

OpenVO: Open-World Visual Odometry with Temporal Dynamics Awareness

Phuc Nguyen* Anh N. Nhu* Ming C. Lin
 University of Maryland, College Park, USA
 {phucnda, anhu, lin}@umd.edu
<https://openvo.github.io/>

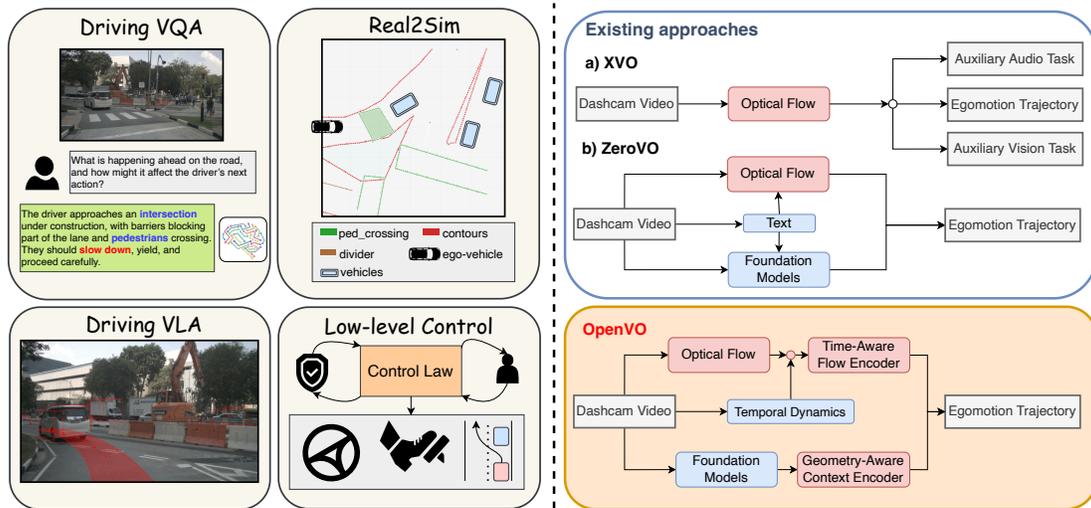


Figure 1. **Left:** Generalized Visual Odometry provides real-world ego-motion and trajectory estimates that bridge perception and control in autonomous driving. It enables scene understanding (Driving VQA [38, 48]), simulation (Real2Sim [26, 39]), action grounding (Driving VLA [17, 25]), and precise motion feedback for low-level control [7, 19, 52]. **Right:** We introduce **OpenVO**, a generalizable visual odometry framework that estimates real-world ego-motion from uncalibrated dashcam videos and remains robust across varying observation rates. Our design integrates our “Time-Aware Flow Encoder” for modeling temporal dynamics and a “Geometry-Aware Context Encoder” for extracting consistent scene geometry, enabling robust and generalizable motion estimation across diverse visual and temporal domains.

Abstract

We introduce *OpenVO*, a novel framework for Open-world Visual Odometry (VO) with temporal awareness under limited input conditions. *OpenVO* effectively estimates real-world-scale ego-motion from monocular dashcam footage with varying observation rates and uncalibrated cameras, enabling robust trajectory dataset construction from rare driving events recorded in dashcam. Existing VO methods are trained on fixed observation frequency (e.g., 10Hz or 12Hz), completely overlooking temporal dynamics information. Many prior methods also require calibrated cameras with known intrinsic parameters. Consequently, their performance degrades when (1) deployed under unseen observation frequencies or (2) applied to uncalibrated cameras. These significantly limit their generalizability to many downstream tasks, such as extracting trajectories from dashcam footage. To address these challenges, *OpenVO* (1)

explicitly encodes temporal dynamics information within a two-frame pose regression framework and (2) leverages 3D geometric priors derived from foundation models. We validate our method on three major autonomous-driving benchmarks – KITTI, nuScenes, and Argoverse 2 – achieving **more than 20%** performance improvement over state-of-the-art approaches. Under varying observation rate settings, our method is significantly more robust, achieving **46%–92%** lower errors across all metrics. These results demonstrate the versatility of *OpenVO* for real-world 3D reconstruction and diverse downstream applications.

1. Introduction

Visual Odometry (VO) [34] system is a fundamental component in autonomous driving and robotics, providing crucial information about an agent’s pose and ego-motion in world coordinates to enable comprehensive situational

*: Equal contribution

awareness and understanding. In many downstream tasks, both online (e.g. onboard perception systems [15, 28, 51]) and offline (e.g. trajectory log analysis and driving dataset construction [2]), VO must be robust to various sources of variations, including but not limited to observation sampling rates and sensor calibrations, which present significant challenges to VO models. One particularly challenging case is trajectory reconstruction from dashcam footage obtained from online platforms such as YouTube. These videos contain rich spatial and dynamical information about rare, long-tail driving events (e.g., crashes) that are otherwise difficult to collect *at scale* due to economic, legal, and ethical constraints. Dashcam videos are typically monocular and uncalibrated, exhibiting large variations in camera calibrations and poses across sources. Moreover, they could be recorded at different frame rates, introducing temporal inconsistencies that degrade model accuracy. The temporal dynamics implicitly encoded in the observation rate could be exploited to enhance VO accuracy and robustness. Prior works in VO largely overlooked this factor. Studies in sequential decision-making [33, 43] have demonstrated that such temporal inconsistencies can significantly degrade model performance when deployed on mismatched observation rates (e.g., trained on 20 Hz, yet deployed on 12 Hz).

Classical geometry-based methods [6, 12, 31, 32, 44] typically require access to accurate camera calibration at inference and struggle with uncalibrated observations. Early learning-based approaches [9, 24, 27, 45, 55] are trained and evaluated on the same datasets under generally consistent sensor calibrations and without explicit awareness of camera geometry, limiting their generalizability to substantially different sensor setups. More recent works explicitly incorporate camera geometric priors into their framework [41, 42, 47] to achieve strong VO estimation accuracy but still assume known camera intrinsics. The latest VO frameworks are more generalized via multi-modal self-training for intrinsic-free estimation [20] or geometry-aware architectures with language priors for zero-shot generalization [21]. However, all existing methods are trained and evaluated on fixed frame rate, overlooking temporal dynamics from varying observation frequencies.

Motivated by these gaps, we aim to address an important question: *How can a VO system generalize to uncalibrated observations under varying observation rates with unknown camera calibrations in open-world scenarios?* To tackle this question, we present **OpenVO**, a generalized visual odometry framework that elegantly leverages *temporal dynamic information* and *3D geometry priors* derived from inferred camera parameters to robustly estimate ego-motion from *uncalibrated monocular observations*. The core contributions of OpenVO are as follows:

1. **Temporal frequency incorporation:** OpenVO encodes video frame-rate information into a time-aware embed-

ding that conditions the optical flow features to adapt to varying temporal dynamics. This explicit modeling of temporal frequency enriches motion representation, enabling accurate and robust ego-motion estimation across diverse frame rates and video sources.

2. **Differentiable 2D-Guided 3D flow estimation:** we construct differentiable 3D flow fields from estimated 2D flow model and metric depth. These 3D flow fields are then fused with time-aware 2D flow features, further enhancing the accuracy of ego-motion estimation in world coordinates.
3. **Geometry context awareness:** we explicitly leverage inferred camera intrinsics and homogeneous projective geometry to provide geometric awareness of the scene, enabling OpenVO to generalize effectively to uncalibrated observations with diverse camera configurations.

Our framework consistently achieves state-of-the-art visual odometry performance on various large-scale benchmarks, including nuScenes [5], KITTI [14], and Argoverse-2 [50], outperforming existing state-of-the-art methods by over 20% global absolute trajectory error. Most notably, while prior methods exhibit substantial performance degradation when tested on unseen observation rates, OpenVO demonstrates significantly greater robustness against temporal dynamic variations. OpenVO demonstrates strong generalization to uncalibrated and out-of-domain observations across varying frame rates, an essential capability for extracting and analyzing high-quality trajectory data from dashcam footage. Because accurate ego-motion estimation is fundamental to robotics and autonomous driving, OpenVO can serve as an important component in a wide range of downstream tasks, including online vectorized mapping, trajectory reconstruction, 3D scene understanding, and motion forecasting.

2. Related Work

OpenVO leverages projective geometry in traditional visual odometry and geometric priors, combined with multi-time-scale learning insights, for generalizable visual odometry.

Monocular Visual Odometry. There are two primary VO approaches: (1) traditional geometry-based and (2) learning-based methods. Traditional geometry-based approaches [34, 40], such as MonoSLAM [10], ORBSLAM [6, 31, 32], SVO [12], DSO [11, 44] employ projective geometry and feature matching pipelines for precise and interpretable motion estimation under calibrated observation settings. These methods heavily depend on known camera parameters, careful initialization, and engineered feature quality or photometric assumptions, which limit their generalization to uncalibrated, texture-poor, or highly dynamic real-world scenarios. To address such limitations, end-to-end learning-based VO methods directly learn the mapping between RGB observations and frame-to-frame

motion [8, 9, 24, 45, 46]. Some hybrid approaches incorporate learned geometric priors into classical pipelines (e.g. using monocular depth modules for constraint or joint optimization), combining geometric priors with the flexibility of deep learning [27, 41, 42, 55, 57, 60]. However, these works train and evaluate their models under similar conditions (e.g., same dataset with fixed camera settings and limited scenario diversity), which restricts their generalizability to in-the-wild scenarios and uncalibrated observations with notable domain gaps (e.g. camera settings and scenarios).

Generalized Visual Odometry. TartanVO [47] incorporates an intrinsic layer and up-to-scale loss to handle scale ambiguity and varying camera configurations, achieving strong cross-dataset performance. However, TartanVO relies on ground-truth intrinsics, limiting its applicability to uncalibrated observations. DINO-VO [1] leverages the pre-trained DINOv2 [35] vision foundation model for more robust and generalizable correspondence matching in challenging scenarios. XVO [20] advances calibration-free generalization through cross-modal self-training, leveraging large-scale pseudo-labeled uncalibrated YouTube dashcam videos to improve robustness. ZeroVO [21] further introduces language and 3D geometric priors via estimated 3D flows for zero-shot generalization. Although XVO and ZeroVO achieve strong performance on unseen domains with varying sensor settings and camera parameters, they overlook the impact of frame rates, leading to degraded performance under unseen observation frequencies. In contrast, our OpenVO achieves state-of-the-art performance across datasets with diverse camera configurations and frame rates, without ground-truth camera intrinsics.

Geometric Prior Models. Geometric priors such as metric depth and camera intrinsics are crucial for 3D reconstruction and improving scale consistency in visual odometry [21, 41, 55]. Monocular foundation metric depth models [3, 4, 16, 36, 37, 37, 53, 54, 56] enable absolute-scale depth prediction with strong zero-shot generalization across diverse domains through large-scale training. Among these, some rely on known camera parameters [16, 56], while others remove this assumption and even achieve competitive performance in estimating focal lengths from monocular observation [4]. However, ground-truth camera parameters are not always available in downstream tasks. To bridge this gap, [18, 22, 61] have proposed dedicated models to estimate camera intrinsics from monocular observations. In OpenVO, we employ the WildCamera [61] and Metric3Dv2 [16] to infer camera intrinsics and metric depth for 3D flow construction, guided by the 2D optical flow from MaskFlowNet [59].

Multi-Time-Scale Learning. Despite being one of the key quantities in temporally-driven problems, including video understanding, visual odometry, sequential decision mak-

ing, and dynamical systems, observation rate (i.e., the time step size $\Delta t = 1/f$) is typically fixed during the training process and *not modeled explicitly* by prior VO methods. Specifically, in frame-to-frame visual odometry, existing state-of-the-art models [20, 21, 47] take a pair of consecutive frames (I_{t-1}, I_t) , optionally along with auxiliary conditioning modality, such as language prior c_{text} , to estimate the ego motion. These models completely overlook the impacts of the temporal interval Δt on the learning process, which might encode important temporal dynamics in visual odometry, such as velocity. As a result, they are trained under a single time scale (e.g., a fixed sampling rate of 20 Hz), leading to a subtle form of *temporal overfitting* [33] – where the learned motion representation is implicitly optimized for that specific sampling rate – and experiencing potential performance degradation when deployed under mismatched observation frequency (e.g., 12 Hz). Although these phenomena are underexplored in visual odometry and computer vision, prior studies in reinforcement learning [43] and world models [33] have shown that temporal overfitting to a fixed observation rate significantly degrades model robustness under unseen time scales. Motivated by these findings, we introduce a novel time-aware flow encoder in OpenVO and train the model across a diverse range of observation frequencies, exposing it to multiple temporal dynamic scales. This strategy effectively improves OpenVO’s accuracy across unseen large-scale autonomous driving benchmarks with varying frame rates.

3. Method

In the open world, dashcam videos are recorded across a vast range of devices with diverse lenses, focal lengths, and frame rates, posing significant challenges due to the absence of calibration information. When attempting to reconstruct camera trajectories from such Internet-scale footage, this lack of calibration becomes a fundamental obstacle. Traditional VO methods [6, 34, 40, 42, 47] assume access to known camera intrinsics to recover metric-scale motion, but this assumption breaks down in many downstream tasks, such as processing unconstrained online dashcam videos. *We aim to develop a generalizable visual odometry system capable of extracting temporally consistent geometry from videos captured by uncalibrated cameras operating at arbitrary frame rates.*

Problem Definition: Given a dashcam video of T RGB frames $\{\mathbf{I}_i\}_{i=1}^T$, with spatial resolution (H, W) , captured under arbitrary frame rates f . The task is to recover a consistent world-coordinate ego-motion trajectory without relying on dataset-specific calibration. Following [20, 21], we focus on the two-frame pose regression setup where we learn a function for mapping two observed image frames $\mathbf{x}_i = \{\mathbf{I}_{i-1}, \mathbf{I}_i\}$ to a relative camera pose with real-world scale $\mathbf{y}_i = [\mathbf{R}_i | \mathbf{t}_i] \in \text{SE}(3)$, where $\mathbf{R}_i \in \text{SO}(3)$ and

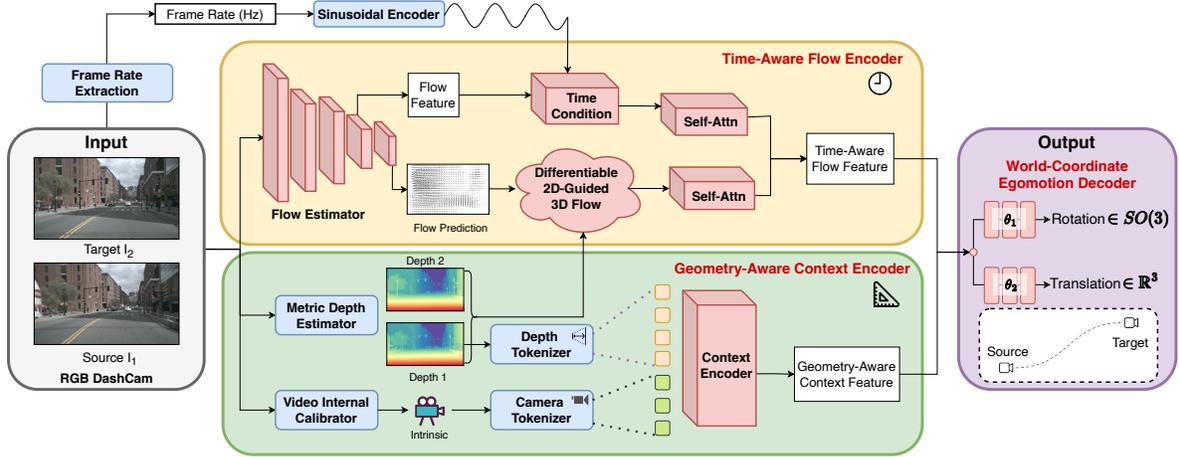


Figure 2. **Overview of OpenVO.** We propose a novel temporal-dynamics-informed, geometry-aware visual odometry system. Our method takes consecutive dashcam frames as input and extracts both temporal and geometric representations for robust egomotion estimation. The **Time-Aware Flow Encoder** (Sec. 3.1) leverages a Differentiable 2D-Guided 3D Flow module and time-conditioned embeddings to model motion dynamics across varying observation rates, while the **Geometry-Aware Context Encoder** (Sec. 3.2) incorporates metric depth and intrinsic priors to build a consistent 3D geometry structure of the scene. Finally, the **World-Coordinate Egomotion Decoder** (Sec. 3.3) predicts accurate world-coordinate egomotion trajectories from the fused dynamic-geometric representation.

$\mathbf{t}_i \in \mathbb{R}^3$. Our approach removes the dependency on pre-calibrated camera intrinsics. Instead, we infer them directly from visual cues through a lightweight video internal calibration module, which dynamically estimates the intrinsic parameters $\{f_U, f_V, c_U, c_V\}$ (denoted as \mathbf{K}) for each video. The estimated intrinsics are jointly leveraged with the optical flow \mathbf{F} , associated correlation feature \mathbf{F}^c under temporal frequency conditions f , and the estimated monocular metric depth sequence $\{\mathbf{D}_i\}_{i=1}^T$ to guide the estimation of *temporally and spatially* consistent ego-motion. This formulation enables our system to reconstruct the ego-motion of any dashcam video found on the Internet, producing metrically consistent world-coordinate trajectories without ground-truth calibration or sensor-specific assumptions.

Overview: OpenVO, shown in Fig. 2, leverages the concept of **temporal and geometric awareness** to recover accurate vehicle ego-motion. In Sec. 3.1, we inject temporal dynamics into the extracted flow features, aggregating them with differentiable 3D flow to obtain a unified time-aware flow representation. In Sec. 3.2, we introduce a transformer-based encoder that fuses multiple complementary modalities for geometry-aware context, which is finally combined with the time-aware features to regress the relative camera pose. Finally, Sec. 3.3 describes world-coordinate egomotion decoder with multi-time-scale training strategy.

3.1. Time-Aware Flow Encoder

The goal of this encoder is to extract time-conditioned optical flow features, enabling the model to encode temporal dynamics within the motion representation. To represent the dynamic motion of the scene, similar to conventional VO frameworks [20, 21, 47], we employ a pretrained Mask-FlowNet [59] encoder to extract pixel displacement fields

from paired image frames. From its intermediate layers, we obtain both the displacement map $\mathbf{F} \in \mathbb{R}^{H \times W \times 2}$ representing horizontal and vertical displacement and the correlation feature $\mathbf{F}^c \in \mathbb{R}^{H_F \times W_F \times C_F}$, which encodes dense 2D correspondences between the two images. However, the obtained displacement map and its corresponding feature only represent the dense correspondences between frames, without explicitly modeling the underlying motion dynamics. Prior approaches [20, 21, 42, 47] are typically trained and evaluated under a fixed temporal sampling rate, an assumption that fails in real-world dashcam data where frame rates vary widely across devices and video compression settings. As a result, these models learn time-agnostic motion features that cannot properly scale or adapt to varying temporal gaps, leading to severe trajectory drift when tested under different frame rates. To address this, we introduce Time Condition Layers that inject frame-rate information into the flow features. This conditioning implicitly encodes pixel-level velocity, thereby enhancing the generalization capability of VO systems in open-world environments.

Time Condition Layers: To incorporate temporal information into the flow feature, we map the frame rate to a time gap representation $\Delta t = \frac{1}{f}$, where f denotes the frame rate. This time interval $\Delta t \in \mathbb{R}$ is then expanded to high-dimensional embedding using a sinusoidal positional encoding. Specifically, for each frequency scale $\omega_i = \pi \cdot 2^i$, for $i = 0, 1, 2, \dots, K - 1$, we compute the encoding as:

$$\text{PE}(\Delta t) = [\Delta t, \sin(\omega_0 \Delta t), \dots, \sin(\omega_{K-1} \Delta t), \cos(\omega_0 \Delta t), \dots, \cos(\omega_{K-1} \Delta t)]. \quad (1)$$

This positional encoding injects smooth periodic variations of different frequencies that capture both fine-grained and coarse temporal scales, allowing the network to perceive

and generalize across different observation rates. We design two lightweight conditioning layers that take the encoded $\text{PE}(\Delta t)$ as input to modulate the flow feature \mathbf{F}^c :

$$\begin{aligned}\alpha &= \text{Linear}_\alpha(\text{PE}(\Delta t)), \\ \beta &= \text{Linear}_\beta(\text{PE}(\Delta t)), \\ \tilde{F}^c &= (1 + \alpha) \odot F^c + \beta,\end{aligned}\tag{2}$$

where \odot denotes the Hadamard product. The adaptive flow features \tilde{F}^c are then processed by four self-attention blocks, which operate on the temporally conditioned feature representation. These attention layers refine the spatial correlations of each pixel, now enriched with temporal cues injected by the Time Condition Layers. This enables the network to reason about motion structures and geometric relationships that are consistent with the underlying temporal dynamics, yielding time-aware flow representations that are more robust to varying observation frequencies.

Differentiable 2D-Guided 3D Flow: We convert the 2D optical flow and per-pixel metric depth into dense 3D motion field using a fully differentiable warping mechanism. First, we back-project each pixel $\mathbf{p}_1 = [u, v, 1]^\top$ with depth $\mathbf{D}_1(u, v)$ into 3D camera coordinates via perspective back-projection $\mathbf{P}_1 = \mathbf{D}_1(u, v)\mathbf{K}^{-1}\mathbf{p}_1$. Then, we use the 2D optical flow to warp pixels from the first frame to their sub-pixel locations in the second frame. Given the base pixel grid (u_1, v_1) and the flow field $\mathbf{F} = (\mathbf{F}_x, \mathbf{F}_y)$, we compute the warped coordinate by $(u_1 + \mathbf{F}_x, v_1 + \mathbf{F}_y)$. We normalize these coordinates and apply bilinear sampling on \mathbf{D}_2 and filters out points that fall outside the image or have invalid depth, yielding the corresponding depth at the flowed locations in a fully differentiable manner. Finally, we back-project the warped pixels to obtain \mathbf{P}_2 and compute a dense 3D flow $(\mathbf{P}_2 - \mathbf{P}_1)$ for all valid pixels. This yields a metric 3D motion field aligned with the camera-1 coordinate system, tightly coupling 2D flow and monocular depth while remaining end-to-end trainable. The resulting 3D flow field encodes per-pixel motion in real-world metric space, providing a geometry-grounded representation of scene dynamics. We further process this 3D flow through four self-attention layers to refine its spatial coherence, and subsequently fuse it with the temporally conditioned flow features produced by the Time Condition Layers. The fused representation, termed the **Time-Aware Flow Feature**, captures both geometric consistency and motion continuity across temporal dynamics, forming a robust foundation for downstream trajectory estimation.

3.2. Geometry-Aware Context Encoder

To ensure metric-scale consistency across diverse camera setups, we introduce a Geometry-Aware Context Encoder that explicitly injects depth and camera intrinsic priors into VO pipeline. Traditional monocular VO approaches often

rely solely on photometric or flow-based cues [20, 45], implicitly assuming that appearance information is sufficient to infer motion. However, such representations are inherently scale-ambiguous and fail to generalize when the testing camera differs from those seen during training. In contrast, our encoder provides a geometry-grounded prior that enables the model to reason about spatial structure and camera projection properties in a unified embedding space.

Camera tokenizer: Different cameras exhibit distinct projection behaviors due to variations in focal length, principal point, and sensor geometry. When training and testing on large-scale dashcam datasets collected from diverse sources, such variations introduce substantial appearance and motion inconsistencies, even for identical underlying 3D motions. To enable camera-aware representation, we leverage a pretrained lightweight internal calibrator Wild-Camera [61] to infer camera intrinsics from unlabeled dashcam video. We obtain \mathbf{K} simply by averaging the result of [61] across the entire video sequence. Then, we define a normalized intrinsic field:

$$r(u, v) \propto \mathbf{K}^{-1}[u, v, 1]^\top = \left[\frac{u - c_U}{f_U}, \frac{v - c_V}{f_V}, 1 \right]^\top \tag{3}$$

which describes the direction of the 3D viewing ray passing through pixel (u, v) in the camera coordinate system. This intrinsic ray field captures the capture the geometric projection pattern across the image plane: pixels near the optical center correspond to narrower ray cones, while those toward the periphery represent wider projection angles.

Depth tokenizer: We further incorporate metric depth to provide scene-scale information. Specifically, we leverage a pretrained metric depth estimator Metric3Dv2 [16] to obtain per-pixel depth \mathbf{D} . Each ray direction from Eq. (3) is then modulated by the corresponding depth value $\mathbf{M}(\mathbf{u}, \mathbf{v}) = \mathbf{D}(u, v) \cdot r(u, v)$. This operation projects each pixel onto its corresponding 3D point relative to the camera origin. The resulting spatial distribution $\mathbf{M}(\mathbf{u}, \mathbf{v})$ effectively reconstructs the scene structure up to metric scale, capturing both the directionality and magnitude of each visual ray. Finally, we assemble the token set $[r, \mathbf{M}, \mathbf{D}]$ representing a geometry-aware encoding of the camera projection behavior and scene structure. These tokens are processed by a Geometry-Aware Context Encoder composed of a stack of eight self-attention layers, which learn to correlate the intrinsic and depth priors with the surrounding visual context. This produces a unified geometric embedding that generalizes across uncalibrated dashcam sources and provides a consistent, metric-scale prior for world-coordinate egomotion estimation.

3.3. World-Coordinate Egomotion Decoder

To estimate world-coordinate ego-motion at step i , we concatenate Time-Aware Flow Feature \mathbf{F}^{TA} and Geometry-

Aware Context Feature \mathbf{F}^{GA} , which is then mapped to (a) ego translation and (b) rotation by two MLP branches:

$$\begin{aligned} \mathcal{F} &= f_{\theta_1}([\mathbf{F}^{\text{TA}}, \mathbf{F}^{\text{GA}}]); \quad \mathcal{F} \in \mathbb{R}^{3 \times 3} \\ \mathbf{R}_i &\sim \text{MF}(\mathcal{F}); \quad \mathbf{R}_i \in \text{SO}(3). \\ \mathbf{t}_i &= f_{\theta_2}([\mathbf{F}^{\text{TA}}, \mathbf{F}^{\text{GA}}]); \quad \mathbf{t}_i \in \mathbb{R}^3. \end{aligned} \quad (4)$$

where \mathbf{R}_i and \mathbf{t}_i are the estimated rotation matrix and translation, respectively, \mathcal{F} is the Fisher matrix, and $\text{MF}(\cdot)$ maps Fisher matrix into an element of $\text{SO}(3)$ [29, 30]. The rotational matrix \mathbf{R}_i has a probabilistic formulation grounded in the Fisher Matrix distribution [20, 21, 29, 30], which models orientation uncertainty to better model ego-motion variations, as rotations heavily influence the estimated trajectory. For translation, we employ a metric-scale regression module [47] to directly predict world-coordinate displacements, enabling scale-consistent motion recovery.

The temporal dynamics embedded in the Time-Aware Flow Feature \mathbf{F}^{TA} inform the decoder about the effective temporal scale, improving its ability to model both short- and long-range motions. Since OpenVO is explicitly conditioned on $f = \frac{1}{\Delta t}$, we perform temporal frequency augmentation during training to expose it to multiple time scales. Given an input video recorded at original frame rate f_0 , we generate lower-frequency observations by sub-sampling frames at a factor k , producing new training pairs at frame rate $\frac{f_0}{k}$ via frame skipping. This allows OpenVO to model motion across short- and long-term temporal scales while enabling the time-aware encoder to adapt its embeddings to different online observation frequencies. To improve training stability, we apply gradient clipping during backpropagation. Using these proposed designs, OpenVO becomes robust to unseen frame-rate variations commonly encountered in Internet-scale dashcam videos.

4. Experiment

4.1. Experimental Setup

Dataset: Following [20, 21], we evaluate the generalization capability of OpenVO under diverse frame rates, camera configurations, and environments using three standard visual-odometry benchmarks: KITTI [14], nuScenes [5], and Argoverse 2 [50]. **KITTI** contains 11 sequences (00–10) recorded at 10 Hz, each lasting about 4–5 minutes with an average of 4,500 front-view images per scene, providing long-baseline trajectories that stress temporal consistency. **nuScenes** comprises 1,000 driving scenes captured at 12 Hz across four regions in Boston and Singapore, each about 20 seconds long with an average of 240 front-view images per scene. It spans diverse conditions, including dense traffic, nighttime, and rainy environments, making it ideal for evaluating temporal adaptability. **Argoverse 2** includes 1,000 sequences recorded at 20 Hz, averaging 380 grayscale stereo front-view images per scene,

collected from six U.S. cities under varying weather and illumination. This dataset provides a strong test of cross-camera and cross-modality generalization. For a fair comparison with prior works [20, 21], we downsample all sequences to 160 stereo front-view images (10 Hz). We also report cross-dataset generalization using the 20 Hz results in our ablation study. Together, these datasets span a broad range of temporal frequencies, camera intrinsics, and scene complexities, forming a rigorous evaluation framework for time-aware open-world visual odometry.

Evaluation metrics: We evaluate our method using three standard metrics: Translation Error (t_{err}), Rotation Error (r_{err}), and Absolute Trajectory Error (ATE), along with a recently introduced Scale Error (s_{err}) [20, 21]. The t_{err} measures the average translational drift (in %) between the predicted and ground-truth trajectories, reflecting how accurately the model estimates displacement over time. The r_{err} quantifies the angular deviation (in $^\circ/100$ m) between predicted and true orientations, indicating how well rotational motion is captured. The ATE assesses global trajectory consistency by aligning the predicted and ground-truth paths and computing the root-mean-square (RMS) positional difference, revealing how local errors accumulate over time. Finally, the s_{err} measures the relative discrepancy in translation magnitudes, evaluating how faithfully the model preserves metric scale. It is computed as the mean difference between L2 norms of translation vectors across timesteps.

Setup details: Following [21], we train OpenVO on the Singapore-OneNorth subset (183 scenes) and evaluate it on unseen regions of nuScenes, as well as KITTI and Argoverse 2 for cross-domain generalization. Testing is performed without access to camera intrinsics. For WildCamera, we use the released nine-dataset weights to estimate intrinsics, and for Metric3Dv2, we adopt DINO2reg-ViT-Large with WildCamera-predicted intrinsics for metric depth estimation. OpenVO is trained for 285.6k iterations (batch = 16) with Random Crop & Resize [47] and Horizontal Flip (seed = 2023), using $K = 8$ and temporal-frequency augmentation at 4, 6, and 12 Hz. Each experiment takes about 96 GPU hours on an A6000. Unlike [20, 21], we use no additional YouTube data to ensure a fair comparison.

4.2. Comparison to prior work

Quantitative comparison: The quantitative evaluation of OpenVO is summarized in Sec. 4.1. We compare OpenVO against prior state-of-the-art methods, including XVO [20], ZeroVO [21], and the SLAM-based baseline Metric3Dv2 [16] + DroidSLAM [41] (M+DS) introduced in [21]. We conduct zero-shot testing on **KITTI 00–10 (10 Hz)** under unknown camera parameters. OpenVO achieves a global trajectory ATE of 93.23, outperforming ZeroVO[‡] by 24% while using the same Depth + Intrinsic priors, and even surpassing ZeroVO⁺ trained with additional text guid-

Method	KITTI 00–10 (10Hz)				nuScenes (12Hz)				Argoverse 2 (10Hz)			
	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
TartanVO [47]	13.85	3.27	103.07	-	10.27	6.35	6.26	-	11.17	5.30	7.03	-
DPVO [42]	8.31	2.37	78.53	-	4.34	2.85	2.66	-	2.66	1.25	1.59	-
ZeroVO ⁺ [21]	6.81	2.69	104.69	0.06	9.74	4.37	6.03	0.12	4.64	2.83	3.05	0.09
XVO [20]	16.82	3.84	168.43	0.17	12.75	5.11	8.30	0.16	9.13	4.86	5.70	0.12
M+DS [16, 41]	14.22	2.72	154.77	0.09	17.08	1.46	10.46	0.18	16.67	1.79	8.51	0.13
ZeroVO [‡] [21]	<u>8.99</u>	<u>2.92</u>	123.42	0.08	12.26	5.23	8.40	0.15	8.62	4.11	5.71	0.11
OpenVO [×]	9.02	3.32	<u>109.01</u>	<u>0.06</u>	<u>9.57</u>	<u>3.71</u>	<u>6.13</u>	<u>0.11</u>	<u>3.30</u>	<u>2.48</u>	<u>2.43</u>	<u>0.07</u>
OpenVO [✓]	8.97	3.43	93.23	0.05	9.04	3.86	5.91	0.10	3.25	2.84	2.39	0.07

Table 1. **Generalizability results.** Comparison with existing methods on 3 large-scale driving benchmarks using standard metrics: translation (t_{err}), rotation (r_{err}), absolute trajectory (ATE), and scale error (s_{err}). OpenVO is **trained exclusively on nuScenes** and evaluated on KITTI and Argoverse 2 with unseen camera setups, as well as unseen regions of nuScenes, to assess cross-domain generalization. **Shaded rows** denote methods not directly comparable to OpenVO (either rely on ground-truth intrinsics or auxiliary modality). “-” indicates unreported values; ⁺ denotes training with additional text guidance and YouTube data, and [‡] suggests the use of the same Depth [16] and Intrinsic [61] priors from foundation models as ours. OpenVO has 2 variants: [✓] with differentiable and [×] with non-differentiable 2D-Guided 3D flow. The **best** and **second best** results are marked as **bold** and underline among intrinsic-free, vision-only models.

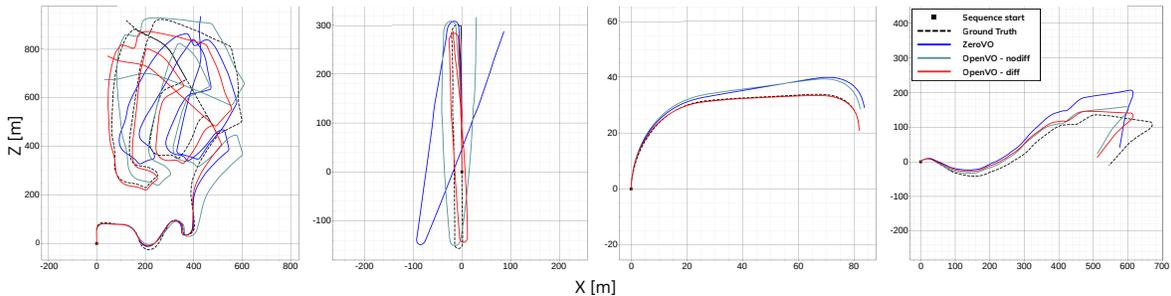


Figure 3. **Qualitative results.** We present trajectory prediction results on the KITTI and nuScenes datasets. Compared to ZeroVO[‡], both variants of our method — differentiable (**OpenVO-diff**) and non-differentiable variants (**OpenVO-nodiff**) of our 2D-guided 3D flow — achieve higher trajectory prediction accuracy and consistency, surpassing the current state-of-the-art.

Setting	KITTI 00–10 (10Hz)				nuScenes (12Hz)			
	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
$K = 2$	9.68	3.55	108.48	0.07	9.47	3.90	6.63	0.11
$K = 4$	9.43	3.54	101.53	0.06	9.95	3.60	6.74	0.12
$K = 8$	8.97	3.43	93.23	0.05	9.04	3.86	5.91	0.10
$K = 16$	9.18	3.72	95.11	0.06	9.85	6.60	6.30	0.10
$K = 32$	8.97	3.07	114.22	0.08	10.21	5.11	7.03	0.12
$K = 64$	10.24	4.00	119.23	0.07	11.15	4.44	6.92	0.13

Table 2. **Ablation on Time Encoding’s Token Size (K).**

ance and preprocessed YouTube data. In terms of real-world scale recovery, our method attains state-of-the-art performance, demonstrating strong generalization to unseen real-world videos. However, slightly higher local trajectory segment errors in t_{err} and r_{err} arise from training with mixed temporal frequencies, which can lead to inconsistent parameter updates despite having applied gradient clipping. This limitation can be mitigated through adaptive temporal sampling, which we leave for future investigation. For **nuScenes (12 Hz)**, we evaluate on unseen regions and achieve state-of-the-art results across all four

Setting	KITTI 00–10 (10Hz)				nuScenes (12Hz)			
	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
{12/6/4}	10.24	4.24	152.42	0.09	9.38	4.02	6.15	0.11
{12}*	12.89	4.40	213.76	0.09	10.23	4.91	6.35	0.12
{12/6}	8.64	2.91	117.58	0.07	8.08	3.40	5.27	0.10
{12/6/4}	8.97	3.43	93.23	0.05	9.04	3.86	5.91	0.10
{12/6/4/3}	9.43	3.57	97.21	0.06	9.06	4.21	6.02	0.11

Table 3. **Ablation on Training Observation Frequency.** **Shaded row** : OpenVO trained *without* proposed Time Condition Layers; Other rows : OpenVO trained *with* proposed Time Condition Layers. * indicates that OpenVO uses 12 Hz as the fixed inference temporal condition, regardless of the dataset.

metrics. OpenVO achieves an ATE of 5.91, effectively capturing scene geometry from limited input and outperforming methods with additional supervision. Under the same setting, both t_{err} and r_{err} also surpass the current state-of-the-art, confirming superior convergence. For **Argoverse 2 (10 Hz)**, we achieve superior performance across all four metrics, even surpassing the YouTube text-guidance self-supervised model from ZeroVO⁺. Since the trajectories

Setting	OpenVO			ZeroVO [‡] [21]		
	t_{err}	r_{err}	ATE	t_{err}	r_{err}	ATE
KITTI (2.5Hz)	28.83	8.22	368.47	52.74	17.46	553.52
KITTI (5Hz)	12.06	3.77	132.72	28.62	8.78	387.10
nuScenes (3Hz)	12.73	3.71	7.60	25.75	8.30	14.64
nuScenes (4Hz)	8.14	3.24	5.00	15.49	6.08	9.33
nuScenes (6Hz)	9.18	4.05	6.07	31.24	10.30	21.55
Argoverse 2 (4Hz)	3.63	1.72	2.09	25.93	19.62	28.38
Argoverse 2 (5Hz)	4.40	1.71	2.47	31.79	16.20	26.83
Argoverse 2 (20Hz)	9.79	4.72	6.47	61.61	27.38	36.14

Table 4. **Ablation on Varying Inference Observation Rates.**

[‡] indicates ZeroVO uses the same Metric Depth [16] and Intrinsic [61] priors from foundation models as OpenVO. Green cells denote OpenVO (left) improved performance of over existing method (right). Our OpenVO is *significantly more robust to observation rate variations, consistently reducing VO errors by 46% to 92%* across different metrics and benchmark settings.

in Argoverse 2 are generally less complex than those in KITTI, most methods – including ours, attain an ATE below 10. Across all three datasets, our differentiable 2D-Guided 3D Flow variant yields more robust trajectory predictions. The differentiable design enables absorption of dynamic 4D geometric cues from foundation priors while keeping the pipeline end-to-end trainable. Even without differentiability, OpenVO[×] achieves state-of-the-art results, highlighting its generalization to unseen dashcam videos through strong foundation priors and context encoding.

Qualitative comparison: We visualize the X-Z projection qualitative trajectory predictions of ZeroVO[‡] and the two variants of our OpenVO in Fig. 3. The differentiable version produces trajectories that align more closely with the ground truth, especially in complex or long-range scenes, highlighting the benefit of gradient-consistent 3D flow modeling and improved geometric stability. *Furthermore, we showcase the versatility of OpenVO across diverse autonomous driving tasks and applications in the Supp. Mat.*

4.3. Ablation Studies

To validate design choices of our method, we conduct a series of ablation studies on different configuration setups.

Study on temporal frequency size is in Tab. 2. The size of the positional encoding from Eq. (1) is $(1 + 2K)$. We observe that *setting $K = 8$ achieves the best performance by providing a balanced range of temporal frequencies* that capture both fine and coarse-scale motion dynamics. Using lower K limits temporal expressiveness, causing the model to under-fit variations in frame rates, while excessively large K introduces unstable high-frequency oscillations. Therefore, $K = 8$ offers an effective trade-off between representational richness and training stability.

Effect of multi-time-scale training demonstrates in Tab. 3. Training solely at 12 Hz (second row) restricts the model to a fixed temporal dynamic, leading to degraded performance on both datasets. On KITTI, testing under the same 12 Hz condition since the model never encountered 10 Hz encodings, resulting in a large ATE of 213.76. Nonetheless, the 12 Hz-only model surpasses ZeroVO[‡] on nuScenes, validating the effectiveness of our differentiable design and geometry-aware context encoder. Incorporating 6 Hz samples (third row) enables the model to exploit time-conditioned features, yielding consistent gains across datasets. Further adding 4 Hz (fourth row) improves ATEs to 93.23 and 5.91 on KITTI and nuScenes, though local consistency slightly decreases due to over-diverse temporal patterns – suggesting adaptive sampling as a remedy. Adding another frequency (last row) yields asymptotic improvement, indicating convergence in temporal generalization. One might argue that mixing multi-frequency data alone could allow the model to implicitly learn temporal relationships. However, *when we remove the Time Condition Layers (first row), forcing the network to self-discover temporal dynamics, performance drops notably* on KITTI as it becomes confused by inconsistent motion patterns – *underscoring the necessity of our temporal awareness for robust multi-time-scale learning.*

Analysis of varying temporal gap is presented in Tab. 4. To set up this configuration, for each dashcam video, we subsample the original framerate by a factor of k that is, we skip $(k - 1)$ frames between consecutive inputs and apply the same subsampling to the corresponding ground-truth camera poses. The original capture rates are 10 Hz for KITTI, 12 Hz for nuScenes, and 20 Hz for Argoverse 2. We compare our OpenVO with the conventional ZeroVO[‡], which disregards temporal frequency during training. As shown, ZeroVO[‡] exhibits substantial degradation when evaluated under varying frame rates (e.g., KITTI 2.5 Hz with 553.52 ATE; nuScenes 6 Hz with 21.55 ATE and Argoverse 2 20 Hz with 27.38 ATE), highlighting its inability to adapt to changes in temporal dynamics. In contrast, OpenVO consistently achieves lower translation, rotation, and trajectory errors across all frame rates by explicitly encoding frame-rate-aware temporal embeddings. This demonstrates that *modeling temporal dynamics is critical for robust egomotion estimation under real-world video conditions*, where frame rates are often inconsistent.

5. Discussion

In summary, we introduce OpenVO, a real-world-scale egomotion estimation framework that models temporal dynamics and geometric consistency from uncalibrated dashcam videos. Through extensive evaluations on KITTI, nuScenes, and Argoverse 2, OpenVO highlights the importance of time-aware, geometry-driven learning for scalable and robust visual odometry in real-world driving scenarios.

Limitation: Despite its generalizability, Depth Estimator and Internal Calibrator in OpenVO function separately, so errors in metric depth or intrinsics may propagate to final VO results. However, this is not unique to OpenVO but is a common limitation across intrinsic-free VO literature. Moreover, the current multi-time-scale settings are set empirically. A more principled strategy, where the model adaptively adjusts the sampling frequency on the fly in a synergistic manner, would be an interesting future direction.

References

- [1] Maulana Bisyr Azhari and David Hyunchul Shim. Dino-vo: A feature-based visual odometry leveraging a visual foundation model. *IEEE Robotics and Automation Letters (RA-L)*, 2025. 3
- [2] Sai Krishna Bashetty, Heni Ben Amor, and Georgios Fainekos. Deepcrashtest: Turning dashcam videos into virtual crash tests for automated driving systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11353–11360, 2020. 2
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3
- [4] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 3
- [5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 2, 6, 13
- [6] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 2, 3
- [7] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022. 1
- [8] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 3
- [9] Gabriele Costante, Michele Mancini, Paolo Valigi, and Thomas A. Ciarfuglia. Exploring representation learning with cnns for frame-to-frame ego-motion estimation. *IEEE Robotics and Automation Letters (RA-L)*, 1(1):18–25, 2016. 2, 3
- [10] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 29(6):1052–1067, 2007. 2
- [11] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 2
- [12] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15–22. IEEE, 2014. 2
- [13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 13, 14, 18
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013. 2, 6
- [15] Andrew Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3946–3952, 2008. 2
- [16] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3, 5, 6, 7, 8, 13
- [17] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 1
- [18] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17307–17316, 2023. 3
- [19] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference*

- on *Computer Vision and Pattern Recognition*, pages 9661–9670, 2020. 1
- [20] Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, and Eshed Ohn-Bar. Xvo: Generalized visual odometry via cross-modal self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10094–10105, 2023. 2, 3, 4, 5, 6, 7
- [21] Lei Lai, Zekai Yin, and Eshed Ohn-Bar. Zerovo: Visual odometry with minimal assumptions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17092–17102, 2025. 2, 3, 4, 6, 7, 8
- [22] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. Ctrl-c: Camera calibration transformer with line-classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16228–16237, 2021. 3
- [23] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022. 14
- [24] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7286–7291. IEEE, 2018. 2, 3
- [25] Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025. 1
- [26] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023. 1, 13, 14, 15, 16
- [27] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In *International conference on robotics and automation (ICRA)*, pages 5218–5223. IEEE, 2019. 2, 3
- [28] Mark Maimone, Yang Cheng, and Larry Matthies. Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics*, 24(3):169–186, 2007. 2
- [29] Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009. 6
- [30] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. *Advances in Neural Information Processing Systems*, 33:4884–4893, 2020. 6
- [31] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 2
- [32] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2
- [33] Anh N Nhu, Sanghyun Son, and Ming Lin. Time-aware world model for adaptive prediction and control. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. 2, 3
- [34] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I, 2004. 1, 2, 3
- [35] 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. *Transactions on Machine Learning Research Journal*, 2024. 3
- [36] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024. 3
- [37] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 3
- [38] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 1
- [39] Anqi Shi, Yuze Cai, Xiangyu Chen, Jian Pu, Zeyu Fu, and Hong Lu. Globalmapnet: An online framework for vectorized global hd map construction. *arXiv preprint arXiv:2409.10063*, 2024. 1, 13
- [40] Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Real-time visual odometry from dense rgb-d images. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 719–722, 2011. 2, 3
- [41] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 2, 3, 6, 7

- [42] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:39033–39051, 2023. 2, 3, 4, 7
- [43] Pierre Thodoroff, Wenyu Li, and Neil D Lawrence. Benchmarking real-time reinforcement learning. In *NeurIPS 2021 Workshop on Pre-registration in Machine Learning*, pages 26–41. PMLR, 2022. 2, 3
- [44] Rui Wang, Martin Schworer, and Daniel Cremers. Stereo dso: Large-scale direct sparse visual odometry with stereo cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3903–3911, 2017. 2
- [45] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2043–2050. IEEE, 2017. 2, 3, 5
- [46] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research (IJRR)*, 37(4-5):513–542, 2018. 3
- [47] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *Conference on Robot Learning*, pages 1761–1772. PMLR, 2021. 2, 3, 4, 6, 7
- [48] Maolin Wei, Wanzhou Liu, and Eshed Ohn-Bar. Driveqa: Passing the driving knowledge test. *arXiv preprint arXiv:2508.21824*, 2025. 1
- [49] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021. 13, 14, 17
- [50] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2, 6
- [51] Yingfu Xu and Guido C. H. E. de Croon. Cnn-based ego-motion estimation for fast mav maneuvers. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7606–7612, 2021. 2
- [52] Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K Wong, Jianqiang Wang, and Hengshuang Zhao. Drivegpt4-v2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17261–17270, 2025. 1
- [53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 3
- [54] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:21875–21911, 2024. 3
- [55] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1281–1292, 2020. 2, 3
- [56] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 3
- [57] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1983–1992, 2018. 3
- [58] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024. 13, 14
- [59] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I-Chao Chang, and Yan Xu. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4
- [60] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017. 3
- [61] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems*, 36:45137–45149, 2023. 3, 5, 7, 8, 13

OpenVO: Open-World Visual Odometry with Temporal Dynamics Awareness

Supplementary Material

In this supplementary material, we first provide a more detailed implementation of our Differentiable 2D-Guided 3D Flow in Sec. 6. We then present additional training and evaluation analysis in Sec. 7. Furthermore, we outline potential applications enabled by OpenVO in Sec. 8. Finally, we include additional qualitative results in Sec. 9, demonstrating the robustness and superior trajectory reconstruction capabilities of OpenVO across diverse scenarios.

6. Differentiable 2D-Guided 3D Flow

Given two consecutive frames with metric depth and a dense optical flow field, our differentiable 2D-guided 3D flow layer computes a per-pixel 3D flow in the coordinate system of the first camera. Let $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{H \times W}$ denote the metric depth maps at times t_0 and t_1 , respectively, expressed in meters. The camera intrinsics are parameterized as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

and the optical flow from $t_0 \rightarrow t_1$ is given by:

$$\mathbf{F}(u, v) = \begin{bmatrix} \Delta u(u, v) \\ \Delta v(u, v) \end{bmatrix}, \quad \mathbf{F} \in \mathbb{R}^{H \times W \times 2} \quad (6)$$

where (u, v) indexes pixel coordinates. We back-project every pixel (u, v) by constructing the homogeneous image coordinate \mathbf{p} into 3D using the depth $\mathbf{D}_1(u, v)$

$$\mathbf{P}_1(u, v) = \mathbf{D}_1(u, v) K^{-1} \mathbf{p} = \begin{bmatrix} X_1(u, v) \\ Y_1(u, v) \\ Z_1(u, v) \end{bmatrix} \in \mathbb{R}^3 \quad (7)$$

In coordinates, this is equivalent to:

$$\begin{bmatrix} X_1(u, v) \\ Y_1(u, v) \\ Z_1(u, v) \end{bmatrix} = \begin{bmatrix} \frac{(u - c_x) \mathbf{D}_1(u, v)}{f_x} \\ \frac{(v - c_y) \mathbf{D}_1(u, v)}{f_y} \\ \mathbf{D}_1(u, v) \end{bmatrix} \quad (8)$$

To establish dense correspondences between the two consecutive frames, we warp each pixel (u, v) from time t_0 to its estimated sub-pixel location at time t_1 using the optical flow field. The warped pixel coordinates are computed as:

$$u' = u + \Delta u(u, v), \quad v' = v + \Delta v(u, v) \quad (9)$$

This operation allows the model to reason about motion at a finer level than integer pixel shifts, which is crucial for handling small object motions, rolling-shutter distortions, and variations in frame rates. Since the predicted coordinates (u', v') are generally non-integer and may lie between pixel centers, we perform differentiable sampling of the target depth map \mathbf{D}_2 . We define the final sampling grid:

$$\mathbf{g}(u, v) = \begin{bmatrix} u'(u, v) \\ v'(u, v) \end{bmatrix} \quad (10)$$

We declare an operation $\sigma(\cdot)$ which takes metric depth \mathbf{D}_2 and sampling grid \mathbf{g} as input and outputs sampled metric depth $\tilde{\mathbf{D}}_2(u, v)$ as:

$$\mathbf{D}_2(u', v') = \sum_{i \in \{0,1\}} \sum_{j \in \{0,1\}} w_{ij} \mathbf{D}_2(u_i, v_j) \quad (11)$$

where pixel neighbors are:

$$u_i = \lfloor u' \rfloor + i, \quad v_j = \lfloor v' \rfloor + j, \quad i, j \in \{0, 1\} \quad (12)$$

and bilinear interpolation weights are:

$$\gamma = u' - u_0, \quad \psi = v' - v_0, \quad (13)$$

$$w = \begin{bmatrix} (1 - \gamma)(1 - \psi) & (1 - \gamma)\psi \\ \gamma(1 - \psi) & \gamma\psi \end{bmatrix} \quad (14)$$

We obtain $\tilde{\mathbf{D}}_2 = \sigma(\mathbf{D}_2, \mathbf{g})$, then following Eq. (7), we back-project $\tilde{\mathbf{D}}_2$ into 3D at time t_1 to obtain $\mathbf{P}_2(u, v)$. Finally, we compute a dense 3D flow for every pixel:

$$\mathbf{S}(u, v) = (\mathbf{P}_2 - \mathbf{P}_1) \quad (15)$$

The entire pipeline used to obtain the warped depth $\tilde{\mathbf{D}}_2(u, v)$ is fully differentiable with respect to the depth map, the warped coordinates (u', v') , and the predicted optical flow $(\Delta u, \Delta v)$. Since the warped position satisfies $u' = u + \Delta u$ and $v' = v + \Delta v$, gradients propagate through the sampling process via the chain rule. For example:

$$\frac{\partial \tilde{\mathbf{D}}_2}{\partial \Delta u} = \frac{\partial \tilde{\mathbf{D}}_2}{\partial u'} \cdot \frac{\partial u'}{\partial \Delta u} = \frac{\partial \tilde{\mathbf{D}}_2}{\partial u'} \quad (16)$$

This smooth interpolation ensures stable gradient flow across sub-pixel displacements, enabling reliable end-to-end optimization for Time-Aware Flow Encoder block.

Method	# params (M)	Runtime (s)
WildCamera [61]	270.447	41.906
Metric3Dv2 [16]	411.941	180.752

Table 5. **Expected latency of foundation priors.**

To ensure that only geometrically valid correspondences contribute to the loss, we construct a binary validity mask

$$\mathbf{m}(u, v) = \mathbf{1}[0 \leq u' \leq W - 1, 0 \leq v' \leq H - 1, \mathbf{D}_1(u, v) > 0, \tilde{\mathbf{D}}_2(u, v) > 0], \quad (17)$$

where $\mathbf{1}$ is an indicator function and apply it element-wise to the 3D flow:

$$\tilde{\mathbf{S}}(u, v) = \mathbf{m}(u, v)\mathbf{S}(u, v) \quad (18)$$

The mask \mathbf{m} is a non-learned, piecewise-constant tensor computed from geometric constraints; during backpropagation we do not differentiate through the indicator function, so the operation remains differentiable w.r.t. \mathbf{S} , and gradients are simply zeroed out at invalid pixels.

7. Analysis

Latency: We report the number of parameters and the total running time for each component of our OpenVO. For the internal calibrator WildCamera [61] and metric depth estimator Metric3Dv2 [16], we report the expected runtime for a *single process on 320 image frames* of ArgoVerse 2 [49] in Tab. 5. For OpenVO, we present the number of parameters and runtime for each components for a single forward for batch of 16 pairs of images in Tab. 6. Because the flow encoder performs convolution operations over full-resolution image grids, while the self-attention blocks operate on patch-level representations, the two modules exhibit distinct runtime behaviors.

Oracle Test: Given that accurate camera intrinsics are pivotal to our pipeline, we further examine an oracle scenario where the method is supplied with ground-truth intrinsic parameters. We conduct these oracle evaluations on nuScenes [5] and KITTI [13] to quantify the potential performance gains obtainable under perfect calibration in Tab. 7.

8. Global High-Definition (HD) Semantic Maps Reconstruction

HD maps play a critical role in autonomous driving and 3D scene understanding, providing precise lane geometry, road topology, traffic elements, and other structural cues that are necessary for enhanced situational awareness and safer control and navigation [26, 39, 58]. Such maps enable downstream tasks—including planning, motion prediction, navigation, and safety validation—to operate with

Method	# params (M)	Runtime (s)
Flow Estimator	20.655	3.089
Time Condition Layers	0.003673	0.0101
Diff. 2D-Guided 3D Flow	-	0.0204
Self-Attn	174.695	0.0339
Context Encoder	56.702	0.0129
Egomotion Decoder	29.988	0.4368

Table 6. **Expected latency of our OpenVO.** Red rows indicate Time-Aware Flow Encoder block. Green row indicates Geometry-Aware Context Encoder block. Purple row indicates Egomotion decoder block.

Setting	KITTI 00–10 (10Hz)				nuScenes (12Hz)			
	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
Ours–Oracle	7.22	2.75	90.85	0.05	6.41	3.44	4.24	0.09

Table 7. **Oracle testing.** We equip OpenVO with ground-truth intrinsic parameters to isolate errors from calibration, revealing the upper-bound performance under perfect camera parameters estimation.

spatial awareness and robust scene priors. However, HD maps are expensive to produce, typically requiring a combination of LiDAR sensors, human annotation, and specialized mapping vehicles. This motivates reconstructing HD maps directly from onboard sensors like cameras as a more scalable alternative. Yet this approach faces substantial technical challenges, including heavy dependency on ego-motion, occlusions, sensor noise, and complex 3D scene structure, which together can lead to compounding errors in the reconstructed maps over the horizon, especially under uncalibrated monocular observation under varying frame rates. Integrating OpenVO into the mapping pipeline helps mitigate these issues by providing accurate frame-to-frame poses that register local map predictions into a coherent global representation in world coordinates. This combination compensates for camera motion and enables consistent multi-view fusion, allowing scalable, camera-only reconstruction of high-quality HD semantic maps.

Following VectorMapNet [26], we reconstruct vectorized local HD maps directly from front-view imagery in Fig. 4. We disregard the LiDAR branch and train the network using only a monocular front-view stream. Ground-truth camera intrinsics and poses are used during training, whereas at inference we rely on the VO estimated by OpenVO and the camera parameters predicted by WildCamera [61]. For reference purpose, we present our quantitative results on local HDMap reconstruction in Tab. 8 and qualitative result in Fig. 6. Since the network predicts local map fragments independently, we apply post-processing heuristics to spatially align and merge these fragments into a coherent global map. Combining the modified architecture

Metrics	Full (6) cams + Lidar	Front cam only
mAP @ 0.5 (↑)	0.187	0.102
mAP @ 1.0 (↑)	0.479	0.318
mAP @ 1.5 (↑)	0.646	0.472
Overall mAP (↑)	0.4375	0.297
mAP ped @ 1.5 (↑)	0.644	0.469
mAP div @ 1.5 (↑)	0.642	0.509
mAP cont @ 1.5 (↑)	0.653	0.437
mAP @ 1.5 (↑)	0.646	0.472

Table 8. Comparison between the full-sensor configuration and our monocular front-camera-only configuration for HDMap reconstruction. **Green cells** highlight our configuration. *ped* denotes pedestrian crossings, *div* denotes dividers, and *cont* denotes contours. Performance is reported in terms of mean Average Precision (mAP).

with our proposed OpenVO, we can obtain the final global map as illustrated in Fig. 5

9. Qualitative Results

We provide **additional qualitative demo videos** in our supplementary materials!

Stereo VO: We showcase the performance of our OpenVO on the challenging stereo benchmark Argoverse 2 [49] in Fig. 7. In Argoverse 2, the stereo image pairs produce loose and noisy metric depth due to limited baseline, challenging lighting, and frequent low-texture or distant regions. This makes the recovered depth unstable across frames, which in turn causes conventional VO systems to become highly sensitive to depth noise and scale fluctuations. As a result, even small stereo-depth errors can propagate and lead to inconsistent trajectory estimates. In contrast, our OpenVO approach remains robust under these conditions and delivers stable, consistent results despite the imperfections of the stereo-derived metric depth.

Long-Range VO: We show the performance of our OpenVO on the challenging long-range KITTI [13] benchmark in Fig. 8. The KITTI odometry dataset contains many long-range highway and suburban scenes where most structures lie far from the camera. In such settings, monocular cues become weak: distant objects provide very small pixel motion, depth becomes highly ambiguous, and small errors in these regions can accumulate into noticeable scale drift. As a result, conventional monocular VO systems tend to be sensitive on KITTI, especially over long trajectories where scale inconsistency quickly compounds. Despite these challenges, OpenVO produces stable and consistent results on KITTI, showing that our time-aware and geometry-aware

design remains reliable even in long-range, low-parallax environments.

Real-World VO: We further assess the generalizability of OpenVO using videos from real-world environments, as illustrated in Fig. 9. Real-world driving scenes are heavily populated with vehicles, pedestrians, and cyclists, and are often dominated by occlusions and rapid appearance changes. Such dynamic factors degrade monocular geometric cues and introduce unstable depth signals, making conventional VO pipelines prone to drift and inconsistency. Despite such challenges, OpenVO produces stable and coherent trajectories, highlighting its robustness to noise, clutter, and complex dynamic activity commonly encountered in real-world driving.

Application beyond VO: OpenVO-enabled Monocular Global Map Reconstruction:

Reconstructing real-world long-tail scenarios is crucial for safety-critical domains such as autonomous driving, where realistic simulations of rare events provide insights into failure modes, vehicle dynamics, and hazardous scene geometry. However, collecting such data directly is extremely challenging due to economic constraints, safety risks, liability concerns, and legal prohibitions. In contrast, dashcam videos offer an abundant source of real-world long-tail footage, but are typically monocular, uncalibrated, and can be captured at different frame rates, which together pose great challenges for accurate 3D reconstruction tasks like mapping. Prior online mapping approaches [23, 26, 58] address local mapping using calibrated multi-camera or LiDAR setups with fixed observation rates, which effectively enhance online situational awareness but cannot reconstruct full HD map trajectories or scene evolution required for simulating rare events. In Fig. 6 and Fig. 5 and our corresponding supplementary video, we show that **OpenVO can bridge the global mapping gap** by generating accurate world-consistent poses and fusing them with local mapping to reconstruct the full observable scene, yielding coherent geometric and dynamic information about rare events. This facilitates scalable, monocular-based reconstruction of complex real-world scenarios from dashcam footage, offering deeper insights and enabling more comprehensive training and validation of autonomous driving algorithms in long-tail settings.

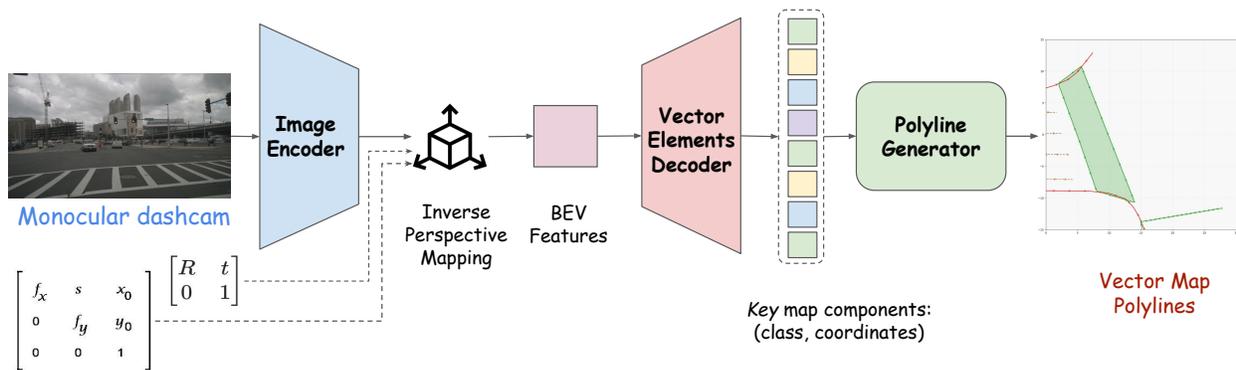


Figure 4. **Modified VectorMapNet** [26]. A front-view input image is first processed by an image encoder to extract semantic and geometric features. These features are then lifted into a bird’s-eye-view (BEV) representation using inverse perspective mapping, which leverages the camera’s intrinsic and extrinsic parameters from **OpenVO** to geometrically project image features onto the ground plane. The resulting BEV feature map is fed into the Vector Map Decoder, which predicts structured map elements in an intermediate representation consisting of key components such as polyline classes, control points, and geometric attributes. Finally, a polyline generator converts these decoded components into continuous vectorized map elements, such as lane boundaries, road dividers, and crosswalks – yielding a high-resolution, topologically meaningful HD map suitable for downstream driving tasks.

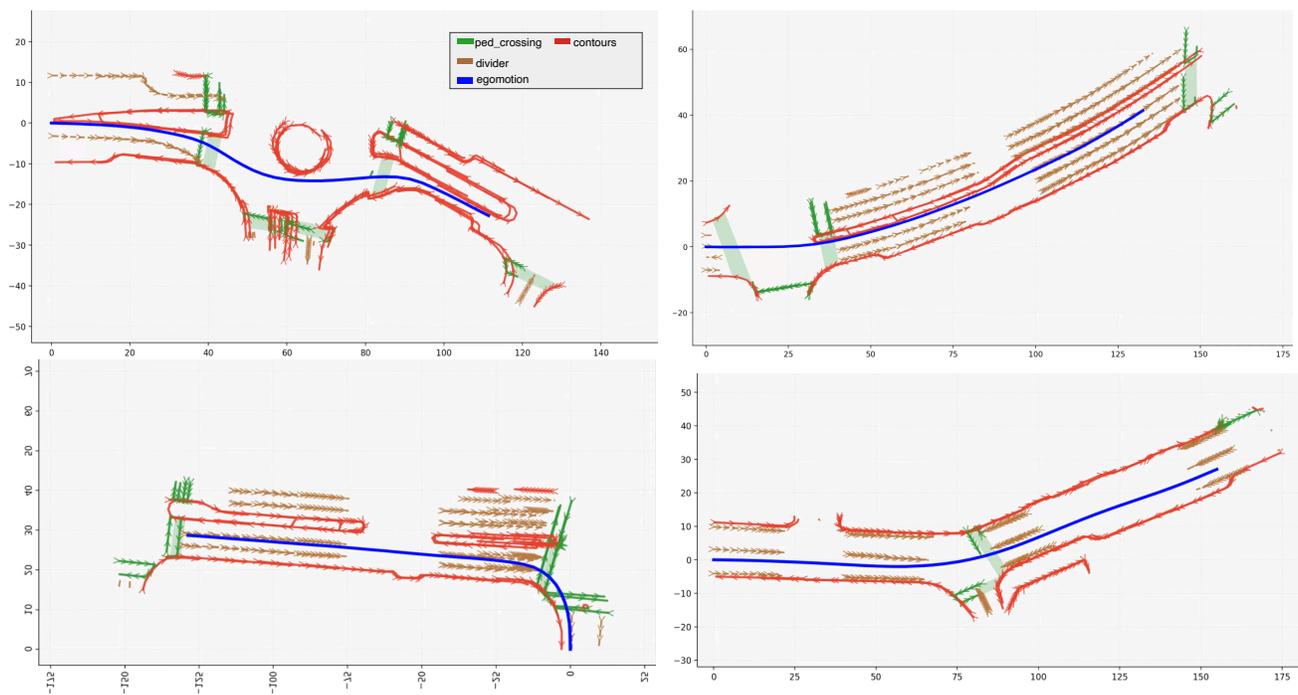


Figure 5. Qualitative results of Global HDMap reconstruction results produced by **OpenVO** + modified monocular VectorMapNet [26]. Local mapping outputs are gradually fused through OpenVO’s ego-to-world pose estimates, producing a coherent global HD-map reconstruction of the full scenario. We would like to refer to **our supplementary videos** for further details of the OpenVO-enabled monocular-based global map reconstruction.

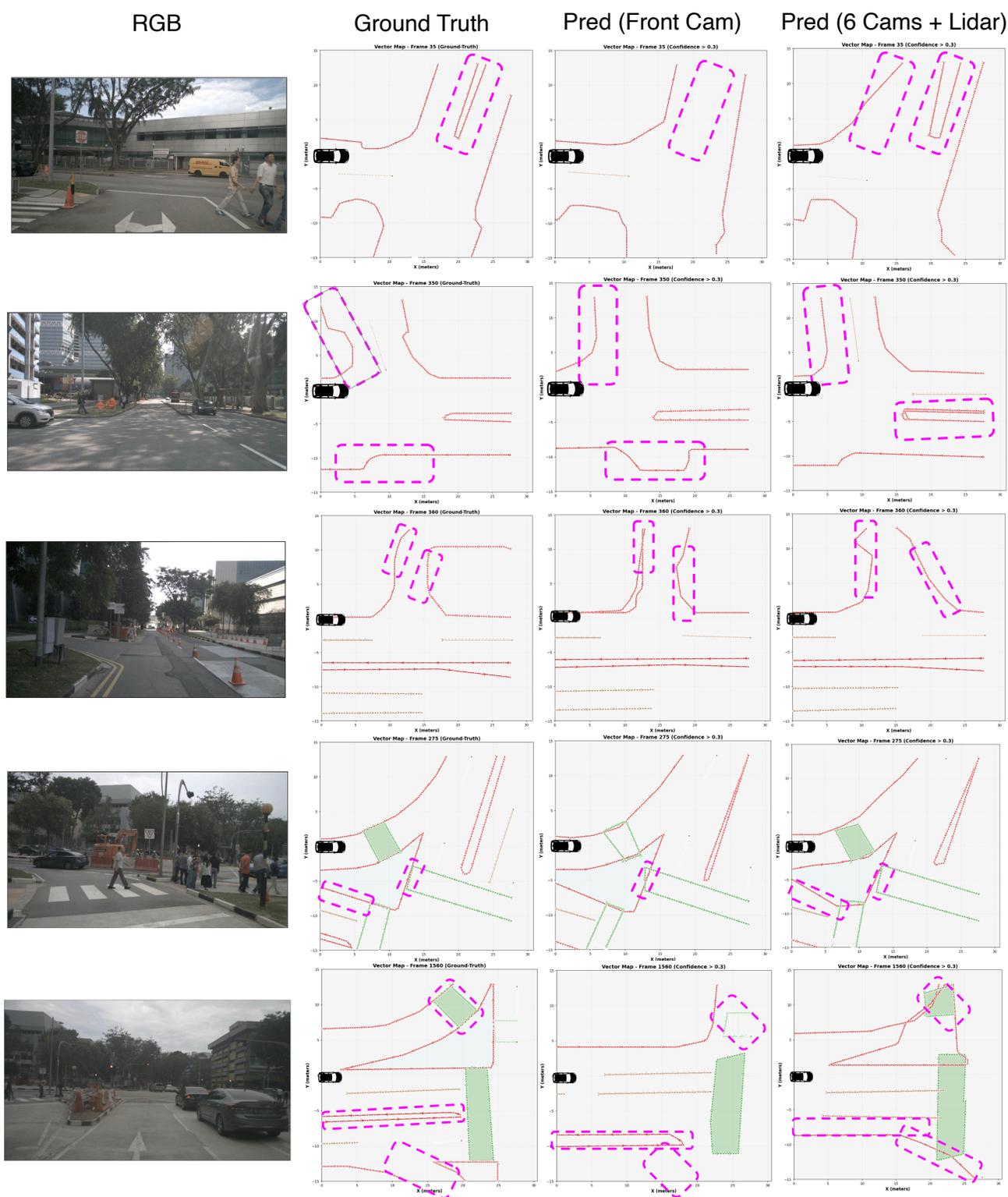


Figure 6. Qualitative HD Map reconstruction results produced by modified monocular VectorMapNet [26]. The leftmost column displays the input RGB frames; the second column shows the ground-truth HDMaps; the third column presents the results from our modified VectorMapNet using only a single front-camera image; and the last column shows the outputs under the six-camera + LiDAR configuration. We highlight the differences in each example using **dashed pink regions**.

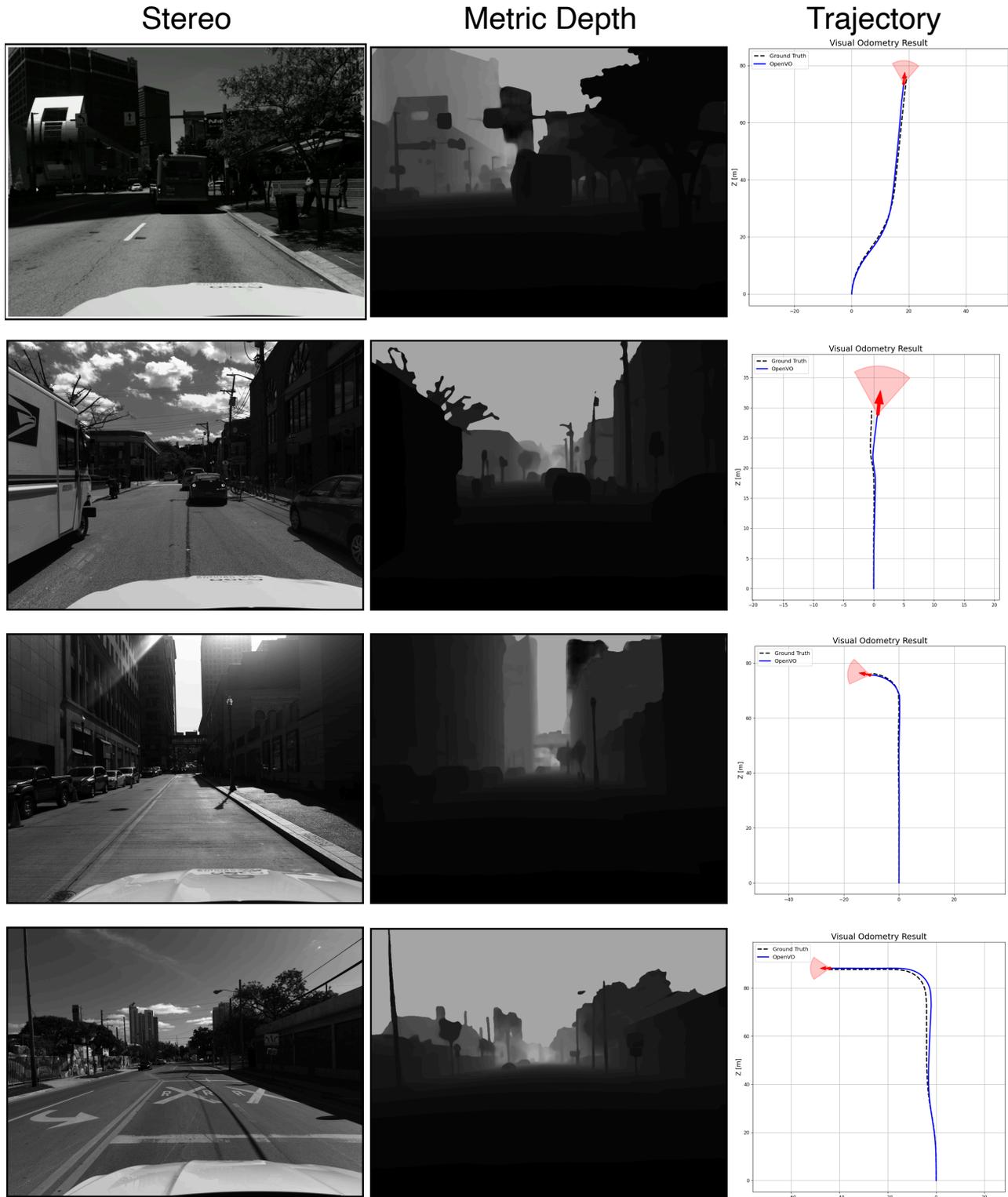


Figure 7. **Qualitative of Stereo benchmark on Argoverse 2 [49].** Each row shows one example, including the input stereo image and the reference metric depth. The stereo images in Argoverse 2 often provides low-quality or weakly constrained metric depth due to limited disparity in long-range regions and visually challenging street scenes. This degradation leads to information loss and introduces uncertainty into downstream VO estimation.

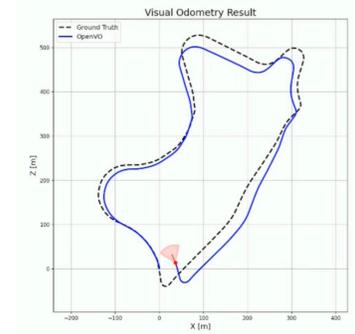
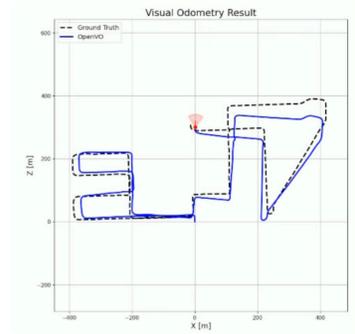


Figure 8. **Qualitative results on KITTI [13] benchmark.** Each row presents one example. The KITTI camera provides a wider field of view than most datasets, allowing it to capture a richer set of dynamic objects while still preserving its long-range odometry characteristics.

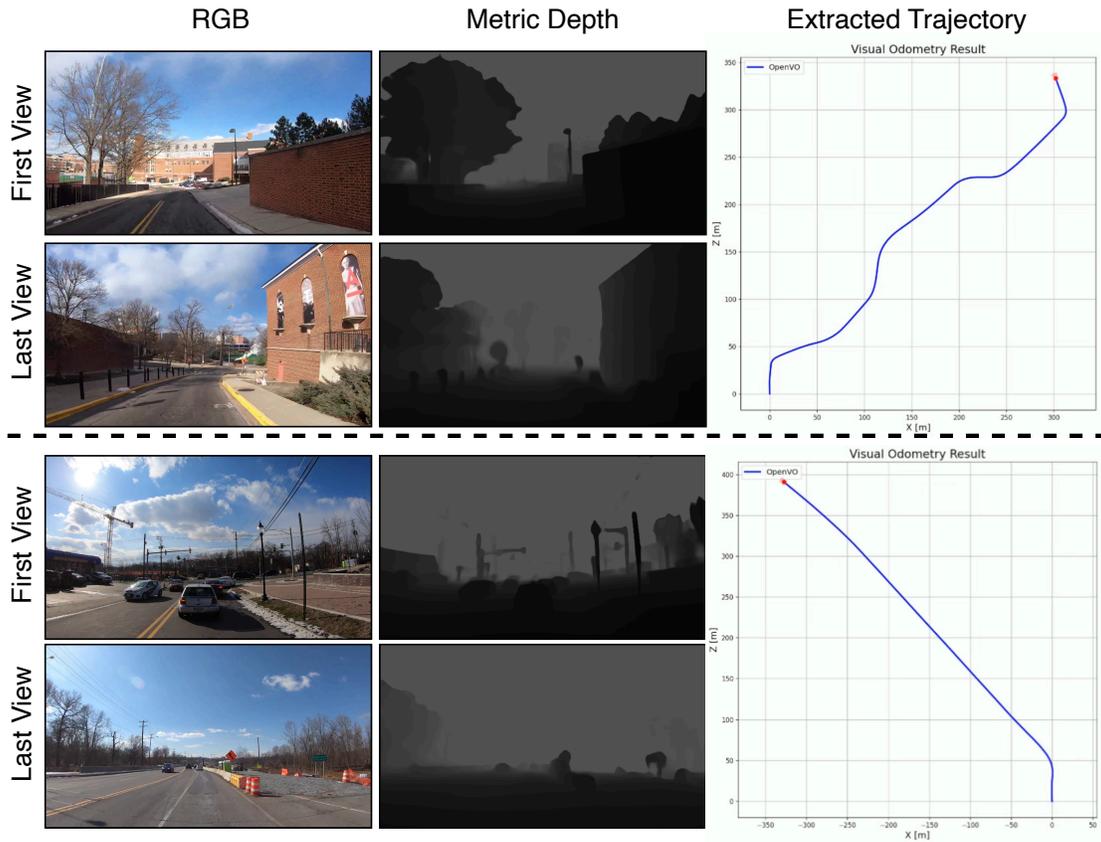


Figure 9. **Qualitative results on real-world captured videos.** We present two examples, each accompanied by the corresponding RGB frames and reference metric-depth images. Real-world videos commonly exhibit numerous environmental artifacts—such as noise, clutter, and dynamic elements—which pose significant challenges for generalizability and real-world performance assessment.