GoViG: Goal-Conditioned Visual Navigation Instruction Generation via Multimodal Reasoning

Fengyi Wu[1,*] **Yifei Dong**[1,*] **Yilong Dai**[1] **Guangyu Chen**[1] **Qifeng Wu**[1]
Huiting Huang[1] **Hang Wang**[2] **Qi Dai**[3] **Alexander G. Hauptmann**[4] **Zhi-Qi Cheng**[1,†]
[1]Tacoma School of Engineering & Technology, University of Washington
[2]Department of Computing, The Hong Kong Polytechnic University [3]Microsoft Research
[4]Language Technologies Institute, Carnegie Mellon University
[*]Equal contribution. [†]Corresponding author.
fyiwu@uw.edu, yfeidong@uw.edu, zhiqics@uw.edu

Abstract

We introduce *Goal-Conditioned Visual Navigation Instruction Generation* (GoViG), a new task that aims to generate contextually coherent navigation instructions solely from egocentric visual observations of initial and goal states. Unlike prior work relying on structured inputs, such as semantic annotations or environmental maps, GoViG exclusively leverages raw egocentric visual data, improving adaptability to unseen and unstructured environments. Our method addresses this task by decomposing it into two interconnected subtasks: (1) navigation visualization, predicting intermediate visual states bridging the initial and goal views; and (2) instruction generation, synthesizing coherent instructions grounded in observed and anticipated visuals. Both subtasks are integrated within an autoregressive multimodal LLM trained with tailored objectives to ensure spatial accuracy and linguistic clarity. Furthermore, we introduce two multimodal reasoning strategies, one-pass and interleaved reasoning, to mimic incremental human navigation cognition. To comprehensively evaluate our method, we propose the *R2R-Goal* dataset, combining diverse synthetic and real-world trajectories. Empirical results demonstrate significant performance improvements over state-of-the-art methods in BLEU-4 and CIDEr scores along with robust cross-domain generalization. Our project is available at https://github.com/F1y1113/GoViG.

1 Introduction

Generating natural language navigation instructions from egocentric visual observations remains a critical yet underexplored area within embodied AI. While Vision-and-Language Navigation (VLN) research has largely focused on language grounding, training agents to interpret and execute human instructions (Anderson et al., 2018; Fried et al., 2018), the inverse challenge of *instruction generation* is relatively understudied. Effective instruction generation is crucial for practical applications such as aiding visually impaired users, facilitating seamless human-agent collaboration, and guiding navigation in hazardous or unfamiliar environments (Zhang et al., 2024).

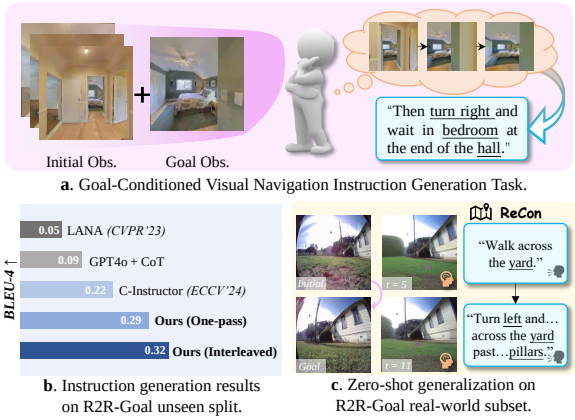


Figure 1: (a) Goal-Conditioned Visual Navigation Instruction Generation (GoViG): generating instructions from egocentric initial and goal views. (b) Results on the R2R-Goal dataset. (c) Zero-shot generalization to real-world scenarios.

Current methods for instruction generation predominantly depend on *privileged inputs*, including semantic maps, landmark annotations, and panoramic views, limiting their applicability beyond controlled and structured scenarios (Fan et al., 2024; Wang et al., 2023; Kong et al., 2024). Alternatively, some approaches simplify visual data into textual summaries, inadvertently discarding essential spatial and semantic information intrinsic to raw visual observations (Fan et al., 2024; Wang et al., 2022b; Zeng et al., 2023). Such oversimplifications impede an agent's capability for accurate reasoning and generalization in novel contexts.

Recent advancements in multimodal large language models (MLLMs), such as LLaVA (Liu et al., 2023), GPT-4o (Hurst et al., 2024), and Gemini (Google, 2024; Comanici et al., 2025), have demonstrated remarkable proficiency in vision-language tasks. Nevertheless, these models generally lack explicit mechanisms for coherent visual

Findings of the Association for Computational Linguistics: ACL 2026, pages 14899–14920
July 2-7, 2026

forecasting and seldom incorporate iterative mental simulation strategies utilized by humans during route planning (Chen et al., 2023; Zhang et al., 2023). Consequently, instructions generated by existing MLLMs frequently suffer from a lack of contextual precision and temporal consistency.

To address these limitations, we introduce *Goal-Conditioned Visual Navigation Instruction Generation* (GoViG), a novel task aiming to generate precise and contextually coherent navigation instructions using only egocentric visual observations from initial and goal viewpoints (Fig.1(a)). Unlike previous approaches, GoViG entirely eliminates reliance on privileged inputs, significantly enhancing the method's generalization capability across diverse and unseen environments (Fig.1(b)-(c)).

Our approach systematically decomposes GoViG into two complementary subtasks: (1) *Navigation Visualization*, predicting intermediate visual states to bridge initial and goal observations; and (2) *Instruction Generation with Visual Cues*, synthesizing instructions grounded in observed and forecasted visual cues (Fig. 2(a)). Both are integrated within an autoregressive MLLM, guided by carefully designed training objectives: a *Token Discrepancy Loss*, which promotes accurate visual predictions, and a *Label Smoothing Loss*, enhancing semantic robustness and linguistic fluency. This methodological synergy aligns closely with human spatial cognition, fostering robust and adaptable instruction generation.

Furthermore, we propose two multimodal reasoning strategies during inference: *One-Pass Multimodal Reasoning*, leveraging global visual context for structured scenarios; and *Interleaved Multimodal Reasoning*, iteratively refining visual predictions and linguistic instructions to emulate human adaptive navigation under uncertainty (Fig. 2(b)). These strategies enhance spatial accuracy, linguistic coherence and cross-domain generalization.

We extensively evaluate GoViG on proposed *R2R-Goal* dataset, integrating synthetic trajectories from R2R-CE (Krantz et al., 2020) and HA-R2R (Dong et al., 2025b) with real-world egocentric videos from GO Stanford (Hirose et al., 2018), ReCon (Shah et al., 2021), and HuRoN (Hirose et al., 2023), each meticulously annotated with natural language instructions. Our interleaved reasoning strategy achieves superior performance with BLEU-4 (0.32) and CIDEr (0.20) scores on validation (Table 1). Moreover, it attains a BLEU-4 of 0.27 in zero-shot cross-domain evaluations (Table 5), showing robust generalization capabilities.

Our contributions can be summarized as follows:

1. We formally propose Goal-Conditioned Visual Navigation Instruction Generation (GoViG), a new task generating precise navigation instructions solely from egocentric initial and goal observations, without privileged inputs (Sec. 3.1).

2. We systematically decompose GoViG into two subtasks: Navigation Visualization and Instruction Generation with Visual Cues, and integrate them within a unified autoregressive MLLM optimized with tailored training objectives (Sec. 3.3).

3. We introduce and evaluate two multimodal reasoning strategies (One-Pass and Interleaved) designed to enhance spatial accuracy and linguistic coherence through global and iterative visual-linguistic reasoning (Sec. 3.4).

4. We release the R2R-Goal dataset, a comprehensive benchmark combining synthetic and real-world navigation scenarios. Extensive empirical evaluations validate our method's superior instruction generation performance and robust cross-domain generalization (Secs. 3.2, 4).

2 Related Work

2.1 Navigation Instruction Generation

Navigation instruction generation originates from cognitive science research examining human spatial cognition and culturally influenced route descriptions (Lynch, 1964; Allen, 1997; Vanetti and Allen, 1988; Hund and Minarik, 2006). Recent advancements in VLN have renewed interest in instruction generation, primarily for data augmentation (Anderson et al., 2018; Zhang et al., 2024). Early computational approaches, such as Speaker-Follower (Fried et al., 2018), employed recurrent neural networks to generate instructions. Subsequent methods (Tan et al., 2019; Wang et al., 2022a, 2025b) enhanced instruction quality but continued relying on structured inputs, such as semantic annotations, panoramic images, and environmental maps (Fan et al., 2024, 2025; Gopinathan et al., 2024; Zeng et al., 2023; Cui et al., 2025; Yan et al., 2024; Zhao et al., 2025b; Wang et al., 2025a), limiting their generalization to novel scenarios.

Moreover, contemporary approaches often preprocess visual inputs into intermediate representations, such as landmarks or commonsense knowledge, unintentionally discarding critical spatial

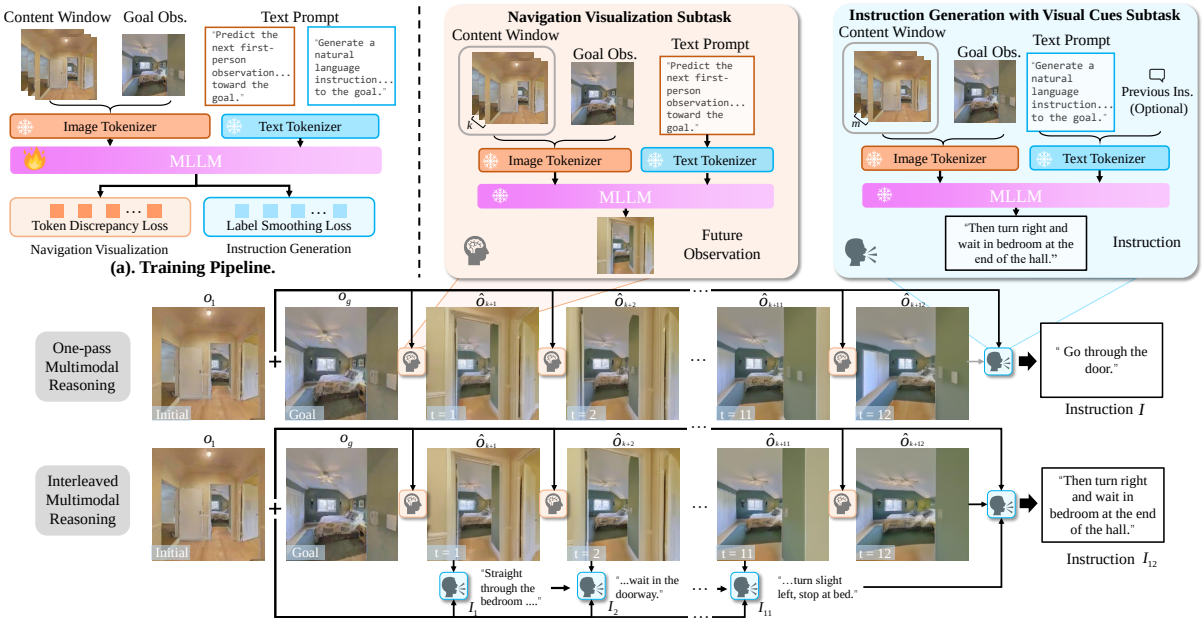


Figure 2: **Overview of our approach** to Goal-Conditioned Visual Navigation Instruction Generation (GoViG): (a) An autoregressive MLLM integrates Navigation Visualization and Instruction Generation subtasks via tailored training objectives. (b) Two inference-time multimodal reasoning strategies: One-pass and Interleaved, which enable coherent visual forecasting and instruction synthesis. The context size is set to one for clarity. [Zoom in for details.]

and semantic details inherent to raw visual observations (Kong et al., 2024; Cui et al., 2025; Gopinathan et al., 2024; Zeng et al., 2023; Wang et al., 2025a). In contrast, our method explicitly leverages raw egocentric visual observations through multimodal reasoning strategies, One-Pass and Interleaved, to directly embed visual cognition into the instruction generation process. Further detailed comparisons are provided in the Appendix.

2.2 Multimodal Reasoning

Recent multimodal large language models (MLLMs) (Yang et al., 2023a; Anthropic, 2024; Google, 2024; Comanici et al., 2025) have advanced visual-textual understanding significantly. Models like LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023b), and VideoChat (Li et al., 2023c) excel at multimodal comprehension, while generative frameworks (Hong et al., 2022; Henschel et al., 2024), have enhanced video synthesis. Integrated architectures such as GPT-4o (Hurst et al., 2024) further showcase sophisticated multimodal reasoning. Concurrently, Chain-of-Thought (CoT) reasoning (Wei et al., 2022) has emerged as essential within multimodal reasoning frameworks. Foundational methods (Chen et al., 2023; Zhang et al., 2023) structured reasoning explicitly around visual data. Subsequent research (Wei et al., 2024; Li et al., 2023a; Zhao et al., 2023)) extended CoT reasoning to zero-shot video understanding and egocentric activities. Recent studies (Shao et al., 2024; Zhou et al., 2024; Wu et al., 2024) advocate spatially coherent visual-textual inference. Inspired by these advances, our method integrates visualization and CoT-based linguistic reasoning, enabling coherent navigation instruction generation directly from egocentric visuals.

2.3 World Models for Visual Generation

World models (Ha and Schmidhuber, 2018) have become a central paradigm for learning predictive representations of environment dynamics (Ding et al., 2024), evolving from compact recurrent structures to large-scale generative and multimodal systems. Early works (Ha and Schmidhuber, 2018; Hafner et al., 2019, 2022, 2024) employed RNN-based latent dynamics to capture temporal transitions. Transformer-based designs (Assran et al., 2023; Bardes et al., 2024; Karypidis et al., 2024; Baldassarre et al., 2025) introduced scalable attention mechanisms for richer spatio-temporal abstraction. Parallel efforts exploit LLMs to simulate dynamics (Zhao et al., 2025a; Xing et al., 2025; Dong et al., 2025a, 2026), but they face modality misalignment, temporal inconsistency, and grounding challenges (Ding et al., 2024; Dong et al., 2025c). Inspired by this paradigm, our Navigation Visualization subtask adopts a lightweight world-model perspective within an autoregressive MLLM: it iteratively predicts intermediate egocentric observa-

tions to bridge the initial and goal states, enabling downstream instruction generation that is grounded in anticipated visual futures.

3 Methodology

3.1 Task Overview & Multimodal Reasoning

Task Formulation. We define Goal-Conditioned Visual Navigation Instruction Generation (GoViG) as the task of generating coherent natural language instructions to guide an agent towards a specified goal using solely egocentric visual observations. Specifically, given an initial visual sequence $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ and a goal observation o_g, where each $o_i, o_g \in \mathbb{R}^{H\times W\times 3}$ denotes an RGB egocentric image, the objective is to produce an accurate navigation instruction I that clearly delineates the necessary steps for reaching the goal. Inspired by the world model paradigm of predicting future states to support downstream decision-making, we systematically approach this task by decomposing it into two interconnected subtasks:

- **Navigation Visualization.** Given a partial visual observation sequence $\mathcal{O}_V = \{o_1, o_2, \ldots, o_k\}$ and the goal observation o_g, the model predicts the next visual observation o_{k+1}, incrementally bridging the gap between the initial and goal states through visual imagination.

- **Instruction Generation with Visual Cues.** Given a visual sequence $\mathcal{O}_I = \{o_1, o_2, \ldots, o_m\}$ (where typically $m = k + 1$), the goal observation o_g, and optionally an intermediate instruction I_{prev}, the model generates a coherent, contextually grounded instruction I that articulates the sequential navigation steps towards the goal.

To facilitate training, we implement an autoregressive MLLM as illustrated in Fig. 2(a), with tailored loss functions specifically designed for each subtask. We present the detailed model architecture and training procedures in Sec. 3.3.

Multimodal Reasoning. Unlike conventional approaches that directly translate visual inputs into textual instructions, our framework explicitly integrates structured visual reasoning to improve robustness and generalization. Specifically, we introduce two distinct multimodal reasoning strategies, illustrated in Fig. 2(b) and detailed in Sec. 3.4:

- **One-Pass Multimodal Reasoning.** Given an initial visual sequence $\mathcal{O}_{\text{init}} = \{o_1, \ldots, o_k\}$ and a goal observation o_g, the model forecasts a complete trajectory $\hat{\mathcal{O}} = \{\hat{o}_{k+1}, \ldots, \hat{o}_{k+t}\}$ toward the goal. Subsequently, the navigation instruction I is generated from selected representative frames in $\hat{\mathcal{O}}$, emphasizing holistic spatial context and global scene awareness.

- **Interleaved Multimodal Reasoning.** Starting from initial observations $\mathcal{O}_{\text{init}}$, the model iteratively alternates between forecasting next visual observation $\hat{o}_{k+t}$ and incrementally updating corresponding instruction I_t. This approach closely mimics incremental human cognitive processes, ensuring precise alignment between visual perception and linguistic instruction generation.

Unlike conventional techniques relying on explicit coordinates, action labels, or semantic maps, our approach solely employs egocentric visual observations, enabling enhanced generalization to diverse, unknown environments and laying the groundwork for cross-domain applications.

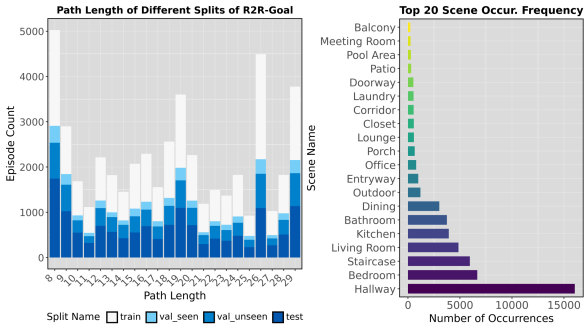


Figure 3: **R2R-Goal dataset statistics:** (*left*) Distribution of trajectory lengths across training, validation and test splits; (*right*) Top 20 scene categories ranked by frequency, showcasing coverage of diverse indoor and outdoor environments.

3.2 Construction of the R2R-Goal Dataset

To support GoViG task, we introduce **R2R-Goal dataset**. This dataset integrates language instructions from existing R2R-CE (Krantz et al., 2020) and HA-R2R (Dong et al., 2025b; Li et al., 2024) datasets and incorporates first-person observations from GO Stanford (Hirose et al., 2018), ReCon (Shah et al., 2021) and HuRoN (Hirose et al., 2023) datasets as a dedicated real-world test subset.

To leverage R2R-CE and HA-R2R, we generate egocentric observation sequences and corresponding navigation paths using an A*-based heuristic search in the HA-VLN simulator (Dong et al., 2025b), using Qwen-VL-2.5 (Bai et al., 2025) to segment both visual observation and corresponding instructions into semantically coherent subscenes. All trajectory-instruction pairs are manually reviewed to verify spatial coherence and semantic alignment. Misaligned or ambiguous pairs are corrected or removed. This human-in-the-loop

validation ensures diverse, high-quality navigation patterns. As a result, this part of the R2R-Goal dataset consists of 74,737 trajectories, partitioned into training (48,490), validation_seen (3,573), validation_unseen (8,361), and testing (14,313) splits, with detailed statistics presented in Fig. 3.

For the real-world subset, we apply the same segmentation strategy to observation sequences obtained from the GO Stanford, ReCon, and HuRoN datasets. We then manually annotate a total of 1080 trajectories with corresponding natural language navigation instructions. Each trajectory in R2R-Goal retains an initial sequence of six egocentric observations and a final goal observation. These visual sequences, combined with the corresponding navigation instructions, constitute the inputs for our proposed task. Additional dataset construction and annotation details are included in the Appendix.

3.3 Autoregressive MLLM Training

To integrate visual and linguistic reasoning within a unified framework, we employ an autoregressive multimodal Transformer based on the Chameleon architecture (Team, 2024). This design simultaneously addresses two complementary subtasks: Navigation Visualization and Instruction Generation with Visual Cues as illustrated in Fig. 2(a) and detailed in Fig. 7, facilitating shared representation learning and robust multimodal interaction.

Data Preparation & Prompt Design. We construct training samples by pairing egocentric RGB image sequences with natural language navigation instructions. Each trajectory generates two types of training instances: (1) **Navigation Visualization**: each instance consists of a sequence of k preceding visual frames alongside the goal frame, with the task being to predict the next frame. Visual observations are denoted by `<image>` tokens in multimodal prompts, with multiple samples extracted via a sliding temporal window. (2) **Instruction Generation with Visual Cues**: inputs consist of the initial frame, goal frame, and up to $m-1$ intermediate frames, embedded within an image prompt. The associated ground-truth instruction serves as the prediction target. Samples from both subtasks are interleaved in batches for joint optimization using a unified Transformer model. Prompt examples are provided in the Appendix.

Multimodal Tokenization. To effectively integrate visual and textual modalities, our model employs two specialized tokenizers. The first is a vector-quantized (VQ) image tokenizer derived from (Team, 2024), discretizing images into sequences of visual tokens via a learned embedding codebook. The second is an optimized byte-pair encoding (BPE) tokenizer, following (Team, 2024; Tan et al., 2019), converting textual navigation instructions into discrete token sequences. Visual and textual token sequences are concatenated and jointly processed by a causal Transformer, promoting coherent multimodal representations.

Subtask-Specific Training Objectives. To optimize our model for distinct characteristics of each subtask, we introduce tailored training objectives. At each iteration, our autoregressive MLLM jointly processes samples from either the navigation visualization or instruction generation subtask, producing logits across the unified vocabulary. The subtask-specific loss functions are detailed as follows:

To construct navigation visualization loss, we utilize Token Discrepancy Loss (Li et al., 2025) to encourage accurate visual forecasting. Given the ground-truth visual embedding emb_i for token i (out of total n tokens in current image) and visual codebook embeddings $\mathcal{C} = \{\text{emb}_1, \ldots, \text{emb}_N\}$ where N denotes the total number of visual token vocabulary, the loss is computed as:

$$\mathcal{L}_{\text{vis}} = \sum_{i=1}^{n} \text{MSE}(\text{emb}_i, \mathcal{C}) \cdot P(t_i), \quad (1)$$

where $\text{MSE}(emb_i, \mathcal{C}) \in \mathbb{R}^{1\times N}$ is similarity vector containing distances between emb_i and all codebook entries. $P(t_i) \in \mathbb{R}^{1\times N}$ denotes predicted probability distribution for visual tokens at position i.

For instruction generation loss design, we apply a label smoothing cross-entropy loss. Let y_i denote the ground-truth text token at position i in the target instruction sequence, the loss is defined as:

$$\mathcal{L}_{\text{ins}} = -\sum_i \sum_{v\in\mathcal{V}} q_v(y_i) \log P_v(y_i), \quad (2)$$

where $P_v(y_i)$ is predicted probability for token v at position i, and $q_v(y_i)$ represents smoothed distribution around y_i within text vocabulary $\mathcal{V}$, applying smoothing factor ϵ to non-ground-truth tokens.

To stabilize training, we implement an input-label concatenation strategy, masking inputs (with -100 labels) so the loss computation focuses exclusively on the prediction targets. During training, tokenizers remain frozen, and only Transformer parameters are updated via an autoregressive next-token prediction objective.

3.4 Multimodal Reasoning Strategies

During inference, we leverage the trained MLLM F_Θ, with fixed parameters, employing two structured multimodal reasoning strategies: One-Pass

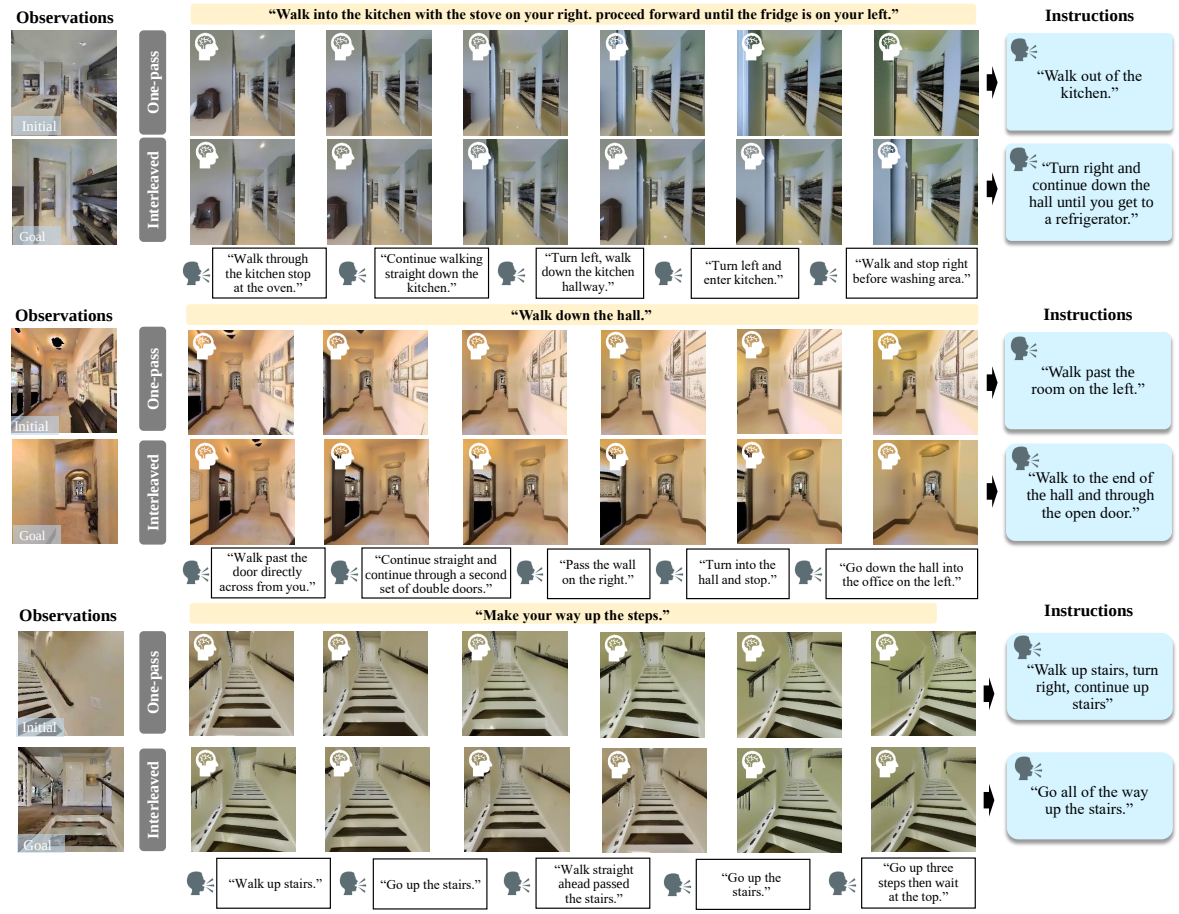


Figure 4: **Qualitative examples** of Navigation Visualization and Instruction Generation results on the R2R-Goal validation unseen split. Both our Multi-modal reasoning strategies perform well on diverse unseen scenes such as stairs and kitchens.

and Interleaved. Both approaches explicitly decompose inference into Navigation Visualization and Instruction Generation with Visual Cues, enhancing interpretability and generalization (Fig. 2(b)).

One-Pass Multimodal Reasoning. This approach employs a sequential visual forecasting strategy to predict a complete trajectory of visual observations from the initial state to the goal. Given an initial visual observation sequence $\mathcal{O}_{\text{init}} = \{o_1, \ldots, o_k\}$ and a goal observation o_g, the model iteratively predicts subsequent visual frames: $\hat{\mathcal{O}} = \{\hat{o}_{k+1}, \ldots, \hat{o}_{k+t}\}$ until a predicted observation $\hat{o}_{k+t}$ satisfies the visual similarity criterion defined by the Structural Similarity Index (SSIM) (Wang et al., 2004): $\text{SSIM}(\hat{o}_{k+t}, o_g) > \tau$. We then strategically select $m-1$ representative intermediate frames:$\{\hat{o}_{i_1}, \ldots, \hat{o}_{i_{m-1}}\}$ from the sequence $\{o_2, \ldots, o_k, \hat{o}_{k+1}, \ldots, \hat{o}_{k+t}\}$ as inputs to generate the final navigation instruction:

$$I = F_\Theta\big(\{o_1, \hat{o}_{i_1}, \ldots, \hat{o}_{i_{m-1}}, o_g\}\big). \quad (3)$$

where $i_1, \ldots, i_{m-1}$ indicate indices of sampled intermediate frames. This method emphasizes holistic visual context and global scene understanding.

Interleaved Multimodal Reasoning. The strategy alternates visualization and instruction generation at each inference step. Initially, the model predicts the immediate next frame $\hat{o}_{k+1}$ based on the init observation sequence $\mathcal{O}_{\text{init}} = \{o_1, \ldots, o_k\}$ and goal o_g, subsequently generating a preliminary instruction I_1 that incorporates updated visual context. Such an iterative cycle of alternating visual predictions and incremental instruction refinements continues until predicted observation aligns visually with goal ($\text{SSIM}(\hat{o}_{k+t}, o_g) > \tau$). Formally, at inference step t, updated instruction I_t is obtained:

$$I_t{=}F_\Theta(\{\, o_t,\ldots,o_k, \hat{o}_{k+1},\ldots,\hat{o}_{k+t},o_g,I_{t-1}\}). \quad (4)$$

This approach effectively mimics human-like cognitive cycles of visual imagining and linguistic description, enhancing the model's spatial reasoning and context adaptability. τ in both reasoning strategies is set to 0.7 in our study. Detailed pseudo-code describing the procedure is provided in Appendix.

Discussion. The proposed multimodal reasoning strategies offer two key advantages: (1) Coordinate-Free and Action-Free Reasoning, enabling robust generalization across diverse visual environments without explicit positional or semantic maps; (2) Explicit Visual Reasoning, simulating mental im-

Method	Validation (Seen)				Validation (Unseen)				Test			
	BL-4 ↑	CI ↑	ME ↑	RO-L ↑	BL-4 ↑	CI ↑	ME ↑	RO-L ↑	BL-4 ↑	CI ↑	ME ↑	RO-L ↑
Speaker-Follower (Fried et al., 2018)	0.10	0.08	0.08	0.12	0.09	0.06	0.07	0.11	0.09	0.06	0.07	0.12
LANA (Wang et al., 2023)	0.05	0.05	0.11	0.10	0.05	0.06	0.10	0.10	0.05	0.03	0.11	0.09
GPT-4o (Hurst et al., 2024)	0.08	0.16	0.19	0.18	0.07	0.16	0.17	0.19	0.07	0.15	0.18	0.18
GPT-4o + CoT (zero-shot CoT)	0.08	0.17	0.17	0.21	0.09	0.16	**0.18**	0.20	0.08	0.17	0.18	0.19
C-Instructor (Kong et al., 2024)	0.21	0.19	0.14	0.23	0.22	0.19	0.14	0.19	0.22	0.18	0.13	0.20
Gemini 2.0 (DeepMind, 2024)	0.08	0.11	0.13	0.12	0.06	0.12	0.15	0.14	0.06	0.10	0.14	0.13
Gemini 3.0 (DeepMind and Research, 2025)	0.09	0.13	0.16	0.13	0.09	0.14	0.15	0.15	0.08	0.12	0.15	0.13
Claude 3 Opus (Anthropic, 2024)	0.07	0.12	0.12	0.12	0.06	0.11	0.13	0.12	0.07	0.11	0.13	0.13
Claude 4 Opus (Anthropic, 2025)	0.10	0.15	0.17	0.14	0.09	0.13	0.16	0.14	0.09	0.14	0.17	0.14
Anole-7B (Direct) (Chern et al., 2024)	0.06	0.10	0.10	0.14	0.06	0.09	0.12	0.13	0.05	0.09	0.10	0.13
Anole-7B + CoT (fine-tuned CoT)	0.10	0.14	0.15	0.17	0.09	0.13	0.12	0.16	0.09	0.10	0.13	0.15
Anole-7B + One-pass (Ours)	0.34	0.20	0.18	0.22	0.29	0.18	0.16	0.20	0.29	**0.19**	0.17	0.18
Anole-7B + Interleaved (Ours)	**0.36**	**0.22**	**0.21**	**0.27**	**0.32**	**0.20**	**0.18**	**0.21**	**0.33**	0.18	**0.20**	**0.22**

Table 1: **Comparison with SOTA Methods** on Goal-Conditioned Visual Navigation Instruction Generation on R2R-Goal validation (seen/unseen) and test splits. (BLEU-4 (BL-4), CIDEr (CI), METEOR (ME), and ROUGE-L (RO-L))

agery in a manner analogous to world model-based forward prediction, providing enhanced transparency and interpretability.

4 Experiments

4.1 Experimental Setup

Evaluation Metrics. We evaluate performance with two sets of metrics: (1) **Instruction Quality.** We measure linguistic accuracy for both the GoViG task and the instruction generation with visual cues subtask using BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). Additionally, we employ navigation-specific metrics Success Rate (SR) and Success weighted by Path Length (SPL) (Anderson et al., 2018) to assess instruction quality from a practical usability perspective. (2) **Visualization Quality.** We assess visual prediction quality for the navigation visualization subtask using structural and perceptual metrics: SSIM (Wang et al., 2004), PSNR (Hore and Ziou, 2010), LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023).

Baselines. We benchmark our method against several leading approaches on R2R-Goal: Speaker-Follower (Fried et al., 2018), LANA (Wang et al., 2023), and C-Instructor (Kong et al., 2024), retrained to accept only egocentric observations consistent with our task setting. We also evaluate SOTA multimodal LLMs including Gemini 2.0 (DeepMind, 2024), Gemini 3.0 (DeepMind and Research, 2025), and Claude 4 Opus (Anthropic, 2025) using direct prompting, as well as GPT-4o (Hurst et al., 2024) using zero-shot direct and CoT prompting (Wei et al., 2022; Yang et al., 2023b), and Anole-7B (Chern et al., 2024) with zero-shot direct prompting and fine-tuned CoT reasoning.

Implementation Details. We fine-tune GAIR Anole-7B (Chern et al., 2024) (4096-token context), freezing both text and image tokenizers. Input images (256×256) are discretized into 784 visual tokens. Only LoRA (Hu et al., 2022) adapters (rank=16) in the transformer's *qkv*-projections are updated during training (Liu et al., 2023). We train for 20 epochs using AdamW (learning rate=2×10^{-4}). Training employs 4×NVIDIA A100 GPUs (80GB), global batch size 8 (per-GPU batch=1, gradient accumulation=2). Please refer to the Appendix for detailed implementation and metrics.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSim ↓
GPT-4o + DALL·E	0.29	9.57	0.72	0.61
Anole-7B (Direct)	0.50	14.98	0.39	0.27
Ours	**0.69**	**20.02**	**0.27**	**0.13**

Table 2: **Navigation Visualization Comparison** on R2R-Goal val unseen. Higher SSIM and PSNR, and lower LPIPS and DreamSim reflect superior visual fidelity.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSim ↓
w/o $\mathcal{L}_{vis}$	0.52	15.35	0.36	0.23
w/ $\mathcal{L}_{vis}$	**0.69**	**20.02**	**0.27**	**0.13**

Table 3: **Ablation** of Token Discrepancy Loss ($\mathcal{L}_{vis}$) on navigation visualization (val unseen), context size = 2.

4.2 Comparison to State-of-the-Art Methods

Goal-conditioned Instruction Generation. Table 1 showcases that our proposed One-pass and Interleaved Multimodal Reasoning strategies outperform all baseline methods in GoViG task. These results demonstrate that incorporating navigation visualization enhances both contextual grounding and linguistic coherence. Notably, interleaved reasoning enables progressive integration of fine-grained visual semantics into instruction generation. Qualitative examples in Fig. 4 further illustrate that both methods generate high-quality, visually-grounded instructions on challenging unseen split.

Navigation Visualization. In Table 2, we compare the performance of our fine-tuned Anole-7B model on the navigation visualization subtask against two baselines: GPT-4o with integrated DALL·E (via the GPT-4o API, where image generation is handled

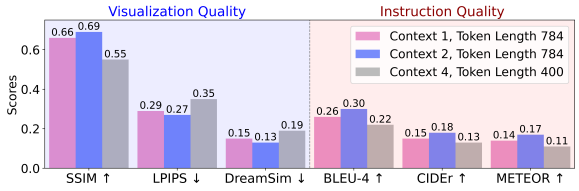


Figure 5: **Trade-off** between context size and image token length on navigation visualization and instruction generation subtasks evaluated on R2R-Goal val unseen split. Context 1 corresponds to context size $k = 1$ for visualization and $m = 2$ for instruction generation. Token Length indicates the number of visual tokens per input or predicted frame.

by the DALL·E module) and Anole-7B with direct prompting. The results show that our approach significantly outperforms both baselines across all evaluation metrics. Notably, our method improves structural fidelity and visual realism (SSIM: 0.69, PSNR: 20.02) in predicted observations.

4.3 Ablation Studies

Impact of Context Size & Image Token Length. We analyze the influence of context size and image token length in Figure 5. Due to Anole-7B's 4096-token constraint, larger contexts (size=4 for visualization, size=5 for instruction generation) necessitate reducing image tokens from 784 to 400 per frame. Results illustrate a clear performance trade-off: moderate context extensions (1→2 frames) enhance temporal coherence and task accuracy, while further expansions at reduced image token length (400 tokens/frame) impair visual fidelity and instruction quality. The results indicate that longer visual histories are effective when each frame retains sufficient token content; otherwise, added context may impede performance.

Effect of Token Discrepancy Loss. Table 3 demonstrates the effectiveness of the token discrepancy loss ($\mathcal{L}_{vis}$). Here, w/o $\mathcal{L}_{vis}$ means using label smoothing loss $\mathcal{L}_{ins}$ instead on image tokens. $\mathcal{L}_{vis}$ substantially improves image quality across all metrics, indicating that explicitly modeling token similarity is crucial for preserving perceptual and structural details in visual predictions.

Computational Efficiency Trade-off. We analyze the computational cost of our two reasoning strategies on the R2R-Goal validation unseen split. One-pass reasoning achieves 1.2× faster inference than Interleaved reasoning. This difference arises from Interleaved's iterative instruction refinement at each step. Besides, this presents a practical trade-off: Interleaved delivers higher accuracy (BLEU-4: 0.32 vs 0.29) with a 20% time cost, suitable for quality-critical applications, while One-pass offers efficient inference when speed is prioritized.

Instruction Generator	ETPNav (An et al., 2024)		BEVBert (An et al., 2023)	
	SR ↑	SPL ↑	SR ↑	SPL ↑
Human Annotation	0.36	0.28	0.34	0.27
LANA	0.18	0.11	0.17	0.12
GPT-4o	0.23	0.16	0.22	0.14
GPT-4o + CoT	0.25	0.17	0.24	0.17
C-Instructor	0.29	0.19	0.27	0.18
Gemini 3.0	0.27	0.18	0.25	0.14
Claude 4 Opus	0.26	0.16	0.25	0.17
Anole-7B + Direct	0.20	0.14	0.18	0.13
Anole-7B + CoT	0.25	0.16	0.23	0.15
Anole-7B + One-pass	0.31	0.20	0.29	0.21
Anole-7B + Interleaved	**0.34**	**0.25**	**0.33**	**0.25**

Table 4: **Instruction Quality Analysis.** Performance of ETPNav (An et al., 2024) and BEVBert (An et al., 2023) in following instructions generated on R2R-Goal val unseen.

Method	BLEU-4 ↑	CIDEr ↑	METEOR ↑	ROUGE-L ↑
LANA	0.05	0.03	0.09	0.09
GPT-4o	0.08	0.11	0.18	0.17
GPT-4o + CoT	0.09	0.13	0.16	0.18
C-Instructor	0.15	0.08	0.12	0.15
Gemini 3.0	0.08	0.11	0.15	0.14
Claude 4 Opus	0.09	0.13	0.16	0.16
Anole-7B + Direct	0.06	0.09	0.10	0.12
Anole-7B + CoT	0.08	0.10	0.13	0.17
Anole-7B + One-pass	0.24	0.14	0.17	0.19
Anole-7B + Interleaved	**0.27**	**0.15**	**0.19**	**0.20**

Table 5: **Zero-shot generalization** on R2R-Goal real-world subset (GO Stanford, ReCon, and HuRoN). All models here are evaluated without additional fine-tuning on this subset.

4.4 Instruction Quality Analysis

While n-gram-based metrics (BLEU-4, CIDEr) provide quantitative assessments of instruction quality, they may not fully capture instruction usefulness or human interpretability. We therefore conduct additional analyzes to evaluate our generated instructions from two complementary perspectives:

Practical Usability. The navigation performance of agents following instructions from different generators serves as an indicator of instruction quality. We regenerate instructions for paths in the R2R-Goal val unseen split and employ two navigators (ETPNav and BEVBert) to evaluate SR and SPL under the regenerated guidance. As shown in Table 4, instructions generated by our methods (one-pass and interleaved) achieve competitive results that exceed those of prior models and closely align with the navigation accuracy obtained using human-annotated instructions.

User Study. To evaluate instruction quality beyond automatic metrics, we recruit 21 anonymous participants from diverse backgrounds to score instructions from 1 to 6 based on semantic alignment with trajectories. We evaluate our two reasoning strategies and SOTA methods on 320 randomly sampled trajectories from R2R-Goal val unseen split, presented in randomized order. Our Interleaved reasoning achieves the highest score of 4.85, followed

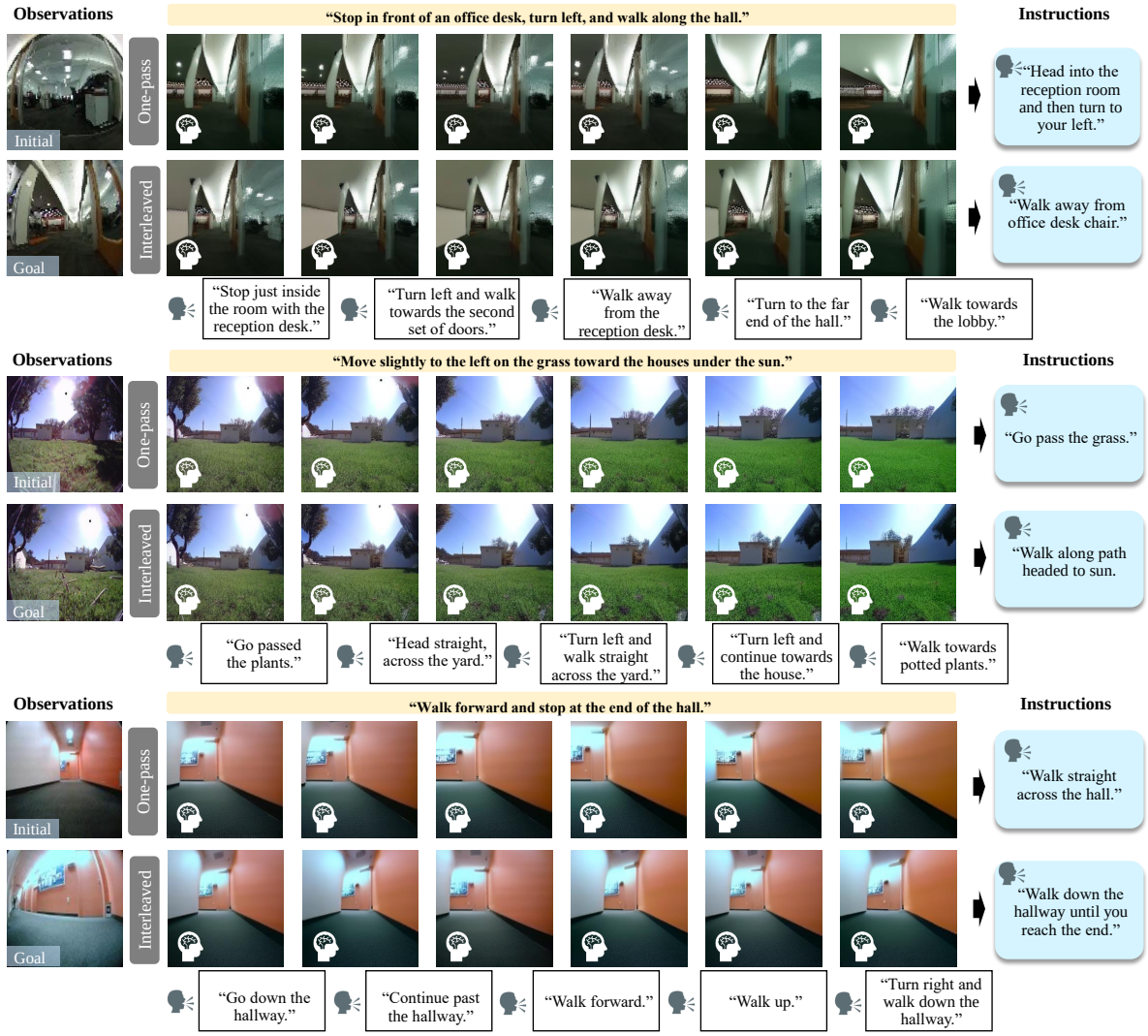


Figure 6: **Qualitative results** on real-world subset of (with ground truth) our One-pass and Interleaved Reasoning.

by One-pass at 4.52, outperforming GPT-4o + CoT (3.76), Gemini 3.0 (3.54), Claude 4 Opus (3.41), C-Instructor (3.08), and LANA (2.67).

4.5 Cross-Domain Generalization

To comprehensively assess cross-domain generalization, we evaluate our method on the real-world subset of R2R-Goal, comprising diverse scenes from GO Stanford, ReCon, and HuRoN. As shown in Table 5, our Interleaved and One-Pass multimodal reasoning strategies notably outperform SOTA approaches under significant domain shifts. Specifically, the Interleaved strategy consistently delivers superior results, underscoring the efficacy of iterative visual-linguistic refinement in improving contextual grounding and instruction coherence. Qualitative examples (Fig. 6) further illustrate how iterative multimodal reasoning mirrors adaptive human cognition, enabling robust instruction generation even in challenging, unseen environments.

5 Conclusion

In this work, we proposed Goal-Conditioned Visual Navigation Instruction Generation (GoViG), a framework to generate precise and context-aware navigation instructions solely from egocentric visual observations, eliminating reliance on privileged data such as maps or semantic annotations. Our approach systematically integrates two interdependent subtasks, Navigation Visualization and Instruction Generation with Visual Cues, into a unified autoregressive MLLM. Furthermore, we developed two multimodal reasoning strategies (One-Pass and Interleaved) that enhance spatial reasoning and linguistic coherence. Comprehensive experiments on our proposed R2R-Goal benchmark demonstrate superior instruction quality and robust cross-domain generalization. Future directions include exploring real-time environmental feedback to advance practical embodied AI.

Limitations

Our approach contributes to advancing research in navigation instruction generation, though several limitations remain. First, achieving the reported performance requires notable computational resources (e.g., 4×A100 GPUs with the Anole-7B base model), which may pose reproducibility challenges for groups with limited hardware. Second, while we expect our multimodal training and reasoning mechanisms to be broadly applicable across architectures and modalities with interleaved multimodal generation, our current evaluation is restricted to Anole-7B due to space considerations. Extending experiments to additional model families would further support claims of generalizability. Finally, although the method could in principle be applied to real-world navigation scenarios, this work is presented as a research study, and deployment in safety-critical contexts would require further validation beyond our current scope.

Responsible NLP Checklist (Filled)

For every item we answer Yes/No and cite a supporting section or justification.

A1 Limitations Section. Yes

A2 Potential Risks. Yes —— Section Limitations.

B Use Or Create Scientific Artifacts. Yes

B1 Cite Creators Of Artifacts. Yes —— Section 3.

B2 Discuss The License For Artifacts. Yes —— Section B.

B3 Artifact Use Consistent With Intended Use. Yes —— Section 3 & B.

B4 Data Contains Personally Identifying Info Or Offensive Content. Yes —— Section B.

B5 Documentation Of Artifacts. Yes —— Section 3 & B.

B6 Statistics For Data. Yes —— Section 3.

C Computational Experiments. Yes

C1 Model Size And Budget Yes —— Section 4.

C2 Experimental Setup And Hyperparameters. Yes —— Section 4.

C3 Descriptive Statistics. Yes —— Section 4.

C4 Parameters For Packages. Yes —— Section C.

D Human Subjects Including Annotators. Yes

D1 Instructions Given To Participants. Yes —— Section B.

D2 Recruitment And Payment. Yes —— Section B.

D3 Data Consent. Yes —— Section B.

D4 Ethics Review Board Approval. N/A.

D5 Characteristics Of Annotators. Yes —— Section B.

E Ai Assistants In Research Or Writing. Yes

E1 Information About Use Of Ai Assistants. Yes —— Section D.

Acknowledgments

This work was supported by the University of Washington Faculty Startup Fund, the Carwein Andrews Fellowship, the UW GSFEI Top Scholar Award, and the U.S. DOT PacTrans sub-center seed funding program. We thank the anonymous reviewers for their helpful comments.

References

Gary L Allen. 1997. From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In *International Conference on Spatial Information Theory*, pages 363–372. Springer.

Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. 2023. Bevbert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2737–2748.

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Preprint.

Anthropic. 2025. Introducing claude 4: Claude opus 4 and claude sonnet 4. https://www.anthropic.com/news/claude-4. Accessed: January 3, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 15619–15629.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. 2025. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. 2023. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yibo Cui, Liang Xie, Yu Zhao, Jiawei Sun, and Erwei Yin. 2025. Generating vision-language navigation instructions incorporated fine-grained alignment annotations. *Preprint*, arXiv:2506.08566.

Google DeepMind. 2024. Introducing gemini 2.0: A new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024. Accessed: January 3, 2026.

Google DeepMind and Google Research. 2025. Gemini 3: Introducing the latest gemini ai model from google. https://blog.google/products/gemini/gemini-3/. Accessed: January 3, 2026.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, and 1 others. 2024. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*.

Yifei Dong, Fengyi Wu, Guangyu Chen, Zhi-Qi Cheng, Qiyu Hu, Yuxuan Zhou, Jingdong Sun, Jun-Yan He, Qi Dai, and Alexander G Hauptmann. 2025a. Unified world models: Memory-augmented planning and foresight for visual navigation. *arXiv preprint arXiv:2510.08713*.

Yifei Dong, Fengyi Wu, Yilong Dai, Lingdong Kong, Guangyu Chen, Xu Zhu, Qiyu Hu, Tianyu Wang, Johnalbert Garnica, Feng Liu, and 1 others. 2026. Language-conditioned world modeling for visual navigation. *arXiv preprint arXiv:2603.26741*.

Yifei Dong, Fengyi Wu, Qi He, Zhi-Qi Cheng, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, and 1 others. 2025b. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. *arXiv preprint arXiv:2503.14229*.

Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025c. Large language model agents in finance: A survey bridging research, practice, and real-world deployment. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:17889–17907.

Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. 2024. Navigation instruction generation with bev perception and large language models. In *European Conference on Computer Vision*, pages 368–387. Springer.

Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. 2025. Scene map-based prompt tuning for navigation instruction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 6898–6908.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. *Preprint*, arXiv:1806.02724.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Google. 2024. Introducing gemini 2.0: our new ai model for the agentic era.

Muraleekrishna Gopinathan, Martin Masek, Jumana Abu-Khalaf, and David Suter. 2024. Spatially-aware speaker for vision-and-language navigation instruction generation. *arXiv preprint arXiv:2409.05583*.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3).

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2022. Mastering atari with discrete world models. *Preprint*, arXiv:2010.02193.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2024. Mastering diverse domains through world models. *Preprint*, arXiv:2301.04104.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2024. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*.

Noriaki Hirose, Amir Sadeghian, Marynel Vázquez, Patrick Goebel, and Silvio Savarese. 2018. Gonet: A semi-supervised deep learning approach for traversability estimation. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3044–3051. IEEE.

Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. 2023. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Alain Hore and Djemel Ziou. 2010. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Alycia M Hund and Jennifer L Minarik. 2006. Getting from here to there: Spatial anxiety, wayfinding strategies, direction type, and wayfinding efficiency. *Spatial cognition and computation*, 6(3):179–201.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. 2024. Dino-foresight: Looking into the future with dino. *arXiv preprint arXiv:2412.11673*.

Xianghao Kong, Jinyu Chen, Wenguan Wang, Hang Su, Xiaolin Hu, Yi Yang, and Si Liu. 2024. *Controllable Navigation Instruction Generation with Chain of Thought Prompting*, page 37–54. Springer Nature Switzerland.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Heng Li, Minghan Li, Zhi-Qi Cheng, Yifei Dong, Yuxuan Zhou, Jun-Yan He, Qi Dai, Teruko Mitamura, and Alexander G Hauptmann. 2024. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems*, 37:119411–119442.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Kevin Lynch. 1964. *The image of the city*. MIT press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. 2021. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Eric J Vanetti and Gary L Allen. 1988. 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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. 2022a. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15450–15460.

Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. 2022b. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15428–15438.

Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. 2023. Lana: A language-capable navigator for instruction following and generation. *Preprint*, arXiv:2303.08409.

Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. 2024. Learning to follow and generate instructions for language-capable navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3334–3350.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. 2025a. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *Preprint*, arXiv:2502.11142.

Zun Wang, Jialu Li, Yicong Hong, Songze Li, Kunchang Li, Shoubin Yu, Yi Wang, Yu Qiao, Yali Wang, Mohit Bansal, and Limin Wang. 2025b. Bootstrapping language-guided navigation learning with self-refining data flywheel. *Preprint*, arXiv:2412.08467.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317.

Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. 2025. Critiques of world models. *Preprint*, arXiv:2507.05169.

Yu Yan, Rongtao Xu, Jiazhao Zhang, Peiyang Li, Xiaodan Liang, and Jianqin Yin. 2024. Instrugen: Automatic instruction generation for vision-and-language navigation via large multimodal models. *Preprint*, arXiv:2411.11394.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023a. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *Preprint*, arXiv:2309.17421.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Haitian Zeng, Xiaohan Wang, Wenguan Wang, and Yi Yang. 2023. Kefa: A knowledge enhanced and fine-grained aligned speaker for navigation instruction generation. *Preprint*, arXiv:2307.13368.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. 2024. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang.

2025a. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10412–10420.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2023. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*.

Yi Zhao, Siqi Wang, and Jing Li. 2025b. Laf-grpo: In-situ navigation instruction generation for the visually impaired via grpo with llm-as-follower reward. *Preprint*, arXiv:2506.04070.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

Appendix

This supplementary material provides expanded details and results that complement the main paper. Section A presents a more detailed analysis of related work and comparisons of navigation instruction generation methods. Section B includes details on the R2R-Goal dataset, pseudo-code for the one-pass and interleaved multimodal reasoning mechanisms, as well as prompt design specifications. Section C reports detailed implementation specifics and additional qualitative results.

A More Related Work

Table 6 categorizes prior work along five orthogonal axes: (i) viewpoint (*ego-centric* vs. *panoramic*); (ii) reliance on privileged inputs (e.g., orientation, GPS, environment labels); (iii) pre-processing pipelines (e.g., landmark vocabularies, BEV encodings, scene graphs); (iv) backbone family (RNN/CNN, Transformer, CLIP/GCN, Vision-Encoder + LLM); and (v) the extent and manner in which LLMs are incorporated. Early "speaker-style" systems, Speaker-Follower (Fried et al., 2018) and CCC-Speaker (Wang et al., 2022a), adopt a non-ego-centric, panoramic observation paradigm with action traces, occasionally augmented by environment labels. These methods typically depend on pre-extracted visual and linguistic features (e.g., ResNet, GloVe) and sequence backbones (CNN/LSTM), without leveraging any large language models. Transformer-based approaches, such as LANA (Wang et al., 2023) and LANA+ (Wang et al., 2024), retain the panoramic setting but incorporate orientation priors and stronger sequence modeling. LANA+ further introduces CLIP-based landmark spotting as an explicit pre-processing signal, improving visual grounding while still assuming privileged panoramic inputs.

Recently, LLM-integrated "instructor" approaches have broadened the modeling toolkit but often at the cost of introducing stronger priors and heavier pre-processing pipelines. C-Instructor (Kong et al., 2024) couples a vision encoder with an LLM and a curated landmark vocabulary, employing Chain-of-Thought prompting to scaffold instruction generation. BEV-Instructor (Fan et al., 2024) moves toward an ego-centric perspective but still depends on multi-view imagery, 3D bounding boxes, and BEV/action-map encodings orchestrated by an MLLM. Retrieval- and map-centric variants—NavRAG (Wang et al., 2025a) and MapInstructor (Fan et al., 2025)—leverage navigable positions, panoramic imagery, GPS, and scene maps to construct hierarchical structures or extract landmarks, then condition an LLM via RAG or map-based prompt tuning.

In contrast, our method operates *exclusively* on ego-centric inputs, free of privileged priors or handcrafted pre-processing. By harnessing the Multimodal LLM *Anole-7B* for unified multi-modal reasoning, it intentionally minimizes task-specific engineering (e.g., curated vocabularies, BEV encodings, retrieval indices) yet preserves strong grounding performance. Our design facilitates practical deployment and promotes robust cross-domain generalization by eliminating dependencies on panoramic sensors, external maps, or GPS signals.

B Methodology Details

B.1 R2R-Goal Dataset Details

To support the GoViG task, we construct the R2R-Goal dataset within the HA-VLN simulation environment (Dong et al., 2025b), using the path start and goal positions provided by the HA-R2R (Dong et al., 2025b) and R2R-CE (Krantz et al., 2020) benchmarks. An A*-based heuristic search identifies the shortest feasible navigation path, with dynamic re-planning triggered in real time upon encountering unexpected obstacles. An egocentric camera mounted on the simulated agent continuously captures observations along each traversed path. Scene-level segmentation is performed in two stages using a frozen Qwen2.5-VL-7B-Instruct model (Bai et al., 2025). First, navigation instructions are segmented into spatially coherent

Method	Ego-centric	Privileged Input	Pre-processed Elements	Backbone	LLM Usage
Speaker-Follower (Fried et al., 2018)	×	Panoramic views, Action history	ResNet features, GloVe embeddings	LSTM-RNN	None
CCC-Speaker (Wang et al., 2022a)	×	Panoramic views, Action, Environment labels	ResNet features	CNN + LSTM	None
LANA (Wang et al., 2023)	×	Panoramic views, Orientation	-	Transformer	None
LANA+ (Wang et al., 2024)	×	Panoramic views, Orientation	Landmark spotting via CLIP	Transformer + CLIP	None
C-Instructor (Kong et al., 2024)	×	Trajectory path, Panoramic views, Action, Orientation	Landmark vocabulary (CoTL)	Vision Encoder + LLM	Chain-of-Thought
BEV-Instructor (Fan et al., 2024)	✗	Multi-view images, Orientation, Action, 3D bounding boxes	BEV encoding, Action map	Vision Encoder + MLLM	Yes
NavRAG (Wang et al., 2025a)	×	Navigable position, Panoramic views, GPS	Hierarchical scene tree	Vision Encoder + LLM	Retrieval-Augmented Generation
MapInstructor (Fan et al., 2025)	×	Panoramic views, Action, Orientation	Landmark extraction via scene map	CLIP + GCN + LLM	Map-based Prompt Tuning
Ours	✓	Not Required	Not Required	Anole-7B	Multi-modal Reasoning

Table 6: Comparison of navigation instruction generation methods. Abbreviations: CoTL = Chain-of-Thought with Landmarks, BEV = Bird's Eye View, GCN = Graph Convolutional Network, MLLM = Multi-modal Large Language Model.

scenes, ensuring each segment corresponds to a navigable space and that all text is uniquely assigned. Post-processing merges consecutive identical scenes and guarantees complete coverage, yielding scene–instruction pairs (e.g., "Kitchen" as an instruction segment). Second, observation frames are aligned with the segmented scenes: the model analyzes the full visual sequence to detect scene transitions based on visual cues and instruction alignment, followed by post-processing to adjust boundaries and eliminate gaps or overlaps.

Annotators were instructed to align their navigation descriptions closely with the visual scenes presented, ensuring that the language reflected the perspective shifts between initial and goal viewpoints. They were encouraged to use diverse expressions when describing actions, environments, and spatial relations among objects, including references to relative positions (e.g., left/right, near/far), motion dynamics (e.g., slow approach, rapid turn), and changes in viewpoint. This emphasis on alignment and variation was intended to capture richer correspondences between instructions and observations, while avoiding repetitive phrasing.

All datasets used in this work (R2R-CE, HA-R2R, GO Stanford, ReCon, HuRoN) are publicly available research datasets released under appropriate licenses. We did not collect any personal data. During dataset construction, we checked that the included visual and textual data do not contain names, faces, or other information that could uniquely identify individuals. Offensive or harmful content was not observed in the sources used. For the real-world subset, only publicly released egocentric videos were used, and all annotations were created to describe navigation trajectories without reference to personal identity. Thus, no anonymization beyond the original dataset release was required.

The annotators were volunteers recruited through university mailing lists and research group announcements. Participation was entirely voluntary, with no monetary compensation provided. Volunteers contributed time to support academic research, and their involvement was recognized in accordance with established ethical guidelines.

B.2 Autoregressive MLLM Training Details

Fig. 7 illustrates the dual training paradigm of our autoregressive multimodal Transformer, which unifies visual and linguistic reasoning based on the Chameleon architecture (Team, 2024). The left panel depicts the **Navigation Visualization Training Phase**, where a structured text prompt (e.g., "Predict the next first-person observation the agent would see if it continues toward the goal.") is tokenized via a BPE tokenizer, and paired with a sequence of k preceding FPV frames plus the goal frame. These visual observations are encoded as `<image>` tokens using a vector-quantized (VQ) image tokenizer. The resulting multimodal prompt is processed by a causal Transformer to predict the next visual token, optimized via the Token Discrepancy Loss $\mathcal{L}_{\text{vis}}$, which compares predicted token distributions $P(t_i)$ against ground-truth embeddings emb_i over the visual codebook $\mathcal{C}$.

The right panel shows the **Instruction Generation Training Phase**, where FPV frames from the initial to goal observation (including $m-1$ intermediate frames) are encoded into an image prompt. This is paired with a text prompt such as "Generate

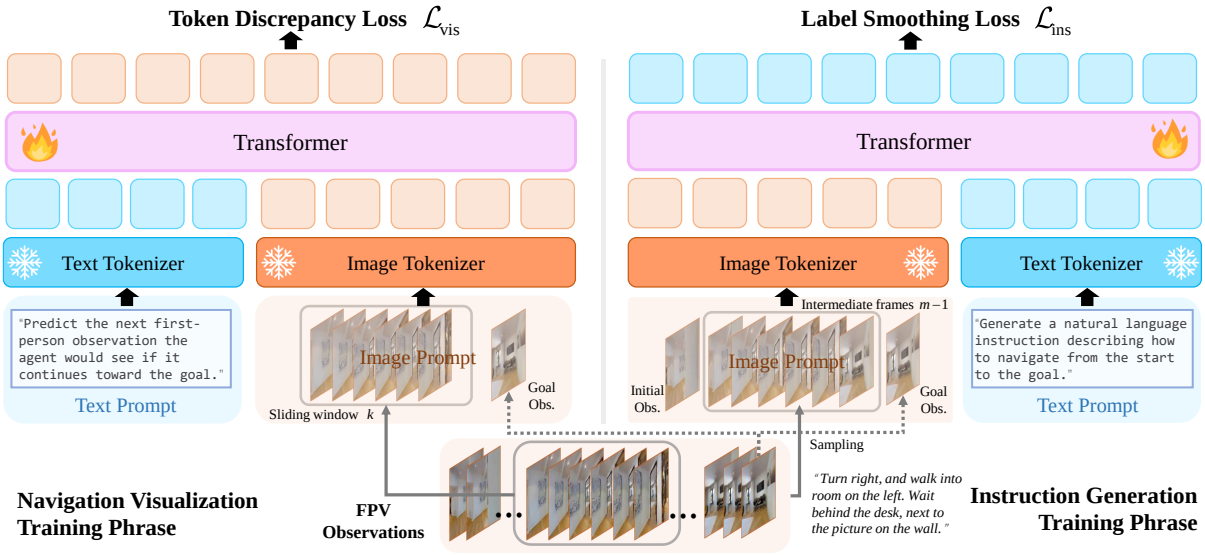


Figure 7: Detailed structure and pipeline of our training procedure.

a natural language instruction describing how to navigate from the start to the goal," and the ground-truth instruction (e.g., "Turn right, and walk into room on the left. Wait behind the desk, next to the picture on the wall."). Both image and text inputs are tokenized and fed into the Transformer, which autoregressively generates the instruction sequence. Training is guided by the label smoothing cross-entropy loss $\mathcal{L}_{\text{ins}}$, computed over the vocabulary $\mathcal{V}$ with smoothed targets $q_v(y_i)$ and predicted probabilities $P_v(y_i)$. Samples from both phases are interleaved during training, enabling joint optimization of visual forecasting and instruction generation within a unified multimodal framework.

B.3 Pseudo-Code of Reasoning Strategies

To generate instructions from an egocentric initial observation and a goal observation, we propose two multimodal reasoning strategies: *One-Pass* and *Interleaved* reasoning, with their pseudo-code provided in Algorithms 1 and 2, respectively. Both strategies employ a frozen multimodal language model F_Θ and iteratively visualize navigation until the predicted frame achieves sufficient visual similarity to the goal observation, measured by an SSIM threshold τ.

One-Pass Multimodal Reasoning (Algorithm 1) first generates the entire future trajectory $\hat{\mathcal{O}} = \{\hat{o}_{k+1}, \ldots, \hat{o}_{k+t}\}$ using a sliding, fixed-size context window, terminating when $\text{SSIM}(\hat{o}_{k+t}, o_g) > \tau$. It then samples $m-1$ representative intermediate frames and produces a final instruction via:

$$I = F_\Theta\big(\{o_1, \hat{o}_{i_1}, \ldots, \hat{o}_{i_{m-1}}, o_g\}\big). \quad (5)$$

Interleaved Multimodal Reasoning (Algorithm 2) alternates between predicting the next visual frame and incrementally refining the instruction. At each step t, the instruction is updated as:

$$I_t = F_\Theta\big(\hat{\mathcal{O}}^{(t+1)} \cup \{o_g, I_{t-1}\}\big), \quad (6)$$

continuing until the SSIM criterion is met. This step-wise refinement allows the agent to progressively incorporate new visual cues, potentially improving instruction grounding in dynamically evolving environments.

Algorithm 1 One-Pass Multimodal Reasoning

Require: Initial observations $\mathcal{O}_{\text{init}} = \{o_1, \ldots, o_k\}$ with visualization context size k; goal observation o_g; SSIM threshold τ; MLLM F_Θ with parameters and tokenizers frozen; instruction context size m

Ensure: Final instruction I

Initialize step $t \leftarrow 1$

Initialize observation context window $\hat{\mathcal{O}}^{(t)} \leftarrow \mathcal{O}_{\text{init}}$

$m = k + 1$

repeat

  $\hat{o}_{k+t} \leftarrow F_\Theta(\hat{\mathcal{O}}^{(t)}, o_g)$

  $\hat{\mathcal{O}}^{(t+1)} \leftarrow \hat{\mathcal{O}}^{(t)}[2{:}] \cup \{\hat{o}_{k+t}\}$

  {Update $\hat{\mathcal{O}}^{(t)}$ by sliding in $\hat{o}_{k+t}$ and keeping most recent k observations}

  $t \leftarrow t + 1$

until $\text{SSIM}(\hat{o}_{k+t}, o_g) > \tau$

Sample $m-1$ intermediate frames $\{\hat{o}_{i_1}, \ldots, \hat{o}_{i_{m-1}}\}$ from $\{o_2, \ldots, o_k, \hat{o}_{k+1}, \ldots, \hat{o}_{k+t}\}$

$I \leftarrow F_\Theta(\{o_1, \hat{o}_{i_1}, \ldots, \hat{o}_{i_{m-1}}, o_g\})$

return I

Algorithm 2 Interleaved Multimodal Reasoning

Require: Initial observations $\mathcal{O}_{\text{init}} = \{o_1, \ldots, o_k\}$ with visualization context size k; goal observation o_g; SSIM threshold τ; MLLM F_Θ with parameters and tokenizers frozen; instruction context size m

Ensure: Final instruction I

Initialize step $t \leftarrow 1$

Initialize observation context window $\hat{\mathcal{O}}^{(t)} \leftarrow \mathcal{O}_{\text{init}}$ {Initial context window}

Initialize instruction $I_0 \leftarrow$ empty string

$m = k + 1$

repeat

$\hat{o}_{k+t} \leftarrow F_\Theta(\hat{\mathcal{O}}^{(t)}, o_g)$ {Predict next observation}

Update $\hat{\mathcal{O}}^{(t+1)} \leftarrow \hat{\mathcal{O}}^{(t)}[2:] \cup \{\hat{o}_{k+t}\}$ {Slide in $\hat{o}_{k+t}$}

$I_t \leftarrow F_\Theta(\hat{\mathcal{O}}^{(t+1)} \cup \{o_g, I_{t-1}\})$ {Update instruction}

$t \leftarrow t + 1$

until $\text{SSIM}(\hat{o}_{k+t}, o_g) > \tau$

$I = I_t$

return I

B.4 Prompt Design Details and Examples

We examine the detailed prompt formulation and response behaviors of two multimodal reasoning strategies—*One-Pass* and *Interleaved*—across two navigation subtasks: *Navigation Visualization* and *Instruction Generation with Visual Cues*. These examples illustrate how multimodal inputs guide both visual prediction and instruction generation in visually grounded navigation.

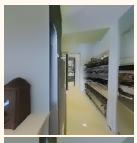
One-Pass Multimodal Reasoning. As shown in Fig. 8, the model begins with an initial observation and iteratively predicts future frames toward the goal, updating the context with each new prediction until the generated frame satisfies the SSIM threshold relative to the goal observation. For instruction generation (Fig. 9), once visual prediction is complete, the model samples key frames—initial, intermediate, and goal—and produces a concise instruction (e.g., "Walk out of the kitchen") summarizing the visual trajectory.

Interleaved Multimodal Reasoning. As shown in Fig. 10, the model conditions each visual prediction on current/past observations, enhancing adaptability in dynamic or ambiguous scenes. For instruction refinement (Fig. 11), it evaluates the previously generated instruction against the updated visual context, revising it accordingly (e.g., "Turn right and continue down the hall until you reach a refrigerator"), thereby maintaining alignment with evolving scene cues.

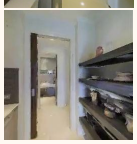
One-pass Multimodal Reasoning for Inference: Navigation Visualization

Input

Task: Navigation Single Step Visualization
Description: Given the previous first-person observation, the current first-person observation, and the goal observation, predict the next first-person observation the agent would see if it continues toward the goal.
Input observations:

Previous obs: 

Current obs: 

Goal obs: 

Response

Predicted obs: 

Figure 8: Prompt design examples on One-pass multimodal reasoning during the inference stage for the Navigation Visualization subtask. (Context size k = 2)

Comparison. One-Pass reasoning prioritizes efficiency and simplicity, whereas Interleaved reasoning offers greater flexibility and robustness. The accompanying figures and outputs highlight how tailored prompt designs can elicit complementary strengths across multimodal navigation tasks.

C Experiments Details

C.1 Evaluation Metrics

We evaluate overall system performance using two complementary categories of metrics:

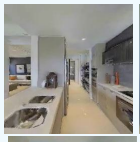
(1) Instruction Quality: Linguistic fidelity is com-

One-pass Multimodal Reasoning for Inference: Instruction Generation with Visual Cues

Input

Task: Scene-level Instruction Generation
Description: Given a sequence of sampled first-person observations along a navigation trajectory, including the starting observation, several intermediate observations, and the goal observation, generate a natural language instruction describing how to navigate from the start to the goal. Focus on essential actions and prominent, easily identifiable landmarks.
Input obs:

Starting obs: 

Intermediate obs:  

Goal obs: 

Response

Predicted Instruction:
Walk out of the kitchen.

Figure 9: Prompt design examples on One-pass multimodal reasoning during inference stage for Instruction Generation with Visual Cues. (Context size $m = 3$)

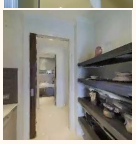
Interleaved Multimodal Reasoning for Inference: Navigation Visualization

Input

Task: Navigation Single Step Visualization
Description: Given the previous first-person observation, the current first-person observation, and the goal observation, predict the next first-person observation the agent would see if it continues toward the goal.
Input observations:

Previous obs: 

Current obs: 

Goal obs: 

Response

Predicted observation: 

Figure 10: Prompt design examples on Interleaved multimodal reasoning during the inference stage for the Navigation Visualization subtask. (Context size $k = 2$)

prehensively assessed for both goal-conditioned and visually grounded instruction generation using widely adopted text-generation metrics: BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). Each generated instruction is compared against the full set of human-authored reference texts to ensure thorough and comprehensive coverage.

(2) Visualization Quality: For the navigation visualization subtask, visual predictions are evaluated with a combination of standard structural and perceptual measures, namely SSIM (Wang et al., 2004), PSNR (Hore and Ziou, 2010), LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023). The latter two are deep perceptual metrics specifically designed to more closely approximate human judgments.

LPIPS: The Learned Perceptual Image Patch Similarity (Zhang et al., 2018) quantifies perceptual resemblance by computing weighted distances between deep feature activations extracted from pretrained vision backbones (e.g., AlexNet, VGG). By operating in a learned feature space, LPIPS better captures perceptually relevant differences than conventional low-level pixel-level measures.

DreamSim: DreamSim extends perceptual evaluation to the multimodal domain by measuring semantic alignment between generated images and a target text description. Given images $\{I_i\}_{i=1}^{N}$ and

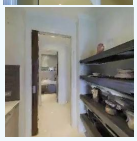
Interleaved Multimodal Reasoning for Inference: Instruction Generation with Visual Cues

Input

Task: Scene-level Instruction Generation
Description: Given a sequence of sampled first-person observations and a previously generated instruction that describes the navigation path — including: the previous instruction (which was originally generated to guide navigation from the start observation to the goal), the current observation (resulting from navigating one step from the previous observation toward the goal), the past observations, and the goal observation. Determine whether the previous instruction needs refinement based on the visual context. Then, generate an updated natural language instruction that accurately guides navigation from the start observation to the goal. Focus on essential actions and prominent, easily identifiable landmarks.
Previous Instruction: Walk and stop right before washing area.
Input obs:

Previous obs: 

Current obs: 

Goal obs: 

Response

Predicted Instruction:
Turn right and continue down the hall until you get to a refrigerator.

Figure 11: Prompt design examples on Interleaved multimodal reasoning during inference stage for Instruction Generation with Visual Cues subtask. (Context size $m = 3$)

a prompt T, it is defined as:

$$\text{DreamSim}(I_{1:N}, T) = \frac{1}{N}\sum_{i=1}^{N}\frac{\langle f_{\text{img}}(I_i),\ f_{\text{text}}(T)\rangle}{\|f_{\text{img}}(I_i)\| \cdot \|f_{\text{text}}(T)\|}. \quad (7)$$

Unlike the standard CLIP score, DreamSim leverages fused or fine-tuned visual–textual features (e.g., CLIP, OpenCLIP, DINO) trained on synthetic human similarity judgments, thereby further enhancing sensitivity to nuanced perceptual and semantic correspondences.

By combining LPIPS and DreamSim, our evaluation jointly accounts for low-level visual fidelity and high-level semantic coherence, offering a balanced and human-aligned assessment across both structural and semantic dimensions.

C.2 Implementation Details for SOTA Methods

In this section, we provide further implementation details on the SOTA navigation instruction generation methods we compare in Table 1 of main text.
Speaker-Follower: The original work (Fried et al., 2018) uses a speaker-follower architecture for vision-and-language navigation, where a follower maps instructions to actions and a speaker generates instructions from routes, enabling data augmentation and pragmatic inference with panoramic action space. We modified it to process sequential egocentric RGB observations, using ResNet-152 to encode input observations ($\mathcal{O}_{full} = \{o_1, \ldots, o_k, o_g\}$) and an LSTM decoder with an attention mechanism for instruction generation. We remove the panoramic action space and adapt the model to work with first-person visual observations only.
LANA: Adapted from (Wang et al., 2023) by extracting its instruction generation module. The original work takes navigation routes (panoramic observations and actions) as input and generates natural language instructions as output, using a unified architecture with shared route/language encoders and cross-attention based decoders for bidirectional translation, jointly trained on both instruction following and generation tasks. We replace the panoramic encoder with ViT-based image encoding. Processes input observations ($\mathcal{O}_{full}$) through cross-attention, removing dependencies on privileged inputs (trajectory coordinates, maps, action labels).
GPT-4o Direct (Zero-shot): Processes input observations ($\mathcal{O}_{full}$) through direct prompting. The model receives explicit instructions that images 1-k represent continuous observations from the starting point along the path, while goal image shows the goal destination. We enforce strict output constraints: (1) no reference to image numbers in the

instruction, as the end user will not have access to these images; (2) pure text output without any markdown formatting, bullet points, or special symbols; (3) single continuous paragraph format; and (4) concise instructions for single-scene navigation. We require concise output because other models and baselines are trained on scene-segmented tasks and naturally produce shorter predictions, while GPT-4o tends to generate longer, more detailed instructions due to the task complexity, which can dilute its true capabilities in certain evaluation metrics. Images are encoded as base64 and resized to a maximum of 512×512 pixels to optimize API usage. The model generates instructions using temperature=0.7 and top_p=0.95 for balanced creativity and coherence.

GPT-4o CoT (Zero-shot): Extends the direct approach with structured chain-of-thought reasoning. The model follows a five-step analysis process: (1) describe the starting position and environment, (2) identify key landmarks and direction changes, (3) describe the path progression, (4) identify the destination, and (5) generate the final navigation instruction. The same output constraints apply as the direct method, with the additional requirement that the final instruction must be prefixed with "FINAL INSTRUCTION:" for automatic extraction. This allows the model to perform detailed visual analysis while ensuring the final output remains concise. The complete reasoning process is preserved for analysis, while only the extracted final instruction is used for evaluation.

C-Instructor: Following (Kong et al., 2024), which takes navigation trajectories with panoramic observations (36 views per step) and actions as input, using Chain-of-Thought with Landmarks (CoTL) to extract critical landmarks before instruction generation and Spatial Topology Modeling Task (STMT) for enhanced spatial understanding. We adapt their method for egocentric observations, using Llama-2-7B with CLIP-ViT-L-14 (36 patches/image) to process input observations ($\mathcal{O}_{full}$). The CoTL mechanism is modified for egocentric views instead of panoramic observations.

Anole-7B Direct (Zero-shot): We employ Anole-7B in a zero-shot setting with sparse observation sampling due to token constraints (4096 tokens total, 1024 tokens per image): $\mathcal{O} = \{o_1, o_k, o_g\}$. As Anole is primarily designed as an image generation model, it requires exceptionally detailed task specifications and comprehensive natural language descriptions to pass the model's regulation. We provide extensive task context explaining: the navigation instruction generation objective, the specific role of each observation (initial position at frame 1, final observation before turning at frame k, and goal destination), and explicit generation requirements for producing clear, actionable instructions. This detailed prompting approach compensates for Anole's architectural expectations without introducing any additional task-specific information beyond what other models receive—the enhancement lies solely in the completeness and clarity of the natural language task description.

Anole-7B CoT (Fine-tuned): We fine-tune Anole-7B using the CoT reasoning approach, wherein the model learns to generate navigation instructions through structured reasoning steps. These include (1) analyzing and describing the visual content of key observations, (2) identifying spatial relationships and environmental changes between the initial and final frames, (3) reasoning about the navigation trajectory from start to goal, and (4) synthesizing these elements into coherent instructions. Unlike the zero-shot setting, the fine-tuned model no longer relies on explicit task specifications or detailed natural language prompts. Through training with ground truth divided instructions, Anole-7B effectively internalizes both the objective and the reasoning patterns required for high-quality instruction generation.

C.3 Implementation Details for SOTA Navigation Visualization Methods

GPT-4o + DALL·E: We implement the two-stage approach for navigation visualization. Given three observations (previous o_{t-1}, current o_t, and goal o_g), GPT-4o analyzes the visual context and generates a text prompt describing the expected next observation $\hat{o}_{t+1}$. This prompt is then passed to DALL·E 2 for image synthesis. The system processes 3 input images and generates 1 output image per prediction.

C.4 Prompt Design and Examples

In this section, we provide prompt examples of our implementation on LLM-related SOTA methods.

GPT-4o Direct Prompt:

```
Task: Navigation Instruction Generation

You are given 7 images from a navigation
    trajectory:
- Images 1-6: Continuous observations
   from the starting point along the
   path
```

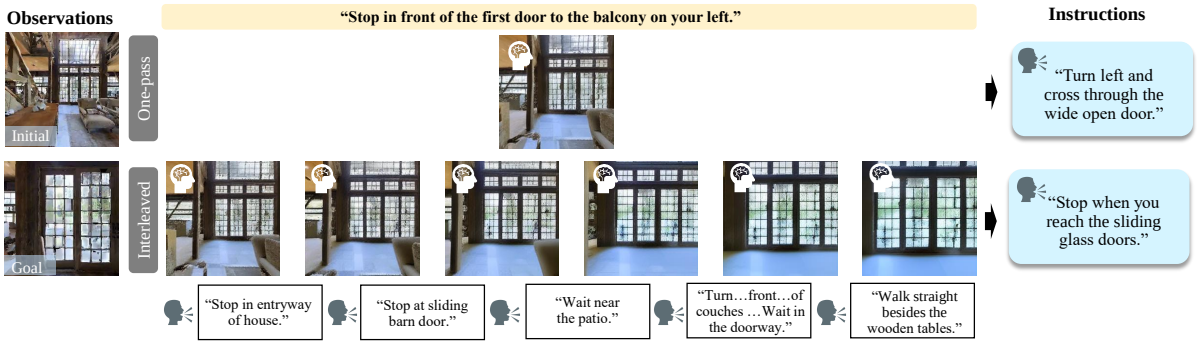


Figure 12: **More qualitative results** on R2R-Goal unseen split (with ground truth) of our One-pass and Interleaved Reasoning.

```
- Image 7: The goal/destination point

Generate a clear navigation instruction
    that guides someone from the
    starting point to the goal.

IMPORTANT REQUIREMENTS:
1. Do not reference image numbers (e.g.,
     'Start at the point shown in Image
    1') in your instruction. The person
    receiving your instruction will not
    have access to these images.
    Describe locations and landmarks
    directly instead.
2. Output ONLY plain text. Do not use
    markdown formatting, bullet points,
    numbered lists, bold text (**text**)
    , headers (#), or any other
    formatting symbols.
3. Write your instruction as a single
    continuous paragraph.
4. Since the navigation target is within
     a single scene, please make your
    instruction more concise.
```

GPT-4o Chain-of-Thought Prompt:

```
Task: Navigation Instruction Generation

You are given 7 images from a navigation
     trajectory:
- Images 1-6: Continuous observations
    from the starting point along the
    path
- Image 7: The goal/destination point

Please analyze step by step:
1. Describe the starting position and
    environment
2. Identify key landmarks and direction
    changes
3. Describe the path progression
4. Identify the destination
5. Generate a clear navigation
    instruction

IMPORTANT REQUIREMENTS:
1. In your final navigation instruction,
     do not reference image numbers (e.g
    ., 'Start at the point shown in
    Image 1'). The person receiving your
     instruction will not have access to
     these images. Describe locations
    and landmarks directly instead.
2. Use ONLY plain text throughout your
    response. Do not use markdown
    formatting, bullet points, numbered
    lists, bold text (**text**), headers
     (#), or any other formatting
    symbols.
3. Write your analysis and final
    instruction as continuous paragraphs
    .
4. For the final navigation instruction
    (step 5), since the navigation
    target is within a single scene,
    please make it more concise.
5. You MUST prefix your final navigation
     instruction with 'FINAL INSTRUCTION
    :' on a new line.
```

C-Instructor Prompt:

```
Based on these 7 navigation images
    showing a path (6 consecutive
    observations + 1 destination),
    analyze the scene step by step:

1. First, identify key objects and
    landmarks in each image:
[IMAGE_TOKEN] [IMAGE_TOKEN] [IMAGE_TOKEN
    ] [IMAGE_TOKEN] [IMAGE_TOKEN] [
    IMAGE_TOKEN] [IMAGE_TOKEN]

2. Next, perceive the spatial
    relationships and transitions
    between consecutive frames:
- How does the viewpoint change from one
     frame to the next?
- What directional movements (forward,
    turn left/right) are implied?
- Which landmarks remain visible across
    multiple frames?

3. Finally, generate a clear and
    complete navigation instruction that
     guides someone from the starting
    point (image 1) to the destination (
    image 7):
```

Anole CoT Finetuned Prompt:

```
Task: Generate navigation instruction
    based on key observations.
You are given three key observations
    from a navigation path:
1. Initial observation at starting point
    : <image>
   Description: [GENERATED_DESCRIPTION_1
       ]
```

```
2. Final observation at starting point (
    before turning): <image>
   Description: [GENERATED_DESCRIPTION_2
       ]
3. Goal observation at destination: <
    image>
   Description: [GENERATED_DESCRIPTION_3
       ]

Based on these observations, analyze
    step-by-step:
1. First, identify the key landmarks and
    spatial layout at the starting
    point.
2. Next, determine the navigation
    direction and movement pattern by
    comparing the initial and final
    observations at the starting point.
3. Then, analyze the goal observation to
    understand the target location and
    its distinguishing features.
4. Finally, synthesize a clear and
    complete navigation instruction that
    guides from the starting point to
    the destination.
Instruction:
```

GPT-4o + DALL·E Navigation Visualization Prompt:

```
Task: Navigation Single Step
    Visualization
Description: Given three observations
    from the previous first-person
    observation, the current first-
    person observation, and the goal
    observation, respectively, predict
    the next first-person observation
    the agent would see if it continues
    toward the goal.
Input observations are Previous
    observation, Current observation,
    and Goal observation
```

C.5 Visual Results

We present further qualitative results to supplement the illustrative examples provided in the main manuscript. As shown in Fig. 12 and Fig. 6, we include complete sequences of observations paired with their corresponding instructions, visualized for both the unseen subset and real-world environments. Compared with ground-truth annotations, the generated instructions reliably capture the majority of salient landmarks and key objects, while producing correct navigational actions. This holds consistently across challenging settings, including unseen scenes and cluttered real-world trajectories, thereby demonstrating the robustness and generalizability of our reasoning strategies.

C.6 Detailed Ablation Results

We analyze the influence of context size and image token length in Figure 5. The detailed values on more metrics are provided in Tables 7 and 8.

Context	Token Len.	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSim ↓
1	784	0.66	19.20	0.29	0.15
2	784	**0.69**	**20.02**	**0.27**	**0.13**
4	400	0.55	16.30	0.35	0.19

Table 7: **Trade-off** between Context Size and Image Token Length on Navigation Visualization (val unseen). Token Length denotes visual token number per frame.

Context	Token Len.	BLEU-4 ↑	CIDEr ↑	METEOR ↑	ROUGE-L ↑
2	784	0.26	0.15	0.14	0.16
3	784	**0.30**	**0.18**	**0.17**	**0.20**
5	400	0.22	0.13	0.11	0.13

Table 8: **Trade-off** between Context Size and Image Token Length on instruction generation with visual cues.(val unseen)

D Use of LLMs

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.