# WalkVLM:Aid Visually Impaired People Walking by Vision Language Model

*Zhiqiang Yuan, Ting Zhang, Yeshuang Zhu, Jiapei Zhang, Ying Deng, Zexi Jia, Peixiang Luo, Xiaoyue Duan, Jie Zhou, Jinchao Zhang**

Tencent WeChat ICCV OCT 19-23, 2025 HONOLULU HAWAII

## Contribution

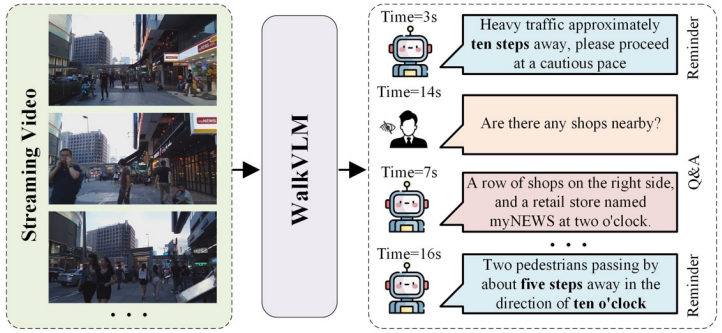



Fig.1 WalkVLM delivers timely, concise, informative walking reminders for visually impaired individuals by utilizing hierarchical planning and temporal-aware adaptive prediction.

- **A diverse and extensive dataset named the Walking Awareness Dataset (WAD), providing extensive data support for the blind walking task.**
- **We propose WalkVLM, a vision-language model for streaming video analysis, which adaptively generates concise yet informative walking guidance for visually impaired people.**
- **We establish a solid baseline to provide timely and reasonable walking guidance for the visually impaired, laying a solid foundation for the practical application of VLM in this field.**

## WalkVLM

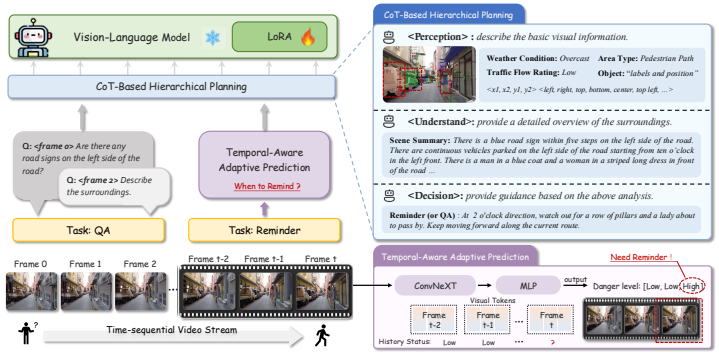



Fig.2 An overview of the proposed WalkVLM framework. WalkVLM employs CoT-based hierarchical planning to summarize the static attributes and understanding of scenes, thereby facilitating the subsequent reminder generation and QA tasks. Furthermore, temporal-aware adaptive prediction is proposed to calculate the trigger state of VLM, so as to reduce the temporal redundancy of outputs.

## Walking Awareness Dataset

- **We conducted a blind test experiment to determine the classification of the six reminder types.**

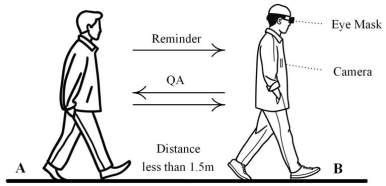



Fig.3 Blind test experiment for analyzing the most critical information needed by users in blind walking. We required two individuals to collaborate as a team, where the participant at the rear provided directions to enable the individual at the front to arrive at a specific location safely in the absence of any visual information.

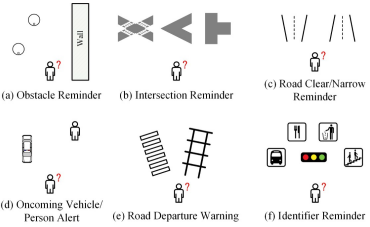



Fig.4 Visualization of six scenarios that require reminders, which were summarized through multiple blind experiments.

- **We provide different levels of annotations for different types of scenes, covering most common objects.**

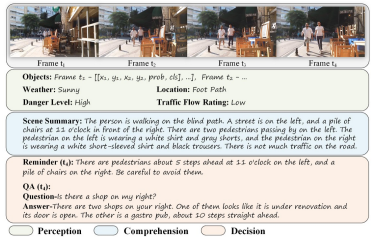
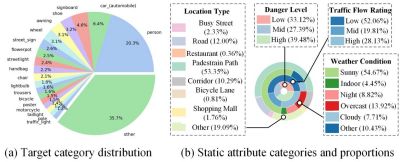



Fig.5 Visualization of the Walking Awareness Dataset. Each sample contains a video clip and multiple annotations, with the hierarchy divided into perception, comprehension, and decision.




(a) Target category distribution (b) Static attribute categories and proportions

Fig.6 Visualization of the proportion of targets and categories in our Walking Awareness Dataset.

- **Compared with other domain datasets, the annotations we provide are more extensive and more granular.**

| Dataset | Type | #Sample | Modality | Bounding Box | Weather | Danger level | Scene Summary | QA | Reminder | Open |
|---|---|---|---|---|---|---|---|---|---|---|
| Obstacle Dataset (2023)[32] | $\mathcal{T}$ | 8k | Image | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| WOTR (2023)[38] | $\mathcal{T}$ | 13k | Image | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| ISLAM *et al.*(2024)[15] | $\mathcal{T}$ | 31 | Image & Video | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Wang *et al.*(2024)[35] | $\mathcal{T}$ | 50 | Video | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| VizWiz (2018)[11] | $\mathcal{S}$ | 31k | Image | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ |
| Zain *et al.*(2024)[27] | $\mathcal{S}$ | 48 | Image | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| WAD (Ours) | $\mathcal{TS}$ | 12k / 120k | Video / Image | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table.1 Static information comparison of different datasets in blind walking. T and S denote the dataset types, representing target-based and semantic-based datasets, respectively. WAD dataset holds a great advantage in terms of sample numbers, categories, and modalities.

## Main Results

| Model | Reminder Task | | | | | QA Task | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TF-IDF | ROUGE-1 | ROUGE-2 | ROUGE-L | GPT Score | TF-IDF | ROUGE-1 | ROUGE-2 | ROUGE-L | GPT Score |
| LLaVa (7B) [23] | 0.061 | 0.062 | 0.005 | 0.070 | 0.013 | 0.074 | 0.084 | 0.011 | 0.072 | 0.012 |
| DeepSeek-VL (1.3B) [26] | 0.073 | 0.098 | 0.015 | 0.090 | 0.148 | 0.182 | 0.103 | 0.020 | 0.095 | 0.336 |
| DeepSeek-VL (7B) [26] | 0.132 | 0.073 | 0.009 | 0.068 | 0.015 | 0.189 | 0.088 | 0.021 | 0.081 | 1.000 |
| Yi-VL (6B) [44] | 0.112 | 0.093 | 0.009 | 0.085 | 0.133 | 0.113 | 0.091 | 0.012 | 0.082 | 0.168 |
| MiniCPM-V2.6 (8B) [43] | 0.111 | 0.071 | 0.007 | 0.064 | 0.025 | 0.192 | 0.139 | 0.025 | 0.120 | 0.832 |
| Qwen2-VL (7B) [36] | 0.106 | 0.107 | 0.010 | 0.097 | 0.044 | 0.232 | 0.182 | 0.037 | 0.162 | 0.504 |
| GPT-4o [13] | 0.116 | 0.078 | 0.008 | 0.072 | - | 0.242 | 0.163 | 0.034 | 0.145 | - |
| *DeepSeek-VL (7B) [26] | 0.129 | 0.152 | 0.043 | 0.141 | 0.969 | 0.193 | 0.166 | 0.037 | 0.149 | 2.496 |
| *MiniCPM-V2.6 (8B) [43] | 0.152 | 0.171 | 0.056 | 0.170 | 1.024 | 0.207 | 0.176 | 0.044 | 0.160 | 3.172 |
| *Qwen2-VL (7B) [36] | 0.147 | 0.163 | 0.054 | 0.165 | 1.018 | **0.246** | 0.196 | 0.047 | 0.167 | 3.246 |
| WalkVLM | **0.166** | **0.191** | **0.062** | **0.173** | **1.103** | 0.189 | **0.202** | **0.051** | **0.174** | **4.168** |

Table.2 Quantitative comparison of different methods on reminder generation and QA tasks. WalkVLM leads in almost all the TF-IDF, ROUGE, and GPT Score metrics.* indicates the fine-tuned model.

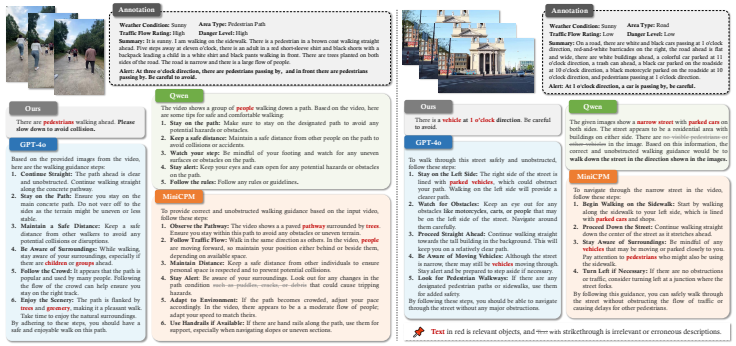



Fig.7 Visualization comparison of different VLM models. Compared to other models, WalkVLM is able to generate concise and informative answers, providing users with a good experience in blind walking.

| Model | Yi-VL | MiniCPM-V2.6 | GPT-4o | Qwen2-VL | WalkVLM |
|---|---|---|---|---|---|
| TRF | 0.341 | 0.396 | 0.430 | 0.449 | **0.505** |

Fig.8 Temporal redundancy assessment of the reminder, our method achieved the highest TRF score.

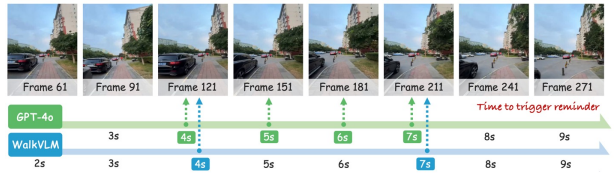



Fig.9 Visualization of triggering moments of GPT-4o and WalkVLM. WalkVLM triggers with less redundancy, providing information to users in a more timely manner.

## Further works

- **Reduce output and temporal redundancy of reminders to improve the informativeness and conciseness**
- **Decouple feature extraction and reminder generation in video stream processing, switching from serial processing to an asynchronous parallel architecture so as to further reduce processing latency.**

*https://walkvlm2024.github.io* *yuanzhiqiang19@mails.ucas.ac.cn* *Pattern Recognition Center, Wechat, Tencent Inc, Beijing, China*