LaF-GRPO: In-Situ Navigation Instruction Generation for the Visually Impaired via GRPO with LLM-as-Follower Reward

Yi Zhao, Siqi Wang, Jing Li*

Department of Computing, The Hong Kong Polytechnic University
{yi-yi-yi.zhao, siqi23.wang}@connect.polyu.hk, jing-amelia.li@polyu.edu.hk

Abstract

Navigation instruction generation for visually impaired (VI) individuals (NIG-VI) is critical yet relatively underexplored. This study focuses on generating precise, in-situ, step-by-step navigation instructions that are practically usable for VI users. Specifically, we propose LaF-GRPO (LLM-as-Follower GRPO), where an LLM simulates VI user responses to navigation instructions, thereby providing feedback rewards to guide the post-training of a Vision-Language Model (VLM). This enhances instruction accuracy and usability while reducing costly real-world data collection needs. To address the scarcity of dedicated benchmarks in this field, we introduce NIG4VI, a 27k-sample open-source dataset to facilitate training and evaluation. It comprises diverse navigation scenarios with accurate spatial coordinates, supporting detailed and open-ended in-situ instruction generation. Experiments on NIG4VI demonstrate the effectiveness of LaF-GRPO through quantitative metrics (e.g., Zero-(LaF-GRPO) boosts BLEU 14%; SFT+(LaF-GRPO) METEOR 0.542 vs. GPT-4o 0.323), and qualitative analysis further confirms that our method yields more intuitive and safer instructions.

Code — https://github.com/YiyiyiZhao/NIG4VI

1 Introduction

The Visually Impaired (VI) community, comprising approximately 2.2 billion (World Health Organization 2019) individuals globally with partial or complete blindness, underscores the significant need for effective assistive technologies. Extensive research in Visually Impaired Assistance (VIA) has emerged to address this need (Zhao et al. 2024; Yuan et al. 2025; Gao et al. 2025; Cai et al. 2024). This paper targets a critical and foundational sub-area: Navigation Instruction Generation for the Visually Impaired (**NIG-VI**). Unlike Navigation Instruction Generation (NIG) for general embodied agents which produces high-level trajectory plans (Dou and Peng 2022; Gopinathan et al. 2024; Fan et al. 2024; Kong et al. 2024), NIG-VI is fundamentally *human-centered*. As illustrated in Figure 1, the task demands in-situ, step-level instructions that (1) incorporate non-visual cues, (2) ensure precise directional and distance guidance, and (3) adapt to obstacles for safety in map coordinates.

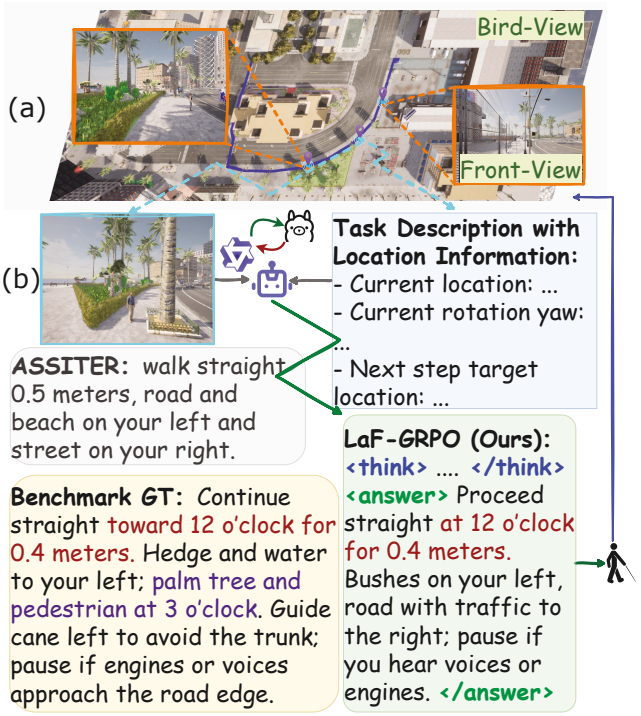

Figure 1: NIG4VI sample. (a) Bird's-eye map and front views. (b) Instructions generated by ASSISTER (Huang et al. 2022), the ground truth, and LaF-GRPO (Ours).

Early attempts, such as ASSISTER (Huang et al. 2022), laid the initial groundwork in this field but were ultimately constrained by the architectural limitations of BERT-based systems (Devlin et al. 2019). The advent of Vision-Language Models (VLMs) has introduced new opportunities due to their multimodal understanding and generation capabilities. Reinforcement Learning (RL) based post-training methods like GRPO (DeepSeek-AI 2025) further enhance reasoning abilities, enabling VLMs to align with human preferences, i.e., the human-centered guidance demanded by NIG-VI. However, this alignment process requires collecting large-scale human feedback data for fine-tuning, which can be costly and often fails to incorporate the interactive user feedback essential for achieving human-centered guidance.

*Corresponding author.

To bridge this gap, we propose **LLM-as-Follower GRPO (LaF-GRPO)**, a novel GRPO-based framework for the NIG-VI task. It features: (1) an LLM that simulates how VI users respond to navigation instructions by interpreting their likely actions, and (2) a VLM post-training procedure for instruction generation guided by an LLM-as-Follower reward. LaF-GRPO mitigates the need for costly trials with VI users while ensuring instruction usability through a *human-in-the-loop* navigation simulation. Furthermore, to address the scarcity of VI navigation benchmarks, we introduce **NIG4VI**—a comprehensive VI navigation instruction benchmark featuring 27k samples. Fully open-sourced and annotated with granular spatial metadata, NIG4VI enables the generation of detailed, open-ended, in-situ instructions.

Evaluation of LaF-GRPO on the NIG4VI benchmark yields three key findings: (1) Qwen2.5-VL models under the **Zero-(LaF-GRPO)** paradigm significantly outperform the standard zero-shot baseline across multiple metrics. (2) Qwen2.5-VL-7B models trained with **SFT+(LaF-GRPO)** achieve state-of-the-art performance, reaching a METEOR score of 0.542 and substantially surpassing strong proprietary models such as GPT-4o. (3) Beyond quantitative gains, qualitative analysis shows that LaF-GRPO produces more human-centered instructions, characterized by greater linguistic diversity, more intuitive directional cues, richer environmental details, and essential safety considerations. In summary, our main contributions are:

- The LaF-GRPO framework, the first to employ GRPO for NIG-VI with a LLM-simulated follower feedback.
- The NIG4VI benchmark, the first open-source comprehensive dataset with precise multi-modal navigation contexts for robust model evaluation for VI navigation.
- Extensive empirical studies across VLMs under various paradigms (Zero-shot, Zero-(LaF-GRPO), SFT, and SFT+(LaF-GRPO)), validating the method's effectiveness.

2 Related Work

VLMs (Liu et al. 2023; Dai et al. 2023; OpenAI 2024a; Anthropic 2024; Team 2024) have gained attention for combining visual perception with language generation. Refining VLMs with Reinforcement Learning (Ouyang et al. 2022) improves alignment with human preferences. Recent success in Group Relative Policy Optimization (GRPO) (DeepSeek-AI 2025) has led to RL fine-tuned VLMs like AlphaDrive (Jiang et al. 2025), VLM-R1 (Shen et al. 2025), Praxis-VLM (Hu et al. 2025), and MedVLM-R1 (Pan et al. 2025), broadening their applications.

Visually Impaired Assistance (VIA) is a broad and diverse field (Gao et al. 2025; Cai et al. 2024; Guan, Xiong, and Fan 2024; Li et al. 2024). VIA with VLMs is closely related to visual captioning and Visual Question Answering (VQA). VIALM (Zhao et al. 2024) frames VIA as a VQA task, generating guidance from environment images and user requests. While VIALM emphasizes environment-grounded guidance with tactile information, it is not specifically designed for navigation. WalkVLM (Yuan et al. 2025) extends this to dynamic walking assistance and introduces the Walking Awareness Dataset (WAD). Though WalkVLM tackles navigation, its focus remains on video captioning rather than precise orientation and mobility guidance.

There are two main branches for NIG studies: NIG for embodied agents and NIG for the visually impaired. Prior research on **Navigation Instruction Generation (NIG) for embodied agents** has primarily focused on visual processing while generating trajectory-level instructions. BEVInstructor (Fan et al. 2024) employs a Bird's-Eye View encoder. SAS (Gopinathan et al. 2024) uses semantic knowledge with adversarial reward learning. C-Instructor (Kong et al. 2024) focuses on style-controlled instruction generation. Our proposed approach differs in two significant ways: (1) it prioritizes navigation feedback for VLM fine-tuning; and (2) it generates step-level in-situ instructions.

In the **NIG-VI** field, ASSISTER (Huang et al. 2022) introduced the UrbanWalk benchmark and developed a navigation assistance model. Our work offers two improvements: (1) we introduce a more detailed evaluation benchmark covering orientation, mobility, scene description, and safety warnings, informed by formative studies from the Human–Computer Interaction field that investigate the needs of VI users (Merchant et al. 2024; Zhao et al. 2025; Chang, Liu, and Guo 2024); and (2) we leverage advanced VLMs within a GRPO framework, leading to more effective instructions.

3 Methodology

Grounded in the **Speaker-Follower paradigm** (Fried et al. 2018), the rationale for LaF-GRPO is to apply **Theory-of-Mind (ToM)** principles (Zhao, Nguyen, and III 2023). Our LLM-as-Follower simulates a user's cognitive mapping, generating feedback from their anticipated interpretation.

NIG-VI Task Formulation

In the NIG-VI task, a VLM-based assistant generates in-situ, step-by-step natural language instructions to guide a visually impaired (VI) user along a pre-planned route $P = [p_1, \ldots, p_K]$. The route P consists of positional waypoints p_i leading to a destination and is generated using the A* algorithm (Huang et al. 2022). At each discrete step i of the navigation, the VLM receives two primary inputs: a front-view camera image $x^{(i)}_{\text{image}}$ and a task description which includes the user's current pose $x^{(i)}_{\text{pose}} = (x^{(i)}_{\text{loc}}, x^{(i)}_{\text{rot}})$ represented by their location $x^{(i)}_{\text{loc}} \in \mathbb{R}^3$ and rotation $x^{(i)}_{\text{rot}} \in \mathbb{R}^3$ within a global map coordinate system, as well as the next target waypoint $p_{i+1} \in P$. Based on these inputs, the VLM π generates a sequence of tokens $y = y^{(i)} = (y^{(i)}_1, y^{(i)}_2, \ldots, y^{(i)}_t)$ of token length t. The generated instruction y might also include details about the current surroundings captured in $x^{(i)}_{\text{image}}$ and any necessary safety alerts:

$$y_j \sim \pi_\theta(y^{(i)}_j | x^{(i)}_{\text{image}}, x^{(i)}_{\text{loc}}, x^{(i)}_{\text{rot}}, p_{i+1}, y^{(i)}_{<j}) \quad (1)$$

where θ denotes the adjustable model parameters.

The LaF-GRPO Framework

We propose LaF-GRPO to address the challenges of generating navigation instructions that are *human-centered and*

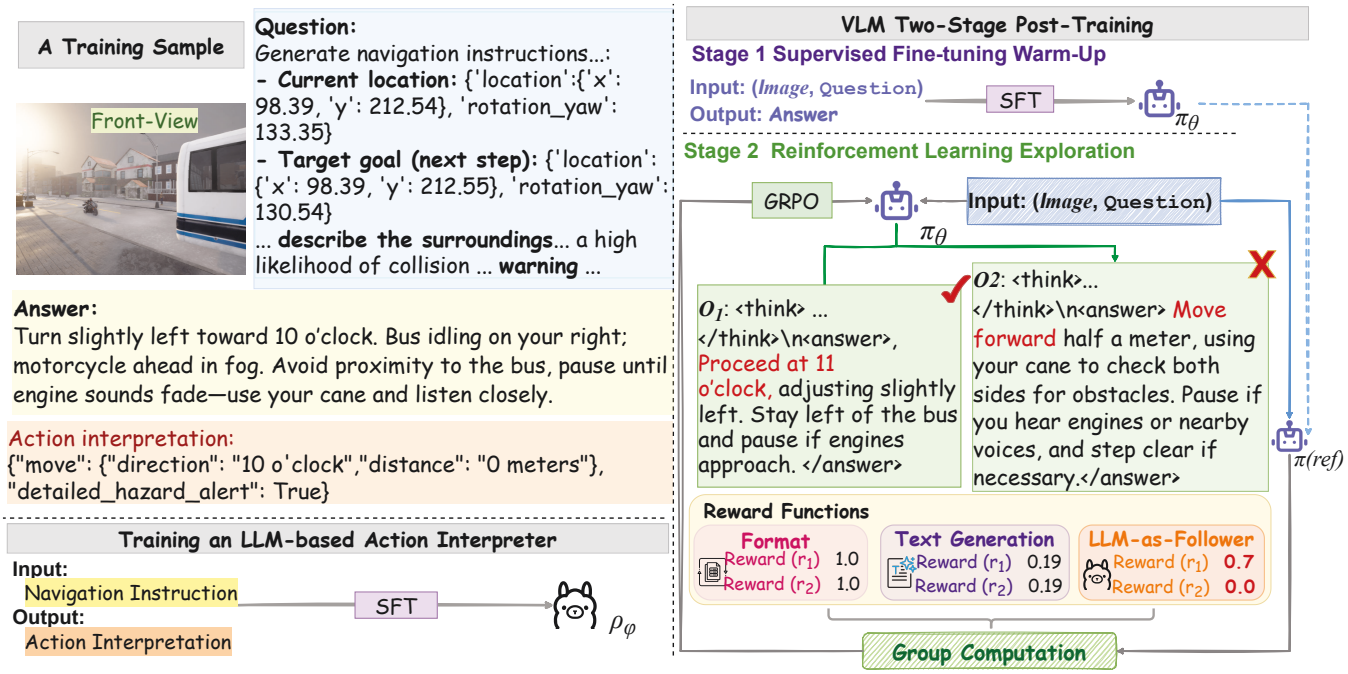

Figure 2: Method Overview. Top left: An training sample with input, target output, and action interpretation. Bottom left: Action interpreter training using `LLaMA-3-8B-Instruct` to simulate VI users' navigation responses. Right: Post-training for VLMs with LaF-GRPO. The LaF reward differentiates outputs when Format and Text Generation rewards are identical.

practically usable for VI users, while *mitigating the need for costly real-world data collection with VI participants*. The overview of LaF-GRPO is illustrated in Figure 2, where the framework comprises two key components: (1) an **action interpreter** LLM (without a visual encoder to *'see'*) that simulates VI users' responses to navigation instructions by interpreting how these users would act upon hearing the instructions, and (2) a VLM post-training procedure that generates these instructions with (1)'s feedback.

Action Interpreter. To simulate a VI user's response to navigation instructions, we use Supervised Fine-tuning (SFT) to train an LLM ρ_φ to predict potential actions. Prior to its deployment, the interpreter is validated to have a parse accuracy above 98% on a held-out set, ensuring its reliability for generating reward signals. Given VLM-generated instruction tokens y, it produces a structured action interpretation $\mathcal{A}$. Formally, we define $\mathcal{A}$ as a structured dictionary containing: (1) a *'move'* action with associated *'direction'* (indicated using clock positions (Yuan et al. 2025)) and *'distance'* parameters, and (2) a *'detailed_hazard_alert'* boolean flag that indicates whether the user perceives warnings about nearby obstacles, as illustrated in Figure 2 Left.

Navigation Instruction Generator. For VI guidance, we use a pre-trained VLM π with parameters θ for in-situ navigation instruction generation. The training of this generator involves two stages: Supervised Fine-tuning (SFT) and Group Relative Policy Optimization (GRPO). Our proposed LaF-GRPO framework enhances standard GRPO (see below) by incorporating a novel LLM-as-Follower reward.

GRPO. The training process of GRPO aims to optimize the policy π_θ by maximizing the objective function $\mathcal{J}_{\text{GRPO}}(\theta)$. For a given query q, GRPO first samples a batch of G outputs $\{o_1, o_2, \ldots, o_G\}$ using an older version of the policy, $\pi_{\theta_{\text{old}}}$. The training process of GRPO aims to optimize the policy π_θ by maximizing the objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \mathcal{L}_i - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right] \tag{2}$$

Here, the term $\mathcal{L}_i$ represents the clipped surrogate objective used in PPO (Schulman et al. 2017):

$$\mathcal{L}_i = \min(w_i A_i, \text{clip}(w_i, 1-\epsilon, 1+\epsilon) A_i) \tag{3}$$

where $w_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the importance sampling ratio, A_i is the estimated advantage for the output o_i, based on relative rewards of the outputs inside each group only, calculated as $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$, and ϵ is a clipping hyperparameter. The second term, $-\beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}})$, regularizes the policy by penalizing divergence from a reference policy π_{ref} with coefficient β. This regularization stabilizes training by keeping the model close to the original effective policy, preventing it from losing previously learned capabilities.

LaF-GRPO Reward Functions

LaF-GRPO utilizes three reward functions as follows.

Format Reward. To encourage controllable generation, we adopt this binary reward ($r_{\text{format}} \in \{0, 1\}$) that evaluates structural compliance with the expected response format. Here, the reward would equal 1 if the output follows the

required format pattern `'<think>.*?</think>\n<answer>.*?</answer>'` in sequence, and 0 otherwise.

Text Generation Reward. The text generation reward (r_{meteor}) is calculated as the METEOR score between the output and the ground-truth reference. METEOR is selected based on its evaluation of semantic overlap, incorporating synonymy and stemming to provide a nuanced assessment.

LLM-as-Follower Reward. The LLM-as-Follower reward (r_{LaF}) assesses the navigational quality of generated instructions by comparing their interpreted actions (*move direction*, *distance*, and *alert flag*) with those of a ground-truth reference. The rationale behind this design is that spatial factors, such as directional accuracy (a_{dir}) and movement distance precision (a_{dist}), play a direct and critical role in determining navigation success. In addition, safety alert flags (a_{alert}) serve as supplementary support for VI navigation by indicating potential hazards, though they are not primary determinants of success (Giudice and Legge 2008; Younis et al. 2019). Accordingly, we formulate the reward as:

$$r_{LaF} = w_{dir}\,\delta(a_{dir}, a_{dir}^{ref}) + w_{dist}\,\delta(a_{dist}, a_{dist}^{ref}) + w_{alert}\,\delta(a_{alert}, a_{alert}^{ref}) \quad (4)$$

$\delta(\cdot)$ denotes an exact match. r_{LaF} is in the range [0, 1].

4 Benchmark: NIG4VI

We introduce the NIG4VI benchmark to address the scarcity of benchmark resources in this field. Inspired by UrbanWalk, NIG4VI utilizes the open-sourced CARLA Simulator (Dosovitskiy et al. 2017) to collect samples from diverse scenarios, including varied environments and weather conditions. Pedestrian trajectories are generated using A* algorithm, with precise geospatial coordinates, orientation, frontal-view images, and semantic segmentation images recorded at each step. NIG4VI offers two advantages: (1) its use of a realistic coordinate system facilitates easier transfer to real-world GPS applications, and (2) it enables the cost-effective generation of accurate and extensive data. Table 1 details NIG4VI's advantages over other datasets.

Dataset Construction. Each question's input includes the user's current location/rotation, the next step's location/orientation, and visual scene data. This information is then structured within a prompt template that includes a detailed task description. The synthesis of the output is a multi-stage process involving both advanced reasoning models and human annotation, as illustrated in Figure 3. Initially, several leading VLMs, specifically GPT-4o, Claude-3.5, and Gemini-2, generate predictions. Following a modality bridging approach, similar to that employed in Vision-R1 (Huang et al. 2025), these outputs are processed through DeepSeek-R1 to enhance blindness-oriented spatial guidance and navigability. Crucially, all instructions undergo rigorous human verification. This task is carried out by two annotators, both proficient in English and holding at least an undergraduate-level education, following a similar practice in (Zhao et al. 2024). The verification involves a two-stage process: first, one annotator performs initial content adjustments, adhering to task requirements. Subsequently, the second annotator independently reviews and verifies this work. Throughout this process, the primary verification criteria include: (1) elimination of visual references (e.g., color-based descriptors), (2) validation of non-visual landmarks, and (3) confirmation of metric precision for mobility-critical parameters.

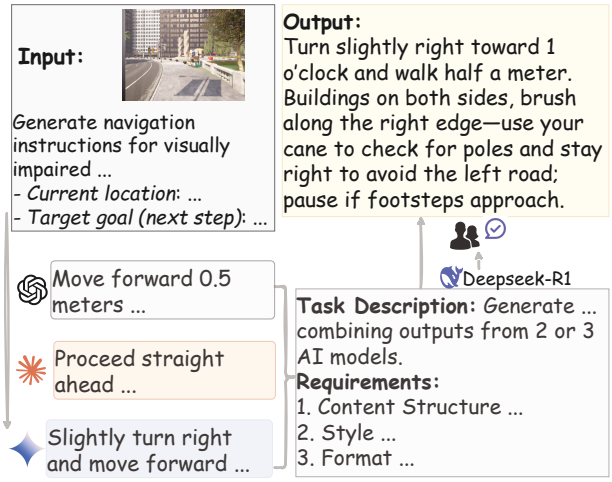

Figure 3: Dataset instances are first generated using Vision-R1's modality bridging method (Huang et al. 2025) and then reviewed and refined by human annotators.

Dataset Statistics. On average, each town contributes approximately 26.2 navigation routes. The average Euclidean distance between the start and end points of these routes is 113.51 units, with an average of 353.8 steps. After deduplication, the dataset yielded an average of 2,222.7 step-level (image, question) samples per town. It is partitioned into a training set of 1,500 samples from Town01 and a test set. The test set comprises the remaining 613 ***intra-town*** samples from Town01, along with all ***inter-town*** samples from Town02 (2,579), Town03 (2,260), Town04 (2,316), Town05 (1,935), and Town10 (2,133). Each data sample is available in two versions: *'with pre-calculation'* and *'without pre-calculation'*. The *'without pre-calculation'* version requires the VLM to independently calculate navigational parameters (e.g., distance, direction), presenting a greater challenge in guidance generation. Conversely, the *'with pre-calculation'* version provides the VLM with basic mathematical movement information. The VLM must validate this data and assess the surroundings to generate the navigation instruction.

5 Experimental Settings

Dataset. Experiments utilized the NIG4VI dataset, with Intra-town ($N = 613$) and Inter-town ($N = 11,223$) test subsets, under *'with/without pre-calculation'* conditions.

Models. Diverse VLMs were evaluated, falling into two main groups: (1) remote models, such as `GPT-4o` (OpenAI 2024b), `Claude-3-5-sonnet-20240620` (Anthropic 2024), and `Gemini-2.0-flash-thinking-exp-01-21` (Google DeepMind 2024); and (2) smaller, locally runnable VLMs, including `DeepSeek-VL-7B` (Lu et al. 2024), `LLaVA-v1.6-Mistral-7B` (Liu et al. 2024), `MiniCPM-o-2.6-8B` (Yao et al.

Benchmark	Level	# Samples	VIA	NIG	Spatial Acc.	Open-ended	Open-sourced
R2R (Anderson et al. 2018)	High	21k	✗	✓	✗	✓	✓
REVERIE (Qi et al. 2020)	High	10k / 6k	✗	✓	✗	✓	✓
UrbanWalk (Huang et al. 2022)	Detailed	2.6k	✓	✓	✓	✗	✗
Merchant et al. (2024)	Detailed	48	✓	✓	✗	✓	✗
VIALM (Zhao et al. 2024)	Detailed	200	✓	✗	✗	✓	✓
WAD (Yuan et al. 2025)	Detailed	12k / 120k	✓	✓	✗	✓	✓
NIG4VI (Ours)	Detailed	3k / 24k	✓	✓	✓	✓	✓
- w/o pre-calculation	Detailed	1.5k / 12k	✓	✓	✓	✓	✓
- with pre-calculation	Detailed	1.5k / 12k	✓	✓	✓	✓	✓

Table 1: Comparison of NIG4VI with existing benchmarks.

2024), `Intern-VL-2.5-8B` (Chen et al. 2024), and `Qwen2.5-VL-3B/7B` (Bai et al. 2025).

Baselines. We compare LaF-GRPO against two primary baseline methods: (1) **Zero-shot**: Models are applied directly to NIG4VI without prior fine-tuning. (2) **Supervised Fine-tuning (SFT)**: Models are fine-tuned to generate instructions. We implement two variants of LaF-GRPO: (a) **Zero-(LaF-GRPO)**: LaF-GRPO is applied directly to the base model without SFT. (b) **SFT+(LaF-GRPO)**: LaF-GRPO is applied to models that have first undergone SFT.

Evaluation Metrics. Following previous studies in NIG (Huang et al. 2022; Fan et al. 2024; Kong et al. 2024), model performance was evaluated using a suite of widely adopted metrics: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), and SPICE (Anderson et al. 2016). Human studies and LLM-as-Judge evaluations were conducted to assess navigational accuracy and user preference for instruction clarity and helpfulness.

Implementation Details. LaF-GRPO training utilized a single NVIDIA H20 GPU (96 GB of memory). This hardware supports loading an 8B-param LLM (`LLaMA-3-8B`) and a 3B/7B-param Qwen2.5-VL model for LoRA (Hu et al. 2022) fine-tuning. The reward weights were configured as $(w_{dir}, w_{dist}, w_{alert}) = (0.4, 0.4, 0.2)$ based on analysis of navigation failure factors, prioritizing spatial parameters over contextual alerts. Training on $3k$ samples took about 15 hours, with the hyperparameter group size G set to 8.

6 Results and Discussions

Main Results

Table 2 summarizes model performance on NIG4VI, categorized by pre-calculation and training paradigms, and evaluated on intra-town and inter-town subsets. Comparing LaF-GRPO with the baselines reveals: (1) **Zero-Shot vs. Zero-(LaF-GRPO):** Zero-(LaF-GRPO) significantly enhances the Zero-Shot performance of VLMs, validating the effectiveness of LaF-GRPO. While the Zero-(LaF-GRPO) results suggest that increased model size (from 3B to 7B) does not necessarily guarantee improved performance across all metrics, it is noteworthy that for METEOR evaluations, specifically in intra-town scenarios, the 7B model achieved the highest scores (i.e., 0.256 and 0.281). This outcome may be attributable to the use of METEOR as a text generation reward during training and to the potentially more refined tuning applied to the 7B models. (2) **SFT & SFT+(LaF-GRPO):** SFT and SFT+(LaF-GRPO) yield significantly superior performance compared to Zero-Shot and Zero-(LaF-GRPO) models across all metrics and subsets, affirming the efficacy of fine-tuning. The SFT+(LaF-GRPO) approach further enhances performance beyond SFT. Moreover, under the SFT+(LaF-GRPO) paradigm, Qwen-VL-3B consistently achieves the highest BLEU and ROUGE scores, while Qwen-VL-7B excels in METEOR and SPICE. This performance pattern is observed for both intra-town and inter-town subsets and holds true regardless of pre-calculation. This may be attributable to 7B models demonstrating enhanced linguistic diversity in their outputs relative to 3B models. (3) **Additional Observations:** Scores improved with pre-calculation, which reduced computational difficulty, whereas the more challenging 'w/o pre-calculation' setting caused more failures. Intra-town scores were higher than inter-town, likely due to a closer data distribution with the training set. We also observed that LaF-GRPO generated significantly more concise, user-friendly instructions (e.g., LaF-GRPO-7B: 34.1 tokens vs. GPT-4o: 117.9 tokens).

Ablation Studies

Reward Types. Table 3 presents an ablation study investigating the impact of different reward types during SFT+(LaF-GRPO) training with the Qwen-VL-3B model. LaF-GRPO, incorporating the LLM-as-Follower reward, consistently achieves the highest BLEU, ROUGE, and METEOR scores. This trend holds true across both intra-town and inter-town evaluations, with or without pre-calculation. This underscores the significant benefit of the LaF reward.

Training Sample Sizes. Table 4 presents an ablation study on the Qwen2.5-VL-7B model trained with SFT+(LaF-GRPO), illustrating the effect of varying training sample sizes ($1k$, $2k$, and $3k$). For comprehensive metrics such as METEOR and SPICE, performance generally scales with the volume of training data. Across the majority of evaluated conditions, scaling up to $3k$ samples typically yields the optimal or near-optimal scores. Nevertheless, training with $2k$ samples also achieves comparable METEOR scores,

Pre-Cal.	Paradigm	Model	Intra-town (N = 613)				Inter-town (N =11,223)			
			BLEU ↑	ROUGE ↑	METEOR ↑	SPICE ↑	BLEU ↑	ROUGE ↑	METEOR ↑	SPICE ↑
No	**Zero-Shot**	DeepSeek-VL-7B	2.179	0.152	0.182	0.116	2.223	0.157	0.196	0.112
		MiniCPM-o-8B	2.009	0.145	0.234	0.131	1.969	0.142	0.233	0.129
		Intern-VL-8B	1.448	0.150	0.215	0.126	1.517	0.149	0.216	0.120
		LLaVA-7B	1.021	0.103	0.201	0.111	1.037	0.107	0.206	0.109
		Qwen-VL-7B	3.204	0.202	0.211	0.166	3.128	0.194	0.210	0.157
		GPT-4o	1.748	0.169	0.249	0.149	1.617	0.165	0.249	0.142
		Claude-3.5	2.803	0.216	0.304	0.211	2.749	0.211	0.301	0.202
		Gemin-2	4.105	0.236	0.232	0.232	4.422	0.252	0.238	0.236
	Zero-(LaF-GRPO)	Qwen-VL-3B	3.292	0.230	0.248	0.230	3.972	0.255	0.259	0.244
		Qwen-VL-7B	3.272	0.234	0.256	0.222	3.566	0.252	0.260	0.227
	SFT	Qwen-VL-3B	9.099	0.282	0.496	0.274	8.949	0.284	0.500	0.276
		Qwen-VL-7B	9.937	0.291	0.518	0.275	9.709	0.294	0.526	**0.281**
	SFT+ (LaF-GRPO)	Qwen-VL-3B	**10.921**	**0.323**	0.528	0.274	**10.157**	**0.309**	0.527	0.276
		Qwen-VL-7B	10.037	0.284	**0.545**	**0.283**	9.002	0.276	**0.535**	0.278
Yes	**Zero-Shot**	DeepSeek-VL-7B	2.517	0.170	0.224	0.161	2.600	0.173	0.237	0.161
		MiniCPM-o-8B	2.349	0.166	0.210	0.136	2.517	0.177	0.220	0.144
		Intern-VL-8B	1.496	0.132	0.233	0.133	1.517	0.134	0.238	0.132
		LLaVA-7B	1.284	0.120	0.222	0.131	1.285	0.121	0.229	0.127
		Qwen-VL-7B	2.903	0.188	0.231	0.178	3.080	0.194	0.243	0.180
		GPT-4o	2.766	0.204	0.302	0.198	2.967	0.213	0.323	0.211
		Claude-3.5	4.124	0.236	0.349	0.257	3.400	0.214	0.326	0.224
		Gemin-2	5.132	0.252	0.266	0.269	6.144	0.276	0.283	0.284
	Zero-(LaF-GRPO)	Qwen-VL-3B	3.798	0.249	0.280	0.261	4.584	0.271	0.288	0.274
		Qwen-VL-7B	3.678	0.241	0.281	0.229	4.284	0.262	0.286	0.230
	SFT	Qwen-VL-3B	9.923	0.308	0.512	0.280	10.724	0.318	0.519	0.280
		Qwen-VL-7B	9.639	0.270	0.521	0.283	9.710	0.272	0.524	0.287
	SFT+ (LaF-GRPO)	Qwen-VL-3B	**11.727**	**0.342**	0.541	0.286	**10.813**	**0.333**	0.535	0.279
		Qwen-VL-7B	10.499	0.285	**0.556**	**0.292**	9.232	0.275	**0.542**	**0.288**

Table 2: Evaluation results on NIG4VI across Intra-town and Inter-town subsets. **Bold** values represent the highest score for each metric under a specific setting (with / without pre-calculation), while underlined values indicate the second-highest score. Boxed values highlight the best performing Qwen2.5-VL model within the Zero-Shot and Zero-(LaF-GRPO) categories.

Pre-Cal.	Reward Types			Intra-town (N = 613)				Inter-town (N = 11,223)			
	Format	**Meteor**	**LLM**	BLEU ↑	ROUGE ↑	METEOR ↑	SPICE ↑	BLEU ↑	ROUGE ↑	METEOR ↑	SPICE ↑
No	✓			10.251	0.318	0.524	0.278	9.401	0.304	0.523	0.279
	✓	✓		10.912	0.317	0.525	**0.279**	10.076	0.306	0.521	**0.279**
	✓	✓	✓	**10.921**	**0.323**	**0.528**	0.274	**10.157**	**0.309**	**0.527**	0.276
Yes	✓			11.269	0.337	0.538	**0.292**	10.217	0.328	0.530	**0.282**
	✓	✓		11.602	0.339	0.539	0.284	10.753	0.331	0.531	0.280
	✓	✓	✓	**11.727**	**0.342**	**0.541**	0.286	**10.813**	**0.333**	**0.535**	0.280

Table 3: Ablation study results for the Qwen2.5-VL-3B model on the NIG4VI dataset with different **reward functions**.

indicating training data efficiency at this sample size.

LaF-GRPO vs. Standard GRPO. We conducted **LLM-as-Judge** and **human evaluations** on the inter-town subset. The LLM-as-Judge evaluation with GPT-4o showed LaF-GRPO's superior navigational accuracy (68.1% vs. 67.3%) and a greater preference rate for instruction clarity and helpfulness (58.3% vs. 41.7%). These findings were corroborated by a human preference study on 100 instruction pairs, evaluated by two trained graduate students acting as VI user simulators. To isolate instruction quality, the evaluators were instructed to navigate relying solely on text and ignoring visual cues. Their assessments achieved substan-

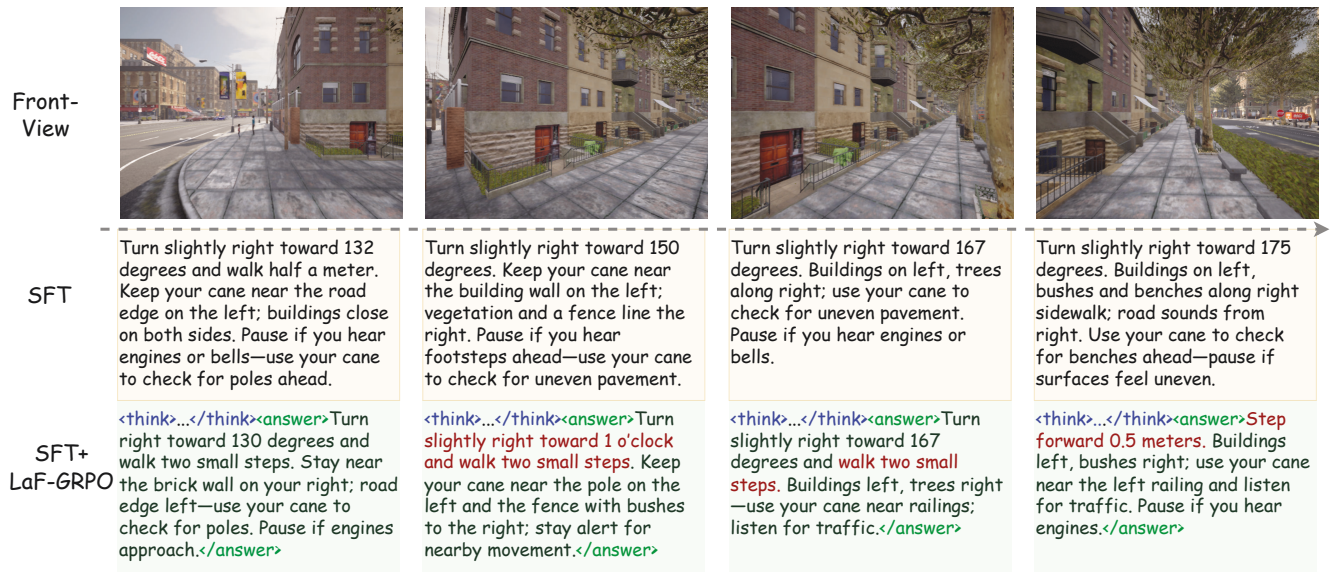

Figure 4: A comparative case study of navigational guidance provided by SFT and SFT+(LaF-GRPO) methods.

Pre-Cal.	Model	Intra-town (N = 613)				Inter-town (N = 11,223)			
		BLEU ↑	ROUGE ↑	METEOR ↑	SPICE ↑	BLEU ↑	Rouge ↑	METEOR ↑	SPICE ↑
No	7B-format+meteor+LLM (**1k**)	9.401	0.283	0.529	0.274	8.963	0.281	0.530	0.275
	7B-format+meteor+LLM (**2k**)	9.657	0.280	0.539	0.276	9.001	**0.276**	**0.535**	0.274
	7B-format+meteor+LLM (**3k**)	**10.037**	**0.284**	**0.545**	**0.283**	**9.002**	**0.276**	**0.535**	**0.278**
Yes	7B-format+meteor+LLM (**1k**)	10.265	0.279	0.543	0.286	9.463	0.271	0.540	0.285
	7B-format+meteor+LLM (**2k**)	10.136	0.284	0.550	**0.292**	**9.245**	**0.276**	0.541	0.284
	7B-format+meteor+LLM (**3k**)	**10.499**	**0.285**	**0.556**	**0.292**	9.232	0.275	**0.542**	**0.288**

Table 4: Ablation study results for the Qwen2.5-VL-7B model on the NIG4VI dataset with varying training **sample sizes**.

tial inter-rater agreement (Cohen's κ= 0.83), and the results confirmed a strong preference for LaF-GRPO instructions (76% vs. 24%) and its higher navigational accuracy (79.0% vs. 77.5%). We identify two main advantages of LaF-GRPO over standard GRPO, which utilizes only format and text generation rewards: (1) **Navigational Accuracy:** LaF-GRPO provides more precise movement and orientation guidance. (2) **Instruction Clarity:** The action interpreter component encourages VLMs to generate instructions that are clearer, more structured, and more comprehensible.

Case Study

Figure 4 provides a qualitative comparison of SFT+(LaF-GRPO) against the SFT baseline. Notably, SFT+(LaF-GRPO) generates instructions with greater linguistic variety and more intuitive directional cues. For instance, in Step 2, SFT+(LaF-GRPO) employs an o'clock direction (*"Turn slightly right toward 1 o'clock"*) and a relatable distance (*"two small steps"*), contrasting with SFT's numerical bearing (*"150 degrees"*). It can yield guidance that is more naturally understood by VI users. Furthermore, SFT+(LaF-GRPO), leveraging its internal reasoning process (i.e. the `<think>...</think>` blocks), frequently incorporates more environmental details and safety considerations. For example, its instruction for Step 4 (*"Step forward 0.5 meters; ...use your cane near the left ... and listen for traffic"*) also emphasizes immediate safety interactions.

Limitations and Future Work

This study's reliance on simulation and proxy users, while necessary for large-scale, controlled, and safe initial testing, introduces limitations. Future work could explore real-world dataset collection and direct engagement with VI users.

7 Conclusion

This study addresses navigation instruction generation for the visually impaired individuals (NIG-VI). We constructed the NIG4VI benchmark. Following this, we developed LaF-GRPO, a novel training paradigm for VLMs that incorporates an LLM-as-Follower reward. Experimental evaluations established LaF-GRPO's superiority over baselines and standard GRPO, with qualitative analysis confirming the practicality of the generated instructions in real-world scenarios. We hope our work and the benchmark inspire the development of more effective aids for the visually impaired.

Acknowledgments

This work was supported by the Innovation and Technology Fund (Project No. PRP/047/22FX), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/25200821), and PolyU Internal Fund from RC-DSAI (Project No. 1-CE1E). We also thank the reviewers for their constructive comments.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. arXiv:1607.08822.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I. D.; Gould, S.; and van den Hengel, A. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA*, 3674–3683. Computer Vision Foundation / IEEE Computer Society.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.anthropic.com/news/claude-3-family. Accessed: 2025-04-29.

Anthropic. 2024. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-04-26.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; and et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Banerjee, S.; and Lavie, A. 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*. Ann Arbor, Michigan: Association for Computational Linguistics.

Cai, S.; Ram, A.; Gou, Z.; and et al. 2024. Navigating real-world challenges: A quadruped robot guiding system for visually impaired people in diverse environments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18. Honolulu, HI, USA: ACM.

Chang, R.; Liu, Y.; and Guo, A. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In Yao, L.; Goel, M.; Ion, A.; and Lopes, P., eds., *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024*, 140:1–140:18. ACM.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; and et al. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Proceeedings of the Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, USA*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16.

Dou, Z.; and Peng, N. 2022. FOAM: A Follower-aware Speaker Model For Vision-and-Language Navigation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States*, 4332–4340. Association for Computational Linguistics.

Fan, S.; Liu, R.; Wang, W.; and Yang, Y. 2024. Navigation Instruction Generation with BEV Perception and Large Language Models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXII*, volume 15080 of *Lecture Notes in Computer Science*, 368–387. Springer.

Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; and et al. 2018. Speaker-Follower Models for Vision-and-Language Navigation. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 3318–3329.

Gao, Y.; Wu, D.; Song, J.; Zhang, X.; and et al. 2025. A wearable obstacle avoidance device for visually impaired individuals with cross-modal learning. *Nature Communications*, 16(1): 2857.

Giudice, N. A.; and Legge, G. E. 2008. Blind navigation and the role of technology. *The engineering handbook of smart technology for aging, disability, and independence*, 479–500.

Google DeepMind. 2024. Introducing Gemini 2.0: Our New AI Model for the Agentic Era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024. Accessed: 2025-04-21.

Gopinathan, M.; Masek, M.; Abu-Khalaf, J.; and Suter, D. 2024. Spatially-Aware Speaker for Vision-and-Language Navigation Instruction Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand*, 13601–13614. Association for Computational Linguistics.

Guan, Z.; Xiong, Z.; and Fan, M. 2024. FetchAid: Making Parcel Lockers More Accessible to Blind and Low Vision People With Deep-learning Enhanced Touchscreen Guidance, Error-Recovery Mechanism, and AR-based Search

Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA*, 39:1–39:15. ACM.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR 2022*. OpenReview.net.

Hu, Z.; Li, J.; Pu, Z.; Chan, H. P.; and Yin, Y. 2025. PraxisVLM: Vision-Grounded Decision Making via Text-Driven Reinforcement Learning. arXiv:2503.16965.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749.

Huang, Z.; Shangguan, Z.; Zhang, J.; and et al. 2022. ASSISTER: Assistive Navigation via Conditional Instruction Generation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, 271–289. Springer.

Jiang, B.; Chen, S.; Zhang, Q.; Liu, W.; and Wang, X. 2025. AlphaDrive: Unleashing the Power of VLMs in Autonomous Driving via Reinforcement Learning and Reasoning. arXiv:2503.07608.

Kong, X.; Chen, J.; Wang, W.; Su, H.; Hu, X.; Yang, Y.; and Liu, S. 2024. Controllable Navigation Instruction Generation with Chain of Thought Prompting. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, volume 15087 of *Lecture Notes in Computer Science*, 37–54. Springer.

Li, F. M.; Liu, M. X.; Kane, S. K.; and Carrington, P. 2024. A Contextual Inquiry of People with Vision Impairments in Cooking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA*, 38:1–38:14. ACM.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA*, 26286–26296. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, USA*.

Lu, H.; Liu, W.; Zhang, B.; and et al. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525.

Merchant, Z.; Anwar, A.; Wang, E.; Chattopadhyay, S.; and Thomason, J. 2024. Generating Contextually-Relevant Navigation Instructions for Blind and Low Vision People. arXiv:2407.08219.

OpenAI. 2024a. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2024b. GPT-4o System Card. arXiv:2410.21276.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; and et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, USA*.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; and et al. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv:2502.19634.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318. ACL.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and van den Hengel, A. 2020. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA*, 9979–9988. Computer Vision Foundation / IEEE.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv:2504.07615.

Team, G. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

World Health Organization. 2019. World report on vision. https://www.who.int/publications/i/item/9789241516570. Accessed: 2025-04-28.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800.

Younis, O.; Al-Nuaimy, W.; Rowe, F.; and Alomari, M. H. 2019. A smart context-aware hazard attention system to help people with peripheral vision loss. *Sensors*, 19(7): 1630.

Yuan, Z.; Zhang, T.; Deng, Y.; and et al. 2025. WalkVLM:Aid Visually Impaired People Walking by Vision Language Model. arXiv:2412.20903.

Zhao, L.; Nguyen, K.; and III, H. D. 2023. Define, Evaluate, and Improve Task-Oriented Cognitive Capabilities for Instruction Generation Models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada*, 3688–3706. Association for Computational Linguistics.

Zhao, Y.; Wang, S.; Geng, Q.; Yu, E.; and Li, J. 2025. "Less is More": Reducing Cognitive Load and Task Drift in Real-Time Multimodal Assistive Agents for the Visually Impaired. arXiv:2511.00945.

Zhao, Y.; Zhang, Y.; Xiang, R.; Li, J.; and Li, H. 2024. VIALM: A Survey and Benchmark of Visually Impaired Assistance with Large Models. arXiv:2402.01735.