing RGB pixels, ScaleDrive predicts latent representations of masked spatiotemporal regions. Given a mask set $\mathcal{M}$, the context encoder E_θ processes only the visible tokens, while a predictor P_ϕ infers the representations at masked positions. The prediction targets are obtained from the complete token sequence using a momentum target encoder $E_{\bar\theta}$:

$$\mathbf{Z}^{\mathrm{con}} = E_\theta(\mathbf{X}_{\setminus\mathcal{M}}), \qquad \hat{\mathbf{Z}}_{\mathcal{M}} = P_\phi(\mathbf{Z}^{\mathrm{con}}, \mathcal{M}), \qquad \mathbf{Z}^{\mathrm{tar}} = E_{\bar\theta}(\mathbf{X}). \tag{9}$$

The target encoder shares the same architecture as the context encoder and is updated by an exponential moving average, $\bar\theta \leftarrow \mu\bar\theta + (1-\mu)\theta$, with gradients stopped.

Patch-level spatial masking. Following V-JEPA2, we sample spatially contiguous groups of spatiotemporal tokens, denoted by $\mathcal{M}_p \subseteq \{1, \dots, N\}$, and remove them from the context encoder input. The predictor then infers their target representations from the remaining visible visual context. This provides dense spatial supervision but, as characterized by Proposition 1, may admit solutions that rely heavily on same-frame spatial redundancy.

Temporal-slice masking. To complement spatial masking, we additionally remove complete temporal slices. Let T' denote the temporal resolution after patchification and $\mathcal{T}_f \subseteq \{1, \dots, T'\}$ the sampled temporal indices. The corresponding temporal mask is

$$\mathcal{M}_f = \{i \mid \tau(i) \in \mathcal{T}_f\}, \tag{10}$$

where $\tau(i)$ maps token i to its post-patchification temporal index. Because all spatial tokens at the selected temporal indices are removed, the predictor cannot directly exploit visible observations from the same temporal slice and must rely more strongly on temporally distributed context. During pretraining, the two masking strategies are combined as

$$\mathcal{M} = \mathcal{M}_p \cup \mathcal{M}_f. \tag{11}$$

Latent prediction objective. For each masked position $i \in \mathcal{M}$, the predicted latent $\hat{\mathbf{z}}_i$ is optimized to match the corresponding target representation $\mathbf{z}_i^{\text{tar}}$. Following V-JEPA2, we minimize an L_1 regression loss over masked tokens:

$$\mathcal{L}_{\text{VLP}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \hat{\mathbf{z}}_i - \text{sg}\left(\mathbf{z}_i^{\text{tar}}\right) \right\|_1, \tag{12}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The context encoder and predictor are optimized jointly, while the target encoder is updated only through the momentum update.

ScaleDrive therefore retains the standard joint-embedding prediction objective and momentum-target mechanism, while modifying the masking strategy: patch-level prediction provides dense spatial supervision, whereas complete temporal-slice prediction reduces reliance on same-frame spatial shortcuts and strengthens temporal predictive learning.

3.3 Scaling and Frozen Representation Transfer

Fixed-compute data scaling. ScaleDrive scales representation learning along the data dimension using raw action-free driving videos. Across different pretraining data scales, we keep the encoder architecture, masking formulation, optimization configuration, and overall training budget fixed, while varying only the size of the sampled driving-video pool. This controlled protocol isolates whether increasing the diversity and coverage of raw driving observations improves the quality of the learned representation under a comparable optimization budget.

Frozen representation transfer. After pretraining, we discard the predictor and retain the target encoder, denoted by E_{vis}, as the visual representation model. Given an observed driving video $\mathbf{V}$, the encoder produces a sequence of spatiotemporal latent representations:

$$\mathbf{Z} = E_{\text{vis}}(\mathbf{V}) \in \mathbb{R}^{N \times D}. \tag{13}$$

The parameters of E_{vis} remain frozen throughout downstream policy training. This design is deliberate: if the visual encoder is jointly fine-tuned with the planner, improvements in downstream planning performance could arise from representation adaptation during policy training. Freezing the encoder makes the downstream performance a more direct measure of the information already learned during action-free video pretraining.

Planning interface. Given the frozen video representation $\mathbf{Z}$ and the current ego status $\mathbf{e}$, the downstream planner models

$$p_\omega\left(\mathbf{Y} \mid \mathbf{Z}, \mathbf{e}\right), \tag{14}$$

where $\mathbf{Y}$ denotes the future ego trajectory and ω contains only the trainable parameters of the planning module. With a set of learnable ego queries $\mathbf{Q}_e \in \mathbb{R}^{M \times D}$, the trajectory decoder predicts

$$\hat{\mathbf{Y}} = D_\omega\left(\mathbf{Q}_e, \mathbf{Z}, \mathbf{e}\right). \tag{15}$$

Table 1: **Planning performance on NAVSIMv1.** "Enc. Frozen" indicates whether the visual encoder is kept frozen during downstream planner training. † denotes planner training without perception-task labels.

Method	Inputs	Enc. Frozen	NC	DAC	EP	TTC	Comf.	PDMS
TransfuserChitta et al. (2022)	L&C	✗	97.7	92.8	79.2	92.8	100.0	84.0
HydraMDPLi et al. (2024a)	L&C	✗	98.3	96.0	78.7	94.6	100.0	86.5
DiffusionDriveLiao et al. (2025a)	L&C	✗	98.2	96.2	82.2	94.7	100.0	88.1
World4Drive† Zheng et al. (2025)	C	✗	97.4	94.3	79.9	92.8	100.0	85.1
Drive-JEPA† Wang et al. (2026b)	C	✗	98.7	96.2	82.9	95.5	100.0	89.0
ScaleDrive†	C	✓	99.0	96.3	82.9	96.0	100.0	89.0
ReCogDriveXiong et al. (2026)	C	✓	97.9	97.3	87.3	94.9	100.0	90.8
SGDriveLi et al. (2026a)	C	✓	98.6	97.8	85.8	96.2	100.0	91.1
iPadGuo et al. (2026)	C	✗	98.6	98.3	88.0	94.9	99.9	91.7
DrivoRKirby et al. (2026)	C	✗	98.9	98.3	89.1	96.2	100.0	93.1
Drive-JEPAWang et al. (2026b)	C	✗	99.1	98.2	**91.4**	96.2	100.0	93.7
ChainFlow-VLAWang et al. (2026c)	C	✗	98.8	**98.6**	90.8	96.1	100.0	93.6
ScaleDrive	C	✓	**99.3**	98.4	90.6	**97.0**	100.0	**93.8**

Table 2: **Planning performance on NAVSIMv2.** ScaleDrive achieves strong overall performance under the more comprehensive closed-loop evaluation metrics.

Method	NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
DiffusionDriveLiao et al. (2025a)	98.2	95.9	99.4	99.8	87.5	97.3	96.8	**98.3**	87.7	84.5
DiffusionDriveV2Zou et al. (2025)	97.7	96.6	99.2	99.8	88.9	97.2	96.0	97.8	**91.0**	87.5
SparseDriveV2Sun et al. (2026)	98.1	98.1	99.6	99.8	**91.1**	97.3	96.9	98.2	78.4	90.1
Latent-WAMWang et al. (2026a)	98.1	97.3	99.6	99.8	87.7	97.3	**97.6**	98.1	87.3	89.3
ScaleDrive	**99.3**	**98.9**	**99.7**	99.6	85.4	**98.9**	95.9	98.2	82.1	**90.5**

The downstream decoder follows established end-to-end planning objectives (Yao et al., 2026; Kirby et al., 2026; Wang et al., 2026c). The architectural and optimization details are provided in Appendix C.2. This deliberately simple interface reduces architectural confounding and allows downstream performance to more directly reflect the quality of the pretrained driving representation.

Heterogeneous planner integration. ScaleDrive is not tied to the primary trajectory decoder. Its frozen visual encoder can serve as a plug-in module and be integrated into heterogeneous planners through simple planner-specific interfaces. BEV-based policies fuse ScaleDrive representation with the native BEV features, whereas camera-view-based policies replace the native visual representation with ScaleDrive. The detailed integration strategies are provided in Appendix C.3.

4 Experiment

4.1 Experimental Setup

We pretrain ScaleDrive with OpenDV-YouTube (Yang et al., 2024), a large-scale collection of internet driving videos without action or task-specific annotations. The raw driving videos are segmented into 4-second clips sampled at 2 Hz, yielding a total of 6M video clips for video latent prediction pretraining. We use NAVSIMv1 (Dauner et al., 2024) and NAVSIMv2 (Cao et al., 2025) as the primary end-to-end planning benchmarks. Beyond the main planning evaluation, we assess the ScaleDrive representation quality through ego-motion and monocular depth probing, evaluate cross-policy transfer with different planner architectures, and study cross-benchmark generalization on HUGSIM (Zhou et al., 2025) and Bench2Drive (Jia et al., 2024). Details on the training configurations and evaluation protocols are provided in Appendix C.

4.2 Main Planning Results

Tables 1 and 2 report the planning performance of ScaleDrive on NAVSIMv1 and NAVSIMv2. ScaleDrive achieves strong planning performance on both benchmarks, reaching a PDMS of 93.8 on NAVSIMv1 and an EPDMS of 90.5 on NAVSIMv2. The pretrained video encoder remains frozen throughout downstream policy training, indicating that the resulting performance primarily reflects the quality and transferability of the ScaleDrive representation.

Table 3: **Transfer across end-to-end driving policies**. Downstream training and evaluation settings are kept fixed within each policy.

Method	Trajectory Modeling	Representation	NC	DAC	TTC	EP	PDMS
LTF	Regression	BEV	97.4	92.8	92.4	79.0	83.8
LTF(V-JEPA2)		w. V-JEPA2	97.8	91.7	92.4	78.8	83.2(-0.6)
LTF(ScaleDrive)		w. ScaleDrive	**98.2**	**94.0**	**93.6**	**81.6**	**86.2**(+2.4)
DiffusionDrive	Diffusion	BEV	98.1	94.8	94.1	81.1	86.6
DiffusionDrive(V-JEPA2)		w. V-JEPA2	**98.6**	95.1	95.1	81.8	87.4(+0.8)
DiffusionDrive(ScaleDrive)		w. ScaleDrive	**98.6**	**95.8**	**95.3**	**82.3**	**88.1**(+1.5)
DrivoR	Regression	DINOv2	98.9	**98.3**	96.2	89.1	93.1
DrivoR(V-JEPA2)		w. V-JEPA2	99.1	97.9	96.5	87.9	92.3(-0.8)
DrivoR(ScaleDrive)		w. ScaleDrive	**99.2**	98.0	**96.7**	**89.8**	**93.4**(+0.3)
ChainFlow	Autoregressive	DINOv2	98.9	98.1	**96.2**	88.9	92.6
ChainFlow(V-JEPA2)		w. V-JEPA2	98.8	97.5	95.6	89.2	92.1(-0.5)
ChainFlow(ScaleDrive)		w. ScaleDrive	**99.0**	**98.3**	**96.2**	**90.5**	**93.2**(+0.6)

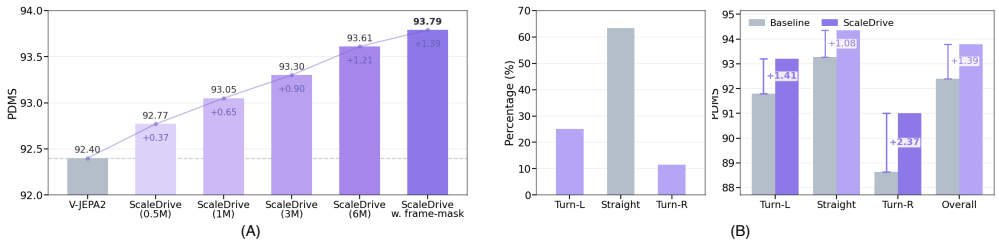

Figure 2: (A) Data scaling performance of ScaleDrive. (B) Command-wise gains under the imbalanced training driving action distribution.

We study how ScaleDrive scales with action-free driving videos for downstream planning. Fig. 2 (A) illustrates how planning performance scales with the amount of action-free driving-video pretraining. As the number of video clips increases from 0.5M to 6M, PDMS consistently improves from 92.77 to 93.61, compared with 92.40 using the official V-JEPA2 representation. The gain trend empirically shows that increasing the diversity and coverage of driving videos improves the quality of the frozen representation under a fixed training budget. Beyond data scaling, we introduce temporal-slice masking to strengthen temporal modeling during pretraining, following Proposition 2. At the same data scale, the additional temporal-slice masking further improves planning performance.

A key motivation for scaling action-free video pretraining is to expose the model to a broader diversity of driving scenes. We investigate whether this broader scene coverage translates into planning gains in long-tail driving scenes. We analyze performance across different driving commands in Fig. 2 (B). The downstream training distribution is highly imbalanced, with turning commands following a clear long-tail distribution. ScaleDrive improves planning performance across all commands, with more pronounced gains on long-tail turning maneuvers. These results suggest that the broader scene coverage provided by large-scale video pretraining is particularly beneficial for long-tail driving situations that are less frequently observed during downstream policy training.

4.3 Does ScaleDrive Transfer Across Driving Policies?

A transferable driving representation should provide consistent benefits across different policy designs. We evaluate the frozen ScaleDrive representation with four end-to-end driving policies while keeping the training settings fixed within each policy. These policies span regression, diffusion, and autoregressive trajectory modeling, with native visual representations ranging from BEV (Chitta et al., 2022) to DINOv2 (Oquab et al., 2023). As shown in Table 3, ScaleDrive consistently improves planning performance across all architectures. In contrast, V-JEPA2 does not yield similarly consistent improvements. This comparison indicates that scaling action-free pretraining on driving videos yields representations that transfer across heterogeneous downstream policies.

4.4 Does ScaleDrive Transfer to Different Driving Scenarios?

A central motivation of ScaleDrive is to investigate whether action-free driving videos can induce transferable driving representations that generalize beyond the downstream training distribution. We evaluate ScaleDrive on HUGSIM (Zhou et al., 2025) and Bench2Drive (Jia et al., 2024), two closed-

Table 4: **Cross-benchmark generalization on HUGSIM.**

Planner	Representation	Route Completion ↑					HD-Score ↑				
		Easy	Medium	Hard	Extreme	Avg.	Easy	Medium	Hard	Extreme	Avg.
LTF	BEV	0.684	0.407	**0.369**	0.255	0.410	0.528	0.246	**0.198**	0.081	0.246
LTF	w. V-JEPA2	0.738	0.400	0.298	**0.279**	0.412	0.630	0.228	0.141	0.113	0.251
LTF	w. ScaleDrive	**0.742**	**0.496**	0.329	0.259	**0.448**	**0.638**	**0.361**	0.180	**0.119**	**0.315**
ChainFlow	DINOv2	0.834	0.433	0.258	0.273	0.430	**0.705**	0.313	0.116	0.122	0.296
ChainFlow	w. V-JEPA2	0.782	0.443	0.315	0.252	0.432	0.640	0.342	0.168	0.108	0.302
ChainFlow	w. ScaleDrive	**0.843**	**0.561**	**0.362**	**0.274**	**0.500**	0.693	**0.442**	**0.210**	**0.123**	**0.361**

Table 5: **Cross-benchmark generalization on Bench2Drive.**

Planner	Representation	Driving Score↑	Success Rate↑	Efficiency↑	Comfort↑
TFv6	BEV	95.20	86.80	-	-
TFv6	w. V-JEPA2	95.00	85.91	202.33	15.45
TFv6	w. ScaleDrive	**96.65**(+1.45)	**90.45**(+3.65)	**210.22**	**18.03**

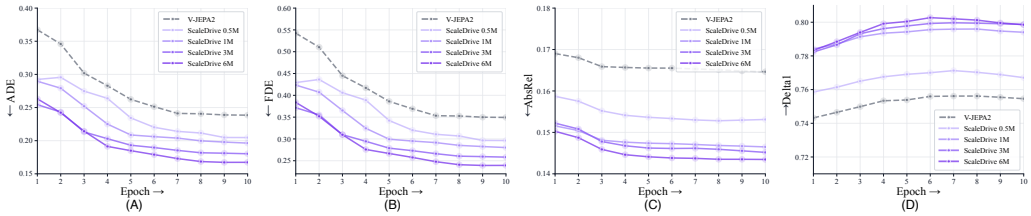

Figure 3: **Video representation probing.** (A–B) Ego-motion estimation with ADE ↓ and FDE ↓; (C–D) monocular depth estimation with AbsRel ↓ and δ_1 ↑.

loop benchmarks with distinct scenario distributions and evaluation protocols. Table 4 evaluates cross-benchmark generalization on HUGSIM. ScaleDrive improves the average Route Completion and HD-Score with different planner architectures and original representations, with clear benefits in the Medium and Hard scenarios. Table 5 further shows consistent improvements over the strong TFv6 (Nguyen et al., 2026) planner across closed-loop metrics on Bench2Drive. These results indicate that the learned representation generalizes across distinct driving environments, scenario distributions, and evaluation protocols.

4.5 Does Action-Free Video Latent Prediction Improve Driving Representations?

Proposition 2 suggests that temporal-slice prediction places stronger requirements on representing temporally evolving scene states. We therefore directly probe whether motion information becomes more accessible from the frozen representation through ego-motion estimation. As a complementary assessment of spatial representation quality, we further use monocular depth probing on KITTI (Geiger et al., 2012) to assess whether scaling action-free pretraining also strengthens spatial-geometric information. In both settings, the pretrained video encoder remains frozen and only the task-specific probe is optimized. Detailed architectures and optimization settings are provided in the Appendix C.5. As shown in Fig. 3, performance consistently improves as the amount of driving-video pretraining increases for both ego-motion and depth estimation. This trend indicates that scaling video latent prediction progressively strengthens the temporal and geometric information encoded in the frozen representation.

4.6 Qualitative Results

Visual representation. Fig. 4 provides qualitative comparisons of the spatiotemporal representations learned by V-JEPA2 and ScaleDrive. In the t-SNE visualizations shown in Fig. 4 (A), ScaleDrive exhibits more clearly separated feature distributions across different driving behaviors. The PCA visualizations in Fig. 4 (B) further show more organized spatial feature patterns across representative driving scenes. These observations suggest that scaling action-free video latent prediction pretraining yields more structured representations.

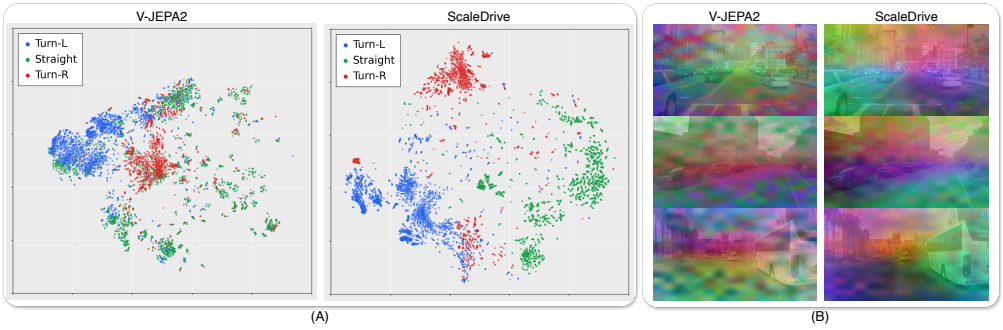

Figure 4: **Qualitative result of representations.** (A) t-SNE visualization. (B) PCA visualization of dense video features across representative driving scenes.



Figure 5: **Planning results on NAVSIM, Bench2Drive and HUGSIM.**

End-to-end Planning. Fig. 5 shows representative planning cases on NAVSIM, HUGSIM, and Bench2Drive. On NAVSIM, the predicted trajectories more closely follow the reference expert behavior across diverse road geometries and intersection scenarios. On HUGSIM and Bench2Drive, ScaleDrive exhibits safer closed-loop behavior, including appropriate responses to approaching vehicles and potentially unsafe interactions. These examples illustrate that stronger spatiotemporal representations can translate into more reliable downstream planning across driving environments.

5 Conclusion

We present ScaleDrive, a scalable representation-learning framework that learns driving representations from action-free videos through video latent prediction. By scaling pretraining data while keeping the downstream encoder frozen, we show that larger collections of driving videos consistently improve representation quality and translate into stronger planning performance. The learned representations capture increasingly useful motion and geometric cues, and generalize across different planning architectures and driving benchmarks. These results suggest that action-free video pretraining provides a promising direction for scaling autonomous driving beyond the availability of action-annotated data.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing the manuscript, including improving grammar, clarity, and readability. We have not used generative AI tools for research ideation or execution, generating synthetic datasets, proving mathematical claims, or drafting substantive sections of the paper. All AI-assisted edits were manually reviewed by the authors to ensure that they preserved the intended technical meaning. Technical claims, experimental results, and references were independently verified by the authors. We take full responsibility for the final content of this work, including any text produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies self-supervised representation learning for autonomous driving using existing research datasets and standard autonomous-driving benchmarks and does not involve the collection of new human-subject data. For large-scale video pretraining, we use OpenDV-YouTube, which is constructed from publicly available YouTube driving videos following the data collection and filtering protocol provided by the dataset authors. Following the official OpenDV release protocol, raw videos are obtained from their original YouTube sources and processed locally rather than redistributed as processed video data. As Internet driving videos may reflect biases or privacy concerns present in their original sources, we encourage careful consideration of data provenance and potential distributional biases when using such data for large-scale pretraining. ScaleDrive is developed and evaluated for research purposes and is not intended for direct deployment in safety-critical autonomous-driving systems without further validation. Although large-scale video pretraining can improve representation quality and downstream planning, it does not eliminate failures arising from rare scenarios, distribution shifts, or imperfect downstream policies. We therefore encourage thorough evaluation of safety and robustness before applying the learned representations in real-world autonomous-driving systems.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the ScaleDrive pretraining procedure, data processing pipeline, masking strategy, and downstream evaluation protocols throughout the main paper and Appendix. The complete pretraining configuration, including model architecture, optimization settings, data scales, and OpenDV video preprocessing, is reported in Appendix C.1. The architectures and training settings for frozen-encoder representation probes are provided in Appendix C.5. We further specify the training and evaluation configurations used for end-to-end planning C.2, cross-policy transfer C.3, and cross-benchmark generalization C.4 in the appendix. These details are provided to facilitate independent reproduction of the reported experiments.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, et al. Vavim and vavam: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672*, 2025.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*, 2025.

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, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Pranav Singh Chib and Pravendra Singh. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*, 9(1):103–118, 2023.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Xingtai Gui, Tengteng Huang, Haonan Shao, Haotian Yao, and Chi Zhang. Fiptr: A simple yet effective transformer framework for future instance prediction in autonomous driving. In *European conference on computer vision*, pp. 19–35. Springer, 2024.

Xingtai Gui, Jianbo Zhao, Wencheng Han, Jikai Wang, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Trajdiff: End-to-end autonomous driving without perception annotation. *arXiv preprint arXiv:2512.00723*, 2025.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. ipad: Iterative proposal-centric end-to-end autonomous driving. *IEEE Robotics and Automation Letters*, 2026.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232. IEEE, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *International Journal of Computer Vision*, 134(9):407, 2026.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, et al. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32058–32069, 2026.

Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Largead: Large-scale cross-sensor data pretraining for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jingyu Li, Junjie Wu, Dongnan Hu, Xiangkai Huang, Bin Sun, Zhihui Hao, Xianpeng Lang, Xiatian Zhu, and Li Zhang. Sgdrive: Scene-to-goal hierarchical world cognition for autonomous driving. *arXiv preprint arXiv:2601.05640*, 2026a.

Yingwei Li, Xin Huang, Yang Liu, Yang Fu, Alex Zihao Zhu, Chen Song, Junwen Yao, Anant Subramanian, Hao Xiang, Weijing Shi, et al. Stellar: Scaling 3d perception large models for autonomous driving. *arXiv preprint arXiv:2605.20390*, 2026b.

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. In *International Conference on Learning Representations*, volume 2026, pp. 7890–7911, 2026c.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024a.

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 lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024b.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047. IEEE, 2025a.

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 133(3):1352–1374, 2025b.

Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. Mgmap: Mask-guided learning for online vectorized hd map construction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14812–14821. IEEE, 2024.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pp. 531–548. Springer, 2022.

Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10978–10997, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2596–2607, 2025.

Long Nguyen, Micha Fauth, Bernhard Jaeger, Daniel Dauner, Maximilian Igl, Andreas Geiger, and Kashyap Chitta. Lead: Minimizing learner-expert asymmetry in end-to-end driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39775–39785, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chen Shi, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. Drivex: Omni scene modeling for learning generalizable world knowledge in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28599–28609. IEEE, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Matthew Strong, Wei-Jer Chang, Quentin Herau, Jiezhi Yang, Yihan Hu, Chensheng Peng, and Wei Zhan. Learning to drive is a free gift: Large-scale label-free autonomy pretraining from unposed in-the-wild videos. *arXiv preprint arXiv:2602.22091*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving. *arXiv preprint arXiv:2603.29163*, 2026.

Haochen Tian, Tianyu Li, Haochen Liu, Jiazhi Yang, Yihang Qiu, Guang Li, Junli Wang, Yinfeng Gao, Zhang Zhang, Liang Wang, et al. Simscale: Learning to drive via real-world simulation at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36365–36374, 2026.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in neural information processing systems*, 36:64318–64330, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, et al. Latent-wam: Latent world action modeling for end-to-end autonomous driving. *arXiv preprint arXiv:2603.24581*, 2026a.

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*, 2026b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xiyang Wang, Xinlin Wang, Tingguang Zhou, Gong Chen, Xingtai Gui, Zhi Xu, Xiaolei Wu, Feiyang Tan, Hangning Zhou, and Mu Yang. Chainflow-vla: Causal flow planning with vision-language models. *arXiv preprint arXiv:2605.23270*, 2026c.

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024b.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21672–21683. IEEE, 2023.

Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026.

Tianyi Yan, Tao Tang, Xingtai Gui, Yongkang Li, Jiasen Zheng, Weiyao Huang, Lingdong Kong, Wencheng Han, Xia Zhou, Xueyang Zhang, et al. Ad-r1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1085–1095, 2026.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17830–17839. IEEE, 2023.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14662–14672. IEEE, 2024.

Jiwei Yang, Zhengxian Chen, Chaosheng Huang, and Jun Li. Auto-jepa: A latent world model of continuous intent for end-to-end autonomous driving. *arXiv preprint arXiv:2607.29031*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11910–11918, 2026.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27220–27230. IEEE, 2025.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9399–9409. IEEE, 2023.

Yupeng Zheng, Pengxuan Yang, Zhongpu Xia, Qichao Zhang, Yuhang Zheng, Songen Gu, Bu Jin, Teng Zhang, Ben Lu, Chao Han, et al. Data scaling laws for imitation learning-based end-to-end autonomous driving. *arXiv preprint arXiv:2412.02689*, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642. IEEE, 2025.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13782–13790, 2026a.

Zewei Zhou, Tianhui Cai, Seth 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. *Advances in Neural Information Processing Systems*, 38: 27920–27956, 2026b.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

A Theoretical Analysis

A.1 Dynamic-System Formulation and Assumptions

The main-text analysis abstracts a driving video as a sequence of observations generated by an underlying latent dynamic system,

$$\mathbf{s}_{t+1} = F(\mathbf{s}_t) + \mathbf{w}_t, \qquad \mathbf{y}_t = G(\mathbf{s}_t) + \mathbf{v}_t, \tag{16}$$

where $\mathbf{s}_t \in \mathbb{R}^{d_s}$ denotes the latent physical state of the driving scene, including ego motion, surrounding-agent states, and local scene geometry and $\mathbf{y}_t$ denotes a temporally indexed latent observation corresponding to the predictive-learning target, and $\mathbf{w}_t$ and $\mathbf{v}_t$ denote process and observation noise, respectively. This formulation is an analytical abstraction rather than an exact model of the neural architecture: in particular, we do not assume that the learned representation explicitly corresponds to $\mathbf{s}_t$ or that the Transformer target representation depends exclusively on the physical state at a single instant.

For a state perturbation $\delta\mathbf{s}_t$ around the observed trajectory, a first-order Taylor expansion gives

$$F(\mathbf{s}_t + \delta\mathbf{s}_t) \approx F(\mathbf{s}_t) + \left.\frac{\partial F}{\partial \mathbf{s}}\right|_{\mathbf{s}_t} \delta\mathbf{s}_t, \qquad G(\mathbf{s}_t + \delta\mathbf{s}_t) \approx G(\mathbf{s}_t) + \left.\frac{\partial G}{\partial \mathbf{s}}\right|_{\mathbf{s}_t} \delta\mathbf{s}_t. \tag{17}$$

Defining

$$\mathbf{A}_t = \left.\frac{\partial F}{\partial \mathbf{s}}\right|_{\mathbf{s}_t}, \qquad \mathbf{H}_t = \left.\frac{\partial G}{\partial \mathbf{s}}\right|_{\mathbf{s}_t}, \tag{18}$$

yields the local linear model used in the main text,

$$\delta\mathbf{s}_{t+1} = \mathbf{A}_t\delta\mathbf{s}_t + \mathbf{w}_t, \qquad \delta\mathbf{y}_t = \mathbf{H}_t\delta\mathbf{s}_t + \mathbf{v}_t. \tag{19}$$

The analysis relies on the following local assumptions.

Assumption A1 (Local smoothness). F and G are sufficiently smooth in a neighborhood of the observed trajectory over a short video clip, such that the first-order Taylor approximation is locally valid. Higher-order approximation residuals are absorbed into the disturbance terms used in the subsequent analysis.

Assumption A2 (Local time invariance). Over the short temporal window considered in Proposition 2, the Jacobians vary sufficiently slowly that $\mathbf{A}_t \approx \mathbf{A}$ and $\mathbf{H}_t \approx \mathbf{H}$. This assumption is introduced only to obtain a compact local observability analysis using powers of $\mathbf{A}$; it does not imply that real driving dynamics are globally linear or time invariant.

Assumption A3 (Local spatial redundancy). Proposition 1 assumes the existence of a mapping $\mathbf{C}$ such that $\|\mathbf{H}_M - \mathbf{C}\mathbf{H}_V\|_2 \le \epsilon_s$ for the analyzed frame. This is a sufficient local condition under which the masked component can be approximately inferred from the visible component of the same temporal observation. It does not assume that every masked region in every driving scene can be reconstructed from same-frame information.

Assumption A4 (Local temporal observability). For the local linear system in Proposition 2, the observability matrix satisfies $\mathrm{rank}(\mathbf{O}_K) = d_s$ over the analyzed temporal window. This does not require global observability of all driving scenes; it only states that the state directions relevant to the local analysis are sufficiently constrained by the available temporal observations.

A.2 Proof of Proposition 1

Let the visible and masked components of the latent observation at time t be

$$\delta\mathbf{y}_t^V = \mathbf{H}_V\delta\mathbf{s}_t + \mathbf{v}_t^V, \qquad \delta\mathbf{y}_t^M = \mathbf{H}_M\delta\mathbf{s}_t + \mathbf{v}_t^M. \tag{20}$$

Consider the same-frame predictor $\widehat{\delta\mathbf{y}_t^M} = \mathbf{C}\delta\mathbf{y}_t^V$. The prediction residual can be written as

$$\delta\mathbf{y}_t^M - \mathbf{C}\delta\mathbf{y}_t^V = (\mathbf{H}_M - \mathbf{C}\mathbf{H}_V)\,\delta\mathbf{s}_t + \left(\mathbf{v}_t^M - \mathbf{C}\mathbf{v}_t^V\right). \tag{21}$$

Applying the triangle inequality and the induced matrix norm gives

$$\left\|\delta\mathbf{y}_t^M - \mathbf{C}\delta\mathbf{y}_t^V\right\| \le \|\mathbf{H}_M - \mathbf{C}\mathbf{H}_V\|_2 \|\delta\mathbf{s}_t\| + \left\|\mathbf{v}_t^M - \mathbf{C}\mathbf{v}_t^V\right\|. \tag{22}$$

Under Assumption A3, $\|\mathbf{H}_M - \mathbf{C}\mathbf{H}_V\|_2 \le \epsilon_s$, and therefore

$$\left\|\delta\mathbf{y}_t^M - \mathbf{C}\delta\mathbf{y}_t^V\right\| \le \epsilon_s \|\delta\mathbf{s}_t\| + \left\|\mathbf{v}_t^M - \mathbf{C}\mathbf{v}_t^V\right\|, \tag{23}$$

which is the bound stated in Proposition 1.

Crucially, neither the same-frame predictor nor its error bound contains the temporal transition operator $\mathbf{A}$. Therefore, whenever the spatial-redundancy condition in Assumption A3 approximately holds, a low masked-prediction error can in principle be achieved without explicitly exploiting temporal transition dynamics. Proposition 1 should therefore be interpreted as establishing the existence of a possible same-frame shortcut under spatial redundancy, rather than claiming that all patch-level prediction is solved through such a shortcut.

A.3 Proof of Proposition 2

Under Assumptions A1 and A2, consider the local linear dynamics over the temporal window $[t-K, t]$,

$$\delta\mathbf{s}_{\ell+1} = \mathbf{A}\delta\mathbf{s}_\ell + \mathbf{w}_\ell, \qquad \delta\mathbf{y}_\ell = \mathbf{H}\delta\mathbf{s}_\ell + \mathbf{v}_\ell. \tag{24}$$

Starting from $\delta\mathbf{s}_{t-K}$, recursively applying the transition dynamics gives, for $j = 0, \dots, K-1$,

$$\delta\mathbf{s}_{t-K+j} = \mathbf{A}^j\delta\mathbf{s}_{t-K} + \boldsymbol{\xi}_j, \tag{25}$$

where $\boldsymbol{\xi}_j$ aggregates the accumulated process disturbances and local approximation residuals up to the j-th step.

The corresponding observations satisfy

$$\delta\mathbf{y}_{t-K+j} = \mathbf{H}\mathbf{A}^j\delta\mathbf{s}_{t-K} + \mathbf{H}\boldsymbol{\xi}_j + \mathbf{v}_{t-K+j}. \tag{26}$$

Stacking the K preceding observations yields

$$\mathbf{Y}_K = \begin{bmatrix} \delta\mathbf{y}_{t-K} \\ \delta\mathbf{y}_{t-K+1} \\ \vdots \\ \delta\mathbf{y}_{t-1} \end{bmatrix} = \mathbf{O}_K\delta\mathbf{s}_{t-K} + \boldsymbol{\eta}, \tag{27}$$

where

$$\mathbf{O}_K = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}\mathbf{A} \\ \vdots \\ \mathbf{H}\mathbf{A}^{K-1} \end{bmatrix}, \tag{28}$$

and $\boldsymbol{\eta}$ aggregates process disturbances, observation noise, and local-linearization residuals.

Under Assumption A4, $\mathrm{rank}(\mathbf{O}_K) = d_s$, so $\mathbf{O}_K$ has full column rank and $\mathbf{O}_K^\dagger\mathbf{O}_K = \mathbf{I}$. The least-squares estimate of the initial state in the local window is therefore

$$\hat{\mathbf{s}}_{t-K} = \mathbf{O}_K^\dagger\mathbf{Y}_K. \tag{29}$$

Substituting the stacked observation model gives

$$\hat{\mathbf{s}}_{t-K} - \delta\mathbf{s}_{t-K} = \mathbf{O}_K^\dagger\boldsymbol{\eta}. \tag{30}$$

Taking norms and using $\|\mathbf{O}_K^\dagger\|_2 = 1/\sigma_{\min}(\mathbf{O}_K)$ yields

$$\|\hat{\mathbf{s}}_{t-K} - \delta\mathbf{s}_{t-K}\| \le \frac{\|\boldsymbol{\eta}\|}{\sigma_{\min}(\mathbf{O}_K)}. \tag{31}$$

For the completely masked target instant t, write the local target observation as

$$\delta\mathbf{y}_t = \mathbf{H}\mathbf{A}^K \delta\mathbf{s}_{t-K} + \mathbf{r}_t, \tag{32}$$

where $\mathbf{r}_t$ collects accumulated process disturbances, target-observation noise, and local-linearization residuals, and satisfies $\|\mathbf{r}_t\| \le \epsilon_{\mathrm{dyn}}$.

Using the recovered state, define

$$\widehat{\delta\mathbf{y}}_t = \mathbf{H}\mathbf{A}^K \hat{\mathbf{s}}_{t-K}. \tag{33}$$

The prediction residual is then

$$\widehat{\delta\mathbf{y}}_t - \delta\mathbf{y}_t = \mathbf{H}\mathbf{A}^K \left(\hat{\mathbf{s}}_{t-K} - \delta\mathbf{s}_{t-K}\right) - \mathbf{r}_t. \tag{34}$$

Applying the triangle inequality and the state-estimation bound gives

$$\left\|\widehat{\delta\mathbf{y}}_t - \delta\mathbf{y}_t\right\| \le \frac{\|\mathbf{H}\mathbf{A}^K\|_2}{\sigma_{\min}(\mathbf{O}_K)}\|\boldsymbol{\eta}\| + \epsilon_{\mathrm{dyn}}, \tag{35}$$

which is the bound stated in Proposition 2.

This result establishes a recoverability statement rather than a learning guarantee. Once the complete observation at the target temporal index is removed, the direct same-frame solution characterized by Proposition 1 is unavailable. Under local observability, the missing observation can nevertheless be inferred from information distributed across preceding temporal observations through the local transition dynamics. The analysis therefore motivates temporal-slice masking rather than guaranteeing that a finite-capacity neural network recovers the complete physical state.

B Extended Related Works

B.1 Vision Representation for Autonomous Driving

Visual representation learning is a central component of end-to-end autonomous driving (Chen et al., 2024a; Chib & Singh, 2023; Ma et al., 2024). Early approaches typically construct task-specific intermediate representations, such as bird's-eye-view (BEV) features (Li et al., 2024b; Huang et al., 2021; Gui et al., 2024; Yang et al., 2023; Liao et al., 2025a; Gui et al., 2025) for spatially aligned scene understanding and occupancy representations (Tian et al., 2023; Zhang et al., 2023; Wei et al., 2023) for modeling dense 3D structure and free-space geometry. More recently, the emergence of large-scale pretrained vision and video foundation models (Oquab et al., 2023; Tschannen et al., 2025; Siméoni et al., 2025; Wan et al., 2025; Yang et al., 2025) and vision-language models (Wang et al., 2024a; Chen et al., 2024b; Bai et al., 2025) have motivated the use of generic foundation-model features as stronger visual priors for downstream driving (Kirby et al., 2026; Gui et al., 2026; Xiong et al., 2026; Yan et al., 2026; Jiang et al., 2026; Zhang et al., 2025). Along this direction, video-based joint-embedding predictive architecture (Bardes et al., 2024; Assran et al., 2025) has recently been adapted to autonomous driving. Drive-JEPA (Wang et al., 2026b) uses self-supervised pretraining on driving videos to learn planning-oriented visual representations, while Auto-JEPA (Yang et al., 2026) extends joint-embedding prediction toward action-oriented future-intent modeling. However, it remains underexplored whether video representation learning with only action-free driving video itself can be systematically scaled and transferred to end-to-end planning.

B.2 Scaling in Autonomous Driving

Recent works have investigated how autonomous driving policies scale with increasing amounts of action-supervised driving data (Naumann et al., 2025; Zheng et al., 2024), showing that larger training sets can yield consistent gains in end-to-end driving performance. Subsequent studies further improve this scaling behavior through large-scale simulation and world-model supervision (Tian et al., 2026; Li et al., 2026c), but the resulting policies still rely on action supervision during training. In parallel, several works explore scaling representation learning for autonomous driving (Kong et al., 2025; Li et al., 2026b), demonstrating the benefits of larger-scale data and models for perception tasks. However, whether action-free driving representations exhibit similar scaling behavior, and whether such gains transfer to downstream planning, remains underexplored. We address this gap by scaling video latent prediction on large-scale action-free driving videos and evaluating the resulting frozen representations across diverse planning architectures and benchmarks.

Table 6: **ScaleDrive video latent prediction pretraining configuration.**

Category	Hyperparameter	Value
Model	Backbone	ViT-Large with 3D RoPE
	Resolution	256×256
	Frames per clip	8
	Patch size	16×16
	Tubelet size	2
	Predictor dimension	256
	Downstream checkpoint	EMA target encoder
Masking	Masking strategy	Multi-group patch + temporal-slice
	Small-patch mask	8 groups, spatial scale 0.15
	Large-patch mask	2 groups, spatial scale 0.70
	Temporal-slice mask	1 frame
Training	Trainable parameters	304.148M
	Optimizer	AdamW
	Start learning rate	5.25×10^{-4}
	Final learning rate	1×10^{-6}
	Number of GPUs	$16 \times$ NVIDIA A800
	Global batch size	4096
	Training epochs	20
	Iterations per epoch	300
	EMA momentum	0.99925
Data	Dataset	OpenDV-YouTube
	Dataset size	0.5M / 1M / 3M / 6M

C Implementation Details

C.1 ScaleDrive Video latent prediction pretraining.

ScaleDrive follows the general video latent prediction framework of V-JEPA2. The encoder adopts a ViT-Large architecture with 3D RoPE and is initialized from the official V-JEPA2 ViT-L checkpoint. Each input consists of eight frames sampled at 2 Hz and resized to 256×256 pixels, with a spatial patch size of 16×16 and a temporal tubelet size of 2. Multi-group patch masking is combined with temporal-slice masking during video latent prediction, while the target encoder is updated by exponential moving average with a momentum coefficient of 0.99925. Pretraining uses AdamW with a starting learning rate of 5.25×10^{-4} and a global batch size of 4096. After pretraining, the EMA target encoder is retained as the visual representation encoder for representation probing and downstream end-to-end planning.

The action-free pretraining data are constructed from the OpenDV-YouTube subset introduced by (Yang et al., 2024). Each raw video is segmented using a 4-second sliding window with a temporal stride of 0.5 seconds, resulting in a 3.5-second overlap between consecutive clips. Each clip is subsequently sampled at 2 Hz to obtain eight input frames. For data-scaling experiments, we construct four pretraining subsets containing 0.5M, 1M, 3M, and 6M video clips. The 6M-clip subset is derived from approximately 1157 hours of raw driving videos, while the 0.5M-, 1M-, and 3M-clip subsets span 72, 131, and 587 raw driving videos, respectively. The model architecture, optimization configuration, and training budget are kept identical across all scales, providing a controlled setting where the amount of pretraining data is the only experimental variable. Table 6 summarizes the detailed pretraining configuration.

C.2 Downstream planning on NAVSIM.

We keep the pretrained video encoder frozen during downstream planning training. The visual input consists of four history frames from the front camera and the current frame from each surround camera, with all images resized to 256×256 pixels. We evaluate two planner variants. The single-

Table 7: **ScaleDrive planning training configuration on NAVSIM.**

Category	Hyperparameter	Value
Model	Backbone	ViT-Large with 3D RoPE
	Resolution	256×256
	Front-view frames	4
	Surround-view frames	1
	Patch size	16×16
	Tubelet size	2
Planning	Predicted poses	8
	Planning horizon	4 s
Training	Trainable parameters	17.575M
	Optimizer	AdamW
	Learning rate	2.0×10^{-4}
	Number of GPUs	$8 \times$ NVIDIA A800
	Global batch size	64
Data	Dataset	NAVSIM navtrain
	Dataset size	85K

Table 8: **Cross-planner configurations.** V-JEPA2 and ScaleDrive use the same integration strategy within each planner.

Planner	Backbone	Decoder	Video representation integration
LTF	TransFuser	Regression	Attention keys and values
DiffusionDrive	TransFuser	Diffusion	Attention keys and values
DrivoR	DINOv2	Regression	Front-view feature replacement
ChainFlow	DINOv2	Autoregressive	Front-view feature replacement

modal variant uses a single learnable trajectory query to predict one trajectory. The multimodal variant uses 64 learnable proposal queries to predict 64 candidate trajectories. Each predicted trajectory comprises eight poses over a 4-second planning horizon. The single-modal and multimodal variants achieve PDMS scores of 89.0 and 93.8 on NAVSIM, respectively. The single-modal planner is trained solely with an L1 trajectory loss:

$$\mathcal{L}_{\text{single}} = \mathcal{L}_{\text{traj}}. \tag{36}$$

The multimodal planner additionally uses a score loss to supervise candidate scoring:

$$\mathcal{L}_{\text{multi}} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{score}} \tag{37}$$

Both variants are trained on 85K NAVSIM navtrain trainset using AdamW with a base learning rate of 2.0×10^{-4}. The single-modal model is trained with 100 epochs, while the multimodal model is trained with 20 epochs. Training is performed on eight NVIDIA A800 GPUs with a global batch size of 64. Table 7 summarizes the training configurations.

C.3 Cross-Planner Configurations

We evaluate video representation transfer with four downstream planners: LTF Chitta et al. (2022), DiffusionDrive Liao et al. (2025a), DrivoR Kirby et al. (2026), and ChainFlow Wang et al. (2026c). All configurations use multi-view images. For each planner architecture, we compare its original baseline with V-JEPA2 and ScaleDrive variants. Table 8 summarizes the native architecture and the shared video representation integration strategy. For each planner, we retain the training losses, optimizer, learning-rate schedule, and training data used in its original paper. These settings are kept identical across the original baseline, V-JEPA2, and ScaleDrive variants within each planner.

Additional feature fusion: LTF and DiffusionDrive. LTF and DiffusionDrive use a Latent TransFuser backbone to encode multi-view images into a BEV representation, followed by regression and diffusion trajectory decoders. For both planners, the video-enhanced variants retain the

Table 9: **Training configuration on Bench2Drive.**

Category	Hyperparameter	Value
Model	Backbone	ViT-Large with 3D RoPE
	Resolution	256×256
	Frames per clip	4
	Patch size	16×16
	Tubelet size	2
Planning	Planner architecture	TFv6
	Image backbone	RegNetY-032
	LiDAR backbone	RegNetY-032
	Input resolution per camera	384×384
Training	Optimizer	AdamW
	Learning rate	3.0×10^{-4}
	Weight decay	0.01
	Global batch size	48
	Training epochs	30
Data	Dataset	LEAD
	Dataset size	960K

native backbone features and introduce video representations as additional attention keys and values in the planner. This allows the planner to attend to the video representation alongside its original visual features.

Front-view feature replacement: DrivoR and ChainFlow. DrivoR and ChainFlow use a fine-tuned DINOv2 backbone to encode multi-view images, with regression and autoregressive trajectory decoders, respectively. For these planners, we replace only the front-view DINOv2 features with corresponding video representations. For these two planners, video features replace the native front-view representation rather than serving as an additional feature branch.

C.4 Cross-Benchmark Configurations

HUGSIM. We directly evaluate the LTF and ChainFlow planners trained with the NAVSIM training set, including their original, V-JEPA2, and ScaleDrive variants, on HUGSIM without additional training or fine-tuning. All variants are evaluated under the same HUGSIM protocol. This setting assesses the cross-benchmark generalization of the trained driving policies without adaptation to the target environment.

Bench2Drive. We evaluate ScaleDrive on Bench2Drive using TFv6 Nguyen et al. (2026) as the downstream planner. The pretrained ScaleDrive encoder remains frozen and processes four-frame clips at a resolution of 256×256. TFv6 uses RegNetY-032 backbones for both image and LiDAR inputs, with camera images resized to 384×384. We train the planner on the LEAD dataset for 30 epochs using AdamW with a learning rate of 3.0×10^{-4}, a weight decay of 0.01, and a global batch size of 48. Table 9 summarizes the detailed training configuration.

C.5 ScaleDrive representation probing

The pretrained visual representation is evaluated through two complementary probing tasks: monocular depth estimation for spatial geometry and inverse-dynamics ego-motion estimation for temporal dynamics. Both tasks use the same pretrained ViT-Large video encoder. The encoder remains frozen throughout probe training, while only the task-specific probe is optimized.

Depth Estimation. We evaluate spatial representation quality through monocular depth estimation on KITTI. The frozen spatiotemporal features are first aggregated along the temporal dimension by mean pooling. The resulting spatial features are passed through a shallow depth head consisting of BatchNorm and a linear projection, followed by bilinear upsampling and a Softplus activation to

Table 10: **Training configuration for ScaleDrive representation probing.**

Category	Hyperparameter	Depth Estimation	Ego-Motion Estimation
Model	Backbone	ViT-Large with 3D RoPE	ViT-Large with 3D RoPE
	Resolution	256×256	256×256
	Tubelet size	2	2
	Encoder token grid	$2 \times 16 \times 16$	$2 \times 16 \times 16$
Probe	Prediction target	Depth map	Waypoints
	Feature aggregation	Temporal mean pooling	3 learnable queries
	Probe architecture	BatchNorm + Linear $1024 \to 1$	Cross-attention block
	Output head	Bilinear upsampling + Softplus	LayerNorm + MLP $1024 \to 1024 \to 2$
Training	Optimizer	AdamW	AdamW
	Learning rate	3.0×10^{-4}	3.0×10^{-4}
	Weight decay	0.05	0.05
	Global batch size	16	16
	Training epochs	10	10
Data	Dataset	KITTI	NAVSIM

Table 11: **Comparison with NAVSIM navtrain dataset pretraining.**

Pretraining Data	NC ↑	DAC ↑	EP ↑	TTC ↑	PDMS ↑
w/o. Pretrain	99.0	98.0	88.4	96.5	92.4
NAVSIM	98.6	98.1	90.4	95.2	92.7
OpenDV	**99.3**	**98.4**	**90.6**	**97.0**	**93.8**

produce dense depth predictions at the input resolution. Table 10 summarizes the probe architectures and optimization settings.

Ego-Motion Estimation. We evaluate temporal dynamics through inverse-dynamics ego-motion estimation on NAVSIM. Learnable queries aggregate the frozen spatiotemporal tokens through a cross-attention block. The resulting query features are normalized and decoded by an MLP to predict the ego trajectory associated with the observed video sequence. This probe assesses how readily motion information can be extracted from the frozen video representation.

D More Experiments

Comparison with NAVSIM pretraining. To further disentangle the effect of large-scale action-free pretraining, ScaleDrive is additionally pretrained on NAVSIM driving videos under the same model architecture, optimization configuration, and downstream planning protocol. This comparison directly contrasts broad open-world driving-video pretraining with benchmark-specific driving data. As shown in Table 11, pretraining on NAVSIM achieves a PDMS of 92.7, whereas scaling pretraining to 6M OpenDV clips improves overall planning performance. These results show that the benefit of ScaleDrive does not depend on pretraining with downstream benchmark data, but instead increases with large-scale action-free driving-video pretraining.

Detailed scaling results Table 12 provides the complete metric breakdown across different pretraining data scales. As the amount of pretraining data increases from 0.5M to 6M clips, PDMS improves monotonically from 92.77 to 93.61. The improvement is accompanied by a consistent increase in EP and generally stronger TTC, while NC and DAC remain at a similarly high level. These detailed results further show that the planning benefit from scaling action-free video pretraining is reflected across multiple components of the closed-loop evaluation rather than a single aggregate metric.

Table 12: **Detailed data-scaling results on NAVSIM.**

Method	Pretraining Data	NC ↑	DAC ↑	EP ↑	TTC ↑	PDMS ↑
V-JEPA2	–	99.03	98.18	88.62	96.10	92.40
ScaleDrive	0.5M	98.99	97.92	89.65	96.12	92.77
	1M	99.10	97.99	89.78	96.57	93.05
	3M	99.14	98.10	90.32	96.51	93.30
	6M	**99.16**	**98.28**	**90.67**	**96.58**	**93.61**

Table 13: **Ablation on temporal-slice masking.** All configurations share the same pretraining data and training settings. Ego ADE/FDE are measured with the frozen-encoder ego-motion probe.

Masking	PDMS↑	EPDMS↑	Ego ADE↓	Ego FDE↓
Patch only	93.61	89.96	0.193	0.284
Patch + Temporal-slice	**93.79**	**90.48**	**0.172**	**0.246**

Ablations on temporal-slice masking. Table 13 compares patch-only masking with the full patch- and temporal-slice masking under identical pretraining data, backbone, training steps, and optimization configuration, evaluated with the same downstream planner. Beyond closed-loop planning (PDMS), we report the frozen-encoder ego-motion probe, since Proposition 2 concerns temporal state inference rather than planning performance alone.



Figure 6: **Visual representation comparison.**

Additional qualitative visualizations Additional qualitative results are provided to complement the visual analyses in the main paper. Fig. 6 presents more PCA visualizations of the learned visual representations across diverse driving scenes, providing a broader view of the feature patterns produced by ScaleDrive. Fig. 7 further presents additional planning examples across different driving environments in NAVSIM and HUGSIM.



(A)



(B)

Figure 7: **ScaleDrive planning results.(A) NAVSIM and (B) HUGSIM.**

E LIMITATIONS

ScaleDrive focuses on learning transferable representations from monocular driving videos and does not explicitly model cross-view correspondences during pretraining. Although the frozen representations can be integrated into multi-camera planners, extending latent prediction to synchronized views may further improve spatial consistency across cameras. In addition, our experiments use a fixed temporal sampling rate and clip length, leaving adaptive temporal sampling and longer-context pretraining unexplored.