This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3546853

LangLoc: Language-Driven Localization via Formatted Spatial Description Generation

Weimin Shi, Changhao Chen, Kaige Li, Yuan Xiong, Xiaochun Cao, *Senior Member, IEEE*, Zhong Zhou

***Abstract*—Existing localization methods commonly employ vision to perceive scene and achieve localization in GNSS-denied areas, yet they often struggle in environments with complex lighting conditions, dynamic objects or privacy-preserving areas. Humans possess the ability to describe various scenes using natural language to help others infer the location by recognizing or recalling the rich semantic information in these descriptions. Harnessing language presents a potential solution for robust localization. Thus, this study introduces a new task, Language-driven Localization, and proposes a novel localization framework, LangLoc, which determines the user's position and orientation through textual descriptions. Given the diversity of natural language descriptions, we first design a Spatial Description Generator (SDG), foundational to LangLoc, which extracts and combines the position and attribute information of objects within a scene to generate uniformly formatted textual descriptions. SDG eliminates the ambiguity of language, detailing the spatial layout and object relations of the scene, providing a reliable basis for localization. With generated descriptions, LangLoc effortlessly achieves language-only localization using text encoder and pose regressor. Furthermore, LangLoc can add one image to text input, achieving mutual optimization and feature adaptive fusion across modalities through two modality-specific encoders, cross-modal fusion, and multimodal joint learning strategies. This enhances the framework's capability to handle complex scenes, achieving more accurate localization. Extensive experiments on the Oxford RobotCar, 4-Seasons, and Virtual Gallery datasets demonstrate LangLoc's effectiveness in both language-only and visual-language localization across various outdoor and indoor scenarios. Notably, LangLoc achieves noticeable performance gains when using both text and image inputs in challenging conditions such as overexposure, low lighting, and occlusions, showcasing its superior robustness.**

***Index Terms*—Language-driven Localization, Visual Localization, Spatial Description, Large-Language Model**

I. Introduction

LOCALIZATION aims to determine the user's position and orientation in a 3D scene, which is crucial for intelligent machines such as robots [1], [2], autonomous vehicles [3], [4], and virtual/augmented reality systems [5], [6]. While traditional Global Navigation Satellite Systems (GNSS) provide global location information, their signals can be attenuated or blocked in underground, densely built urban areas or tunnels [7]. Intuitively, humans possess the ability to describe and comprehend various scenes through natural language. As shown in Fig. 1, in GNSS-denied environments such as downtown streets with high buildings or underground facilities, humans could localize themselves and share location information by verbally describing notable scene components, without relying on localization sensors. Similarly, by integrating language, intelligent machines can more precisely capture the high-level semantics of scenes, such as specific functions, behavioral patterns, and event backgrounds of objects in the scene [8]. This enhances their spatial perception of the scene and introduces a novel approach to practical localization applications.

Currently, these intelligent machines normally leverage visual information for localization in GNSS-denied regions. Integrating deep learning techniques into this domain has witnessed remarkable progress, particularly in pose regression using deep neural networks directly. Pioneering works, PoseNet [9] shows the ability to train deep neural networks on extensive datasets to map images directly to poses. Building upon this foundation, AtLoc [10] and MapNet [11] further introduce attention mechanisms or geometric constraints for improved accuracy. Similarly, AD-PoseNet [12] refines localization performance by filtering dynamic objects. Following these advancements,

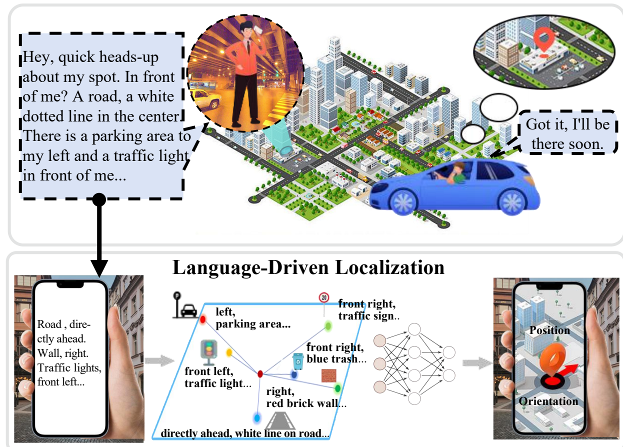


Fig. 1. **Language-driven localization.** Humans can naturally describe their surroundings using language to localize themselves and share their location with others. This work aims to impart machines with a comparable capability by proposing a language-driven localization method, involving spatial textual descriptions generation and deep neural networks based pose regression.

This work was supported in part by the Science and Technology Project of Hainan Provincial Department of Transportation under Grant HNJTT-KXC-2024-3-22-02 and in part by the National Natural Science Foundation of China under Grant 62272018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alessio Del Bue. (Corresponding author: Zhong Zhou.)

Weimin Shi, Kaige Li are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China (shiwm@buaa.edu.cn; lkg@buaa.edu.cn).

Changhao Chen is with the Thrust of Intelligent Transportation and Thrust of Artificial Intelligence, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China (changhaochen@hkust-gz.edu.cn).

Yuan Xiong, Xiaochun Cao are with the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China (xiongy89@mail.sysu.edu.cn; caoxiaochun@mail.sysu.edu.cn).

Zhong Zhou is with the Zhongguancun Laboratory, Beijing, and the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China.

c2f-MS-Trans [13] introduces a mixed classification-regression architecture, achieving precise cross-scene localization.

Despite vision-based localization performing well in controlled environments, it often fails under adverse conditions such as changes in illumination and the presence of dynamic objects in the scene. In contrast, language can provide more abstract and robust cues for the scene, offering a potential solution for localization. However, research into developing techniques for understanding spatial scenes and localization based on language is still relatively limited [14]. Against this backdrop, the emergence of Large Language Models (LLM) presents new possibilities for understanding complex scenes [15]. These models have made notable strides in handling the diversity and complexity of natural language, demonstrating their potential in spatial description and localization tasks [16]. However, the inherent ambiguity and randomness of natural language, combined with the dynamic complexity of scenes, continue to make language-driven localization a challenging endeavor.

To tackle these challenges, we introduce a new task: language-driven localization, which determines a user's position and orientation in a scene through language descriptions. Our solution, a novel **Lang**uage-driven **Loc**alization framework, **LangLoc**, mimics human abilities to infer location using language, enabling localization under diverse scenes with either language-only or vision-language. Given the inherent ambiguity and randomness of language, there is a scarcity of language data for accurate localization. Thus, we propose a Spatial Description Generator (SDG), comprising two modules: Spatial Scene Description (SSD) and Formatted Text Generation (FTG). Considering the distinct roles of objects in localization tasks, SSD specifically extracts and combines the position and key attributes of each object to generate a detailed spatial scene description. Subsequently, FTG guides the LLM (e.g., GPT-3 [17]) in excluding dynamic objects from the descriptions generated by SSD, organizing them into a unified format. This reduces ambiguity and precisely conveys the spatial layout and object relationships, providing a reliable basis for localization. Based on these generated descriptions, LangLoc effortlessly achieves language-only localization using just two components: a text encoder and a pose regressor. Further, when visual data is available, LangLoc can also adaptively integrate linguistic semantics with visual spatial cues through two modality-specific encoders, cross-modal fusion, and multimodal joint learning strategies. This enhances independent learning and mutual supplementation between modalities, thereby improving the accuracy and robustness of localization.

Experiments on the Oxford RobotCar dataset [18] demonstrate that LangLoc achieves a median localization error of 29.48m and 6.79° in language-only localization. This performance meets the benchmark commonly accepted in large-scale localization studies, where an error of less than 50m is considered effective in city-scale [19]–[21]. Furthermore, even with solely human natural language input, LangLoc demonstrates effective localization capabilities. Finally, by integrating both image and text inputs, LangLoc achieves significant performance gains on the Oxford RobotCar, 4-Seasons, and Virtual Gallery datasets, across both indoor and outdoor scenarios in vision-language localization mode. Notably, LangLoc also exhibits stronger robustness in image degradations and missing modalities, showcasing a promising performance advantage.

In summary, our main contributions are as follows:

- We introduce a new task: language-driven localization, aiming to determine the user's position and orientation via natural language.
- We propose a Spatial Description Generator to generate formatted textual descriptions of scenes, facilitating effective language-driven localization.
- We propose LangLoc, a novel localization framework, supporting both language-only and vision-language localization, accommodating various input data types.
- Extensive experiments conducted on public datasets demonstrate the effectiveness of LangLoc in both language-only and vision-language localization.

II. Related Work

Vision-based localization remains an active area of research. Existing works leverage images for global-scale geolocalization through visual-geographic matching, such as Translocator [22], ISNs [23], CPlaNet [24], and others [25]–[28]. Building upon geolocalization, visual localization estimates the camera's 6-DoF pose within a known environment using images. However, changes in seasons, weather, and environment make accurate visual localization challenging. Recently, advances in deep learning offer new ways to address this issue by learning from large-scale datasets. This paper reviews deep learning-based visual localization methods and language-driven approaches, highlighting the differences between existing methods and our proposed approach for more effective visual localization.

A. Deep Learning based visual Localization

A pioneering work in this field is PoseNet [9], which integrates a GoogLeNet [29] backbone with a multilayer perceptron (MLP) for end-to-end supervised learning. GeoPoseNet [30] and c2f-MS-Trans [13] concurrently optimize position and orientation learning, refining the accuracy of spatial information through balanced parameter adjustments. Atloc [10] introduces a self-attention mechanism for focused key information processing, facilitating precise camera pose regression through an MLP head. Building on these frameworks, some studies explore techniques for extracting robust visual features to handle scene variations. For instance, Translocator [22] creates stable feature representations under changing appearances through semantic segmentation. Similarly, LT-Loc [31] employs semantic segmentation images to tackle the challenges of long-term visual localization. To mitigate the impact of dynamic objects on visual localization, AD-PoseNet [12] enhances accuracy by quantifying uncertainty in pose estimation, enabling CNN to ignore interference from dynamic objects. CoordiNet [32] adopts a joint training approach for pose prediction and uncertainty estimation, effectively removing outliers of the trajectory and achieving robust performance in single-view localization. Lens [33] heightens accuracy through novel view synthesis. ImPosing [34] efficiently connects query images to implicit maps, offering precise real-time localization in large urban scenarios. EffLoc [35] designs an efficient visual transformer via diversified inputs, redundancy reduction, and capacity expansion, enhancing efficiency in outdoor urban localization. In addition, multi-frame methods improve localization by incorporating temporal context. MapNet [11] incorporates visual odometry and multi-frame data alongside visual relocalization for refined pose estimates. Atloc+ [10] also improves localization by extending the network to support multi-view inputs. GNNMapNet [36] enriches environmental understanding using graph neural networks for feature extraction from multi-view images. To handle environmental changes effectively, RobustLoc [37] combines graph neural networks with a neural graph diffusion model, providing robust multi-view representations to boost localization performance.

Besides vision-based methods, LiDAR-based methods, such as HypLiLoc [38] and DiffLoc [39], achieve centimeter-level localization accuracy by reconstructing 3D scenes using LiDAR sensors. However, the high resource demands of dense point cloud processing restrict their scalability in urban environments. In contrast, visual methods show broader applicability due to their lower computational and storage costs. However, visual methods struggle with image degradation caused by dynamic elements or environmental changes, especially in complex scenes [40]–[42]. In this paper, we propose to leverage the stability of language descriptions to assist localization. By effectively integrating visual and language data, our method shows high spatial localization accuracy and robustness.

B. Language-Driven Applications

In recent years, language-driven applications have attracted widespread attention in artificial intelligence. Large Language Models (LLM) like GPT-3 [17], PaLM [43], and OPT [44], ChatGPT [45] and LLaMA [46] show remarkable capabilities in complex text tasks. These advances have motivated researchers to explore combining visual input with language models, leading to the development of multimodal large language models (MLLM). For instance, MiniGPT-4 [47] and MiniGPT-V2 [48] align cross-modal encoders with language models, offering advanced functions like generating website code from handwritten text. Ferret [49] enhances MLLMs with referencing and grounding, while GLaMM [50] enables user interaction across different levels of granularity in both textual and visual domains. As a result, LLM and MLLM become powerful tools for a range of language-driven tasks [51]–[53]. Some studies utilize MLLMs to create general-purpose visual understanding systems, capable of handling diverse vision-language tasks through unified instructions, such as VistaLLM [54], XGen-MM [55], and InternLM [56], among others [57], [58].

Recent studies explore language-driven spatial intelligence tasks. For instance, CMG-AAL [59] trains agents to understand the correspondence between vision and language, enabling them to navigate to target locations using textual instructions. VoxPoser [8] utilizes LLM to facilitate 3D robotic manipulation responsive to human language. LP-SLAM [60] and TextSLAM [61] integrate textual information into the SLAM system, allowing machines to locate positions using text labels. Text2Pos [62] and Text2Loc [63] are pioneering efforts to tackle large-scale urban localization based on language, yet these methods rely on pre-built databases, locating by querying corresponding image information, and have not yet achieved effective localization directly through language. To improve language efficiency in spatial intelligence, some research [64], [65] explores generating appropriate language descriptions to convey spatial semantics. However, they rely on pre-extracted 3D scene features and extra training, and their descriptions lack effective validation in spatial intelligence tasks.

In contrast, our work leverages LLM to generate spatial descriptions by precisely extracting key spatial attributes from scenes, without the additional training. Utilizing these generated descriptions, our framework can achieve effective language-only localization via an end-to-end strategy, without relying on pre-built localization databases.

III. Task Formulation

In this work, we introduce a new task: language-driven localization, aiming to determine the user's pose, including a position vector $\boldsymbol{p} \in \mathbb{R}^3$ and an orientation vector $\boldsymbol{q} \in \mathbb{R}^4$, via textual descriptions $\boldsymbol{T}$. This task encompasses two modes:

1) Language-only Localization: in this mode, the objective is to achieve localization solely through language. The ambiguity and randomness of natural language pose challenges in parsing spatial layouts and key features. To address this challenge, the primary goal is to generate efficient textual descriptions $\boldsymbol{T}$, using clear semantics to accurately indicate the spatial locations of objects. Then, based on these generated descriptions, the user's pose is precisely regressed:

$$\min_{\phi} \mathbb{E}_{(\boldsymbol{p},\boldsymbol{q},\boldsymbol{T})\sim D}\left[\|(\boldsymbol{p},\boldsymbol{q}) - \phi(\boldsymbol{T})\|_1\right], \quad (1)$$

where D is the dataset, ϕ denotes a neural network trained to process text inputs $\boldsymbol{T}$ and produce the pose $(\boldsymbol{p},\boldsymbol{q})$.

2) Vision-Language Localization: in this mode, we extend the language-only localization to support multimodal inputs, fusion text $\boldsymbol{T}$ and image I inputs to learn the joint feature, thus enabling more accurate and robust pose regression:

$$\min_{\theta,\psi} \mathbb{E}_{(\boldsymbol{p},\boldsymbol{q},\boldsymbol{T},\boldsymbol{I})\sim D}\left[\|(\boldsymbol{p},\boldsymbol{q}) - \theta\left(\psi(\boldsymbol{T},\boldsymbol{I})\right)\|_1\right], \quad (2)$$

where ψ denotes a neural network trained to generate joint feature. θ represents a neural network utilized to predict the pose $(\boldsymbol{p},\boldsymbol{q})$ based on joint feature.

IV. Methodology

To effectively address the challenge of language-driven localization introduced in the preceding section, this section presents a novel localization framework, LangLoc. It offers support for both language-only localization mode and vision-language localization mode, catering to diverse input data types. As shown in Fig. 2, LangLoc starts with the Spatial Description Generator (SDG). SDG extracts spatial information from either images $\boldsymbol{I}$ or human language $\boldsymbol{L}$ and generates formatted text $\boldsymbol{T}$ to precisely describe the spatial scene (Sec. IV-A). In the language-only localization mode, the LangLoc framework utilizes the spatial textual descriptions produced by the SDG for localization (Sec. IV-B). In the vision-language localization mode, the LangLoc framework leverages both text and image as inputs for localization (Sec. IV-C).

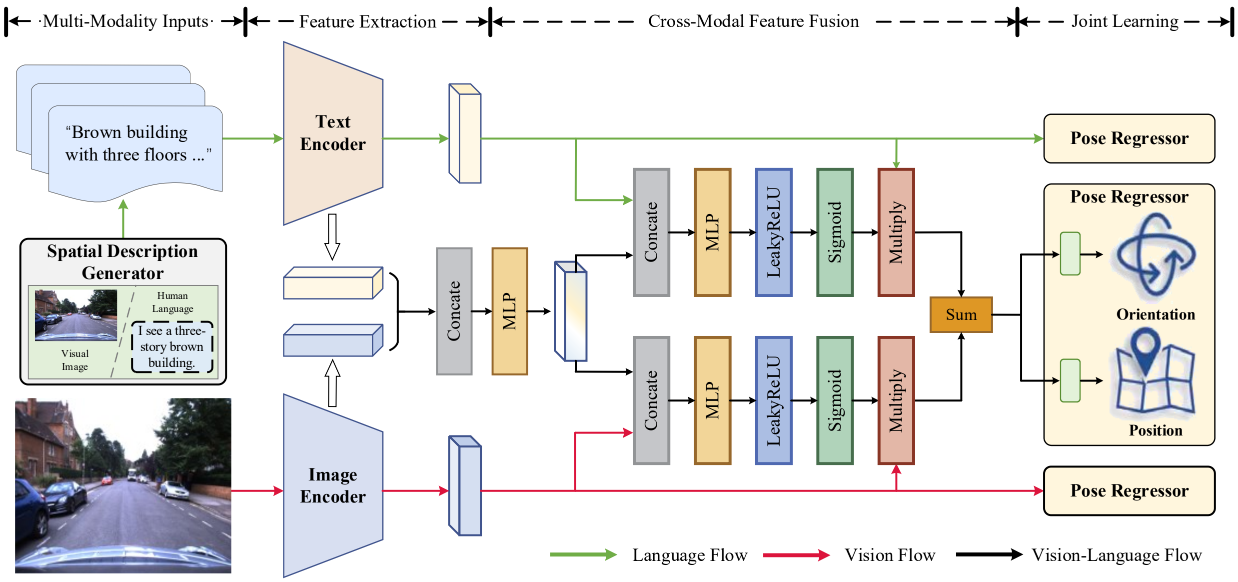


Fig. 2. An overview of our proposed LangLoc framework. LangLoc supports two modes: 1) Language-only localization, which relies solely on text input for localization. In this mode, the input data is processed through the framework's **Language Flow**, involving the SDG, text encoder, and pose regressor, to achieve precise localization. 2) Visual-language localization, which utilizes both image and text inputs for localization. In this mode, input data is processed through **Language**, **Vision**, and **Vision-Language Flows**, utilizing cross-modal feature fusion and joint learning strategies to generate joint features that combine linguistic semantics and visual spatial cues, thereby achieving more precise and robust localization.

A. Spatial Description Generator

Due to the randomness in natural language expression, achieving precise localization directly from either raw language descriptions generated by LLM or humans is challenging. To tackle this issue, we introduce SDG to capture the key spatial information of scenes, which combines spatial information extraction with the reasoning capabilities of LLMs to effectively capture a scene's geometric details and spatial layout. It consists of two components: Spatial Scene Description (SSD) and Formatted Text Generation (FTG). As depicted in Fig. 4, SSD provides detailed spatial data, and FTG translates this into formatted text $\boldsymbol{T}$. This process mitigates the ambiguity in descriptions, enhancing the effectiveness of expressions for spatial features valuable to localization.

*1) **Spatial Scene Description**:* To accurately determine the user's location within a 3D scene, it is crucial to comprehend and extract the vital spatial information from scene objects relevant to the localization task. We conceive the image I as a combination of detected objects $O_i^{j_i} = \{O_1^{j_1}, O_2^{j_2}, \dots, O_i^{j_i}\}$, where $O_i^{j_i}$ represents each object in the image, and i signifies the number of detected objects, j_i denotes the category of the object. Our SSD extracts the spatial position $POS_i^{j_i}$ and specific attributes $A_i^{j_i}$ from objects $O_i^{j_i}$. By using the concatenation operation "+", it synthesizes the spatial information $S_i^{j_i}$. This approach effectively captures both the category information C_i and spatial information S_i of scene:

$$\begin{aligned}\{C_i : S_i\} &= \mathrm{SSD}\left\{O_1^{j_1}, O_2^{j_2}, \dots, O_i^{j_i}\right\} \\ &= \left\{\left(C_1^{j_1} : POS_1^{j_1} + A_1^{j_1}\right) \dots + \left(C_i^{j_i} : POS_i^{j_i} + A_i^{j_i}\right)\right\}\end{aligned} \tag{3}$$

In practice, we initially employ a Multimodal Large-Language Model (MLLM), such as MiniGPT-v2 [48], to obtain the category labels $C_i^{j_i}$ and position bounding boxes B_i for these objects. To determine the position $POS_i^{j_i}$ of the objects within an image, we map each object's bounding box B_i to predefined position descriptions. This mapping is based on the relationship between the geometric center of the object and the image center, following the guidelines outlined in Fig. 3. This procedure replicates human perspective by using the image center as a reference, uniformly indicating objects' relative positions. For example, an object's geometric center in the top 60% and between 10%-40% to the left of the image center is labeled "front left"; if it extends beyond the front 60% but remains within 10% to the left, it is described as "left".

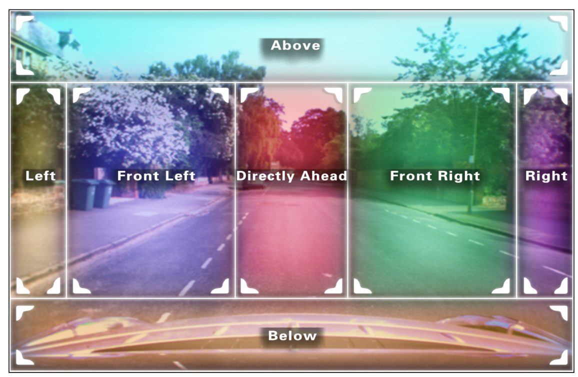


Fig. 3. Transforming Image Regions to Positional Descriptions: Translating object detection bounding box coordinates into textual descriptions. When an object's geometric center falls within a defined region, the corresponding positional description is generated.

Subsequently, to acquire the key attributes $A_i^{j_i}$ of different objects, we guide the MLLM to focus on extracting specific attributes by using prompts related to the categories of objects. In particular, since various objects fulfill different roles in understanding the scene and meeting localization demands, we categorize the objects into key objects and other objects,

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3546853

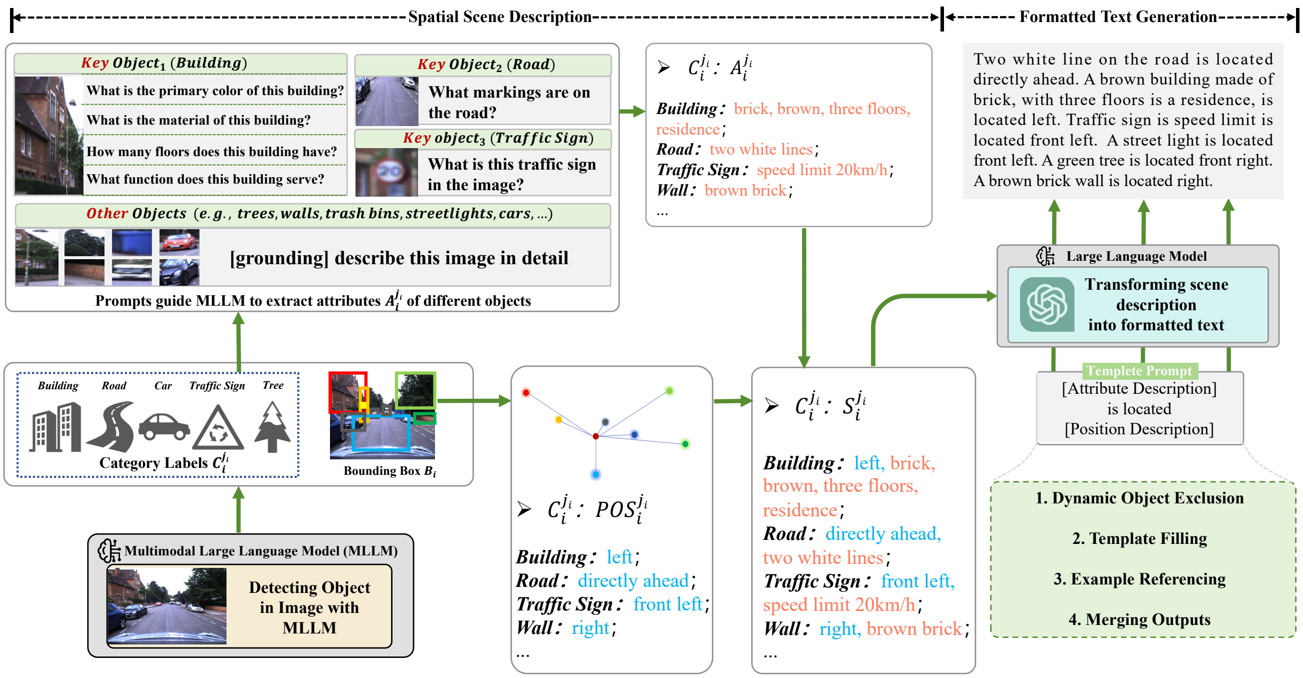


Fig. 4. Spatial Description Generator (SDG) consists of two modules: Spatial Scene Description (SSD) and Formatted Text Generation (FTG). SSD uses a MLLM to identify objects $O_i^{j_i}$ and their bounding boxes B_i, converting these into positional descriptions $POS_i^{j_i}$. It also extracts key attributes $A_i^{j_i}$ using category-based prompts $C_i^{j_i}$, detailing the spatial information of scene objects. FTG then transforms SSD's outputs into uniformly formatted text, ensuring consistency and uniformity in the descriptions. These precise textual descriptions provide a foundation for subsequent language-driven localization task.

as illustrated in Fig. 4. For key objects such as buildings, traffic signs, and streets, we tailor question prompts based on the distinctive features of each object. For instance, we employ a question prompt that concentrates on the building's material (*"What is the primary color of this building?"*), color (*"What is the material of this building?"*), the number of floors (*"How many floors does this building have?"*), and its function (*"What function does this building serve?"*). The responses to these questions, encapsulated as the specific attributes A_i^{building} of the building, in conjunction with its position POS_i^{building}, collectively form the spatial information S_i^{building} of the building:

$$
\begin{aligned}
C_i^{\text{building}} : S_i^{\text{building}} &= \{\text{“building”} : \mathrm{POS}_i^{\text{building}} + \mathrm{A}_i^{\text{building}}\} \\
&= \{\text{“building”} : \text{“front left”}, \text{“brick”}, \text{“brown”}, \\
&\quad \text{“three floors”}, \text{“school”}\}
\end{aligned} \tag{4}
$$

For other objects, we employ a unified prompt, namely, *"[grounding] describe this image in detail"*. This facilitates the MLLM to conduct grounded caption [48], generating a phrase that describes the attributes of detected objects, such as, *"a brown brick wall"*.

As shown in Fig. 5 (in the Example Referencing), SSD systematically extract spatial information from objects, forming a comprehensive spatial scene description. These descriptions are then input into FTG, providing a foundation for accurately expressing key localization features.

*2) **Formatted Text Generation**:* To ensure consistent formatting in language descriptions across scenes and facilitate more efficient extraction of key semantic features for downstream pose estimation, we introduce a Formatted Text Generation module (FTG). This module transforms scene descriptions $\{C_i : S_i\}$ generated by SSD into formatted text $\boldsymbol{T}$:

$$
T = \mathrm{FTG}\left(\{C_i : S_i\}, \mathrm{Template}\right), \tag{5}
$$

where *Template* denotes a template prompt containing multiple operation instructions, guiding the LLM (e.g., GPT-3.5) to perform dynamic object exclusion, template filling, example referencing, and merging outputs, as illustrated in Fig. 5.

Specifically, static objects (such as buildings, roads, traffic signs, etc.) provide more stable and reliable features for localization, while dynamic objects (such as cars, people, etc.) pose challenges due to their impacts on scene appearance and occlusions. Therefore, we first exclude textual descriptions related to dynamic objects to enhance the stability and consistency of the descriptions. In particular, we guide LLM to automatically identify and filter out descriptions related to a predefined set of categories for dynamic objects, such as "Red bus parked under a streetlight" and "Woman wearing skirt walking by the roadside".

Then, we process the remaining object descriptions based on a predefined template. In this process, LLM fills scene descriptions into the template "$[X_i^{j_i}]$ is located at $[Y_i^{j_i}]$", where $X_i^{j_i}$ represents the attribute description of the object, and $Y_i^{j_i}$ refers to the position description. This uniform output format clearly conveys scene features, effectively reducing the ambiguity of language descriptions. Moreover, the designed template guides the LLM to generate object descriptions in a predetermined order, enabling the model to establish an intuitive comparison benchmark between different scene descriptions. From our observations, even minor scene changes, such as the addition, movement, or removal of objects, are reflected in

Template Prompt In Formatted Language Generation

Operation Guide:

1. Dynamic Object Exclusion: First, identify and exclude all information related to persons and cars.

2. Template Filling: Next, process the elements in the scene description according to a predetermined order and template:

- Road: If applicable, output: "[A^{street}] on the road **is located** [POS^{street}]."
- Building: If applicable, output: "A [A_1^{building}] building made of [A_2^{building}], with [A_3^{building}] is [A_4^{building}], **is located** [POS^{building}]."
- Traffic Sign: If applicable, output: "A traffic sign is [A^{sign}] **is located** [POS^{sign}]."
- Other Objects: For other objects, output: "[A^{other}] **is located** [POS^{other}]."

3. Example Referencing:

Example1

Input: ***building***: front left, brick, brown, three floors, school; ***road:*** directly ahead, two white lines; ***A traffic sign***: front left, speed limit 20km/h; ***car***: directly ahead, a white car on a street; ***car***: front left, a blue car on a street; ***wall***: right, a brown brick; ***tree***: front right, a green tree; ***bushes***: directly ahead, blue bushe in front of the wall; ***street light***: front left, a tall street light.

Output: Two white line on the road is located directly ahead. A brown building made of brick, with three floors is a residence, is located left. Traffic sign is speed limit is located front left. A street light is located front left. A green tree is located front right. A red brick wall is located right.

Example2

...

4. Merging Outputs:

Please strictly adhere to the above Operation Guide, first identify and exclude dynamic objects, then organize the static objects according to the template, and finally, referencing the provided examples, output coherent natural language without extra descriptions.

Fig. 5. A Template Prompt in the Formatted Text Generation module (FTG), guides the Large-Language Model (LLM) to exclude dynamic objects from the SSD-generated scene descriptions, transforming them into Formatted Text.

the order and content of the descriptions, thereby accurately describing the changes in scene structure.

Finally, to enhance the LLM's comprehension of these operations, we include specific examples in the prompts, each consisting of complete input-output pairs. After the template filling process, by referencing the given examples, LLM integrates all processed object descriptions into uniformly formatted text descriptions. As depicted in Fig. 5 (in the output section of Example Referencing), the FTG module excludes descriptions of dynamic objects (e.g., cars), describes static scene components (such as streets, buildings, traffic signs, and other objects) in a fixed order, and generates a cohesive, formatted text description. By leveraging the LLM's ability to interpret varied language patterns via prompts, FTG overcomes the limitations of traditional manually defined text-matching rules and can handle diverse scene descriptions, including unformatted language provided by humans (Sec. V-B.2).

B. Language-Only Localization

Based on the formatted text descriptions $\boldsymbol{T}$ generated by SDG, we can further train our LangLoc framework end-to-end to achieve language-only localization, precisely mapping these descriptions to pose.

Specifically, we first apply a pre-trained text encoder f_{enc_t} (e.g., the text encoder of CLIP [66]) to encode the text $\boldsymbol{T}$:

$$\boldsymbol{x}_t = f_{\text{enc}_t}(\boldsymbol{T}), \tag{6}$$

where the dimensionality of $\boldsymbol{x}_t \in \mathbb{R}^C$ is set to $C = 2048$. Subsequently, we assign the encoded feature vector $\boldsymbol{x}_t$ to a pose $\boldsymbol{y} = (\boldsymbol{p}, \boldsymbol{q})$ using a two-layer MLP:

$$[\boldsymbol{p}, \boldsymbol{q}] = \text{MLP}(\boldsymbol{x}_t) \tag{7}$$

During the training process, we optimize the model parameters to minimize the difference between the estimated and actual poses using the L1 loss function:

$$L(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_1 e^{-\beta} + \beta + \|\log \boldsymbol{q} - \log \hat{\boldsymbol{q}}\|_1 e^{-\gamma} + \gamma, \tag{8}$$

where $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})$ represents the ground-truth label of position and orientation. Utilizing the logarithmic form of quaternions, $\log \boldsymbol{q}$, enables us to accurately describe continuous changes in orientation. To address the issue of quaternion non-uniqueness in rotation representation, we ensure all quaternions fall within the same hemisphere during training, thereby assigning a unique quaternion to each rotation:

$$\log \boldsymbol{q} = \begin{cases} \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \cos^{-1}(u), & \text{if} \|\boldsymbol{v}\| \neq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $\boldsymbol{u}$ denotes the real part of the quaternion and $\boldsymbol{v}$ represents its imaginary component. Particularly, to enhance pose estimation accuracy, we further optimize the weights for both position and rotation loss (β and γ) during training, ensuring a balance between position and rotation loss.

By end-to-end training on datasets, LangLoc framework can effectively infer localization information solely from natural language, even in the absence of direct visual inputs. To the best of our knowledge, this is the first work to achieve localization solely using natural language.

C. Vision-Language Localization

We further introduce LangLoc in the vision-language localization mode, as depicted in Fig. 2 (2). This mode extends the input of the language-only localization mode to integrate language with vision, aiming to achieve more precise and robust localization. In this mode, LangLoc initially employs two modality-specific encoders to process text and image inputs, respectively, capturing distinct modality features. Subsequently, it combines these features using cross-modal fusion for a comprehensive latent representation. Finally, multimodal joint learning is utilized to enhance the learning of pose by leveraging the individual capacities of different modalities.

Modality-Specific Encoders: We use a pre-trained text encoder consistent with the language-only localization for extracting text features, and a corresponding image encoder (e.g., the image encoder of CLIP) for image feature extraction:

$$\boldsymbol{x}_v = f_{\text{enc}_v}(\boldsymbol{I}), \tag{10}$$

Cross-Modal Fusion: With text and image features, we introduce a fusion strategy to evaluate feature significance from each modality. Specifically, we first concatenate $\boldsymbol{x}_v$ and $\boldsymbol{x}_t$ along channels to generate $\boldsymbol{x}_c$, followed by convolution. However, although $\boldsymbol{x}_c$ encodes both text and image information, it may introduce redundant noise from each modality for localization. Hence, we apply a scoring function f_{score} to $\boldsymbol{x}_c$

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3546853

to measure each modality's contribution. As shown in Fig. 2, f_{score} concatenates $\boldsymbol{x}_t$ or $\boldsymbol{x}_v$ with $\boldsymbol{x}_z$, producing weights $\boldsymbol{W}_r$ for each modality:

$$\boldsymbol{W}_r = \sigma\left(f_{\text{score}}([\boldsymbol{x}_c;\boldsymbol{x}_r];\boldsymbol{\theta})\right), \quad r \in \{v,t\} \tag{11}$$

where σ denotes the sigmoid function, and θ represents the parameters of $fscore$, which consists of sequential linear layers, each succeeded by a Leaky ReLU activation function.

Finally, we apply weights $\boldsymbol{W}_r$ to corresponding modal features $\boldsymbol{x}_r$ through element-wise multiplication, followed by summing these weighted features:

$$\boldsymbol{x}_z = \sum_{r\in\{v,t\}} \boldsymbol{W}_r \odot \boldsymbol{x}_r \tag{12}$$

Thus, we obtain an effective joint feature representation $\boldsymbol{x}_z$ for downstream pose regression.

Multimodal Joint Learning: Image features excel in capturing scene details and structures, whereas text features offer abstract scene semantics [67]–[69]. To exploit their complementarity, we design a joint learning strategy for vision-language localization. This strategy enables features from different modalities to learn both independently and jointly.

Specifically, we allocate three pose regressors MLP_v, MLP_t, and MLP_z to the visual, language, and fused modalities, respectively, entrusting them with mapping their respective modalities features to poses:

$$[\boldsymbol{p},\boldsymbol{q}] = \text{MLP}_n(\boldsymbol{x}_n), \quad n \in \{v,t,z\} \tag{13}$$

To facilitate the learning process, we introduce a loss function that balances individual and joint learning:

$$L = \lambda \sum_{r\in\{\text{v,t}\}} L_{\text{intra}}(\boldsymbol{y}_r,\hat{\boldsymbol{y}}_r) + L_{\text{joint}}(\boldsymbol{y}_z,\hat{\boldsymbol{y}}_z), \tag{14}$$

where λ is a hyperparameter governing the trade-off between individual and joint learning. Particularly, by minimizing the discrepancy between predicted and ground-truth poses, L_{intra} encourages modality-specific feature learning, while L_{joint} promotes intra-modal and cross-modal learning.

This dual-objective approach ensures that each modality refines its predictions independently through L_{intra}, while the joint learning objective L_{joint} fosters a synergistic improvement across modalities, leveraging the complementary information inherent in each. Consequently, LangLoc becomes adept at extracting and utilizing modality-specific cues, enhancing its ability to integrate these cues effectively across different modalities, thereby demonstrating superior localization performance.

V. Experiments

In this section, we extensively test the LangLoc framework on public datasets. Specifically, we first evaluate the effectiveness of Spatial Description Generator (SDG) (Sec. V-A) and explore the feasibility of using human natural language for language-only localization (Sec. V-B). Subsequently, we evaluate the vision-language localization mode through quantitative (Sec. V-C) and qualitative experiments (Sec. V-D), offering a comprehensive comparison with existing visual localization approaches. Finally, we analyze the robustness of LangLoc in several challenging scenarios (Sec. V-E).

TABLE I
Evaluating the Impact of Various Description Generation Methods on Language-Only Localization Performance, Using the Oxford RobotCar Loop Dataset. The bold values indicate the best results.

Methods	Localization Error	
	Mean	Median
MLLM with SP	144.93m, 80.76°	141.43m, 78.74°
MLLM with SP and TP	123.23m, 71.91°	122.81m, 56.83°
MLLM with MC and TP	83.46m, 42.19°	73.37m, 20.16°
SSD (Ours)	68.26m, 27.84°	47.06m, 13.01°
SSD + FTG (SDG, Ours)	**47.25m, 19.85°**	**29.48m, 6.79°**

Datasets: We use Oxford RobotCar Dataset [18], 4-Seasons Dataset [70] and Virtual Gallery Dataset [71] in experiments. The Oxford RobotCar dataset includes diverse urban driving data under varying weather, time, and seasonal conditions. Following the experimental setup of AtLoc [10], we conduct experiments with the LOOP and FULL subsets. The 4-Seasons Dataset, notable for its scale and diversity over 350 kilometers and nine environment types. We specifically examined business and neighborhood scenarios to test the robustness of our localization method in different urban environments. The Virtual Gallery Dataset is a large indoor dataset consisting of 3 to 4 rooms, with 42 publicly available paintings displayed on the walls. It includes looped data across five distinct paths, with camera positions and orientations randomly sampled. We train on loops 2 and 3 and evaluate on the occlusion 1 path.

Implementation: LangLoc framework uses the ResNet-50 image encoder and its corresponding text encoder from CLIP [66]. Training uses the Adam optimizer at a learning rate of 5×10^{-5}, resizing images to 224×224 pixels and processing them in 64 batches over 350 epochs on an NVIDIA RTX 3090 GPU. Additionally, we utilize MiniGPT-V2 [48] as the MLLM, and GPT-3 [17] (gpt-3.5-turbo) as the LLM component.

Evaluation Metrics: To ensure a fair comparison, we employ the evaluation metrics commonly utilized in prior research [9], [10], [13]: the mean and median errors of both position and orientation. The mean error offers a comprehensive assessment of algorithm performance by averaging errors across all samples. Conversely, the median error, less susceptible to outliers owing to its emphasis on the midpoint of errors, provides a more accurate indication of typical performance [32].

A. The Evaluation of Spatial Description Generation

In this subsection, we analyze the impact of various language description methods on language-only localization, which is crucial in determining localization accuracy.

1) Quantitative Results. We explore various approaches using the Multimodal Large-Language Model (MLLM) to translate images into textual descriptions, as outlined in Tab. I, which includes:

- MLLM with SP (Scene Description Prompt): it employs SP *"Describe textures, color, position, visual depth and spatial relations of objects in the image"* to guide

Input	MLLM with MC (Multiple Choice) and TP (Template Prompt)	SSD (Spatial Scene Description)+FTG (Formatted Text Generation)
(a_1) 	The camera captures the middle of the road, and on the road, there are double yellow lines and pedestrian crosswalk markings. The buildings in the image are primarily located at the front, characterized by their brick construction and multiple stories. The trees in the image are mainly distributed at the front right, providing a natural element to the urban setting.	Two white dashed lines on a paved street **is located** directly ahead. Red building made of brick with three floors is a house **is located** front left. Brown building made of brick with three floors is a residential building **is located** right. Traffic sign is no parking **is located** directly ahead. Green hedge along sidewalk **is located** directly ahead.
(a_2) 	The camera captures the middle of the road, and on the road, there are pedestrian crosswalk markings. The buildings in the image located at the front left and front right, with characteristics of traditional residential architecture. The image includes a pedestrian crossing sign traffic sign, located at the front left. The trees in the image are mainly distributed at the front right.	A part of a roadway **is located** directly ahead. Red building made of brick with two floors is a house **is located** left. Brown building made of brick with three floors is a residential building **is located** right. Traffic sign is no parking **is located** front left. Green hedge along sidewalk **is located** directly ahead.
(b_1) 	The camera captures the left side of the road, and on the road, there are white center line and traffic island marking. The buildings in the image are primarily located at the front left, with characteristics including a brick facade and windows visible from the perspective. The trees in the image are mainly distributed at the front right providing lush greenery to the scene.	Two white lines on a paved city street **is located** directly ahead. A fence on the side of the road **is located** directly ahead. Red building made of brick with three floors is a house **is located** front left. A chimney on a building **is located** above. A tall street light **is located** left. A brown brick wall **is located** directly ahead. Trees lining the street **is located** directly ahead.
(b_2) 	The camera captures the left side of the road, there are white center line on the road. The building in the image is primarily located at the front left, with characteristics of a brick structure with visible windows and greenery around it. The trees in the image are mainly distributed at the front right and left side, providing a lush backdrop.	A paved city street **is located** directly ahead. Red building made of brick with three floors is a house **is located** above. A brown brick wall **is located** directly ahead. Trees lining the street **is located** directly ahead.

Fig. 6. Visualize the comparison results of descriptions between MLLM with MC and TP, and SSD + FTG. Figures (a_1) vs (a_2), (b_1) vs (b_2) present descriptions from different viewpoints of the same scene. Text highlighted in color marks the changes in descriptions of the same object across viewpoints (a_1 vs a_2, b_1 vs b_2), such as streets (pink), buildings (blue), traffic signs (yellow), and others (**green**). Horizontal lines emphasize the contrast in descriptions of the same object by different methods (MLLM with MC and TP vs SSD + FTG).

MiniGPT-4 [47] to generate descriptions that include specific information relevant to localization.

- MLLM with SP (Scene Description Prompt) and TP (Template Prompt): building on MLLM with SP, it guides MiniGPT-4 to fill the generated description into the designated template with prompt *"extract information from the description to fill in the template. Template is "The street is [ ]..."*, thus producing formatted descriptions.
- MLLM with MC (Multiple Choice) and TP (Template Prompt): it adds a multiple-choice prompt *"Answer questions based on image, fill template for summary."*, guiding MiniGPT-4 to select answers related to localization, which are then filled into a template for formatted descriptions.
- SSD (Spatial Scene Description Module): our SSD accurately depicts the positions and specific attributes of objects within a scene, emphasizing key features through language expression.
- SSD (Spatial Scene Description Module) + FTG (Formatted Text Generation Module): it utilizes FTG to transform SSD outputs into formatted text via a well-designed template, while excluding descriptions of dynamic objects.

As depicted in Tab. I, in language-only localization, MLLM with SP shows larger mean and median position and orientation errors than other methods, specifically at 144.93m, 80.76° and 141.43m, 78.74°, respectively. This could be attributed to the non-specific and irregular language descriptions directly generated by MLLM [47], which are ambiguous and imprecise in expressing scenes, thereby posing challenges to localization. Incorporating TP into MLLM with SP improves performance, highlighting the importance of formatted output for enhancing description effectiveness in localization. Additionally, MLLM with MC and TP, which generates language descriptions for specific key objects, further enhances performance.

Despite these performance improvements, the generated descriptions still constrain localization accuracy, due to imprecise descriptions of location-relevant features and inconsistent descriptions across similar scenes. In contrast, our method employs the SSD to precisely describe the positions and attributes of various objects, obviously reducing the mean and median errors of the method. Furthermore, our SDG incorporates FTG with SSD to generate uniform textual descriptions, excluding dynamic objects and further reducing the mean and median errors to 47.25m, 19.85° and 29.48m, 6.79°, respectively, lower than other methods. This shows that only language descriptions that can reflect key object attributes and maintain consistent format can be used for localization, because they can provide stable scene semantics and present scene layout through regular description changes. It is also noteworthy that the median errors of all methods are typically smaller than the mean errors, indicating the presence of outliers solely relying on textual descriptions.

2) Qualitative Results. As shown in Fig. 6, we compare descriptions from the MLLM with MC and TP, and our SSD + FTG (SDG), across different viewpoints of the same scene (Figures (a_1 vs a_2) and (b_1 vs b_2)). The MLLM with MC and TP provides formatted text but shows inconsistencies in linguistic expression across different viewpoints. For instance, descriptions of buildings in Figures (a_1) and (a_2) change from

TABLE II
IN THE "LANGUAGE-ONLY LOCALIZATION" MODE, WE EVALUATE THE INFLUENCE OF VARIOUS OBJECT DESCRIPTIONS ON THE PERFORMANCE OF LANGLOC FRAMEWORK USING THE OXFORD ROBOTCAR LOOP DATASET. "POSITION" DENOTES DESCRIPTIONS CONTAINING SOLELY LOCATION ATTRIBUTES. "GENERAL" ENTAILS UNIFORMLY ASSIGNING ATTRIBUTE INFORMATION TO EACH OBJECT VIA GROUNDED CAPTION. "BUILDINGS", "SIGNS", AND "STREETS" PERTAIN TO DESCRIPTIONS SPECIFICALLY TARGETING THESE OBJECTS, ACQUIRED THROUGH SPECIALIZED QUESTIONING PROMPTS. EACH DESCRIPTION IS PROCESSED BY THE FTG MODULE AND THEN INPUT INTO THE POSE REGRESSION NETWORK. THE BOLD VALUES INDICATE THE BEST RESULTS.

Strategies					**Localization Error**	
Position	**General**	**Buildings**	**Signs**	**Streets**	**Mean**	**Median**
✓	-	-	-	-	59.53m, 23.11°	39.17m, 11.83°
✓	✓	-	-	-	54.42m, 20.56°	36.92m, 10.47°
✓	✓	✓	-	-	51.71m, 20.39°	35.15m, 9.08°
✓	✓	✓	✓	-	48.92m, 20.77°	31.38m, 7.63°
✓	✓	✓	✓	✓	**47.25m, 19.85°**	**29.48m, 6.79°**

"characterized by their brick construction and multiple stories" to "characteristics of traditional residential architecture". Although the text conveys similar observations, this variability can lead to differences in feature vector encoding, complicating the model's learning and generalization processes.

In contrast, our method, i.e., SSD + FTG not only maintains the consistency of the textual format but also accurately captures changes in scene viewpoints through subtle variations in text. For example, in Figures (a_1) and (a_2), while the attributes description of traffic signs remains unchanged, the position description shifts from "directly ahead" to "front left", accurately reflecting the change in viewpoints. The transition from Figures (b_1) to (b_2) accurately documents the appearance and disappearance of objects (e.g., "street light" and "A chimney"), enhancing the accuracy and reliability of descriptions. Moreover, SSD + FTG can also eliminate information about dynamic objects from descriptions. Such as, in Figure (a_2), the description "pedestrian crossing" appears when using the MLLM with MC and TP, whereas SSD + FTG removes this description, displaying only "traffic signs".

By employing a fixed text format and systematic changes in descriptions, our SSD + FTG enables the model to effectively identify and learn spatial relationships between images. This highlights the importance of choosing suitable description-generation methods for language-driven localization and provides valuable insights and implications for related research.

B. The Evaluation of Language-only Localization

In this subsection, we validate LangLoc's effectiveness in language-only localization. We analyze how different key object attributes affect performance, identifying which are more relevant to localization. Additionally, we test human language-driven localization, assessing its feasibility using natural human language inputs instead of LLM-generated language from images. This highlights LangLoc's potential in real-world scenarios that involve human interaction.

1) Component Analysis. We assess how textual descriptions of object attributes affect language-only localization accuracy. As in Tab. II, position-only descriptions yield a mean error of 59.53m and 23.11°, with a median error of 39.17m and 11.83°. Adding general attributes via grounded caption [48] reduces mean error by 5.11m and 2.55°, and median error by 2.25m and 1.36°. This improvement shows that combining object position with general attributes enhances the model's spatial understanding, enabling it to effectively localize objects in typical street scenes even without focusing on specific objects.

Notably, localization accuracy is further enhanced when descriptions include specific attributes of key objects. Describing building attributes, for instance, lowers the mean error to 51.71m and 20.39°, with a median error of 35.15m and 9.08°. Adding descriptions of traffic signs and streets further decreases the mean error by 4.46m and 0.54°, while reducing the median error by 5.67m and 2.29°. These results indicate that enriching descriptions with additional key object attributes provides clearer spatial references, thereby improving localization accuracy within the scene.

2) Localization Using Human Natural Language. We further explore the feasibility of localization using natural language descriptions provided by human participants. In this experiment, several participants were invited to describe the scenes they observed, and localization was accomplished solely based on these descriptions, using LangLoc.

As illustrated in Fig. 7, LangLoc first transforms colloquial human natural language into formatted textual descriptions using SDG. For example, given the human input "I'm situated in a car, looking directly ahead at a two-lane road," our method reformats this using a fixed structure to produce "A two-lane road is located directly ahead," ensuring consistency and accuracy in the description. Additionally, for dynamic objects mentioned in human language (e.g., a bus in row 2), our method effectively excludes them, thereby enhancing localization performance. As we can see, based on the language expressions of five participants, LangLoc achieves an average localization error of 18.74m, 1.29°, illustrating that our method can effectively process human natural language inputs.

This real-world experiment shows that our method tackles a novel task of using human natural language for localization. With the LangLoc framework, users can determine their location by describing landmarks or features from memory, without requiring specialized geographic knowledge. Furthermore, this localization approach implies that users need not share personal images or other sensitive information for location sharing, providing a privacy-secure localization solution.

Human Natural Language	Formatted Textual Descriptions	Image of Scene	Localization Error
Facing a straight road ahead that's marked with lane dividers; to the side, there's a white car parked. Directly ahead, a three floors building resembling a school can be seen. To the front left is a three floors residential building, and to the right, there stands a brown wall.	Lane division lines on the road **is located** directly ahead. A building with three floors is a school **is located** directly ahead. A building with three floors is a house **is located** front left. A brown wall **is located** right.		21.45m, 0.52°
I'm looking down a paved city street that stretches out directly in front of me. To the front right, there's a brown residence, a three floors building with a distinctive appearance. On the front left, there's another three floors residential building. Directly ahead, I can see a red bus.	A paved street **is located** directly ahead. A brown building with three floors is residence **is located** front right. A building with three floors **is located** to front left.		20.92m, 1.36°
I am observing a road marked with double white lines directly ahead. On the right, there's a street light situated on the sidewalk. Directly ahead, there is a brown brick building is two floors is a home. To front left, lush green trees line the street.	Double white lines on the road **is located** directly ahead. A brown building made of brick with two floors is a home **is located** directly ahead. A street light **is located** right. Lush green trees lining the street **is located** front left.		16.56m, 0.91°
I'm situated in a car, looking directly ahead at a two-lane road. To the front left, the curb of the sidewalk is visible. There's a red sign attached to a fence, also to the front left, and lush green trees are present in the same direction.	A two lane road **is located** directly ahead. The curb of a sidewalk **is located** front left. A red sign on the fence **is located** front left. The green trees **is located** front left.		28.27m, 1.77°
From the viewpoint within the car, I see a street directly in front, marked with a white line. On the right, a street light is visible. Directly ahead, there is a white brick building with two floors, possibly a shop. To the front left, there are dense green trees.	A white line on a street **is located** directly ahead. A white building made of brick with two floors is a shop **is located** directly ahead. A street light **is located** right. Dense green trees **is located** front left.		6.51m, 1.91°

Fig. 7. Localization results using unformatted Human Natural Language inputs, where text highlighted in color, marks the transformation between two types of descriptions for the same object. "Human Natural Language" pertains to unformatted, narrative scene descriptions provided by humans. "Formatted Textual Descriptions" denotes the formatted text generated from human natural language inputs through SDG. "Image of scene" denotes the image associated with the description. "Localization Error" indicates the discrepancy between the predicted pose and the ground truth (GT).

C. The Evaluation of vision-language Localization

In this subsection, we evaluate the performance of LangLoc in the vision-language localization mode by integrating both image and text inputs. Initially, we compare the performance of LangLoc with vision-based localization methods on the Oxford RobotCar [18] and the 4-Seasons datasets [70]. Subsequently, we conduct an ablation study to visually compare the performance of LangLoc with and without language input, analyzing the factors contributing to performance improvement.

1) Quantitative Results on the Oxford RobotCar Dataset: We compare LangLoc with representative visual localization methods on the Oxford RobotCar dataset to demonstrate the effectiveness of our approach. As shown in Tab. III, LangLoc achieves promising localization accuracy on the Loop trajectory. This trajectory was collected on a different date than the training data to evaluate localization performance in cross-day scenarios. Compared to the baseline method AtLoc [10], LangLoc shows improvements of 3.15m and 1.83° in mean localization accuracy, and 1.94m and 0.68° in median accuracy. When compared with the SOTA single-view visual localization method, CoordiNet [32], LangLoc also reduces the median error by 1.06m and 0.64°. Moreover, by incorporating time constraints, LangLoc+ supports multi-view inputs and demonstrates enhanced localization performance on the Loop trajectory, with smaller localization errors compared to AtLoc+ [10] and RobustLoc [37].

On the Full trajectory, LangLoc also exhibits obvious improvements in mean and median errors compared to baseline methods AtLoc and AtLoc+. Given the extensive road coverage in the Full trajectory, which often leads to more outliers, existing SOTA methods like RobustLoc [37] use outlier removal modules, resulting in smaller mean errors. In contrast, LangLoc+ leverage language descriptions to achieve competitive localization results, reducing the median error by 0.71m, 0.04° compared to RobustLoc. These results highlight the effectiveness of our method, as it better captures key and stable scene features through the integration of language descriptions. Compared to methods that rely solely on visual information, our method achieves superior performance, even in cross-day scenes or across a wider range of trajectories.

2) Quantitative Results on the 4-Seasons Dataset: We further assess the performance of LangLoc on the 4-Seasons dataset. As shown in Tab. IV, compared to the AtLoc, LangLoc reduces the mean error by 0.93m in Neighborhood scene. Besides, in the challenging Business scene, LangLoc achieves notable improvements, with the mean error reduced by 4.01m, 2.82°, and the median error reduced by 3.07m and 0.49°. When compared to CoordiNet [32], LangLoc also exhibits substantial reductions in both mean and median localization errors in the Business scenes. These findings underscore the generalization capability of LangLoc across various urban scenarios. Furthermore, compared to multi-view input-based methods, LangLoc+ outperforms AtLoc+ in both scenes. In the Business scene, LangLoc+ reduces the median localization

TABLE III
THE PERFORMANCE COMPARISON OF DIFFERENT LOCALIZATION METHODS ON THE OXFORD ROBOTCAR DATASET. THE BOLD VALUES INDICATE THE BEST RESULTS.

Oxford RobotCar Dataset		Mean error			Median error		
Methods	Input	LOOP	FULL	Average	LOOP	FULL	Average
PoseNet [9]	Single-view	7.9m, 3.53°	46.61m, 10.45°	27.26m, 6.99°	-	-	-
AD-PoseNet [12]	Single-view	6.40m, 3.09°	33.82m, 6.77°	20.11m, 4.93°	-	-	-
PoseNet+ [11]	Single-view	28.81m, 19.62°	125.6m, 27.10°	77.21m, 23.36°	5.80m, 2.05°	28.81m, 19.62°	17.31m, 10.84°
AtLoc [10]	Single-view	8.86m, 4.67°	29.6m, 12.4°	19.23m, 8.54°	5.05m, 2.01°	11.1m, 5.28°	8.08m, 3.65°
EffLoc [35]	Single-view	7.89m, 4.19°	27.23m, 11.41°	17.56m, 7.80°	4.76m, 2.06°	10.28m, 4.98°	7.52m, 3.52°
CoordiNet* [32]	Single-view	6.03m, 1.81°	11.99m, 6.15°	9.01m, 3.98°	4.17m, 1.97°	**4.21m**, 1.06°	4.19m, 1.52°
LangLoc (ours)	Single-view	5.71m, 2.84°	26.82m, 4.01°	16.27m, 3.43°	3.11m, 1.33°	6.68m, 1.55°	4.90m, 1.44°
MapNet [11]	Multi-view	9.84m, 3.96°	41.4m, 12.5°	25.62m, 8.23°	4.91m, 1.67°	17.94m, 6.68°	11.43m, 4.18°
AD-MapNet [12]	Multi-view	6.45m, 2.98°	19.18m, 4.60°	12.82m, 3.79°	-	-	-
AtLoc+ [10]	Multi-view	7.24m, 3.60°	21.0m, 6.15°	14.12m, 4.88°	3.78m, 2.04°	6.40m, 1.50°	5.09m, 1.77°
RobustLoc [37]	Multi-view	4.46m, 2.77°	**9.37m, 2.47°**	**6.91m, 2.62°**	4.04m, 1.41°	5.93m, 1.06°	4.99m, 1.24°
LangLoc+ (ours)	Multi-view	**4.19m, 1.74°**	15.7m, 2.85°	9.95m, 2.30°	**2.85m, 1.07°**	5.22m, **1.02°**	**4.04m, 1.05°**

*Implementation according to source code. https://github.com/dawnzyt/coordinet-pytorch

TABLE IV
THE PERFORMANCE COMPARISON OF DIFFERENT LOCALIZATION METHODS ON THE 4-SEASONS DATASET. THE BOLD VALUES INDICATE THE BEST RESULTS.

4-Seasons dataset		Mean error			Median error		
Methods	Input	Business	Neighborhood	Average	Business	Neighborhood	Average
GeoPoseNet [30]	Single-view	11.04m, 5.78°	2.87m, 1.30°	6.96m, 3.54°	5.93m, 2.03°	1.92m, 0.88°	3.93m, 1.46°
AtLoc [10]	Single-view	11.53m, 4.84°	2.80m, 1.16°	7.17m, 3.00°	5.81m, 1.50°	1.83m, 0.93°	3.82m, 1.22°
IRPNet [72]	Single-view	10.95m, 5.38°	3.17m, 2.85°	7.06m, 4.12°	5.91m, 1.82°	1.98m, 0.90°	3.95m, 1.36°
CoordiNet [32]	Single-view	11.52m, 3.44°	1.72m, 0.86°	6.62m, 2.15°	6.44m, 1.38°	1.37m, 0.69°	3.91m, 1.04°
LangLoc (ours)	Single-view	7.52m, 2.02°	1.87m, 1.17°	4.70m, 1.60°	2.74m, 1.01°	1.17m, 0.51°	1.96m, 0.76°
MapNet [11]	Multi-view	10.35m, 3.78°	2.81m, 1.05°	6.58m, 2.42°	5.66m, 1.83°	1.89m, 0.92°	3.78m, 1.38°
GNNMapNet [36]	Multi-view	7.69m, 4.34°	3.02m, 2.92°	5.36m, 3.63°	5.52m, 2.16°	2.14m, 1.45°	3.83m, 1.81°
AtLoc+ [10]	Multi-view	13.70m, 6.41°	2.33m, 1.39°	8.02m, 3.90°	5.58m, 1.94°	1.61m, 0.88°	3.60m, 1.41°
RobustLoc [37]	Multi-view	**4.28m**, 2.04°	**1.36m, 0.83°**	**2.82m**, 1.44°	2.55m, 1.50°	1.00m, 0.65°	1.78m, 1.08°
LangLoc+ (ours)	Multi-view	4.83m, **1.32°**	1.68m, 1.39°	3.26m, **1.36°**	**1.98m, 0.81°**	**0.93m, 0.55°**	**1.45m, 0.68°**

error by 0.57m and 0.69° compared to RobustLoc. Moreover, given that the 4-Seasons dataset encompasses a wide range of seasonal changes, weather conditions, and lighting variations in urban settings, LangLoc consistently maintains high localization accuracy under these conditions. These experiments further demonstrate the effectiveness and superiority of the proposed language-driven localization method.

3) Quantitative Results on the Virtual Gallery Dataset: To validate the generalization ability of LangLoc, we assess its performance in a large indoor scene. In these experiments, we first employ the MLLM to detect all objects present in the images. Then, we employ a uniform prompt, *"[grounding] describe this image in detail"* to guide the MLLM in describing attributes of detected objects, while instructing the LLM to output spatial descriptions in a consistent format: *"[Attribute] is located [Position]."* The results are shown in Tab. V. Our method outperforms other vision-only methods, with LangLoc demonstrating large improvements. Specifically, compared to the baseline method AtLoc, the mean localization error is reduced by 1.12m and 1.26°, and the median error is reduced by 1.16m and 0.83°. This improvement is attributed to the rich linguistic semantics embedded in the descriptions generated by SDG. For instance, the description "a painting of a garden with flowers and trees is located left" provides both the position and detailed content of the painting. These experimental results highlight that our language-driven localization framework benefits from the flexibility and scalability of language, enabling it to easily adapt to diverse application scenarios.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT LOCALIZATION METHODS ON THE VIRTUAL GALLERY DATASET. BOLD VALUES REPRESENT THE BEST RESULTS.

Methods	Localization Error	
	Mean	**Median**
Atloc [10]	2.47m, 7.31°	2.03m, 6.74°
Coordinet [32]	1.87m, 6.91°	1.69m, 6.55°
LangLoc (ours)	**1.35m, 6.05°**	**0.87m, 5.91°**

4) Ablation Study: In the ablation study, we explore the role of language in enhancing the performance of LangLoc. As shown in Fig. 8, LangLoc, when integrating both vision and language inputs, notably outperforms the vision-only approach in scenarios with illumination changes, shadow occlusion, and prominent key objects. For instance, in Figure (a1), exposure and shadow issues obscure building details and some road features. In comparison, textual descriptions covering the building's function, material, color, and road features are less affected by these visual changes. Therefore, with vision-language, LangLoc's localization accuracy improves by

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3546853

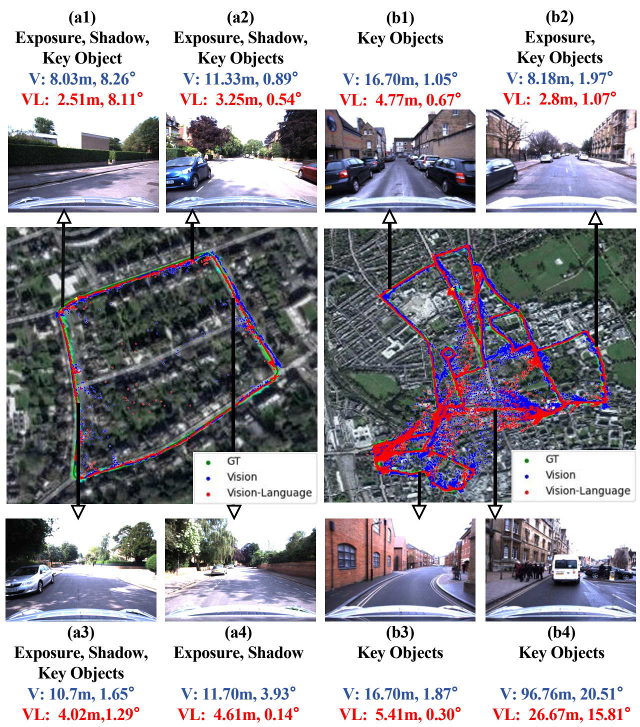


Fig. 8. Visualization of pose regression results for the loop trajectory (left) and Full trajectory (right) on Oxford RobotCar dataset. The ground truth is represented by green dots, while blue dots and red dots respectively illustrate LangLoc's predictions based solely on vision and on vision-language. In the images, Exposure, Shadow, and Key Object indicate the presence of exposure, shadow occlusion, and prominent key objects, respectively.

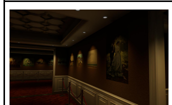
Image	Formatted Text Generation	Localization Error
	A yellow line on the road **is located** directly ahead. A brown building made of brick with one floors is house **is located** front left. Dense Trees with foliage **is located** front left.	**Atloc:** 17.24m, 17.24° **Coordinet:** 18.19m, 13.51° **Langloc: 7.91m, 4.33°**
	Two white lines on the road **is located** directly ahead. A red building made of brick with two floors is residential **is located** left. A white building made of glass with three floors is office **is located** front right. A brick wall **is located** right.	**Atloc:** 33.66m, 2.65° **Coordinet:** 24.58m, 2.96° **Langloc: 8.02m, 2.07°**
	White wall **is located** below. A painting of a man and woman in a long dress are walking through the woods **is located** front right. A painting of a vase with white flowers in it **is located** front left. A red carpet **is located** front left.	**Atloc:** 2.03m, 5.74° **Coordinet:** 1.61m, 5.33° **Langloc: 0.80m, 3.31°**
	A brown wooden ceiling with two white lights **is located** above. A painting of a woman in a yellow dress and white collar is reading a book **is located** front left. A painting of a girl in a blue dress stands in front of a garden **is located** directly ahead.	**Atloc:** 1.99m, 3.54° **Coordinet:** 1.35m, 3.07° **Langloc: 0.71m, 2.37°**

Fig. 9. Qualitative Comparison of Various Localization Methods. Here, Formatted Text Generation represents the output of SDG in Langloc, and Localization Error indicates method performance.

5.52m, 0.15° compared to only vision input. Further, in more complex scenes like Figure (b4), where images are disrupted by pedestrians and vehicles, vision-only LangLoc faces higher errors. In contrast, vision-language LangLoc, through precise descriptions of key objects, effectively enhances localization accuracy, achieving an improvement of 70.09m, 4.7°.

These findings suggest that relying solely on visual information may not accurately capture key features in complex environments, particularly when visual cues are unstable due to lighting variations or obstructions. By integrating language and vision data, LangLoc introduces additional semantic information through textual descriptions, enhancing the framework's recognition of important landmarks and features within the scene. Consequently, this integration improves the accuracy and robustness of localization in complex environments.

D. Qualitative Analysis in Challenging Scenarios

To further reveal the superiority of our method, we compare the localization results of different methods under different environmental conditions. As shown in Fig. 9, LangLoc shows better localization performance when dealing with challenges of environmental changes. For example, in row 1, even if low lighting causes blurred image details, LangLoc can still utilize stable language semantics (e.g., "A yellow line on the road is located directly ahead") to represent spatial clues, thereby improving localization accuracy. In particular, with generated descriptions, LangLoc enhances the expression of key features in the scene, such as the description in row 2, "A white building made of glass". Finally, LangLoc demonstrates performance advantages even in closed indoor environments with low light levels. As shown in row 3, LangLoc can also achieve more accurate localization by using the additional semantic information of a rough description of the content of the painting, "A painting of a man and woman in a long dress". Overall, LangLoc demonstrates superior localization performance across various challenging environments by leveraging stable semantics of language descriptions.

E. Robustness Analysis

In this subsection, we analyze LangLoc's robustness, by showing its localization performance under image degradation and scenarios with partial modality data missing.

1) Robustness to Image Degradations: We validate the robustness of LangLoc, using data constructed under image degradation conditions. Specifically, following the Robust-Mat [73], we generate degraded images based on the Loop trajectory of the Oxford RobotCar and use these images as visual inputs for LangLoc and other models. As shown in Fig. 10, these data include extreme weather conditions such as rain, snow, fog, and complex illumination Conditions including exposure and dim. As shown in Tab. VI, LangLoc notably outperforms representative visual localization methods such as PoseNet+ and AtLoc in two types of conditions. This result illustrates the robustness of LangLoc under image degradation conditions, which can be attributed to LangLoc's integration of vision with natural language. The natural language provide additional semantic information for localization, particularly crucial when visual data quality degrades due to poor weather or lighting variations.

We further evaluate the effectiveness of LangLoc using language descriptions generated from different images (i.e., "degraded" and "standard" images), while maintaining the "degraded" image as input. The results show that LangLoc's performance varies minimally between the two language inputs, with the median localization error differing by no more than 1m. This consistency highlights the advantage of language

TABLE VI

PERFORMANCE COMPARISON OF LANGLOC WITH DIFFERENT LANGUAGE DESCRIPTIONS UNDER IMAGE DEGRADATION CONDITIONS. HERE, $\mathrm{I_D}$ REPRESENTS A "DEGRADED" IMAGE INPUT, WHILE $\mathrm{L_{I_D}}$ AND $\mathrm{L_{I_S}}$ DENOTE LANGUAGE DESCRIPTIONS GENERATED FROM "DEGRADED" AND "STANDARD" IMAGES, RESPECTIVELY.

Method	Inputs	Extreme Weather		Complex Illumination	
		Mean	Median	Mean	Median
PoseNet+ [11]	$\mathrm{I_D}$	31.74m, 12.13°	11.67m, 4.18°	41.53m, 20.94°	17.66m, 20.94°
AtLoc [10]	$\mathrm{I_D}$	26.68m, 10.05°	9.84m, 2.45°	36.87m, 15.56°	11.95m, 3.15°
LangLoc (ours)	$\mathrm{I_D}$+$\mathrm{L_{I_D}}$	23.14m, 8.80°	6.73m, 1.69°	29.97m, 12.18°	7.58m, 1.98°
LangLoc (ours)	$\mathrm{I_D}$+$\mathrm{L_{I_S}}$	**20.73m, 8.05°**	**6.15m, 1.31°**	**25.13m, 11.41°**	**6.81m, 1.25°**

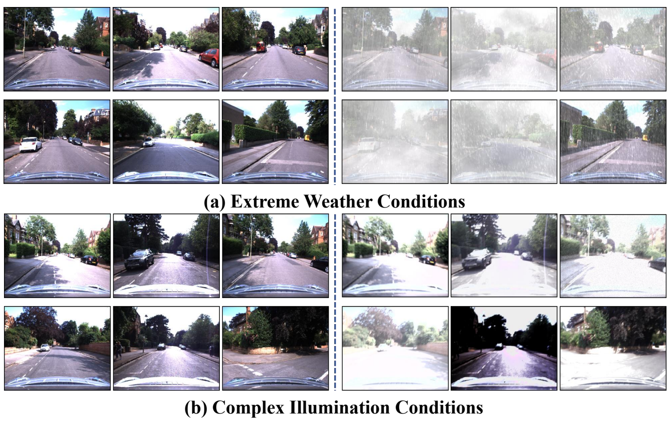


Fig. 10. Showcasing examples of data constructed under Image Degradation Conditions, based on the Loop trajectory of the Oxford RobotCar dataset: Clean Image (Left) vs. Degraded Image (Right).

descriptions in providing stable semantic information, enabling LangLoc to maintain robust localization performance even in challenging environments.

2) Robustness to Missing Modalities: In practical applications, the occurrence of missing modalities is common. Therefore, we evaluate the performance of LangLoc in handling situations where partial modality data is lost. During training, LangLoc receives complete visual and textual data; however, during testing, we input different modalities to assess the method's performance. As shown in Tab. VII, when only visual data is used, the median error is 4.84m and 2.45°, lower than training with visual data alone. This improvement is due to the additional semantic information provided by language descriptions in multimodal training, which enhances the model's understanding of scene structure and object attributes, allowing it to achieve better localization even with only visual input. However, when only language input is used, the model's performance is not as strong as when it is trained and tested with only language data. This discrepancy arises because multimodal training often leads the model to prioritize visual features, which are typically more intuitive for localization tasks and offer richer scene details. In contrast, models trained solely with language data focus more on linguistic features, leading to better performance with language input alone. Nevertheless, in both scenarios, effective localization accuracy is achieved.

The results demonstrate that LangLoc is highly robust and adaptable following multimodal joint learning. Even when visual information is limited or unavailable (e.g., in privacy-sensitive areas or overexposed environments), the language-driven LangLoc provides a reliable alternative or complementary solution for localization.

TABLE VII

THE LOCALIZATION RESULTS OF LANGLOC IN HANDLING MISSING MODALITIES. V DENOTES VISION, L DENOTES LANGUAGE.

Input Type		Localization Error	
Training	Testing	Mean	Median
V	V	13.67m, 6.38°	7.49m, 3.63°
L	L	47.25m, 19.85°	29.48m, 6.79°
V + L	V + L	**5.71m, 2.84°**	**3.11m, 1.33°**
V + L	L	72.44m, 32.45°	39.11m, 12.19°
V + L	V	9.68m, 5.05°	4.84m, 2.45°

VI. CONCLUSION AND FUTURE WORK

This work introduces a new task - language-driven localization, and proposes the LangLoc framework, capable of achieving localization using either language alone or in combination with visual cues. LangLoc first leverages the proposed spatial description generator to accurately characterize a scene by generating formatted text descriptions, enabling language-based localization. Further, through a joint-learning strategy, LangLoc enhances localization accuracy and robustness by fusing visual cues with linguistic semantics. Experiments on Oxford RobotCar, 4-Seasons and Virtual Gallery datasets show LangLoc's advantages, particularly in localizing complex and dynamic environmental conditions.

However, LangLoc currently depends on multiple models working together, which may impact real-time performance, especially on resource-limited devices or in applications demanding high responsiveness. In the future, we will optimize the algorithm's structure and efficiency to improve end-to-end multimodal reasoning, enhancing real-time performance. Additionally, we plan to expand the capabilities of LangLoc by integrating not only visual and language data but also other sensor inputs, such as depth sensors and LiDAR, to enable more accurate and robust localization.

REFERENCES

[1] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," in *ICML*, 2023.

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3546853

[2] H. Wei and L. Wang, "Visual navigation using projection of spatial right-angle in indoor environment," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3164–3177, 2018.

[3] Z. Zheng, X. Li, Q. Xu *et al.*, "Deep inference networks for reliable vehicle lateral position estimation in congested urban environments," *IEEE Trans. Image Process.*, vol. 30, pp. 8368–8383, 2021.

[4] C. Pan, B. Yaman, T. Nesti, A. Mallik *et al.*, "Vlp: Vision language planning for autonomous driving," in *CVPR*, 2024.

[5] S. Gupta, J. Chakareski, and P. Popovski, "mmwave networking and edge computing for scalable 360° video multi-user virtual reality," *IEEE Trans. Image Process.*, vol. 32, pp. 377–391, 2022.

[6] X. Ma, J. Su, C. Wang, W. Zhu, and Y. Wang, "3d human mesh estimation from virtual markers," in *CVPR*, 2023, pp. 534–543.

[7] Y. Wang, Z. Feng, H. Zhang *et al.*, "Angle robustness unmanned aerial vehicle navigation in gnss-denied scenarios," in *AAAI*, 2024.

[8] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[9] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *ICCV*, 2015, pp. 2938–2946.

[10] B. Wang, C. Chen, C. X. Lu, P. Zhao *et al.*, "Atloc: Attention guided camera localization," in *AAAI*, 2020, pp. 10 393–10 401.

[11] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *CVPR*, 2018, pp. 2616–2625.

[12] Z. Huang, Y. Xu, J. Shi, X. Zhou, H. Bao, and G. Zhang, "Prior guided dropout for robust visual localization in dynamic environments," in *CVPR*, 2019, pp. 2791–2800.

[13] Y. Shavit, R. Ferens *et al.*, "Coarse-to-fine multi-scene pose regression with transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[14] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," *arXiv preprint arXiv:2401.12168*, 2024.

[15] Y. Wang, H. Xie, S. Fang, M. Xing, J. Wang, S. Zhu, and Y. Zhang, "Petr: Rethinking the capability of transformer-based language model in scene text recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 5585–5598, 2022.

[16] S. Park, H. Kim, and Y. M. Ro, "Integrating language-derived appearance elements with visual cues in pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020, pp. 1877–1901.

[18] W. Maddern, G. Pascoe, C. Linegar *et al.*, "1 year, 1000 km: The oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[19] A. R. Zamir and M. Shah, "Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1546–1558, 2014.

[20] L. Weng, V. Gouet-Brunet, and B. Soheilian, "Semantic signatures for large-scale visual localization," *Multimedia Tools Appl.*, vol. 80, no. 15, pp. 22 347–22 372, 2021.

[21] M. Geppert, V. Larsson, J. L. Schönberger, and M. Pollefeys, "Privacy preserving partial localization," in *CVPR*, 2022, pp. 17 337–17 347.

[22] S. Pramanick, E. M. Nowara, J. Gleason, C. D. Castillo, and R. Chellappa, "Where in the world is this image? transformer-based geo-localization in the wild," in *ECCV*. Springer, 2022, pp. 196–215.

[23] E. Muller-Budack, K. Pustu-Iren, and R. Ewerth, "Geolocation estimation of photos using a hierarchical model and scene classification," in *ECCV*, 2018, pp. 563–579.

[24] P. H. Seo, T. Weyand, J. Sim, and B. Han, "Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps," in *ECCV*, 2018, pp. 536–551.

[25] L. Haas, M. Skreta, S. Alberti, and C. Finn, "Pigeon: Predicting image geolocations," in *CVPR*, 2024, pp. 12 893–12 902.

[26] B. Clark, A. Kerrigan, P. P. Kulkarni, V. V. Cepeda, and M. Shah, "Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes," in *CVPR*, 2023, pp. 23 182–23 190.

[27] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *ECCV*. Springer, 2016, pp. 37–55.

[28] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," in *WACV*, 2022, pp. 750–760.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[30] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *CVPR*, 2017, pp. 5974–5983.

[31] J. Kabalar, S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Towards long-term retrieval-based visual localization in indoor environments with changes," *IEEE Robotics Autom. Lett.*, vol. 8, no. 4, pp. 1975–1982, 2023.

[32] A. Moreau, N. Piasco, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "Coordinet: uncertainty-aware pose regressor for reliable vehicle localization," in *WACV*, 2022, pp. 2229–2238.

[33] Moreau, Arthur and Piasco, Nathan and Tsishkou, Dzmitry and Stanciulescu, Bogdan and de La Fortelle, Arnaud, "Lens: Localization enhanced by nerf synthesis," in *CoRL*, 2022, pp. 1347–1356.

[34] A. Moreau, T. Gilles, N. Piasco, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "Imposing: Implicit pose encoding for efficient visual localization," in *WACV*, 2023, pp. 2892–2902.

[35] Z. Xiao, C. Chen, S. Yang, and W. Wei, "Effloc: Lightweight vision transformer for efficient 6-dof camera relocalization," in *ICRA*, 2024, pp. 8529–8536.

[36] F. Xue, X. Wu, S. Cai *et al.*, "Learning multi-view camera relocalization with graph neural networks," in *CVPR*, 2020, pp. 11 372–11 381.

[37] S. Wang, Q. Kang, R. She, W. P. Tay, A. Hartmannsgruber, and D. N. Navarro, "Robustloc: Robust camera pose regression in challenging driving environments," in *AAAI*, 2023, pp. 6209–6216.

[38] S. Wang, Q. Kang, R. She, W. Wang, K. Zhao, Y. Song, and W. P. Tay, "Hypliloc: Towards effective lidar pose regression with hyperbolic fusion," in *CVPR*, 2023, pp. 5176–5185.

[39] W. Li, Y. Yang, S. Yu, G. Hu, C. Wen, M. Cheng, and C. Wang, "Diffloc: Diffusion model for outdoor lidar localization," in *CVPR*, 2024, pp. 15 045–15 054.

[40] J. Cheng, H. Zhang, and M. Q.-H. Meng, "Improving visual localization accuracy in dynamic environments based on dynamic region removal," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1585–1596, 2020.

[41] K. Zhou, C. Chen, B. Wang, M. R. U. Saputra, N. Trigoni, and A. Markham, "Vmloc: Variational fusion for learning-based multimodal camera localization," in *AAAI*, vol. 35, no. 7, 2021, pp. 6165–6173.

[42] H. Xie, T. Deng, J. Wang, and W. Chen, "Robust incremental long-term visual topological localization in changing environments," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2022.

[43] A. Chowdhery, S. Narang, J. Devlin *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[44] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[46] H. Touvron, T. Lavril, G. Izacard *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[47] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[48] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[49] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," in *ICLR*, 2023.

[50] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan, "Glamm: Pixel grounding large multimodal model," in *CVPR*, 2024, pp. 13 009–13 018.

[51] C. Li, C. Wong, S. Zhang, N. Usuyama *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[52] K. Niu, L. Huang, Y. Long, Y. Huang, L. Wang, and Y. Zhang, "Comprehensive attribute prediction learning for person search by language," *IEEE Trans. Image Process.*, 2024.

[53] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang, "Editable scene simulation for autonomous driving via collaborative llm-agents," in *CVPR*, 2024.

[54] S. Pramanick, G. Han, R. Hou, S. Nag, S.-N. Lim, N. Ballas, Q. Wang, R. Chellappa, and A. Almahairi, "Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model," in *CVPR*, 2024, pp. 14 076–14 088.

[55] L. Xue, M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo *et al.*, "xgen-mm (blip-3): A family of open large multimodal models," *arXiv preprint arXiv:2408.08872*, 2024.

[56] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao *et al.*, "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[57] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[58] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.

[59] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3469–3481, 2020.

[60] W. Zhang, Y. Guo, L. Niu, P. Li, C. Zhang, Z. Wan, J. Yan, F. U. D. Farrukh, and D. Zhang, "Lp-slam: Language-perceptive rgb-d slam system based on large language model," *arXiv preprint arXiv:2303.10089*, 2023.

[61] B. Li, D. Zou, Y. Huang *et al.*, "Textslam: Visual slam with semantic planar text features," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[62] M. Kolmet, Q. Zhou, A. Ošep, and L. Leal-Taixé, "Text2pos: Text-to-point-cloud cross-modal localization," in *CVPR*, 2022, pp. 6687–6696.

[63] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers, "Text2loc: 3d point cloud localization from natural language," in *CVPR*, 2024.

[64] Y. Zhao, J. Wei, Z. Lin *et al.*, "Visual spatial description: Controlled spatial-oriented image-to-text generation," in *EMNLP*, 2022.

[65] Y. Zhao, H. Fei, W. Ji *et al.*, "Generating visual spatial description via holistic 3d scene understanding," in *ACL*, 2023, pp. 639–657.

[66] A. Radford, J. W. Kim *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[67] W. Jin, Z. Zhao, X. Cao, J. Zhu, X. He, and Y. Zhuang, "Adaptive spatio-temporal graph enhanced vision-language representation for video qa," *IEEE Trans. Image Process.*, vol. 30, pp. 5477–5489, 2021.

[68] C. Liu, H. Ding, Y. Zhang, and X. Jiang, "Multi-modal mutual attention and iterative interaction for referring image segmentation," *IEEE Trans. Image Process.*, 2023.

[69] J. Yu, X. Jiang, Z. Qin, W. Zhang, Y. Hu, and Q. Wu, "Learning dual encoding model for adaptive visual understanding in visual dialogue," *IEEE Trans. Image Process.*, vol. 30, pp. 220–233, 2020.

[70] P. Wenzel, R. Wang, N. Yang, Q. Cheng, Q. Khan, L. von Stumberg, N. Zeller, and D. Cremers, "4seasons: A cross-season dataset for multi-weather slam in autonomous driving," in *GCPR*, 2021, pp. 404–417.

[71] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger, "Visual localization by learning objects-of-interest dense match regression," in *CVPR*, 2019, pp. 5634–5643.

[72] Y. Shavit and R. Ferens, "Do we really need scene-specific pose encoders?" in *ICPR*, 2021, pp. 3186–3192.

[73] R. She, Q. Kang, S. Wang, Y.-R. Yáng, K. Zhao, Y. Song, and W. P. Tay, "Robustmat: Neural diffusion for street landmark patch matching under challenging environments," *IEEE Trans. Image Process.*, 2023.

Weimin Shi received the M.S. degree from Beijing University of Chemical Technology in 2021. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. His current research interests include deep learning, multi-modal learning, computer vision, and smart city.

Changhao Chen obtained his Ph.D. degree at University of Oxford (UK), M.Eng. degree at the National University of Defense Technology (China), and B.Eng. degree at Tongji University (China). Now he is an Assistant Professor at the Thrust of Intelligent Transportation and Thrust of Artificial Intelligence, the Hong Kong University of Science and Technology (Guangzhou). His research interest lies in robotics, embodied AI and cyber-physical systems.

Kaige Li received the M.S. degree from the Ocean University of China, Qingdao, China, in 2020. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. His current research interests include deep learning, image semantic segmentation, computer vision, and smart city.

Yuan Xiong received the M.S. degree in computer science from Clemson University in 2014 and the Ph.D. degree from Beihang University, Beijing, China, in 2024. He is currently a Post-Doctoral Researcher with the School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen Campus, Shenzhen, China. His research interests include machine learning, computer vision and virtual reality.

Xiaochun Cao is a Professor at the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University. He received his B.E. and M.E. degrees, both in Computer Science, from Beihang University (BUAA), China, and his Ph.D. degree in Computer Science from the University of Central Florida, USA. His dissertation was nominated for the university-level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. Before joining SYSU, he was a professor at the Institute of Information Engineering, Chinese Academy of Sciences. He has authored and coauthored over 200 journal and conference papers. In 2004 and 2010, he received the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He serves on the editorial boards of IEEE Transactions on Image Processing and IEEE Transactions on Multimedia and was previously on the editorial board of IEEE Transactions on Circuits and Systems for Video Technology.

Zhong Zhou is a Professor of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, and also of Zhongguancun Laboratory, Beijing, China. He got B.S. degree from Nanjing University and Ph.D. degree from Beihang University in 1999 and 2005 respectively. His main research interests include Virtual Reality/Augmented Reality/Mixed Reality, Computer Vision, Artificial Intelligence and Cognitive Security.