Learning to Drive is a Free Gift: Large-Scale Label-Free Autonomy Pretraining from Unposed In-The-Wild Videos

Matthew Strong[1,2] Wei-Jer Chang[1,3] Quentin Herau[1] Jiezhi Yang[1] Yihan Hu[1]
Chensheng Peng[1,3] Wei Zhan[1,3]
[1]Applied Intuition [2]Stanford University [3]UC Berkeley
https://lfg-ai.github.io/

Abstract

*Ego-centric driving videos available online provide an abundant source of visual data for autonomous driving, yet their lack of annotations makes it difficult to learn representations that capture both semantic structure and 3D geometry. We propose **LFG** (**L**earning to drive is a **F**ree **G**ift), a label-free, teacher-guided framework that learns geometry-, motion-, and semantics-aware representations directly from unposed, single-view YouTube driving videos. LFG extends a feedforward 3D reconstruction backbone with a lightweight causal autoregressive module and multi-modal teacher supervision to jointly predict current* and *short-horizon future point maps, camera poses, semantic segmentation, confidence maps, and motion masks—forming a unified pseudo-4D representation learned entirely without poses, labels, or LiDAR. On the NAVSIM planning benchmark, LFG achieves state-of-the-art performance using only a single front camera, outperforming multi-camera and LiDAR baselines while exhibiting strong data efficiency: with only 10% labeled data, LFG matches the full-data performance of DINOv3. It further transfers effectively to semantic segmentation, depth estimation, and trajectory prediction, positioning LFG as a compelling video-centric foundation model for autonomous driving.*

1. Introduction

In-the-wild, ego-centric driving videos are abundantly available online, yet their lack of annotations makes it challenging to learn representations that encode semantic, temporal, and geometric structure. Inspired by the success of large-scale self-supervised pretraining in vision and language [11, 13], a natural question arises: *can we leverage raw driving video to learn geometry- and motion-aware features for autonomy?*

Most autonomy approaches rely heavily on labeled data—expert trajectories, LiDAR scans, and semantic annotations [2, 5]. Prior self-supervised methods [16, 19] use frame-to-frame consistency losses that implicitly assume static scenes, limiting their ability to model dynamic agents. Feedforward 3D reconstruction models [14, 15] show that point maps and ego-motion can be regressed from unposed sequences in a single forward pass, but focus on present-frame reconstruction without modeling future dynamics.

We introduce **LFG**, a label-free, teacher-guided pretraining framework that addresses these limitations. Motivated by findings that humans make low-level driving decisions from only a short motion history, LFG extends the π^3 [15] feedforward backbone with a causal autoregressive transformer to predict future geometry, motion, and semantics. Multi-modal teachers—π^3 for geometry, SegFormer [17] for semantics, and Grounded SAM2 [12] + CoTracker3 [7] for motion—provide pseudo-supervision on unlabeled OpenDV [18] YouTube videos, enabling LFG to learn a unified *pseudo-4D* representation without poses, labels, or LiDAR.

Unlike large world models requiring supervised labels [3, 4], LFG uses a short-horizon, feedforward formulation producing features directly useful for planning. On NAVSIM [2], LFG achieves state-of-the-art planning with *only a single front-camera view*, outperforming multi-view and BEV-based methods such as UniAD [6] and Hydra-MDP [8]. With 10% labeled data it matches full-data DINOv3, and it transfers effectively to semantic, geometric, and motion tasks.

Contributions.

- We propose **LFG**, a label-free video-centric pretraining framework that learns geometry-, motion-, and semantics-aware representations from unposed, single-view driving videos.
- We design a unified architecture combining a pretrained

π^3 encoder with a causal autoregressive module for short-horizon prediction of point maps, camera poses, semantic layouts, confidence maps, and motion masks.
- We demonstrate state-of-the-art planning on NAVSIM with a single front camera, compelling data efficiency, and strong transfer across semantic, geometric, and motion tasks.

2. Method

2.1. Problem Formulation

Given N ego-centric RGB frames $(I_t)_{t=1}^{N}$, LFG predicts outputs for all $N{+}M$ frames (observed + future): **point maps** $P_t \in \mathbb{R}^{3\times H\times W}$, **camera poses** $T_t \in \mathbb{R}^{4\times 4}$, **semantic segmentation** $S_t \in \mathbb{R}^{7\times H\times W}$ (road, vehicle, pedestrian, building, vegetation, sky, background), **confidence maps** $C_t \in [0,1]^{H\times W}$, and **motion masks** $M_t \in [0,1]^{H\times W}$. The full output is $\mathcal{O} = \{(P_t, T_t, S_t, C_t, M_t)_{t=1}^{N+M}\}$. All modalities are learned jointly from video with teacher-guided supervision, promoting shared representations of geometry, semantics, and motion relevant to autonomous driving.

2.2. Architecture

LFG builds on π^3 [15], a $\sim$1B-parameter feedforward model with a DINOv2-pretrained backbone that predicts point maps, confidence maps, and camera poses from unposed images, and is trained on dynamic data. We add three components:

(1) Causal autoregressive transformer. After the π^3 alternating attention encoder, we insert a 4-layer, 8-head causal transformer $\mathcal{T}_{\text{AR}}$ (dropout 0.1). Given encoder tokens $\mathbf{Z}_{1:N}$, it autoregressively rolls out future tokens: $\mathbf{Z}_{1:N+M} = \mathcal{T}_{\text{AR}}(\mathbf{Z}_{1:N})$. Each future token attends only to past and observed tokens, enforcing forward-only information flow. The newly generated $\mathbf{Z}_{N+1:N+M}$ represent latent scene features for unobserved frames, decoded into all output modalities.

(2) Semantic head. Initialized from the point decoder, it predicts dense per-pixel class probabilities for all $N{+}M$ frames, allowing the model to leverage shared geometric features for semantic prediction.

(3) Motion head. Also initialized from the point decoder, it predicts per-pixel dynamic/static classification to disentangle moving objects from the static environment.

The total model has $\sim$1.45B parameters and runs at 5–6 Hz on an NVIDIA RTX 5090.

2.3. Teacher-Guided Supervision

Geometry teacher. A frozen π^3 teacher receives all $N{+}M$ frames and outputs pseudo-GT point maps, confidence maps, and camera poses for the full sequence. The student observes only the first N frames and must predict all $N{+}M$ outputs, forcing it to learn temporal extrapolation of geometry and ego-motion from partial observations.

Semantic teacher. A pretrained SegFormer [17] (trained on Cityscapes [1]) generates soft per-pixel pseudo-labels $\hat{S}_t = \Phi_{\text{seg}}(I_t)$ for each frame. The teacher receives all frames while LFG must predict current and future semantics from only the first N inputs.

Motion pseudo-labels. We generate motion annotations in a fully label-free pipeline. Grounded SAM2 [12] segments vehicle and pedestrian instances in the first frame; CoTracker3 [7] tracks their 2D keypoints across time. Using teacher π^3 point maps, tracked pixels are backprojected into 3D and per-instance 3D displacements are measured: $d_t^{(i)} = \|\bar{\mathbf{p}}_{t+1}^{(i)} - \bar{\mathbf{p}}_t^{(i)}\|_2$. Objects with $d_t^{(i)} > \tau_{\text{motion}}$ for at least $K_{\min}$ frames are labeled dynamic, and their masks are rasterized into dense supervision targets $\mathbf{M}_t$.

2.4. Losses and Training

The total loss balances current and future objectives: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{current}} + \lambda_{\text{future}}\,\mathcal{L}_{\text{future}}$, where each term aggregates task-specific losses: weighted BCE for segmentation (with class-specific weights for imbalance), geodesic rotation + Huber translation for pose, scaled L_1 for point maps ($\alpha\|\mathbf{P} - \hat{\mathbf{P}}\|_1$), and BCE for motion masks. A temporal weight $\omega = 10$ is applied to all future-frame losses.

Training proceeds in three stages: (1) future geometry/pose, (2) semantics, (3) motion masks, each training end-to-end. We use $\sim$2M samples from OpenDV [18] across varied driving conditions, sampled at 2, 5, and 10 Hz without frequency conditioning. Training uses 32 A100 GPUs for 40k iterations with AdamW (lr $= 10^{-4}$), linear warmup (500 steps), cosine annealing, and gradient clipping at 1.0. Augmentations (color jitter, blur, grayscale) are applied to student inputs; the teacher receives clean images.

2.5. Distillation

We distill the output LFG tokens into an encoder similar to Depth Anything 3[10], with alternating local and global attention, adding in an autoregressive transformer to produce additional tokens. We only supervise on a combination of cosine and L1 loss on the tokens from the version of LFG with a longer prediction horizon, running it for 35k steps with only 16 A100 GPUs, with inference at 24Hz, 4 times faster than LFG.

2.6. Planning Fine-Tuning

The frozen LFG encoder outputs high-dimensional **autonomy tokens** from 3 consecutive front-view frames on the NAVSIM benchmark [2], including future tokens from the autoregressive module. A lightweight **anchor-based trajectory decoder** (20 anchors, 8 waypoints, 4 s horizon) uses cross-attention from autonomy features to trajectory

Table 1. **Data efficiency (PDMS↑) on NAVSIM.** All pretrained encoders use 1 front camera (3 frames) and the same decoder. **L** = LiDAR.

Method	Input	1%	10%	100%
DiffusionDrive [9]	3Cam+L	64.9	72.6	88.1
DINOv3 [13]	1Cam	60.0	75.8	81.4
PPGeo [16]	1Cam	61.5	65.6	74.6
π^3 [15]	1Cam	56.2	77.5	82.8
LFG (Ours)	1Cam	**66.3**	**81.4**	**85.2**

Table 2. **NAVSIM benchmark: LFG vs. BEV baselines.** Higher is better. **L** = LiDAR. LFG uses only a single front camera (3 frames).

Method	Input	NC	DAC	TTC	EP	PDMS
UniAD [6]	6Cam	97.8	91.9	92.9	78.8	83.4
TransFuser	3Cam+L	97.7	92.8	92.0	79.2	84.0
Hydra-MDP [8]	3Cam+L	96.9	94.0	94.0	78.7	84.7
DiffDrive [9]	3Cam+L	96.8	**95.4**	**94.7**	**82.0**	**88.1**
LFG (Ours)	1Cam	**98.2**	93.7	94.4	79.1	85.2

anchors and self-attention across modes, outputting confidence scores and coordinate offsets in a single forward pass—no diffusion or iterative refinement is needed. The highest-confidence trajectory is selected as the final plan.

3. Experiments

3.1. NAVSIM Planning

Table 1 evaluates data efficiency. LFG achieves the best PDMS across all label fractions: 66.3 at 1% (surpassing even multi-sensor DiffusionDrive at 64.9), 81.4 at 10% (*matching full-data DINOv3*), and 85.2 at 100%. This strong data efficiency stems from the encoder's temporal understanding: learning to predict future geometry, semantics, and motion produces features inherently suited for planning from short past-frame sequences. LFG surpasses both its π^3 teacher (82.8) and PPGeo [16] (74.6) at full data, showing that powerful feedforward architectures require semantic and temporal understanding of the future.

Table 2 compares LFG to BEV baselines with multi-view cameras and/or LiDAR. Despite using only one front camera, LFG achieves the best NC (98.2), competitive TTC (94.4) and EP (79.1), and overall PDMS of 85.2, demonstrating that a single-camera encoder pretrained on large-scale video can rival specialized multi-sensor systems.

3.2. Ablations

Table 3 validates each component, including the distilled encoder. Distilled *from scratch* on a smaller amount of data, the encoder performs competitively with LFG and its ablations, and outperforms other baselines in both the 1% and 10% data regime. Removing semantic and motion supervision degrades 10% PDMS by 4.3 points (81.4→77.1), confirming that multi-modal tasks provide complementary learning signals. Removing the autoregressive head drops it to 77.7, validating that future prediction is critical for temporally aware features. Doubling pretraining data boosts 1% PDMS, and extending the prediction horizon further improves both regimes, confirming that LFG benefits from data scale and temporal reasoning.

Table 3. **Component and scaling ablations (PDMS↑).**

Variant	1%	10%
Distilled Long Horizon LFG **24Hz**	72.7	78.4
LFG (default) **6Hz**	66.3	81.4
+ 2× pretraining data	76.6	82.3
+ longer prediction horizon	80.5	84.4
– Seg. & motion supervision	–	77.1
– Autoregressive head	–	77.7

Table 4. **Semantic segmentation on KITTI-360.** LFG receives 3 frames and predicts for 6; baselines receive all 6 frames.

Method	PA	mIoU	mDice	FW
SegFormer [17]	0.926	0.677	0.744	0.723
MaskFormer	0.922	0.760	0.829	0.760
LFG (Ours)	**0.947**	**0.768**	0.827	**0.770**

Table 5. **Depth (RMSE↓, m) and trajectory (ATE↓, m).** LFG uses 3 input frames; π^3/DA3 use all 6.

Method	KITTI-360	Waymo	Metric
π^3 [15]	4.37	6.68	RMSE
DA3	4.43	–	
LFG	4.38	6.87	
π^3 [15]	0.43	0.02	ATE
LFG	1.00	0.08	

3.3. Semantic Segmentation

On KITTI-360 (Table 4), LFG surpasses its SegFormer teacher in PA (0.947 vs. 0.926) and mIoU (0.768 vs. 0.677), and outperforms MaskFormer (0.760 mIoU). LFG also excels on *future* frames predicted without seeing RGB images (pred-frame mIoU: 0.751), showing that geometric pretraining provides a strong inductive bias for temporally consistent semantics.

3.4. Depth Estimation and Trajectory

Table 5 shows that for monocular depth, LFG achieves 4.38 m RMSE on KITTI-360—on par with its π^3 teacher

(4.37 m) and better than VGGT (4.46 m) and DA3 (4.43 m). On Waymo, depth remains competitive (6.87 m vs. 6.68 m). For trajectory, LFG achieves ATE of 1.00 m (KITTI-360) and 0.08 m (Waymo), slightly above π^3 which receives all 6 frames, but strong given LFG must extrapolate future camera motion from only 3 observed frames.

4. Conclusion

LFG demonstrates that large-scale, label-free pretraining on in-the-wild driving videos produces powerful representations for autonomous driving. By combining a feedforward 3D reconstruction backbone with a causal autoregressive module and multi-modal teacher supervision, LFG learns a unified pseudo-4D representation of geometry, semantics, and motion from raw YouTube videos. The resulting encoder achieves state-of-the-art NAVSIM planning with a single front camera and transfers effectively to perception tasks. Future work includes extending the temporal horizon, incorporating multi-view cues from emerging multi-camera datasets, and progressive self-bootstrapping to reduce teacher reliance.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[3] Danijar Hafner et al. Mastering diverse domains through world models. *arXiv:2301.04104 [cs.AI]*, 2023.

[4] Jonathan Ho and Tim Salimans. Video diffusion models. *arXiv:2204.03458 [cs.CV]*, 2022.

[5] Yihan Hu, Jiazhi Li, Li Chen, Chonghao Sima, Xizhou Zhu, Siqi Wang, Guan_Heng Lin, Sen Zhang, X. H. Geng, Yihang Liu, Chen Jiang, Lewei Lin, Hongyang Li, Yu Qiao, and Jifeng Dai. Planning-oriented autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page —, 2023.

[7] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.10659*, 2024.

[8] Zhenxin Li, Kailin Li, Kevin Ziglar, Sergio Zuniga, Jiachen Chen, and Jose M. Alvarez. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.07122*, 2024.

[9] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12037–12047, 2025.

[10] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[11] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[12] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[13] Oriane Siméoni et al. Dinov3: Self-supervised learning for vision at unprecedented scale. *arXiv preprint arXiv:2508.10104*, 2025.

[14] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025.

[15] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347 [cs.CV]*, 2025. v2 revised Sept. 9 2025.

[16] Penghao Wu, Li Chen, Hongyang Li, Xiaosong Jia, Junchi Yan, and Yu Qiao. Policy pre-training for autonomous driving via self-supervised geometric modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[18] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Dataset: OpenDV–YouTube.

[19] Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. Selfd: Self-learning large-scale driving policies from the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17316–17326, 2022.