

DiffVL: Diffusion-Based Visual Localization on 2D Maps via BEV-Conditioned GPS Denoising

Li Gao^{1*} Hongyang Sun^{2*†} Liu Liu^{1‡} Yunhao Li^{1†} Yang Cai¹
 {liangliang.gl, diana.ll, huaimu.lyh}@alibaba-inc.com
 sunhongyang10@zju.edu.cn
 yangcai.cy@autonavi.com

Abstract—Accurate visual localization is crucial for autonomous driving, yet existing methods face a fundamental dilemma: While high-definition (HD) maps provide high-precision localization references, their costly construction and maintenance hinder scalability, which drives research toward standard-definition (SD) maps like OpenStreetMap. Current SD-map-based approaches primarily focus on Bird’s-Eye View (BEV) matching between images and maps, overlooking a ubiquitous signal-noisy GPS. Although GPS is readily available, it suffers from multipath errors in urban environments. We propose DiffVL, the first framework to reformulate visual localization as a GPS denoising task using diffusion models. Our key insight is that noisy GPS trajectory, when conditioned on visual BEV features and SD maps, implicitly encode the true pose distribution, which can be recovered through iterative diffusion refinement. DiffVL, unlike prior BEV-matching methods (e.g., OrienterNet) or transformer-based registration approaches, learns to reverse GPS noise perturbations by jointly modeling GPS, SD map, and visual signals, achieving sub-meter accuracy without relying on HD maps. Experiments on multiple datasets demonstrate that our method achieves state-of-the-art accuracy compared to BEV-matching baselines. Crucially, our work proves that diffusion models can enable scalable localization by treating noisy GPS as a generative prior—making a paradigm shift from traditional matching-based methods. Code and models will be open-sourced.

I. INTRODUCTION

Visual localization is a critical technology for applications like autonomous driving [1], [2], augmented reality, and robotics, where precise and reliable pose estimation is paramount for safe navigation [3] and decision-making [4]. The core task involves estimating a 3-DoF pose (position and orientation) from visual imagery against a 2D map. To meet the stringent demands of autonomous systems, traditional methods [5] have heavily relied on High-Definition (HD) maps. However, the high costs of creating, annotating, and frequently updating these maps severely limit their scalability and widespread adoption, creating a significant bottleneck for deploying autonomous technology at a global scale.

In response, recent research has shifted towards localization with low-cost, globally available Standard-Definition

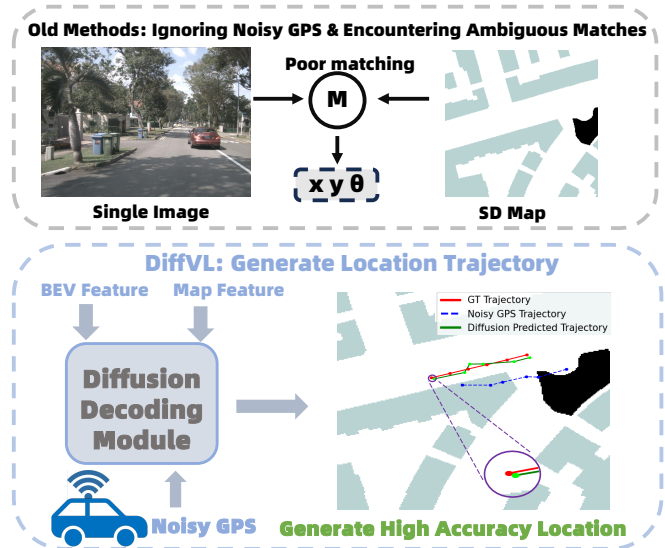


Fig. 1. Overview of the proposed DiffVL. Most existing SD map-based localization methods rely on exhaustive geometric matching between Bird’s-Eye View (BEV) features and map elements to compute the pose. In contrast, our approach fundamentally reformulates visual localization as a generative modeling task.

(SD) maps, such as OpenStreetMap [6]. Representative works [7], [8] leverage deep learning [9], [10], [11], [12] to infer the 3-DoF pose by aligning a Bird’s-Eye-View (BEV) representation, derived from the input image, with an SD map. While these approaches have shown promise, they are susceptible to challenges like perceptual aliasing in visually repetitive areas and typically overlook a crucial, ubiquitous source of information: noisy GPS data. This omission inherently limits the upper bound of their localization accuracy and robustness, particularly in challenging scenarios.

In recent years, diffusion models [13], [14], [15], [16], [17], have emerged as powerful generative techniques, achieving breakthrough performance by learning to reverse gradual noising processes. Such techniques have been successfully applied in embodied navigation[18] and autonomous driving[19] for predicting future trajectories from motion anchors. Inspired by these advances, we present a paradigm shift for visual localization: we reformulate it as a

* Equal Contribution

†Work done during the internship at Amap, Alibaba Group

‡Corresponding Author

¹AMAP, Alibaba Group

²Zhejiang University

conditional denoising process of GPS trajectories. Our key insight is that noisy GPS signals, often dismissed as unreliable, actually encode the true pose distribution and can be transformed into precise localization through diffusion-based denoising. Specifically, we reframe visual localization from a traditional image-to-map matching task [20], [21], [22] into a conditional generation problem, where a diffusion model learns to reverse the noise corruption in GPS trajectories conditioned on visual observations.

We propose DiffVL, a novel diffusion-based framework that synergistically integrates sequential GPS signals and visual cues. To imbue the diffusion model with geometric and semantic awareness, we introduce a dual-objective training strategy. The primary trajectory refinement loss ($\mathcal{L}_{\text{diff}}$) ensures kinematic and temporal coherence in the denoised trajectory. Simultaneously, an auxiliary localization prior loss (\mathcal{L}_{loc}), computed by aligning Bird’s-Eye View (BEV) visual features with map elements, provides strong geometric regularization. This dual loss compels shared feature encoders to learn representations that are both visually discriminative and geometrically consistent with the map’s coordinate system. Through joint optimization, our model achieves a robust balance between motion-based prediction and appearance-based matching, enabling high-precision pose estimation from noisy GPS and significantly enhancing localization robustness.

Our main contributions are summarized as follows:

- We introduce DiffVL, a novel visual localization paradigm that, to the best of our knowledge, is the first to successfully and principally apply diffusion models to denoise noisy GPS trajectories for this task. It reframes localization from a matching problem to conditional generation problem.
- We redefine the role of noisy GPS signals in visual localization. Whereas prior research often ignored or filtered these signals, we are the first to frame them as a ‘Noisy Observation’ of the true pose distribution, providing a methodological basis for leveraging generative models to recover high-precision poses.
- We design a dual-objective training framework where a trajectory refinement loss and a localization prior loss work in tandem. This ensures the model not only recovers a coherent trajectory but also learns a powerful, geometrically consistent feature representation from visual and map data.
- We conduct rigorous and comprehensive evaluations on multiple large-scale autonomous driving datasets (KITTI[23], nuScenes[24], and MGL[7]). Our experimental results demonstrate that DiffVL significantly outperforms existing methods across all datasets, achieving state-of-the-art performance. We plan to release our code and models to facilitate future research and contribute to the community.

II. RELATED WORK

A. Map-based Visual Localization

Map-based visual localization is a long-standing research task central to robotics [3] and autonomous systems[25], [26], [27]. Traditional methods rely on matching sensor data against meticulously pre-built 3D maps[28], [29]. These maps are typically composed of dense point clouds from LiDAR sensors or are reconstructed via Structure-from-Motion (SfM) from multiple image views [30]. They enable high-precision pose estimation through the alignment of handcrafted or learned local features between the query image and the 3D model[31], [32], or through direct geometric registration of point clouds, often using variants of the Iterative Closest Point (ICP) algorithm [33]. However, these High-Definition (HD) maps come with significant practical drawbacks. Their creation requires specialized vehicles and extensive surveying, their maintenance is a continuous and costly effort, and their large memory footprint poses a major barrier to on-board storage and large-scale deployment.

To overcome these scalability issues, recent research has shifted towards Standard-Definition (SD) maps. These maps, often derived from crowd-sourced data like OpenStreetMap (OSM) [6], are lightweight, globally-available, and semantically rich, making them a highly attractive alternative. A prominent line of work involves generating a neural Bird’s-Eye-View (BEV)[34], [35] representation from a monocular camera and then aligning it with a corresponding rasterized map tile [7], [8]. These methods effectively tackle the cross-view matching problem between a ground-level perspective and a top-down map. Despite their promising performance, these approaches face two key challenges. First, they must bridge the significant domain gap between rendered map semantics and complex, real-world visual features. Second, they almost universally overlook an important and readily accessible signal: noisy GPS data. This omission of a critical information source, often due to the difficulty of handling its inherent noise, inherently limits their robustness and accuracy, particularly in ambiguous environments.

B. Diffusion Models

In recent years, Denoising Diffusion Probabilistic Models (DDPMs) [13], [14], [36] have emerged as a disruptive technology in generative AI, achieving state-of-the-art results in diverse domains such as high-fidelity image and video generation [37], [38], robotic policy learning [39], and motion prediction [40]. The core idea is to learn to reverse a fixed Markov chain that gradually adds noise to data, allowing the model to generate new samples by progressively denoising from pure Gaussian noise. Their ability to capture complex, multi-modal distributions makes them particularly well-suited for robotics and autonomous systems. For instance, Diffusion Policy [39], [41] has proven highly effective at learning multi-modal action distributions for complex manipulation tasks. In motion prediction, works like MotionDiffuser [40], [42] leverage conditional diffusion models to generate kinematically coherent and socially-aware multi-agent trajectories.

Despite their demonstrated potential, the application of diffusion models to visual localization remains an unexplored frontier. Inspired by these pioneering works[4], [43], [44], we are the first to introduce the diffusion model paradigm to this task. Our key insight is to reframe the problem: instead of treating localization as a deterministic matching problem[7], [8], we embrace the inherent uncertainty of sensor data by formulating it as a conditional denoising task. In our DiffVL framework, the diffusion model learns to recover a kinematically smooth trajectory from a noisy raw GPS sequence. Crucially, this denoising process is not performed in isolation; it is intelligently conditioned on rich, multi-modal features extracted from both the camera image and the SD map[6], providing essential environmental context. To ensure these conditional features are geometrically grounded and semantically meaningful, we employ a dual-objective training strategy. A primary trajectory refinement loss is complemented by an auxiliary BEV-map matching loss, which acts as a powerful geometric regularizer, forcing the shared encoders to learn a spatially-aware representation. This approach directly integrates the generative power of diffusion models with the discriminative task of visual-map matching at the feature level, significantly enhancing both localization robustness and accuracy.

III. METHOD

A. Problem Formulation

Given a single front-view image \mathcal{I} and a historical sequence of noisy GPS measurements $\mathbf{P}^{\text{gps}} = \{\mathbf{p}_t^{\text{gps}}\}_{t=1}^T = \{x_t^{\text{gps}}, y_t^{\text{gps}}\}_{t=1}^T$ where T is the time horizon and $\mathbf{p}_t^{\text{gps}} = (x_t^{\text{gps}}, y_t^{\text{gps}})$ denotes the East-North-Up (ENU) coordinates at timestep t , along with an SD map M representing the local map tile encompassing the trajectory \mathbf{P}^{gps} . Our task is to estimate a 3-DoF pose $\hat{\mathbf{p}} = (x, y, \theta) \in \mathbb{R}^3$, where (x, y) represents the ENU position and $\theta \in (-\pi, \pi]$ denotes the heading angle around the vertical \mathbf{z} -axis.

The pose estimation process is formalized as a conditional diffusion model:

$$\hat{\mathbf{p}} = \mathcal{A}_\theta \left(\{\mathbf{P}_t^{\text{gps}}\}_{i=1}^T \mid \mathbf{z} \right) \quad (1)$$

where:

- $\mathcal{A}_\theta(\cdot)$ denotes the diffusion model head
- θ represents trainable parameters
- $\mathbf{z} = f(\mathcal{I}, M)$ is the conditional latent vector encoding image-map features

B. Overview

The DiffVL framework comprises four core modules with the following architecture, as shown in Fig.2.

Image Encoder: Performs environmental perception through front-view images, extracting multi-scale visual features. Utilizes depth estimation to analyze scene geometry, combined with view transformation that convert perspective representations into geometrically consistent BEV(bird’s-eye view) feature maps. This process effectively preserves spatial relationships and semantic content while establishing

correspondence with real-world coordinate systems, enabling robust understanding of the surroundings from an overhead perspective.

Map Encoder: Leverages OpenStreetMap [6] to construct rasterized map representations through grid-based processing, encoding critical prior knowledge about the environment.

Diffusion Guidance Generator: Implements cross-modal fusion between environmental perception and prior map knowledge through attention mechanisms. Generates conditional guidance embeddings that integrate visual observations with map semantics, constructing global contextual representations that guide subsequent diffusion denoising processes.

Diffusion Head: Achieves probabilistic localization refinement through iterative diffusion-based correction. Formulates pose estimation as a progressive noise reduction problem, where initial position hypotheses undergo multi-stage optimization guided by multimodal contextual features. By jointly optimizing positional and orientation parameters, it systematically reduces localization uncertainty, ultimately achieving meter-level accuracy on the Standard Definition (SD) Map.

Detailed parameter configurations of the system architecture and multi-task loss function designs will be comprehensively presented in subsequent chapters.

C. Image Encoding Module

This module extracts structured environmental features from a single front-view image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ and transforms them into a Bird’s-Eye View (BEV) representation. The implementation consists of three key stages:

Multi-scale Feature Extraction: A ResNet-101[45] backbone network is employed to extract multi-scale feature pyramids in the perspective view (PV):

$$\{\mathbf{F}_{\text{pv}}^1, \mathbf{F}_{\text{pv}}^2, \mathbf{F}_{\text{pv}}^3, \mathbf{F}_{\text{pv}}^4\} = \Phi_{\text{img}}(\mathcal{I}) \quad (2)$$

where $\mathbf{F}_{\text{pv}}^i \in \mathbb{R}^{H_i \times W_i \times C_i}$ denotes the feature map at the i -th level. Multi-resolution information is fused across different layers via skip connections, enhancing the model’s ability to perceive complex scenes.

Depth Probability Distribution Prediction: To improve geometric modeling accuracy, a parallel depth estimation branch is introduced to predict per-pixel depth distributions from the feature pyramid:

$$\mathcal{D} = \Psi_{\text{depth}} \left(\bigoplus_{i=1}^4 \text{UpSample}(\mathbf{F}_{\text{pv}}^i) \right) \quad (3)$$

Here, $\mathcal{D} \in \mathbb{R}^{H \times W \times D}$ is the depth distribution tensor, where D denotes the number of discrete depth bins, Ψ_{depth} is the depth prediction subnetwork, and \bigoplus represents feature concatenation.

Differentiable View Transformation: Employs a polar-Cartesian dual projection like OrienterNet[7] and LSS[46]:

$$\mathbf{F}_{\text{bev}} = \mathcal{P}_{\text{cart}}(\mathcal{P}_{\text{polar}}(\mathbf{F}_{\text{pv}}, \mathcal{D}), \mathcal{C}) \quad (4)$$

where:

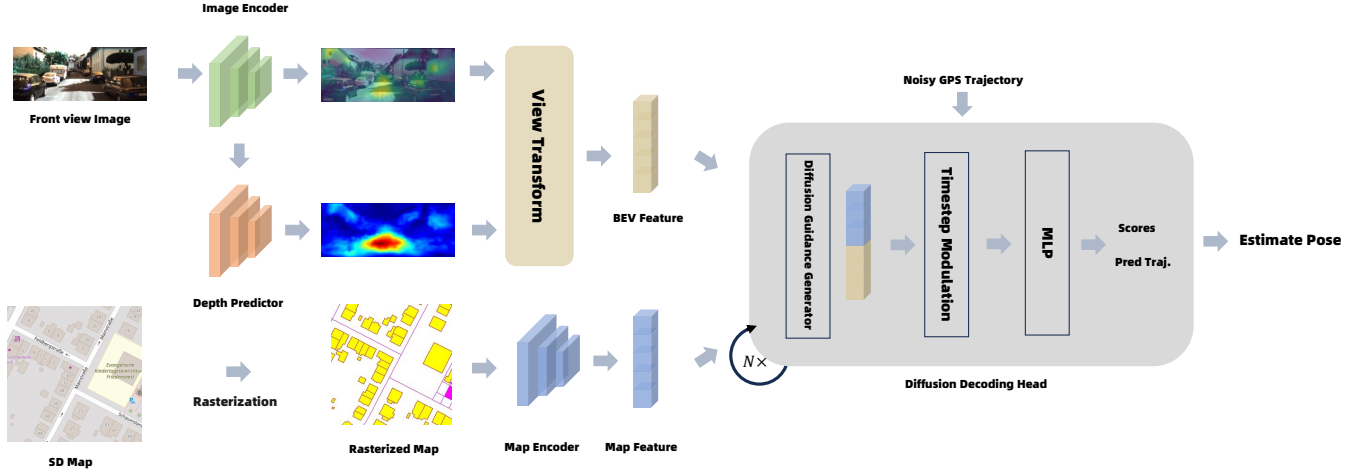


Fig. 2. Architecture of DiffVL. As the first visual localization framework built upon diffusion models, our system pioneers a paradigm shift from traditional matching-based approaches to a generative formulation. The architecture accepts three critical inputs: (i) a monocular front-view RGB image capturing immediate scene context, (ii) standard-definition (SD) map data providing structural priors, and (iii) a noisy GPS trajectory offering coarse positional cues. Central to our innovation, the Image Encoding Module transforms perspective views into geometrically consistent Bird’s-Eye-View (BEV) features, while the Map Encoding Module extracts topological representations from SDmaps. These complementary features undergo multi-modal fusion to generate conditioning features for our novel diffusion module—the core component that fundamentally redefines visual localization as a conditional generation task. Through iterative reverse diffusion steps, this module progressively denoises the corrupted GPS input, transforming unreliable sensor measurements into precise 3-DoF pose estimates. This generative approach marks the first successful application of diffusion models to visual localization, establishing a new trajectory refinement paradigm.

- $\mathcal{P}_{\text{polar}}$: Polar coordinate projection with scale priors \mathcal{D}
- $\mathcal{P}_{\text{cart}}$: Polar-to-Cartesian coordinate transformation
- \mathcal{C} : Camera intrinsic/extrinsic parameters

yielding BEV features $\mathbf{F}_{\text{bev}} \in \mathbb{R}^{B \times C \times H_b \times W_b}$.

D. Map Encoding Module

This module constructs structured environmental priors from open-source map data through four processing stages:

Map Data Acquisition: A bounding box is computed based on the spatial distribution of historical GPS trajectory \mathbf{P}^{GPS} , and the corresponding vector map data for the region of interest is retrieved from OpenStreetMap (OSM). This ensures spatial consistency between the map information and the vehicle’s current location, forming the foundation for global prior construction.

Semantic Rasterization: The vector map data is converted into a three-channel RGB image: Channel 1 encodes the road network (including highways, arterials, and local roads), channel 2 represents building footprints, and channel 3 encodes natural features (such as vegetation, water bodies, and terrain). The rasterization resolution aligns with that of the BEV features (0.5m/pixel), resulting in a map image $\mathbf{M}_{\text{rgb}} \in \mathbb{R}^{X \times Y \times 3}$. This approach is inspired by previous modeling techniques in [7], with the objective of improving the system’s understanding of static environmental structures.

Hierarchical Feature Extraction: A VGG16[47] architecture is used to extract features from the rasterized map:

$$\mathbf{F}_{\text{map}} = \Psi_{\text{MapEnc}}(\mathbf{M}_{\text{rgb}}) \quad (5)$$

The resulting \mathbf{F}_{map} is a compressed feature map that captures key priors such as road topology and traversability

constraints that emphasize effective representation learning of structured map information.

E. Diffusion Guidance Generator

This module achieves deep fusion of visual perception and map priors to generate multi-scale contextual features for the diffusion model, comprising of multimodal feature fusion and diffusion decoding head:

Multimodal Feature Fusion: The BEV and map features undergo dimensional alignment and contextual integration:

$$\mathbf{F}_{\text{cond}} = \Gamma(\phi_{\text{bev}}(\mathbf{F}_{\text{bev}}), \psi_{\text{map}}(\mathbf{F}_{\text{map}})) \quad (6)$$

where:

- ψ_{map} : Map feature compression and spatial restructuring
- ϕ_{bev} : BEV feature projection and spatial alignment
- Γ : Cross-modal fusion operator

yielding a unified representation $\mathbf{F}_{\text{cond}} \in \mathbb{R}^{B \times C_f \times H \times W}$ that synthesizes visual perception and semantic priors.

Diffusion Decoding Head: This module implements conditional trajectory generation and pose refinement through diffusion modeling. Firstly, historical trajectories are normalized and noise-injected through a linear diffusion process. We normalize historical GPS trajectories $\mathbf{P}^{\text{GPS}} \in \mathbb{R}^{T \times 2}$ to $[-1, 1]$ range:

$$\mathbf{P}_{\text{norm}} = \text{norm}_{\text{odo}}(\mathbf{P}^{\text{GPS}}) \quad (7)$$

then injects Gaussian noise:

$$\mathbf{P}_t = \sqrt{\bar{\alpha}_t} \mathbf{P}_{\text{norm}} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (8)$$

Inspired by DiffusionDrive[48], during training, the diffusion decoder takes as input N_{anchor} noisy trajectories



Fig. 3. The localization results of our method on the KITTI dataset. In these visualizations, the red trajectory represents the ground truth (GT) GPS trajectory from the dataset, while the blue trajectory is the noisy GPS trajectory we synthetically generate. Given the noisy blue trajectory and a single image as input, our method produces the refined green "Generated Location" trajectory.

$\{\mathbf{p}_t^{\text{gps}}\}_{k=1}^{\text{anchor}}$ and predicts classification scores $\{\hat{s}_k\}_{k=1}^{N_{\text{anchor}}}$ and denoised trajectories $\hat{\mathbf{p}}_{k=1}^{N_{\text{anchor}}}$:

$$\{\hat{s}_k, \hat{\mathbf{p}}_k\}_{k=1}^{N_{\text{anchor}}} = f_{\theta} \left(\{\mathbf{p}^k\}_{k=1}^{N_{\text{anchor}}}, \mathbf{F}_{\text{cond}} \right) \quad (9)$$

where \mathbf{F}_{cond} represents the conditional information. We assign the noisy trajectory around the closest anchor to the ground truth trajectory τ_{gt} as positive sample ($y_k = 1$) and others as negative samples ($y_k = 0$).

F. Loss Formulation

The total training objective combines trajectory refinement and localization losses:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{diff}}}_{\text{Trajectory Refinement}} + \alpha \underbrace{\mathcal{L}_{\text{loc}}}_{\text{Localization Prior}} \quad (10)$$

where α balances the contribution of the localization loss.

1) *Trajectory Refinement Loss*: Guides diffusion-based trajectory generation through multi-modal optimization:

$$\mathcal{L}_{\text{diff}} = \sum_{k=1}^{N_{\text{anchor}}} [y_k \|\hat{\mathbf{p}}_k - \tau_{\text{gt}}\|_1 + \lambda \mathcal{L}_{\text{BCE}}(\hat{s}_k, y_k)] \quad (11)$$

with $y_k = \mathbb{I} \left[k = \underset{j}{\text{argmin}} \|\mathbf{p}_j - \tau_{\text{gt}}\|_2 \right]$, enforcing mode selection around ground truth anchors.

2) *Localization Loss*: Provides spatial regularization through BEV-map matching:

$$\mathcal{L}_{\text{loc}} = -\log \mathbf{P}(\tau_{\text{gt}} | \mathbf{S}) \quad (12)$$

$$\mathbf{S} = \text{Match}(\mathbf{F}_{\text{bev}}, \mathbf{F}_{\text{map}}) \quad (13)$$

implemented via label-smoothed negative log-likelihood that accounts for annotation uncertainties. This multi-objective design enables *Precise trajectory generation* via $\mathcal{L}_{\text{diff}}$ and *Geometric consistency* via \mathcal{L}_{loc} .

IV. EXPERIMENT

A. Implementation Details

Input Representation. Our method follows the setup of previous work for a fair comparison. We use a single front-view image as the visual input. The map input is a $128\text{m} \times 128\text{m}$ tile extracted from a rasterized navigation map, centered on the ego-vehicle’s noisy GPS position. The map tile has a resolution of 0.5 meters per pixel (mpp).

Training Setup. To simulate real-world GPS errors and train a robust model, we apply random perturbations to the ground truth pose for each training sample. These perturbations are sampled uniformly from a rotation range of $\theta \in [-30^\circ, 30^\circ]$ and a translation range of $t \in [-30\text{m}, 30\text{m}]$. Our model is trained end-to-end using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-2} . The implementation is based on PyTorch. We train the DiffVL model on a single NVIDIA RTX 2080 GPU.

B. Datasets

We conduct extensive experiments on three large-scale and diverse datasets: KITTI [23], MGL [7], and nuScenes [24]. Below, we provide a brief introduction to each and highlight the specific challenges they pose.

KITTI. The KITTI dataset [23] is an authoritative benchmark widely used in the autonomous driving domain. Collected in and around Karlsruhe, Germany, it covers diverse real-world traffic scenarios, ranging from dense urban streets and rural roads to high-speed highways. Its challenging sequences with dynamic objects and varied lighting conditions make it ideal for evaluating localization robustness. KITTI provides precisely synchronized and calibrated multi-modal sensor data, accompanied by ground truth poses generated by a high-grade GPS/IMU system. We adhere to the official training and testing splits for our experiments.

MGL. The MGL dataset was introduced by OrienterNet [7] to facilitate research in large-scale visual geo-localization. It was collected from the Mapillary platform and comprises over 760k images from 12 cities across Europe and the US. This dataset is particularly challenging due to its immense diversity; images were captured by various cameras (hand-held, mounted on cars, bikes) under a wide range of weather and lighting conditions. This diversity tests the generalization capability of a model. All images are accompanied by ground truth (GT) poses and OpenStreetMap (OSM) data. As of this writing, data from two cities (Amsterdam and Vilnius) are no longer accessible. Therefore, we utilize the data from the remaining 10 cities for both training and evaluation.

nuScenes. The nuScenes [24] dataset is a widely-used large-scale autonomous driving dataset, known for its comprehensive sensor suite and complex urban environments. It consists of 1,000 driving scenarios captured in Boston and Singapore, with each scene being 20s long. The dataset

features dense traffic, complex intersections, and significant pedestrian activity, making it an excellent testbed for evaluating performance in safety-critical situations. We follow the official data splits, using the training set of 750 sequences (approx. 28,000 frames) and the validation set of 150 sequences (approx. 6,000 frames) for our experiments.

C. Localization Results

Quantitative Analysis on KITTI. We first evaluate our method on the KITTI dataset, comparing it with state-of-the-art methods including OrienterNet[7], DSM[49], VIGOR[50], and BeyondRetrieval[29]. Following standard evaluation protocols, we use Lateral Recall@Xm, Longitudinal Recall@Xm, and Orientation Recall@X° as our primary metrics. As presented in Table 1, DiffVL significantly outperforms all baseline methods across every metric.

Performance on Large-Scale MGL and nuScenes Datasets. To demonstrate the scalability and generalization of our approach, we conduct further comparisons on the MGL and nuScenes datasets. The evaluation metrics are Recall Accuracy (RA) at distance thresholds of 1, 2, 5, 10 meters and Orientation Recall Accuracy at thresholds of 1, 2, 5, 10 degrees. The experimental results, summarized in Table 2, confirm the superiority of our method. On the highly diverse MGL dataset, DiffVL’s consistent lead suggests that our model learns a generalizable representation rather than overfitting to a specific city or camera type. On the complex urban nuScenes dataset, our method’s strong performance underscores its robustness in dense traffic and perceptually challenging scenarios.

Visualization of Localization Results. Fig.3 visualizes the localization results of our method on the KITTI dataset. In these visualizations, the red trajectory represents the ground truth (GT) GPS trajectory from the dataset, while the blue trajectory is the noisy GPS trajectory we synthetically generate. Given the noisy blue trajectory and a single image as input, our method produces the refined green "Generated Location" trajectory.

As shown in the Fig.3, our method’s output trajectory closely aligns with the ground truth in the purple region (which corresponds to the location of the input image), achieving high-quality localization. It is worth noting that for GPS points outside the purple region, the accuracy of the generated trajectory shows a slight decrease. This is expected, as our inference process only utilizes the single image from the purple region; the visual information corresponding to the other GPS points is unknown, thus precluding accurate visual localization for those points. However, this demonstrates that in the region where visual information is available, our method can achieve high-quality visual localization performance.

D. Ablation Study

To isolate and verify the contribution of our core Trajectory Refinement module, we performed a targeted ablation study on the KITTI dataset. We configured a baseline variant, denoted as w/o Trajectory Refinement, by removing the

TABLE I

QUANTITATIVE COMPARISON ON THE KITTI DATASET. ALL METRICS ARE RECALL (%), WHERE HIGHER IS BETTER. EACH CELL IS COLORED TO INDICATE THE BEST PERFORMANCE IN EACH COLUMN.

Method	Lateral Recall (%)			Longitudinal Recall (%)			Orientation Recall (%)		
	1m	3m	5m	1m	3m	5m	1°	3°	5°
DSM	10.77	31.37	48.24	3.87	11.73	19.50	3.53	14.09	23.95
VIGOR	17.38	48.20	70.79	4.07	12.52	20.14	-	-	-
BeyondRetrieval	27.82	59.79	72.89	5.75	16.36	26.48	18.42	49.72	71.00
OrienterNet	51.26	84.77	91.81	22.39	46.79	57.81	20.41	52.24	73.53
Ours	65.95	90.08	94.67	23.51	51.72	63.03	26.00	66.07	84.27

TABLE II

QUANTITATIVE COMPARISON ON THE MGL AND nuScenes DATASETS. ALL METRICS ARE RECALL ACCURACY (%), WHERE HIGHER IS BETTER. THE BEST PERFORMANCE FOR EACH COLUMN WITHIN EACH DATASET GROUP IS HIGHLIGHTED.

Dataset	Method	Position Recall (%)				Orientation Recall (%)			
		1m	2m	5m	10m	1°	2°	5°	10°
MGL	OrienterNet	10.78	29.88	54.72	67.25	18.98	35.15	63.03	76.63
	Ours	11.07	31.46	57.23	69.30	19.57	35.74	64.79	77.91
nuScenes	OrienterNet	2.89	6.01	18.57	38.49	9.30	16.86	35.40	55.81
	Ours	15.70	28.83	56.74	79.20	19.32	37.46	70.41	86.91

TABLE III

ABLATION STUDY OF THE TRAJECTORY REFINEMENT MODULE ON THE KITTI DATASET. THE METRIC IS POSITION RECALL ACCURACY, WHERE HIGHER IS BETTER. THE BEST RESULTS ARE HIGHLIGHTED.

Method	Position Recall			
	1m	2m	5m	10m
w/o Trajectory Refinement	0.0978	0.2871	0.5325	0.6591
Ours	0.1554	0.3530	0.5935	0.7075

diffusion head and its associated loss ($\mathcal{L}_{\text{diff}}$). This variant relies solely on the BEV-map matching mechanism for localization, similar to conventional approaches.

The results, summarized in Table 3, clearly showcase the module’s critical role. Removing the trajectory refinement leads to a significant degradation in performance across all position recall metrics. It is the diffusion model’s ability to denoise the temporal GPS sequence and impose a kinematically coherent prior that is essential for achieving high-precision results. This ablation provides compelling evidence that our proposed conditional denoising paradigm is the key driver of DiffVL’s state-of-the-art performance.

V. CONCLUSION

In this paper, we introduced DiffVL, a novel framework that pioneers a new paradigm for visual localization by reformulating the task as a conditional GPS denoising problem. Our core contribution is to reframe noisy GPS signals as a valuable generative prior, providing a methodological basis for recovering high-precision poses using a diffusion model. We actualize this through a dual-objective training strategy

that synergistically guides the model to produce kinematically coherent trajectories while simultaneously learning a geometrically-grounded representation via visual-to-map alignment. Extensive experiments on large-scale datasets, including KITTI, MGL, and nuScenes, validate our approach, demonstrating that DiffVL consistently achieves state-of-the-art performance.

REFERENCES

- [1] I. Yaqoob, L. U. Khan, S. A. Kazmi, M. Imran, N. Guizani, and C. S. Hong, “Autonomous driving cars in smart cities: Recent advances, requirements, and challenges,” *IEEE Network*, vol. 34, no. 1, pp. 174–181, 2019.
- [2] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, *et al.*, “Towards fully autonomous driving: Systems and algorithms,” in *2011 IEEE intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 163–168.
- [3] W. Cai, J. Peng, Y. Yang, Y. Zhang, M. Wei, H. Wang, Y. Chen, T. Wang, and J. Pang, “Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance,” *arXiv preprint arXiv:2505.08712*, 2025.
- [4] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, *et al.*, “Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 037–12 047.
- [5] I. A. Barsan, S. Wang, A. Pokrovsky, and R. Urtasun, “Learning to localize using a lidar intensity map,” *arXiv preprint arXiv:2012.10902*, 2020.
- [6] M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.
- [7] P.-E. Sarlin, D. DeTone, T.-Y. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulo, R. Newcombe, P. Kotschieder, and V. Balntas, “OrienterNet: Visual Localization in 2D Public Maps with Neural Matching,” in *CVPR*, 2023.
- [8] Z. Zhou, Z. Qi, L. Cheng, and G. Xiong, “Seglocnet: Multimodal localization network for autonomous driving via bird’s-eye-view segmentation,” *arXiv preprint arXiv:2502.20077*, 2025.

- [9] L. Gao, J. Zhang, L. Zhang, and D. Tao, "Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2825–2833.
- [10] L. Gao, L. Zhang, and Q. Zhang, "Addressing domain gap via content invariant representation for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7528–7536.
- [11] L. Gao, D. Nie, B. Li, and X. Ren, "Doubly-fused vit: Fuse information from vision transformer doubly with local representation," in *European Conference on Computer Vision*. Springer, 2022, pp. 744–761.
- [12] C. Xing, G. Li, and L. Zhang, "Bsam: Bidirectional scene-aware mixup for unsupervised domain adaptation in semantic segmentation," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 54–66.
- [13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [15] M. Jiang, Y. Bai, A. Cornman, C. Davis, X. Huang, H. Jeon, S. Kulshrestha, J. Lambert, S. Li, X. Zhou, et al., "Scenediffuser: Efficient and controllable driving simulation initialization and rollout," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 729–55 760, 2024.
- [16] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [17] J. Li, M. Zhang, N. Li, D. Weyns, Z. Jin, and K. Tei, "Generative ai for self-adaptive systems: State of the art and research roadmap," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 19, no. 3, pp. 1–60, 2024.
- [18] S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang, "Trackvla: Embodied visual tracking in the wild," *arXiv preprint arXiv:2505.23189*, 2025.
- [19] X. Jiang, Y. Ma, P. Li, L. Xu, X. Wen, K. Zhan, Z. Xia, P. Jia, X. Lang, and S. Sun, "Transdiffuser: End-to-end trajectory generation with decorrelated multi-modal representation for autonomous driving," *arXiv preprint arXiv:2505.09315*, 2025.
- [20] M. A. Qudus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," *Transportation research part c: Emerging technologies*, vol. 15, no. 5, pp. 312–328, 2007.
- [21] O. Pink, "Visual map matching and localization using a global feature map," in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008, pp. 1–7.
- [22] Z. Huang, S. Qiao, N. Han, C.-a. Yuan, X. Song, and Y. Xiao, "Survey on vehicle map matching techniques," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 55–71, 2021.
- [23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [24] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [25] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong, et al., "Aads: Augmented autonomous driving simulation using data-driven algorithms," *Science robotics*, vol. 4, no. 28, p. eaaw0863, 2019.
- [26] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, et al., "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.
- [27] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians: Modeling dynamic urban scenes with gaussian splatting," in *ECCV*, 2024.
- [28] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European conference on computer vision*. Springer, 2012, pp. 15–29.
- [29] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 010–17 020.
- [30] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.
- [31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [32] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.
- [33] J. Zhang, Y. Yao, and B. Deng, "Fast and robust iterative closest point," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3450–3466, 2021.
- [34] Z. Zhang, M. Xu, W. Zhou, T. Peng, L. Li, and S. Poslad, "Bev-locator: An end-to-end visual semantic localization network using multi-view images," *Science China Information Sciences*, vol. 68, no. 2, p. 122106, 2025.
- [35] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bev-former: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [36] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.
- [37] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.
- [38] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou, et al., "Streetcrafter: Street view synthesis with controllable video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 822–832.
- [39] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [40] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, et al., "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9644–9653.
- [41] C. R. Shipan and C. Volden, "The mechanisms of policy diffusion," *American journal of political science*, vol. 52, no. 4, pp. 840–857, 2008.
- [42] G. Barquero, S. Escalera, and C. Palmero, "Belfusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2317–2327.
- [43] T. Wang, C. Zhang, X. Qu, K. Li, W. Liu, and C. Huang, "Diffad: A unified diffusion modeling approach for autonomous driving," *arXiv preprint arXiv:2503.12170*, 2025.
- [44] R. Zhao, Y. Fan, Z. Chen, F. Gao, and Z. Gao, "Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy," *arXiv preprint arXiv:2505.19516*, 2025.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [46] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European conference on computer vision*. Springer, 2020, pp. 194–210.
- [47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [48] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, et al., "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 037–12 047.
- [49] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.
- [50] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3640–3649.