Navigation Instruction Generation with BEV Perception and Large Language Models

Sheng Fan, Rui Liu, Wenguan Wang[†], and Yi Yang

ReLER, CCAI, Zhejiang University
https://github.com/FanScy/BEVInstructor

Abstract. Navigation instruction generation, which requires embodied agents to describe the navigation routes, has been of great interest in robotics and human-computer interaction. Existing studies directly map the sequence of 2D perspective observations to route descriptions. Though straightforward, they overlook the geometric information and object semantics of the 3D environment. To address these challenges, we propose BEVInstructor, which incorporates Bird's Eye View (BEV) features into Multi-Modal Large Language Models (MLLMs) for instruction generation. Specifically, BEVInstructor constructs a Perspective-BEV Visual Encoder for the comprehension of 3D environments through fusing BEV and perspective features. To leverage the powerful language capabilities of MLLMs, the fused representations are used as visual prompts for MLLMs, and perspective-BEV prompt tuning is proposed for parameter-efficient updating. Based on the perspective-BEV prompts, BEVInstructor further adopts an instance-guided iterative refinement pipeline, which improves the instructions in a progressive manner. BEVInstructor achieves impressive performance across diverse datasets (*i.e.*, R2R, REVERIE, and UrbanWalk).

Keywords: Navigation Instruction Generation · Bird's Eye View · Multi-Modal Large Language Model

1 Introduction

Navigation instruction generation, serving as a crucial interface between intelligent robotics and human interaction, has garnered significant attention in various fields, including robotics [29], psychology [77], and cognitive science [24,39]. This research aims to describe the navigation route precisely based on the observations. The process involves analyzing a series of visual inputs and subsequently converting them into natural language instructions. The generated instructions are required to incorporate the details for accurate replication of the navigated path. Navigation instruction generation plays a crucial role in fostering trust between humans and machines. It provides intuitive feedback to humans and guides

[†] Corresponding author: *Wenguan Wang*.

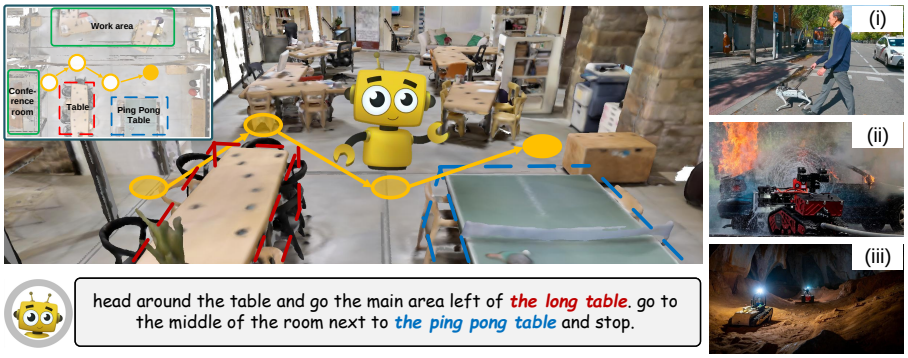

Fig. 1: BEVInstructor verbalizes concise navigation instructions. Navigation instruction generation is of great value to a wide range of tasks, (i) assisting in navigation for blind individuals, (ii) executing long-term tasks with automatic progress reporting, and (iii) conducting autonomous search and rescue operations in disaster areas.

them to accomplish goals, such as aiding visually impaired individuals [37] and explaining the agent's plan [91, 92] (see Fig. 1). Furthermore, embodied agents are expected to communicate with humans for efficient collaboration, instead of executing instructions only [4, 12, 37, 81–86, 91].

Early solutions [21, 29, 55] in instruction generation utilize hand-crafted rules or templates that fill a predetermined format with specific details. While straightforward, these approaches lack flexibility. Subsequent studies employ neural networks to facilitate end-to-end learning for instruction generation, *e.g.*, LSTM [27, 75, 82, 100] or Transformer [87, 91, 92]. Recent Multi-Modal Large Language Models (MLLMs) showcase immense capabilities of vision-language understanding and generation [44, 64, 76, 90, 98, 105]. MLLMs take advantage of cross-modal transfer, allowing knowledge to be shared between multi-modal domains [9, 22, 36, 96, 97]. Despite their promising performance on various vision-language tasks [64, 76], they cannot fully satisfy the requirements of navigation instruction generation in a zero-shot manner (see Table 5). Specifically, MLLMs are pre-trained on extensive image-text pairs, primarily involving isolated images from a third-person view. In contrast, navigation instruction generation relies on a sequence of egocentric observations from an embodied agent [30]. This poses challenges for MLLMs in understanding spatial context from navigation trajectories, especially in complex 3D environments. More importantly, such embodied (egocentric) perception requires a comprehensive scene understanding of the 3D physical world, interpreting objects and actions to generate detailed instructions. However, existing studies [27, 75, 82, 91, 92, 104] often rely on 2D perspective features as visual representations, ignoring the 3D scene geometry and object semantics [42, 60]. This underscores the need for more advanced solutions capable of integrating 3D spatial understanding to improve the accuracy and relevance of navigation instructions.

As a response, we propose BEVInstructor, an iterative instruction generator driven by BEV perception and MLLMs. BEVInstructor develops a *BEV encoder* to reconstruct 3D information from perspective features under the supervision of 3D detection. This allows to preserve 3D geometry and object semantics of the environment. The encoded BEV features are combined with the perspective features, thereby enriching the visual representations. Then BEVInstructor enhances the capability of the MLLMs by finetuning on navigation instruction-specific data through a *parameter-efficient update* strategy. In addition, *iterative refinement* is proposed to progressively enhance instruction generation, leveraging the powerful language capabilities of MLLMs.

Specifically, BEVInstructor processes a sequence of embodied observations. It adopts a *BEV encoder* to aggregate the multi-view image features into the BEV grid features through a 2D-3D view transformation (§3.2). Then it uses a *Perspective-BEV fusion* module to fuse the BEV features with perspective features, converting the fused embeddings into shorter tokens to prevent excessively long inputs for the MLLM. We also devise perspective-BEV prompt tuning for parameter-efficient updating (§3.3), with trainable parameters constituting only **7.2**% of the entire framework. Prior research in cognitive science [59] has validated the significance of key instances and landmarks in human route description. This motivates us to propose *instance-guided iterative refinement* (§3.4). Initially, BEVInstructor identifies the key instances and generates the corresponding landmark tokens along the path, leveraging the rich object semantics encoded in the perspective-BEV embeddings. Subsequently, it organizes the complete instructions conditioned on these landmark drafts. After multi-turn refinement, BEVInstructor produces high-quality instructions that include more concise details about the 3D environment.

We conduct extensive experiments on indoor R2R [6], REVERIE [69] and outdoor UrbanWalk [37] datasets. Compared with the state-of-the-art navigation instruction algorithms [37, 82, 91], our BEVInstructor attains better performances across all datasets. Especially, BEVInstructor achieves **12.6**% and **8.3**% CIDEr gains on REVERIE val seen and unseen splits, respectively, compared with previous best methods. This suggests that the BEV features effectively integrate the 3D scene information into the MLLM, thereby establishing a connection between real-world perceptions and human languages.

2 Related Work

Navigation Instruction Generation. The study of navigation instruction generation can date back to the 1960s [59]. Instruction generation [18] has garnered significant interest across various fields, *e.g.*, robotics [29], psychology [77], and cognitive science [24, 39] for a long time, yet attained far less attention in embodied vision. Early studies predominantly rely on hand-crafted rules [21] or templates [29, 55], binding the format of generated instructions. While these methods are effective in producing high-quality sentences tailored to specific scenes, they demand significant linguistic expertise and extensive manual ef-

fort. To alleviate this inflexibility, several studies employ neural network [26,27] to facilitate end-to-end learning. Subsequent efforts [27, 75, 82] utilize LSTM-based speakers integrated with instruction-following agents, enabling simultaneous training on pairs of path navigation and instructions navigation. They leverage the sequential processing strengths of LSTMs to better understand and generate navigation instructions that accurately reflect the temporal process of navigating. Additionally, motivated by the success of Transformer [78] in the natural language processing domain, a new wave of research [87,91,92] has emerged to leverage the advanced capabilities of Transformer to further improve generation performance. Existing efforts delve into understanding the foundational principles of how humans construct route descriptions [2,57,95] and explore the qualities that make instructions easy to follow [55,71,80]. These studies emphasize that crucial landmarks and concise topological descriptions play a crucial role in the description of wayfinding. In light of these, recent studies [1,17,63,88] lean on the process of landmark grounding to improve the instructions.

From the perspective of network architectures, a text encoder with powerful representation capacity significantly enhances output quality. However, there is a lack of dedicated research on the application of MLLMs for creating navigation instructions. In this work, we explore the potential of incorporating MLLMs endowed with superior linguistics for navigation instruction generation. Furthermore, BEVInstructor is designed to initially identify critical landmarks as drafts, aiding in forming comprehensive instructions. We devise an instance-guided iterative refinement process, which decomposes the generation into two stages. This allows for iterative refinement and enrichment of the instructions.

Scene Understanding. Scene understanding has emerged as a pivotal aspect of perception, navigation, and interaction with humans and environments [8,19,42]. Traditional SLAM systems [23] leverage data from different sensors, such as LiDAR and cameras, to build maps. They facilitate the robot to perceive depth and structure but exhibit limitations in comprehending the scene semantics. Several efforts develop semantic spatial representations [3, 10, 13, 32, 33, 52, 53, 94] or neural scene representations [40, 62], showing effectiveness across diverse scenes. Recently, BEV perception [46, 47, 67, 70] is proposed to infer the 3D geometry by projecting the multi-view images onto the BEV plane.

Current research [27,75,82,91,92] in instruction generation relies on perspective features. While this provides a foundational understanding of the environment, it ignores critical aspects like scene geometry and object semantics, often resulting in suboptimal performance in complex environments [75, 82]. Our approach aims to enhance the instruction generation process by integrating both perspective and BEV features. This fusion achieves a more holistic understanding of the scene, thereby facilitating the generation of higher-quality instructions.

Multi-Modal Large Language Models (MLLMs). MLLMs have surged in popularity and application. Although primarily trained on text data, initial studies [43, 51, 105] have demonstrated that pre-trained MLLMs can process visual information by fine-tuning the vision encoder via a learnable interface. The profound impact of this simple yet efficient approach drives advancements

in MLLMs. Existing open-source MLLMs can be broadly classified into three categories based on the approach of vision fusion: query-based, projection-based and parameter-efficient tuning-based. Motivated by the success of BLIP-2 [43], numerous efforts [20,44,90,101,105] investigate the use of a lightweight Q-Former to efficiently extract vision information. Albeit simple, [41,51,61,68,73,99] adopt the linear layers to project the vision embeddings into the language embedding space. Several solutions [28, 58, 102] train on multi-modal data simultaneously by parameter-efficient tuning, *i.e.*, LoRA [35] and Adapter [34, 74].

Though impressive, these methods only address the alignment of individual image-text pairs, neglecting the interaction between the egocentric observation and 3D spatial scene understanding in the task of navigation instruction generation, *i.e.*, fine-grained vision signals about landmarks and objects. Consequently, our research also delves into the effectiveness of scene understanding in improving the generation of navigation instructions by MLLMs.

3 Methodology

3.1 Overview

Problem Definition. The goal of the instruction generation task is to verbalize the navigation trajectory of an embodied agent using natural language. Here we formulate the task under R2R [6] setting. The agent observes a navigation path and collects a sequence of perspective observations $\mathcal{O} = \{O_t\}_{t=1}^{T}$ along with actions $\mathcal{A} = \{a_t\}_{t=1}^{T}$. Each observation O_t contains K multi-view images of its surroundings with the orientation angles $\{\delta_{t,k}\}_{k=1}^{K}$. These RGB images are encoded as $\{\boldsymbol{F}_{t,k} \in \mathbb{R}^{D_p \times HW}\}_{k=1}^{K}$. H and W are the spatial shape of image features, D_p is the channel dimension. The action embedding $\boldsymbol{a}_t \in \mathbb{R}^D$ is represented by the corresponding feature $\boldsymbol{F}_{t,a}$ of the action view $\delta_{t,a}$ ($0 \le a \le K$). Based on the observation-action sequence $\{O_t, a_t\}_{t=1}^{T}$, the agent is required to produce an instruction $\mathcal{X} = \{\boldsymbol{x}_l \in \mathbb{R}^D\}_{l=1}^{L}$ with L words in an autoregressive style (D is the embedding dimension and Θ is the model parameters):

$$\max_{\Theta} \sum\nolimits_{l=1}^{L} \log P_{\Theta}(\boldsymbol{x}_l | \boldsymbol{x}_{<l}, \mathcal{O}, \mathcal{A}). \tag{1}$$

Core Idea. We propose BEVInstructor, a new navigation instruction generator built upon LLaMA [76] (Fig. 2). To encode the semantic and geometric information of the 3D environment, BEV features are introduced and combined with 2D perspective features in *Perspective-BEV Visual Encoder* (§3.2). To exploit the capacity of cross-modal alignment in MLLMs, the visual embeddings are considered as visual prompts and fed into the *Perspective-BEV Prompt Tuning* (§3.3). Furthermore, we devise *Instance-Guided Iterative Refinement* (§3.4) to improve the quality of the generated instructions in a progressive manner.

3.2 Perspective-BEV Visual Encoder

Perspective Embedding. The perspective embedding $\boldsymbol{P}_t = \{\boldsymbol{p}_{t,k} \in \mathbb{R}^D\}$ is built upon the multi-view features $\{\boldsymbol{F}_{t,k}\}_{k=1}^{K}$ of the surroundings with different

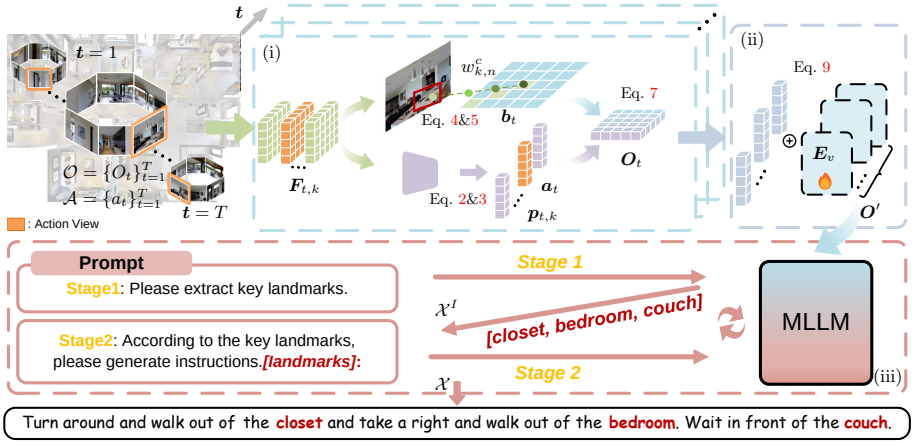

Fig. 2: Overview of BEVInstructor for navigation instruction generation. (i) BEV incorporates perspective embeddings by querying for 3D scene understanding (§3.2), (ii) we adopt BEV-Perspective prompt tuning for the cross-modal alignment with MLLMs (§3.3), (iii) the instructions are generated and improved progressively through multiple refinements (§3.4). Please refer to §3 for more details.

view angles. To maintain the direction indication information, the orientation representation is incorporated into the perspective embedding of each view:

$$\boldsymbol{p}_{t,k} = \mathcal{E}^p(\boldsymbol{F}_{t,k}) + \mathcal{E}^\delta(\delta_{t,k}) + \boldsymbol{E}_t + \boldsymbol{E}_o \in \mathbb{R}^D, \tag{2}$$

where $\mathcal{E}^p$ and $\mathcal{E}^\delta$ represent a linear layer, $\boldsymbol{E}_t \in \mathbb{R}^D$ and $\boldsymbol{E}_o \in \mathbb{R}^D$ denote the learnable embeddings of time step t and observation token type, respectively. Analogously, the embedding of action (view) is formulated as:

$$\boldsymbol{a}_t = \mathcal{E}^a(\boldsymbol{F}_{t,a}) + \mathcal{E}^\delta(\delta_{t,a}) + \boldsymbol{E}_t + \boldsymbol{E}_a \in \mathbb{R}^D, \tag{3}$$

where $\mathcal{E}^a$ indicates a linear layer and $\boldsymbol{E}_a \in \mathbb{R}^D$ is the learnable embedding of action token type.

BEV Embedding. Previous studies [27, 75, 82, 91] adopt the 2D perspective features as visual representations of the observations $\mathcal{O}$ in the 3D environment. However, these 2D features only capture limited semantic information and geometry, easily leading to ambiguous path descriptions. To further enhance the visual representations for the 3D environment, BEV features are introduced to encode the spatial scene understanding. The BEV encoder assigns each BEV query $\boldsymbol{Q}(x, y) \in \mathbb{R}^D$ located at (x, y) on the BEV plane $(H_b \times W_b)$ with a set of 3D reference points $\mathcal{P}_k(x, y, z_n)$. The BEV encoder projects them to sample the feature $\boldsymbol{F}_{t,k}$. Then the multi-view features are aggregated into the BEV grid features $\{\boldsymbol{b}_t(x, y) \in \mathbb{R}^{D_b}\}_{x=1,y=1}^{H_b,W_b}$ as (the subscripts x, y are omitted for simplicity):

$$\boldsymbol{b}_t = \sum_{k=1}^{K} \sum_{n=1}^{N_{\text{ref}}} \mathcal{F}^d\big(\boldsymbol{Q}(x, y), \mathcal{P}_k(x, y, z_n), \boldsymbol{F}_{t,k}\big) \cdot w_{k,n}^c \in \mathbb{R}^{D_b}, \tag{4}$$

where $\mathcal{F}^d$ is the BEV encoder with deformable attention layers [106]. $\mathcal{F}^d$ uses $\boldsymbol{Q}(x, y)$ to sample the corresponding image feature $\boldsymbol{F}_{t,k}$ and N_{ref} represents the number of 3D reference points. Since different reference points may be projected on the same image pixels to sample the feature, a depth consistency weight $w^c_{k,i}$ is introduced to distinguish them by predicting the weights of different depths:

$$w^c_{k,n} = \frac{\mathcal{E}^c(\boldsymbol{F}_{t,k}) \cdot \mathcal{D}(\mathcal{P}_k(x, y, z_n))}{\left\|\mathcal{E}^c(\boldsymbol{F}_{t,k})\right\| \cdot \left\|\mathcal{D}(\mathcal{P}_k(x, y, z_n))\right\|} \in [0, 1], \tag{5}$$

where $\mathcal{D}$ denotes a parameter-free operation that converts the 3D reference point $\mathcal{P}_k(x, y, z_n)$ into a depth distribution vector, and $\mathcal{E}^c$ is a depth network to predict the depth distribution vector of the projected image pixel based on $\boldsymbol{F}_{t,k}$. By calculating the cosine similarity between them, the depth consistency can guarantee the sampling quality in the 3D space [45, 47]. The BEV encoder is trained under the supervision of 3D detection. The detection heads [46, 93] with ℓ_1 loss and cross-entropy loss are used to supervise 3D bounding box regression and semantic classification, respectively (more detailed in Appendix). Then the frozen BEV encoder $\mathcal{F}^d$ is adopted to sample the image features into the BEV plane as $\boldsymbol{B}_t = \{\boldsymbol{b}_t(x, y)\}^{H_b, W_b}_{x=1, y=1}$.

Perspective-BEV Fusion. The perspective embedding $\boldsymbol{P}_t$ preserves rich visual cues in multi-view images. It is complementary to the BEV embedding $\boldsymbol{B}_t$ that mainly represents the 3D geometric information. Hence, *perspective-BEV fusion* is proposed for comprehensive scene understanding. This module consists of several standard transformer layers $\mathcal{F}^o$ for attending to the spatial relationships between $\boldsymbol{B}_t$ and $[\boldsymbol{P}_t, \boldsymbol{a}_t]$ ($[,]$ denotes the *concatenation* operation):

$$\boldsymbol{O}_t = \mathcal{F}^o(\boldsymbol{B}_t, [\boldsymbol{P}_t, \boldsymbol{a}_t]) \in \mathbb{R}^{D \times H_b W_b}. \tag{6}$$

Through the *perspective-BEV fusion* module, the fused embedding, *i.e.*, the complete observation embedding $\boldsymbol{O}_t$, is served as the visual token and fed into MLLMs (Eq. 1) for instruction generation. However, given a large number of visual tokens (*e.g.* $H_b W_b$ tokens for each step), aggregating such long tokens poses a significant challenge for MLLMs. In Table 4, directly feeding all visual tokens into MLLMs results in excessive computational burden, and makes it difficult to capture critical semantic information. Therefore, we design a lightweight transformer $\mathcal{Q}$ with N_q learnable queries to map $\boldsymbol{B}_t$ into a fixed number of tokens:

$$\boldsymbol{O}_t = \mathcal{Q}(\mathcal{F}^o(\boldsymbol{B}_t, [\boldsymbol{P}_t, \boldsymbol{a}_t])) \in \mathbb{R}^{D \times N_q}, \tag{7}$$

where N_q is independent of the BEV dimension ($N_q \ll H_b W_b$). It reduces the number of visual tokens to N_q and feeds the most useful tokens for instruction generation [43]. We conduct extensive experiments to confirm the effectiveness of our proposed perspective-BEV fusion (§4). In Table 6a, we compare our fusion method with other approaches and find that our fusion method is notably more effective in boosting performance.

3.3 Perspective-BEV Prompt Tuning

The MLLMs [64, 76] typically utilize extensive corpora of vision-language pairs and project visual and linguistic information into a unified representation space. They encode numerous world knowledge acquired from massive data and possess strong capabilities in multi-modal tasks [51, 61, 105]. However, directly using general-purpose MLLMs for instruction generation fails to capture intricate details due to the complexity of scenarios (Table 5). Additionally, the large size of MLLMs makes them costly to train from scratch. Thus, we propose perspective-BEV prompt tuning to exploit the scene geometry and unleash the cross-modal potential of MLLMs. Our perspective-BEV prompt tuning is parameter-efficient and incorporates 3D geometry into prompts. While our proposal is agnostic to the model, *i.e.*, P_Θ (Eq. 1) is a generic multi-modal learner, we formulate it with LLaMA [76] in a parameter-efficient updating manner:

$$\max_{\Theta^* \cup \Psi} \sum\nolimits_{l=1}^{L} \log P_{\Theta \cup \Psi}(\boldsymbol{x}_l | \boldsymbol{x}_{<l}, \boldsymbol{O}_{1:T}), \tag{8}$$

where Ψ is the additional parameters for prompt tuning ($|\Psi| \ll |\Theta|$), Θ^* is the finetuned parameters of Θ ($|\Theta^*| \ll |\Theta|$), and $\boldsymbol{O}_{1:T}$ is the visual embedding sequence of $\mathcal{O}$ and $\mathcal{A}$ in Eq. 1.

Perspective-BEV Prompt. To overcomes the issues of catastrophic forgetting, N_p learnable embeddings $\boldsymbol{E}_v \in \mathbb{R}^{D \times N_p}$ are inserted into the visual embedding $\boldsymbol{O}_{1:T}$ as perspective-BEV prompts $\boldsymbol{O}'$:

$$\boldsymbol{O}' = \boldsymbol{O}_{1:T} \oplus \boldsymbol{E}_v \in \mathbb{R}^{D \times T N_p}, \tag{9}$$

where $\oplus$ indicates broadcast and addition on the sequence length dimension. Then they are fed into the transformer layer with the text tokens $\boldsymbol{x}_{<l} \in \mathbb{R}^{D \times (l-1)}$:

$$[\boldsymbol{O}', \boldsymbol{x}_{<l}]_{m+1} = \mathcal{F}_m^{\mathrm{LM}}([\boldsymbol{O}', \boldsymbol{x}_{<l}]_m), \tag{10}$$

where $\mathcal{F}_m^{\mathrm{LM}}$ is the m-th transformer layer in LLaMA.

Parameter-Efficient Updating. To stabilize the training process and modulate the deep features, we modify the vanilla attention mechanism with the self-attention and linear layers at the last N_a layers of LLaMA. Specifically, for the self-attention part, zero-initialized attention [102] is adopted to adaptively control the importance of $\boldsymbol{O}'$ for instruction generation at the early stage. For the linear layers, we introduce the learnable scale vectors and use the dot product between the scale factors and the weight/bias tensor of each layer, respectively. In this way, we simplify the training process by retaining the parameters of LLaMA to stress the scene-instruction alignment and eliminate the potential risk of impairing the capacity of text generation [38, 48]. The number of added parameters (*i.e.*, Ψ) only accounts for **7.2**% of the entire model, demonstrating that BEVInstructor is a parameter-efficient framework.

3.4 Instance-Guided Iterative Refinement

Given the complexity of 3D environments, it is difficult to generate precise instructions that align with the scene layout. Humans usually describe a route by conceiving a rough draft based on the landmarks and then improve it [25,65]. Motivated by how humans refine their descriptions, we devise an instance-guided refinement strategy to learn from the generated landmarks and optimize the instructions. In the initial stage, BEVInstructor outputs a series of candidate instance words as initial landmark tokens $\mathcal{X}^I = \{\boldsymbol{x}_1^I, \boldsymbol{x}_2^I, \cdots \in \mathbb{R}^D\}$, *i.e.*, $\mathcal{O} \times \mathcal{A} \rightarrow \mathcal{X}^I$. Next, the instance-guided draft is incorporated into the model to refine the instructions, *i.e.*, $\mathcal{O} \times \mathcal{A} \times \mathcal{X}^I \rightarrow \mathcal{X}$. The optimization objective (Eq. 1) is reformulated as:

$$\max_{\Theta} \Big(\sum\nolimits_{l=1}^{L} \log P_\Theta(\boldsymbol{x}_l | \boldsymbol{x}_{<l}, \mathcal{O}, \mathcal{A}, \mathcal{I}) + \sum\nolimits_{l=1}^{|\mathcal{X}^I|} \log P_\Theta(\boldsymbol{x}_l^I | \boldsymbol{x}_{<l}^I, \mathcal{O}, \mathcal{A}) \Big). \tag{11}$$

Parameter-efficient finetuning is omitted here for the sake of clarity. We implement multi-turn refinement in the process of generation (see Table 6b). Compared with existing methods [75,82], BEVInstructor can identify crucial objects based on informative perspective-BEV prompts. In this way, BEVInstructor enriches the object semantics in the generated instructions.

3.5 Implementation Details

BEVInstructor encodes the perspective embeddings and BEV embeddings by the visual encoder (§3.2). Then the fused embeddings are served as the visual prompts for *perspective-BEV prompt tuning* (§3.3). Moreover, BEVInstructor employs a two-stage strategy for progressive instruction generation (§3.4)

Visual Embedding. The spatial shape of multi-view features is $H = W = 14$. The orientation representation $\boldsymbol{\delta}_{t,k}$ of view $\boldsymbol{F}_{t,k}$ is defined as $(\cos\varphi_{t,k}, \sin\varphi_{t,k}, \cos\phi_{t,k}, \sin\phi_{t,k})$, where φ and ϕ are the angles of heading and elevation, respectively. The embedding dimension is set as $D = 768$. For BEV embeddings, the shape of the BEV plane $(H_b \times W_b)$ is set as 15×15. The corresponding perception range is [-5.0 m, 5.0 m]. $N_{\text{ref}} = 4$ reference points uniformly distributed over the height [-1.2 m, 2.0 m] are used for each BEV query. The relative 2D coordinates are used for position encoding $\boldsymbol{E}_p$. For BEV encoding, there are six deformable attention blocks in $\mathcal{F}^d$. The depth prediction network $\mathcal{E}^c$ employs discrete convolutions [45,67] (More detailed in Appendix).

Word Embedding. We follow the default tokenizer of LLaMA [76]. The word embedding dimension is set as 4096. The max sequence token length is 128.

Network Architecture. The transformer $\mathcal{F}^o$ has six blocks for perspective-BEV fusion. The lightweight transformer $\mathcal{Q}$ is used to map the BEV embedding into $N_q = 10$ tokens, consisting of eight blocks. The backbone of LLM $\mathcal{F}^{\text{LM}}$ is initialized from LLaMA-7B [76] with 32 layers. All experiments follow default training configurations on R2R [6], REVERIE [69], and UrbanWalk [37]. Perspective-BEV prompts $\boldsymbol{O}'$ are inserted into the topmost $N_a = 31$ layers.

Finetuning. We use the AdamW [56] as the optimizer with the learning rate of $1e^{-4}$ and overall batch size of eight with 20k iterations. We only finetune the small-scale parameters (*i.e.*, $\Theta^* \cup \Psi$ in Eq. 8, <500M) while freezing most parameters (6.68B) of BEVInstructor.
Inference. During inference, BEVInstructor processes a sequence of multi-view images to generate navigation instructions. This autoregressive method involves an iterative instruction generation process that builds upon previously generated instructions. Each new instruction is refined progressively using instance tokens until the <end> token.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on three datasets with instructions: R2R [6] and REVERIE [69] for indoor scenes, and UrbanWalk [37] for outdoor scenes.

- **R2R** [6] builds upon Matterport3D [11], including diverse photo-realistic house scenes. There are three splits for the experiment, *i.e.*, `train` (61 scenes, 14,039 instructions), `val seen` (61 scenes, 1,021 instructions), and `val unseen` (11 scenes, 2,349 instructions). There are three human-annotated navigation instructions for each path and the average length is approximately 29 words. There are no overlapping scenes between `train` and `unseen` splits. Following previous studies [82,91], 6,482K synthesized navigation route-instruction pairs from PREVALENT [31] are also involved for training.
- **REVERIE** [69] extends the Matterport3D [11] simulator to incorporate object annotations. It comprises indoor scenes with 4,140 target objects and 21,702 instructions with an average length of 18 words. There are three splits for our experiment, *i.e.*, `train` (61 scenes, 10,466 instructions), `val seen` (61 scenes, 1,371 instructions), and `val unseen` (10 scenes, 3,753 instructions).
- **UrbanWalk** [37] involves outdoor scenes from the simulator with 26,808 naturalistic instructions. On average, there are 21.7 words per instruction.

Evaluation Metrics. Following previous studies [1,82,91], we employ five standard metrics: **1) BLEU** [66] refers to the geometric mean of n-gram precision scores computed over reference and candidate descriptions. **2) CIDEr** [79] represents the average cosine similarity between n-grams of the reference and candidate descriptions, weighted by their corresponding term frequency-inverse document frequency values. **3) METEOR** [7] is defined as the harmonic mean of precision and recall of unigram matches between sentences. **4) ROUGE** [50] is a measure of correspondence between the reference and candidate texts by computing the recall and precision scores for each n-gram size, word sequences and word pairs, and thus averaging them by a weighted F-measure. **5) SPICE** [5] is the F-score of the scenegraph [72] tuples of the candidate sentence and all reference sentences. **SPICE** is considered as the primary indicator.
Reproducibility. BEVInstructor is implemented in PyTorch and all models are trained and tested using a single machine with 2 NVIDIA A40 GPUs.

Table 1: Quantitative comparison results for **instruction generation** on R2R [6] `val seen` and `val unseen`. See §4.2 for more details.

Methods	R2R `val seen`						R2R `val unseen`					
	SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑	SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
BT-speaker [27] NeurIPS2018	0.182	0.685	0.253	0.483	0.227	0.473	0.178	0.658	0.250	0.391	0.209	0.440
EDrop-speaker [75] NAACL2019	0.195	0.701	0.265	0.486	0.224	0.463	0.184	0.660	0.260	0.413	0.215	0.455
CCC-speaker [82] CVPR2022	0.196	0.698	0.267	0.498	0.233	0.467	0.183	0.679	0.254	0.401	0.226	0.456
Lana [91] CVPR2023	0.201	0.694	0.270	0.503	0.230	0.473	0.194	0.689	0.260	0.419	0.219	0.463
BEVInstructor **(Ours)**	**0.220**	**0.731**	**0.285**	**0.549**	**0.238**	**0.480**	**0.208**	**0.699**	**0.264**	**0.449**	**0.230**	**0.467**

Table 2: Quantitative comparison results for **instruction generation** on REVERIE [69] `val seen` and `val unseen`. See §4.2 for more details.

Methods	REVERIE `val seen`						REVERIE `val unseen`					
	SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑	SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
BT-speaker [27] NeurIPS2018	0.121	0.693	0.347	0.269	0.223	0.602	0.103	0.664	0.302	0.190	0.200	0.519
EDrop-speaker [75] NAACL2019	0.133	0.666	0.353	0.517	0.237	0.589	0.114	0.655	0.312	0.252	0.222	0.534
CCC-speaker [82] CVPR2022	0.138	0.680	0.377	0.549	0.244	0.593	0.117	0.671	0.329	0.280	0.233	0.533
Lana [91] CVPR2023	0.137	0.714	0.408	0.619	0.280	0.615	0.108	0.701	0.332	0.406	0.237	0.542
BEVInstructor **(Ours)**	**0.208**	**0.773**	**0.425**	**0.745**	**0.324**	**0.635**	**0.159**	**0.732**	**0.335**	**0.489**	**0.267**	**0.560**

4.2 Quantitative Results

R2R. As depicted in Table 1, BEVInstructor achieves the best performance across all metrics on both `val` splits. Notably, on the primary metrics — **SPICE**, it achieves an improvement of **1.9%** on `val seen` and **1.4%** on `val unseen`. Additionally, our model surpasses the leading benchmarks of prior studies, achieving a **4.6%** improvement on `CIDEr` of R2R `val seen` split and a **3.0%** enhancement on `CIDEr` of R2R `val unseen` split. This verifies the efficacy of BEVInstructor in generating navigation instructions for indoor scenarios.

REVERIE. Table 2 compares BEVInstructor with the recent state-of-the-art instruction generation models [27, 75, 82, 91] on REVERIE dataset. BEVInstructor outperforms previous approaches across all the evaluation metrics on the `val` split. Specifically, on the `val seen` split, BEVInstructor exceeds the previous best model by **7.0%** on `SPICE`, **12.6%** on `CIDEr`, and **4.4%** on `Meteor`. On the `val unseen` split, BEVInstructor improves the performance by **4.2%** on `SPICE`, **8.3%** on `CIDEr`, and **3.0%** on `Meteor`. These results underscore the versatility and generality of our architectural design for goal-based tasks.

UrbanWalk. Table 3 presents comparison results on UrbanWalk dataset. On UrbanWalk test split, BEVInstructor outperforms previous methods by a significant advancement, **11.3%** on `SPICE`, **12.5%** on `Bleu-4`, **7.3%** on `Meteor` and **13.1%** on `Rouge`. This suggests BEVInstructor is also capable of handling more challenging outdoor scenes.

Table 3: Quantitative comparison results for **instruction generation** on UrbanWalk [37] `test`. See §4.2 for more details.

Methods	UrbanWalk `test`			
	SPICE↑	Bleu-4↑	Meteor↑	Rouge↑
BT-speaker [27] NeurIPS2018	0.524	0.408	0.350	0.620
EDrop-speaker [75] NAACL2019	0.531	0.435	0.358	0.634
ASSISTER [37] ECCV2022	0.451	0.164	0.319	0.557
Kefa-speaker [100] Arxiv2023	0.566	0.450	0.378	0.655
BEVInstructor **(Ours)**	**0.679**	**0.575**	**0.451**	**0.786**

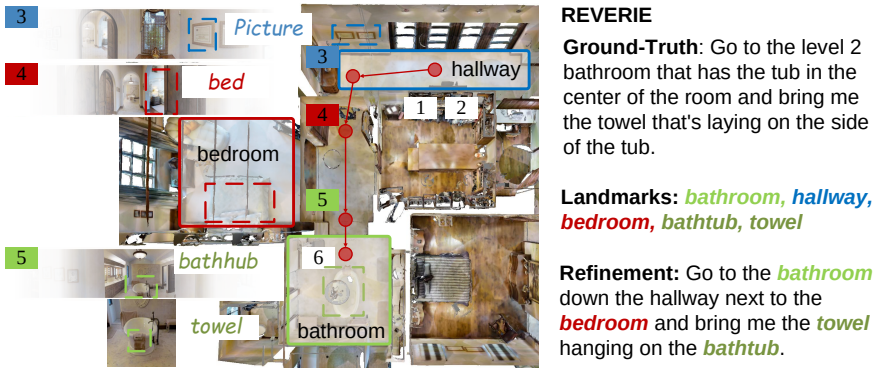

Fig. 3: Visual comparison results between ground truth and BEVInstructor for instruction generation on REVERIE [69]. See §4.3 for more details.

Table 4: Ablation study on R2R [6] `val unseen`. See §4.4 for more details.

#	Perspective	BEV	Fusion	Refinement	R2R val unseen					
					SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
1	✔				0.154	0.625	0.170	0.209	0.198	0.392
2		✔			0.172	0.653	0.184	0.281	0.206	0.405
3	✔	✔			0.180	0.673	0.217	0.342	0.224	0.442
4	✔	✔	✔		0.190	0.683	0.238	0.373	0.224	0.453
5	✔	✔		✔	0.192	0.676	0.242	0.419	0.220	0.455
6	✔	✔	✔	✔	**0.208**	**0.699**	**0.264**	**0.449**	**0.230**	**0.467**

4.3 Qualitative Results

Fig. 3 provides qualitative comparisons of BEVInstructor against the ground truth on the REVERIE. BEVInstructor shows an enhanced capability in identifying scenes and objects related to action views, and explicitly incorporates these elements into the instructions in the refinement stage.

4.4 Diagnostic Experiment

To assess the efficacy of essential modules of BEVInstructor, we conduct a series of detailed ablation studies on `val unseen` split of R2R [6].

Overall Design. We first study the efficacy of the core components of BEVInstructor in Table 4. Row #1 illustrates the impact of fine-tuning MLLMs. This shows competitive performance, demonstrating its potential by elevating language capabilities. Row #2 and #3 indicate that the integration of BEV features alongside perspective features yields notable performance improvements by **6.1%** on `CIDEr`. From row #3 and #4, compared with simply concatenating features, fusing BEV and perspective features through the transformer module results in a greater performance improvement by **1.0%** on `SPICE`. Comparisons between row #3 and #5, as well as row #4 and #6, underscore the efficacy of the

Table 5: Ablation study of other transformer-based algorithms on R2R val unseen. See §4.4 for more details.

Methods	R2R val unseen					
	SPICE↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
GPT-4V [64] [Arxiv2023]	0.098	0.403	0.079	0.076	0.130	0.296
AutoVLN(GPT2) [15] [ECCV2022]	0.145	0.613	0.181	0.248	0.188	0.398
PASTS [87] [EAAI2024]	0.151	0.645	0.195	0.258	0.186	0.415
InstructBLIP [20] [NeurIPS2023]	0.163	0.666	0.220	0.321	0.201	0.418
BEVInstructor **(Ours)**	**0.208**	**0.699**	**0.264**	**0.449**	**0.230**	**0.467**

Table 6: A set of ablation studies on R2R [6] `val unseen`. See §4.4 for more details.

Fusion	SPICE↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
Addition	0.185	0.226	0.366	0.214	0.450
Concat	0.184	0.192	0.310	0.213	0.436
Ours	**0.208**	**0.264**	**0.449**	**0.230**	**0.467**

(a) Ablation study of different fusion of Perspective-BEV features on R2R [6] `val unseen`.

steps	SPICE↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
Base	0.190	0.238	0.373	0.224	0.453
One	**0.208**	**0.264**	0.449	**0.230**	0.467
Two	0.204	**0.264**	**0.456**	**0.230**	**0.474**

(b) Ablation study of different steps of refinement on R2R [6] `val unseen`.

instance-guided iterative refinement module. In row #6, we combine all the components, and obtain the best performance. This suggests that these modules are complementary to each other, and confirms the effectiveness of our whole design.

Language Architectures. Table 5 presents the performance comparison of various transformer-based algorithms on R2R val unseen split. Except for GPT-4V, the other methods are fine-tuned on R2R train split. The results show that more advanced language architectures can effectively adapt to the task of generating navigation instructions with just fine-tuning, achieving competitive performance. This confirms the potential of MLLMs to enhance instruction generation.

Fusion of Perspective-BEV Features. We compare the performance of three different approaches for fusing perspective and BEV features: **i)** addition of perspective and BEV features, **ii)** concatenation of perspective features and BEV features along the token dimension, and **iii)** fusion by transformer modules. The results are summarized in Table 6a. Note that, the fusion design of BEVInstructor outperforms the other two simpler fusion approaches. This robustly validates the architecture design of BEVInstructor.

Refinement. Table 6b presents the performance comparison of different refinement steps on the R2R `val unseen` split. As shown in row #1 and #2, the instance-guided iterative refinement improves the instructions through multi-step reasoning. However, from row #2 and #3, further increasing the steps of refinements only brings limited improvement.

4.5 Instruction Quality Analysis

The above captioning metrics reflect the word alignment quality of the generated instructions. To further demonstrate the effectiveness of the instructions, we

Table 7: A set of analysis of instruction quality. See §4.5 for more details.

Method	HAMT [14]		DUET [16]	
	SR↑	SPL↑	SR↑	SPL↑
Human annotation [69] [CVPR2020]	32.95	30.20	46.98	33.73
BT-speaker [27] [NeurIPS2018]	22.48	19.47	28.41	15.30
EDrop-speaker [75] [NAACL2019]	23.74	20.98	30.66	19.27
CCC-speaker [82] [CVPR2022]	23.80	21.18	28.84	14.36
Lana [91] [CVPR2023]	23.94	21.34	31.61	21.26
BEVInstructor **(Ours)**	**25.68**	**22.48**	**33.81**	**23.23**

(a) The performance of HAMT [14] and DUET [16] guided by instructions generated on REVERIE [69] `val unseen`.

Data Source	R2R `val unseen`			
	TL	NE↓	SR↑	SPL↑
Original [69] [CVPR2020]	9.62	5.86	45.4	41.8
+ BT-speaker [27] [NeurIPS2018]	9.81	5.95	45.1	41.5
+ EDrop-speaker [75] [NAACL2019]	9.33	5.68	45.5	42.2
+ CCC-speaker [82] [CVPR2022]	9.43	5.73	45.3	42.0
+ Lana [91] [CVPR2023]	9.48	5.75	45.6	42.1
BEVInstructor **(Ours)**	9.96	**5.66**	**47.1**	**43.6**

(b) The results of EDrop-follower [75] using different generators for data augmentation on R2R [6] `val unseen`.

conduct the following experiments to evaluate the alignment of instructions and trajectories in actual vision-language navigation tasks. We compare two aspects, *i.e.*, Path Guiding Proficiency, and Data Augmentation, with previous methods to further validate the performance of BEVInstructor.

Path Guiding Proficiency. Table 7a presents the comparison performance of HAMT [14] and DUET [16] guided by instructions from different generators on REVERIE `val unseen`. Concretely, we reproduce **one** instruction for each path on REVERIE `val unseen` and replace the original instructions with them. We follow [14,16,82,91] and adopt the Success Rate (SR) and Success rate weighted by Path Length (SPL) of navigation as metrics. BEVInstructor outperforms others across all metrics of HAMT and DUET. These results confirm the fidelity and generalization of our instructions.

Data Augmentation. One application of the navigation instruction generator is to create diverse instructions for data augmentation in Vision-Language Navigation (VLN) agents. We randomly sample paths in the R2R `train` split and employ different generators to produce instructions. The generated instructions combined with the original ones are used to train the EDrop-follower [75]. In Table 7b, the instructions generated by BEVInstructor for data augmentation enhance the performance of VLN, leading to promising improvements.

5 Conclusion

Navigation instruction generation has been of great importance in many disciplines. However, existing studies encounter the following challenges: **i)** they rely exclusively on perspective features, ignoring the geometric prior and object semantics inherent in 3D scenes, and **ii)** their language decoders are limited by a lack of extensive prior world knowledge and the scale of the model. In light of this, we propose BEVInstructor that integrates BEV features with MLLMs to jointly improve 3D perception and linguistic capabilities. BEVInstructor exhibits superior performance in comparison to previous studies. This work brings us closer to developing interactive and trustworthy navigation robots.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62293554), the Fundamental Research Funds for the Central Universities (No. 226-2022-00051), National Natural Science Foundation of China (No. 62372405), and CCF-Tencent Open Fund.

References

1. Agarwal, S., Parikh, D., Batra, D., Anderson, P., Lee, S.: Visual landmark selection for generating grounded and interpretable navigation instructions. In: CVPR workshop (2019) 4, 10
2. Allen, G.L.: From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In: International Conference on Spatial Information Theory. pp. 363–372. Springer (1997) 4
3. An, D., Qi, Y., Li, Y., Huang, Y., Wang, L., Tan, T., Shao, J.: Bevbert: Multimodal map pre-training for language-guided navigation. In: ICCV (2023) 4
4. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. IEEE Transactions on PAMI (2024) 2
5. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV (2016) 10
6. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018) 3, 5, 9, 10, 11, 12, 13, 14
7. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL Workshop (2005) 10
8. Baruch, G., Chen, Z., Dehghan, A., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In: NeurIPS (2021) 4
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023) 2
10. Cartillier, V., Ren, Z., Jain, N., Lee, S., Essa, I., Batra, D.: Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In: AAAI (2021) 4
11. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017) 10
12. Chen, J., Wang, W., Liu, S., Li, H., Yang, Y.: Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In: ICCV (2023) 2
13. Chen, P., Ji, D., Lin, K., Zeng, R., Li, T., Tan, M., Gan, C.: Weakly-supervised multi-granularity map learning for vision-and-language navigation. In: NIPS (2022) 4
14. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. In: NeurIPS (2021) 14
15. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Learning from unlabeled 3d environments for vision-and-language navigation. In: ECCV (2022) 13

16. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: CVPR (2022) 14
17. Cui, Y., Xie, L., Zhang, Y., Zhang, M., Yan, Y., Yin, E.: Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In: ICCV (2023) 4
18. Curry, A.C., Gkatzia, D., Rieser, V.: Generating and evaluating landmark-based navigation instructions in virtual environments. In: ENLG (2015) 3
19. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017) 4
20. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023) 5, 13
21. Dale, R., Geldof, S., Prost, J.: Using natural language generation in automatic route. Journal of Research and practice in Information Technology **36**(3), 23 (2004) 2, 3
22. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. In: ICML (2023) 2
23. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE robotics & automation magazine **13**(2) (2006) 4
24. Evans, G.W., Marrero, D.G., Butler, P.A.: Environmental learning and cognitive mapping. Environment and behavior **13**(1), 83–104 (1981) 1, 3
25. Fernandes, P., Madaan, A., Liu, E., Farinhas, A., Martins, P.H., Bertsch, A., de Souza, J.G., Zhou, S., Wu, T., Neubig, G., et al.: Bridging the gap: A survey on integrating (human) feedback for natural language generation. arXiv preprint arXiv:2305.00955 (2023) 9
26. Fried, D., Andreas, J., Klein, D.: Unified pragmatic models for generating and following instructions. In: NAACL (2018) 4
27. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: NeurIPS (2018) 2, 4, 6, 11, 14
28. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023) 5
29. Goeddel, R., Olson, E.: Dart: A particle-based method for generating easy-to-follow directions. In: International Conference on Intelligent Robots and Systems (2012) 1, 2, 3
30. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022) 2
31. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. In: CVPR (2020) 10
32. Henriques, J.F., Vedaldi, A.: Mapnet: An allocentric spatial memory for mapping environments. In: CVPR (2018) 4
33. Hong, Y., Zhou, Y., Zhang, R., Dernoncourt, F., Bui, T., Gould, S., Tan, H.: Learning navigational visual representations with semantic map supervision. In: CVPR (2023) 4
34. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML (2019) 5

35. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 5
36. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. In: CoRL (2023) 2
37. Huang, Z., Shangguan, Z., Zhang, J., Bar, G., Boyd, M., Ohn-Bar, E.: Assister: Assistive navigation via conditional instruction generation. In: ECCV (2022) 2, 3, 9, 10, 11
38. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV (2022) 8
39. Kuipers, B.: Modeling spatial knowledge. Cognitive science **2**(2), 129–153 (1978) 1, 3
40. Kwon, O., Park, J., Oh, S.: Renderable neural radiance map for visual navigation. In: CVPR (2023) 4
41. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2024) 5
42. Li, H., Sima, C., Dai, J., Wang, W., Lu, L., Wang, H., Zeng, J., Li, Z., Yang, J., Deng, H., et al.: Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 2, 4
43. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) 4, 5, 7
44. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 2, 5
45. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: AAAI (2023) 7, 9
46. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022) 4, 7
47. Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., Alvarez, J.M.: FB-BEV: BEV representation from forward-backward view transformations. In: ICCV (2023) 4, 7
48. Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & shifting your features: A new baseline for efficient model tuning. In: NeurIPS (2022) 8
49. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS (2022)
50. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text summarization branches out pp. 74–81 (2004) 10
51. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2024) 4, 5, 8
52. Liu, R., Wang, W., Yang, Y.: Volumetric environment representation for vision-language navigation. In: CVPR (2024) 4
53. Liu, R., Wang, X., Wang, W., Yang, Y.: Bird's-eye-view scene graph for vision-language navigation. In: ICCV (2023) 4

54. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA (2023)
55. Look, G., Kottahachchi, B., Laddaga, R., Shrobe, H.: A location representation for generating descriptive walking directions. In: IUI (2005) 2, 3, 4
56. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 10
57. Lovelace, K.L., Hegarty, M., Montello, D.R.: Elements of good route directions in familiar and unfamiliar environments. In: Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science: International Conference COSIT'99 Stade, Germany, August 25–29, 1999 Proceedings 4. pp. 65–82. Springer (1999) 4
58. Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models. In: NeurIPS (2023) 5
59. Lynch, K.: The image of the city. MIT press (1964) 3
60. Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., Manocha, D., Zhu, X.: Vision-centric bev perception: A survey. arXiv preprint arXiv:2208.02797 (2022) 2
61. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 5, 8
62. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 4
63. Moudgil, A., Majumdar, A., Agrawal, H., Lee, S., Batra, D.: Soat: A scene-and object-aware transformer for vision-and-language navigation. In: NeurIPS (2021) 4
64. OpenAI: Gpt-4 technical report (2023) 2, 8, 13
65. Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., Wang, W.Y.: Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. arXiv preprint arXiv:2308.03188 (2023) 9
66. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002) 10
67. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV (2020) 4, 9
68. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. In: EMNLP (2023) 5
69. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., van den Hengel, A.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR (2020) 3, 9, 10, 11, 12, 14
70. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distributionnetwork for monocular 3d object detection. In: CVPR (2021) 4
71. Richter, K.F., Duckham, M.: Simplest instructions: Finding easy-to-describe routes for navigation. In: Geographic Information Science: 5th International Conference, GIScience 2008, Park City, UT, USA, September 23-26, 2008. Proceedings 5 (2008) 4
72. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image

retrieval. In: Proceedings of the fourth workshop on vision and language (2015) 10
73. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. In: Taming Large Language Models (TLLM) (2023) 5
74. Sung, Y.L., Cho, J., Bansal, M.: Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In: CVPR (2022) 5
75. Tan, H., Yu, L., Bansal, M.: Learning to navigate unseen environments: Back translation with environmental dropout. In: NAACL (2019) 2, 4, 6, 9, 11, 14
76. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2, 5, 8, 9
77. Vanetti, E.J., Allen, G.L.: Communicating environmental knowledge: The impact of verbal and spatial abilities on the production and comprehension of route directions. Environment and Behavior **20**(6), 667–682 (1988) 1, 3
78. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 4
79. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015) 10
80. Waller, D., Lippa, Y.: Landmarks as beacons and associative cues: their role in route learning. Memory & Cognition **35**(5), 910–924 (2007) 4
81. Wang, H., Liang, W., Gool, L.V., Wang, W.: Towards versatile embodied navigation. In: NeurIPS (2022) 2
82. Wang, H., Liang, W., Shen, J., Van Gool, L., Wang, W.: Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In: CVPR (2022) 2, 3, 4, 6, 9, 10, 11, 14
83. Wang, H., Liang, W., Van Gool, L., Wang, W.: Dreamwalker: Mental planning for continuous vision-language navigation. In: ICCV (2023) 2
84. Wang, H., Wang, W., Liang, W., Hoi, S.C., Shen, J., Gool, L.V.: Active perception for visual-language navigation. IJCV **131**(3), 607–625 (2023) 2
85. Wang, H., Wang, W., Liang, W., Xiong, C., Shen, J.: Structured scene memory for vision-language navigation. In: CVPR (2021) 2
86. Wang, H., Wang, W., Shu, T., Liang, W., Shen, J.: Active visual information gathering for vision-language navigation. In: ECCV (2020) 2
87. Wang, L., Liu, C., He, Z., Li, S., Yan, Q., Chen, H., Chen, Q.: Pasts: Progress-aware spatio-temporal transformer speaker for vision-and-language navigation. Engineering Applications of Artificial Intelligence **128**, 107487 (2024) 2, 4, 13
88. Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., Anderson, P.: Less is more: Generating grounded navigation instructions from landmarks. In: CVPR (2022) 4
89. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., Liu, X., Lu, C., Lin, D., Pang, J.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: CVPR (2024)
90. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In: NeurIPS (2024) 2, 5
91. Wang, X., Wang, W., Shao, J., Yang, Y.: Lana: A language-capable navigator for instruction following and generation. In: CVPR (2023) 2, 3, 4, 6, 10, 11, 14
92. Wang, X., Wang, W., Shao, J., Yang, Y.: Learning to follow and generate instructions for language-capable navigation. IEEE Transactions on PAMI (2023) 2, 4

93. Wang, Y., Guizilini, V., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning (2021) 7
94. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Gridmm: Grid memory map for vision-and-language navigation. In: ICCV (2023) 4
95. Ward, S.L., Newcombe, N., Overton, W.F.: Turn left at the church, or three miles north: A study of direction giving and sex differences. Environment and Behavior **18**(2), 192–213 (1986) 4
96. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: Personalized robot assistance with large language models. Autonomous Robots **47**(8), 1087–1102 (2023) 2
97. Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Jiang, C., Tan, H., Kang, J., Zhang, Y., Zhou, K., et al.: Octopus: Embodied vision-language programmer from environmental feedback. arXiv preprint arXiv:2310.08588 (2023) 2
98. Yang, Z., Chen, G., Li, X., Wang, W., Yang, Y.: Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In: ICML (2024) 2
99. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In: NeurIPS (2024) 5
100. Zeng, H., Wang, X., Wang, W., Yang, Y.: Kefa: A knowledge enhanced and fine-grained aligned speaker for navigation instruction generation. arXiv preprint arXiv:2307.13368 (2023) 2, 11
101. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: EMNLP (2023) 5
102. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In: ICLR (2024) 5, 8
103. Zhao, M., Anderson, P., Jain, V., Wang, S., Ku, A., Baldridge, J., Ie, E.: On the evaluation of vision-and-language navigation instructions. In: EACL (2021)
104. Zheng, Z., Wang, W., Qi, S., Zhu, S.C.: Reasoning visual dialogs with structural and partial observations. In: CVPR (2019) 2
105. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024) 2, 4, 5, 8
106. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020) 7