WalkVLM : Aid Visually Impaired People Walking by Vision Language Model

Anonymous CVPR submission

Paper ID 12248

Abstract

Approximately 200 million individuals around the world suffer from varying degrees of visual impairment, making it crucial to leverage AI technology to offer walking assistance for these people. With the recent progress of vision-language models (VLMs), employing VLMs to improve this field has emerged as a popular research topic. However, most existing methods are studied on self-built question-answering datasets, lacking a unified training and testing benchmark for walk guidance. Moreover, in blind walking task, it is necessary to perform real-time streaming video parsing and generate concise yet informative reminders, which poses a great challenge for VLMs that suffer from redundant responses and low inference efficiency. In this paper, we firstly release a diverse, extensive, and unbiased walking awareness dataset, containing 12k video-manual annotation pairs from Europe and Asia to provide a fair training and testing benchmark for blind walking task. Furthermore, a WalkVLM model is proposed, which employs chain of thought for hierarchical planning to generate concise but informative reminders and utilizes temporal-aware adaptive prediction to reduce the temporal redundancy of reminders. Finally, we have established a solid benchmark for blind walking task and verified the advantages of WalkVLM in stream video processing for this task compared to other VLMs. Our dataset and code will be released at anonymous link https://walkvlm2024.github.io.

1. Introduction

Approximately 200 million people worldwide suffer from varying degrees of visual impairment, with 36 million completely blind [1, 2]. These visually impaired people (VIPs) are facing severe challenges in daily activities such as walking, which may be alleviated by contemporary artificial intelligence technologies [3, 4].

The current walking assistance works primarily concentrate on electronic assistive devices, sensory substitution devices, and computer vision-based assistive systems [5–7]. Among them, vision-based assistive systems can be roughly divided into detection-based methods and semantic-based methods [8–10]. Detection-based methods have been studied for a long time, aiming to detect potential obstacles in the field of view, so as to let VIPs avoid them [11, 12]. Semantic-based methods utilize vision-language models (VLMs) to analyze images, thereby generating responses to VIPs' questions [13, 14]. In recent days, with the development of VLMs [15, 16], semantic-based methods have gained significant attention. Some studies have tested VLMs in a zero-shot manner to analyze the effectiveness of these models in blind walking [11, 14]. Moreover, some studies have fine-tuned VLMs using traditional visual question-answer (QA) datasets in this field or a small quantity of self-built datasets, so that the model can better answer user questions [3, 17]. These studies have empowered blind walking tasks with VLMs and already achieved attractive application results.

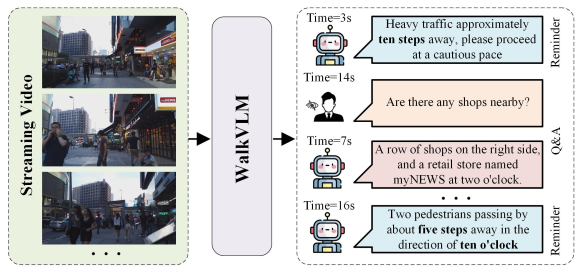


Figure 1. WalkVLM provides opportune, concise, informative walking reminders and answers for visually impaired people based on hierarchical planning and temporal-aware adaptive prediction.

Although some VLM-based models for blind walking have been developed, these models still face challenges before they can be applied in practice. Firstly, most current research relies on a small number of self-collected image-text pairs and lacks a consistent and extensive benchmark [17, 18]. Moreover, the images and text in traditional datasets are predominantly in a question-and-answer paradigm, which makes it challenging for VLMs to proactively generate guided responses rather than specific answers to questions [13, 19]. Secondly, in blind walking task, it is necessary to perform real-time streaming video parsing and generate concise yet informative reminders, which poses a great challenge for VLMs that suffer from redundant responses and low inference efficiency[20, 21].

In this paper, we propose a WalkVLM for the blind walking task and establish a new benchmark to promote the development of this field. Specifically, we first introduce a diverse, extensive, and unbiased Walking Awareness Dataset (WAD), which contains 12k video-manual annotation pairs from Europe and Asia to provide a fair training and testing baseline. After that, as shown in Figure 1, we introduce the WalkVLM model to interpret video streaming, which employs a chain of thought to hierarchically direct VLM in generating concise yet informative reminders, and achieves opportune reminders by the proposed temporal-aware adaptive prediction. Comprehensive experiments show that, compared to other VLM models, WalkVLM can generate more concise reminders and has better temporal adaptability when handling video streaming in blind walking task. The main contributions of our work are as follows:

- We construct a diverse, extensive, and unbiased walking awareness dataset, providing extensive data support for blind walking task.
- A WalkVLM model for streaming video parsing has been proposed to adaptively provide concise yet informative walking reminder for visually impaired people.
- To the best of our knowledge, this is the first work to utilize VLM to provide opportune walking guidance for visually impaired individuals, laying a solid foundation for the practical application of VLM in this field.

2. Related Work

Vision Datasets for Blind Walking. Existing datasets for blind walking can be roughly divided into two types: detection-based [8, 22–24] and semantic-based [9, 13]. Detection-based datasets have been extensively studied in the blind walking, where researchers utilize these datasets to train the obstacle detection model, thereby reducing the accident rate of VIPs in this task. For example, Zhang *et al.* [22] recently developed a TP-Dataset for detecting visual tactile paving surfaces and offered guidance for the visually impaired through provide walking routes. Islam *et al.* [23] introduced a dataset for improving real-time object recognition systems to aid VIPs in navigation tasks, which contains 90 object annotations from 31 video clips. Compared with detection-based datasets, semantic-based datasets are relatively rare, which contain question-answering properties and provide an enhanced human-computer interaction experience. Gurari *et al.* [9] constructed a VQA dataset for VIPs, which contains 31k visual questions, each with 10 crowdsourced answers. In addition, some researchers have constructed several self-built question-answer datasets with specific attributes during their studies [3, 13], however, these self-built datasets are not open-sourced and are relatively small in scale, making them unsuitable for large-scale and unified benchmarking.

Vision-based Methods for Blind Walking. Similar to the division of datasets, the vision-based methods that help VIPs walking can also be divided into detection-based methods [11, 12] and semantic-based methods [13]. Detection-based methods typically use detectors to obtain obstacles during walking, thereby providing users with specific object locations. Liu *et al.* [12] proposed an open scene understanding system, which improves detection performance by using SAM [25] to generate pixel-level dense segmentation masks. Tian *et al.* [26] proposed a system for understanding dynamic crosswalk scenes, including crosswalks, vehicles, and pedestrians, thereby providing VIPs with indications of when and where to cross the road. The semantic-based approach provides VIPs with the scene understanding in the form of question-answer. Merchant *et al.* [17] verified that vision-language models can generate correct and useful instructions for VIPs, and studied methods to provide users with context-related guidance. Yang *et al.* [3] explored how to utilize VLMs to provide reliable visual question answers for VIPs, and they fine-tuned the VLMs by LoRA on a small amount of self-built dataset to generate detailed and practical suggestions. Moreover, a few applications such as Be My AI [1] have also adopted semantic-based methods to enable VIPs to take photos for answering questions. However, these applications also only support the question-and-answer paradigm and struggle to provide concise and opportune reminders during walking.

Vision-language Models. With the popularity of large language models (LMM), vision-language models have also begun to receive significant attention [18, 27, 28]. Liu *et al.* [29] proposed the LLaVa, which employ the ViT visual encoder to encode images, follow by mapping them through an MLP to the LLM, yields favorable outcomes in benchmark tests when answering pertinent questions. Subsequently, a plethora of studies emerged based on LLaVa, which greatly impacted various fields [30–33]. Furthermore, multimodal models like Qwen, Gemini, and MiniCPM-V [34–36] have progressively adopted support for multi-frame image inputs and have undergone optimizations for scenarios such as edge devices, significantly enhancing the usability of VLMs in a wide range of applications. Despite the existing studies validating the viability of multimodal large-scale models [37], there remains a dearth of related applications within specific vertical sectors. For instance, only a limited number of studies [3, 13, 17] have focused on the applicability of VLMs in the blind walking task, with a notable absence of unified and systematic modeling approaches.

3. Walking Awareness Dataset

In this section, we have constructed a walking awareness dataset to provide open data support for blind walking task.

[1] https://www.bemyeyes.com

CVPR #12248

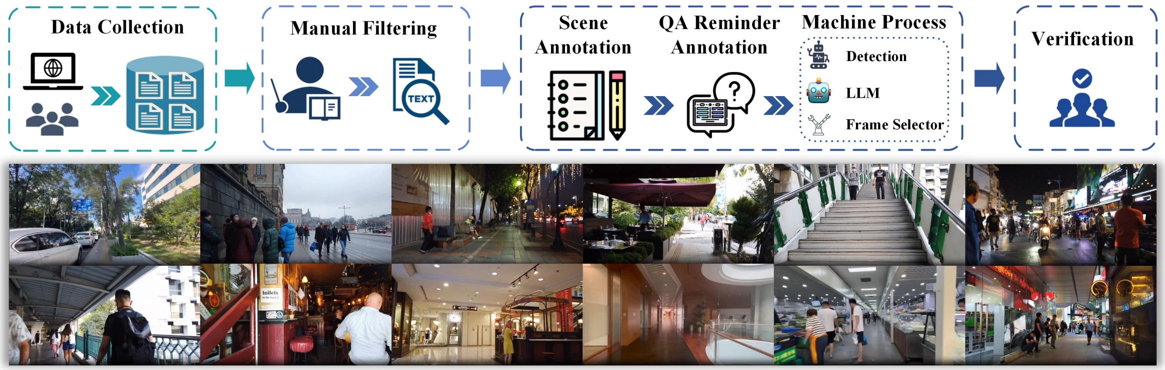


Figure 2. The data annotation pipeline for constructing the walking awareness dataset. Appendix A.5 provides more random sampling examples to observe the diversity and complexity of WAD dataset.

3.1. Data Collection

The WAD dataset has a wide range of geographical sources, which originate from 10 different locations in Europe and Asia. 20% of the original data in the WAD dataset comes from the annotators' recordings, and the rest comes from YouTube[2]. During the recording, six recorders positioned the camera at a height corresponding to chest level, employing focal lengths of 13mm, 20mm, and 26mm, as well as resolutions ranging from 1080p to 4k at 60fps, to enhance the variability of the data. Lastly, we have amassed approximately 13 hours of walking video, and see Appendix A for the duration of data gathered from various regions.

3.2. Annotation Strategy

Figure 2 shows the overall annotation pipeline of walking awareness dataset. Next, we will elaborate from two aspects: scene annotation and response annotation.

Scene annotation. Scene annotation aims to label the inherent attributes of the current scene. We requested nine annotators to label the video scene in terms of weather conditions, location type, traffic flow rating, danger level, and scene description. When outdoors, weather conditions are divided into six categories such as sunny and rainy, while the status is empty when indoors. The location type is divided into eight categories, such as corridors and pedestrian walkway. The traffic flow rating is divided into three levels, which are defined based on the person number in the video stream. The danger level is defined as the walking hazard in the current scene, which is qualitatively divided by the traffic flow rating and road smoothness. The scene description is an overview of the current environment, including an expansion on factors such as pedestrian flow, vehicle traffic, road conditions, and the surrounding environment. Subsequently, we employed the open-world detection model [39] for the preliminary detection of targets, and carried out a corresponding human review to uphold the result accuracy.

[2]https://www.youtube.com/@poptravelorg

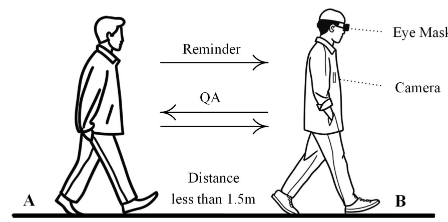


Figure 3. Blind test experiment for analyzing the most critical information needed by users in blind walking. We required two individuals to collaborate as a team, where the participant at the rear provided directions to enable the individual at the front to arrive at a specific location safely in the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

Response annotation. Response denotes the concise reminders that the model is required to generate, as well as the answer that reply to user's question in blind walking task. In order to analyze the most critical information needed by users in blind walking, we conducted a blind test experiment as shown in Figure 3. In the experiment, we requested two people to collaborate in pairs, with the person A behind giving directions, so that the person B in front with eye mask could reach a certain destination without any collisions. In such a scenario, the instructions received by person B during walking come entirely from person A, and the route priors possessed by real blind people are avoided, which can help us analyze what types of information are necessary for the blind walking task. In a large number of such experiments, we have verified that such guidance can guide visually impaired people to walk safely, indicating that the information provided by person A is sufficiently effective for person B. We recorded the video and audio that occurred during this process, analyzed the information interaction between the subjects, and thus provided the following valid information types that need to be marked for subsequent *reminder* and *QA* annotations:

- *Reminder type.* Based on the blind test experiment, as shown in Figure 4, we divided the reminders during walking into six types. (a) Obstacle reminder: Trigger a re-

Dataset	Type	#Sample	Modality	Bounding Box	Weather	Danger level	Scene Summary	QA	Reminder	Open
Obstacle Dataset (2023)[24]	$\mathcal{T}$	8k	Image	✓	✗	✗	✗	✗	✗	✓
WOTR (2023)[8]	$\mathcal{T}$	13k	Image	✓	✗	✗	✗	✗	✗	✓
ISLAM *et al*.(2024)[38]	$\mathcal{T}$	31	Image & Video	✓	✗	✗	✗	✗	✗	✓
Wang *et al*.(2024)[11]	$\mathcal{T}$	50	Video	✓	✓	✗	✗	✗	✗	✗
VizWiz (2018)[9]	$\mathcal{S}$	31k	Image	✗	✗	✗	✗	✓	✗	✓
Zain *et al*.(2024)[17]	$\mathcal{S}$	48	Image	✗	✗	✗	✗	✗	✓	✗
WAD (Ours)	$\mathcal{TS}$	12k / 120k	Video / Image	✓	✓	✓	✓	✓	✓	✓

Table 1. Static information comparison of different datasets in blind walking task. For dataset types, $\mathcal{T}$ and $\mathcal{S}$ denote the target-based and semantic-based dataset, respectively. WAD dataset holds a significant advantage in terms of sample numbers, categories, and modalities.

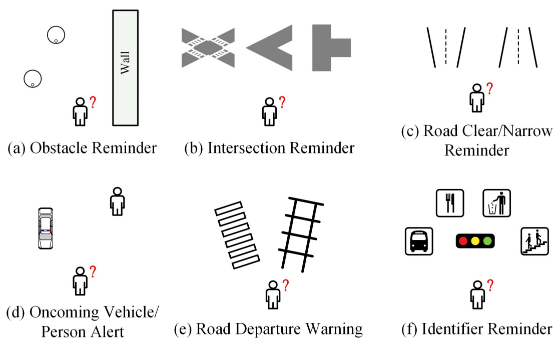


Figure 4. Visualization of six scenarios that require reminders, which were summarized through multiple blind experiments.

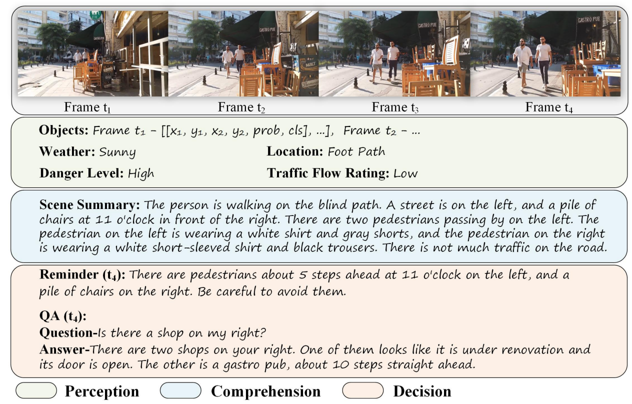


Figure 5. Visualization of the walking awareness dataset. Each sample contains a video clip and multiple labels, with the label hierarchy divided into perception, comprehension, and decision.

minder when there is a non-moving obstacle on the walking route. (b) Intersection reminder: Trigger a reminder when the current road has intersections, turns, *etc*. (c) Road clear/narrow reminder: Provide reminders about the width and pass ability of the road. (d) Oncoming vehicle/person reminder: When there are moving obstacles on the walking route, trigger a reminder for potential dangers. (e) Road departure warning: Issue a warning when there is an angular offset between the walking route and the current road. (f) Identifier reminder: Provide reminders for prominent landmarks in the scene, such as road signs and traffic lights.

- *QA type.* For QA type, we proceed from three aspects: scene perception, road inquiry, and detailed consultation. (a) Scene perception: Macro-level insights such as the understanding of the scene. (b) Road inquiry: Route planning to reach a certain location within visible range. (c) Detailed consultation: Knowledge QA on local details, such as road sign content, shop names, *etc*.

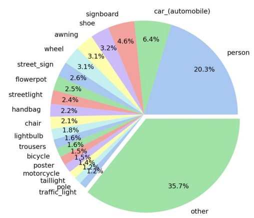
When marking reminder and QA, we require annotators to indicate the specific location of obstacles in the video. In the annotations, the distances are represented by steps on a scale of 5, the directions are indicated by clock positions, so as to reduce the offset caused by the camera perspective. We require nine annotators to annotate the above content, and the relevant annotation interface is shown in Appendix A.3. After the annotation is completed, in order to further standardize the annotation content to remove potential bias, we used GPT [40] to rephrase the annotated content and conducted manual verification.

3.3. Dataset Analysis

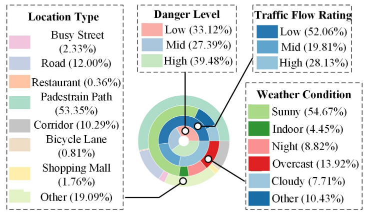
Figure 5 shows a sample of the WAD dataset, and we divide the annotations into three parts following lower to higher levels: perception, comprehension, and decision. The perception label reflects the basic attributes of the video, such as obstacle location, weather conditions, *etc*., while the comprehension label reflects the model's understanding of the entire scene. The decision label contains reminder and QA, reflecting the model's decision on the user's walking based on its understanding of the current scenario.

Table 1 illustrates the comparison between the WAD dataset and other prevalent datasets utilized in blind walking tasks, with $\mathcal{T}$ representing the detection-based dataset and $\mathcal{S}$ indicating the semantic-based dataset. Compared to other different types of datasets, WAD has a larger data size while containing more static attributes of the environment, scene summaries, QA, and reminder, thus providing more supervision to train the model. It is worth emphasizing that the samples we furnish are exclusively video clips, which possess a greater volume of information in comparison to the images supplied by other datasets. Moreover, for each video clip, we have extracted 10 keyframes to streamline researchers' use. The walking awareness dataset contains 3.47 million instances, with categories and the respective proportions shown in Figure 6(a). The category-related dis-

CVPR #12248



(a) Target category distribution



(b) Static attribute categories and proportions

Figure 6. Visualization of the proportion of targets and categories in our walking awareness dataset.

tribution in the WAD dataset is shown in Figure 6(b). We have selected 1.5k samples as a test set based on different static tag types, different reminder types, and different QA types to ensure the diversity and completeness in evaluation.

3.4. Possible Sources of Bias

Although the WAD dataset is collected from a wide range of geographical sources, we are aware of a few biases in our dataset. The regions are still limited, which is still a long way from complete coverage of the globe. The position of the camera and the divergence of focal length are also concerns for us, which need to obtain more general data to compensate for this. In addition, the linguistic preferences of the annotators can introduce specific biases into the generated reminder, which implies that during the walking process, the model might provide information that are more appropriate for the area where the annotation was made.

4. WalkVLM

This section proposes WalkVLM, attempting to empower the blind walking task using a vision-language model based on the WAD dataset. The overall architecture of WalkVLM is shown in Figure 7. We will start with problem formulation and proceed with hierarchical planning and temporal-aware adaptive prediction to generate concise and opportune walking reminders.

4.1. Problem Formulation

We aim to steer a VLM to process video streams, enabling it to provide walking reminders that include temporal attributes, and to enable the model to answer specific questions in human-machine interactions. Specifically, at time t_0, given the newly appeared frames $[I_{t_{-N}}..., I_{t_{-1}}, I_{t_0}]$, categlory and obstacle position in the image $[O_{t_{-N}}..., O_{t_{-1}}, O_{t_0}]$, VLM is hoped to generate a concise and informative reminder $T^R_{t_0}$ based on visual information. During walking, VIPs can also raise a question Q_{t_0} to communicate with the VLM at any time, so as to inquire about information such as the current scene and route. Additionally, since generating reminders at every frame may lead to a poor walking guidance experience and impose significant real-time processing pressure on hardware, WalkVLM needs to be able to predict the current VLM trigger state s_{t_0} based on historical states $[s_{t_{-N}}..., s_{t_{-1}}, s_{t_0}]$ and the previous N frames, so as to choose specific moments to output reminders.

4.2. CoT-Based Hierarchical Planning

We attempt to make VLM conduct step-by-step derivation by a Chain of Thought (CoT) [41], enabling it to summarize from comprehensive information such as the static attributes and the summary of the scene, thereby refining out concise and informative reminders. The model architecture integrates a vision transformer encoder and a large language model (LLM). The vision encoder generates image tokens, while an attention-based extractor aligns these tokens with the LLM, enabling comprehensive understanding and information processing. WalkVLM combines multi-frame information to make reminders, ensuring that the model has a comprehensive perception of the environment.

We divide the process of reminder generation into three levels: perception, comprehension, and decision. At the **perception level**, the model extracts static visual attributes from the current frame, such as location type, weather conditions, and traffic flow rating. To enhance the VLM model's focus on significant elements and improve visual perception accuracy, we incorporate a priori-object location module (POLM). The POLM initially uses a generic object detector [39] to identify and locate objects in the scene, then filters them based on size and confidence scores to highlight crucial items that reflect road conditions and potential danger. The filtered information and basic environmental attributes provide the necessary input for the model to perceive the external world. At the **comprehension level**, the model integrates all outputs from the perception layer, merging local detection results and fragmented scene information into a comprehensive global summary. Relying on the capabilities of the VLM and the detailed attributes from the perception stage, this stage ensures that the model has a clear understanding of the current environment. At the **decision level**, we focus on training the WalkVLM model to achieve visual QA and reminder. At this stage, the model already possesses an understanding of the static attributes and overall situation of the environment. Therefore, with appropriate guidance, the model is expected to briefly analyze potential hazards in the scene.

During training, we adopted a CoT approach to gradually feed information from three levels into the VLM, and during testing, we let the model predict the aforementioned attributes and generate the corresponding responses.

4.3. Temporal-Aware Adaptive Prediction

Although VLMs are capable of scene parsing across multiple frames and generating the required output, directly applying them to video streaming will lead to unavoidable issues. For instance, when utilizing VLM to generate walking

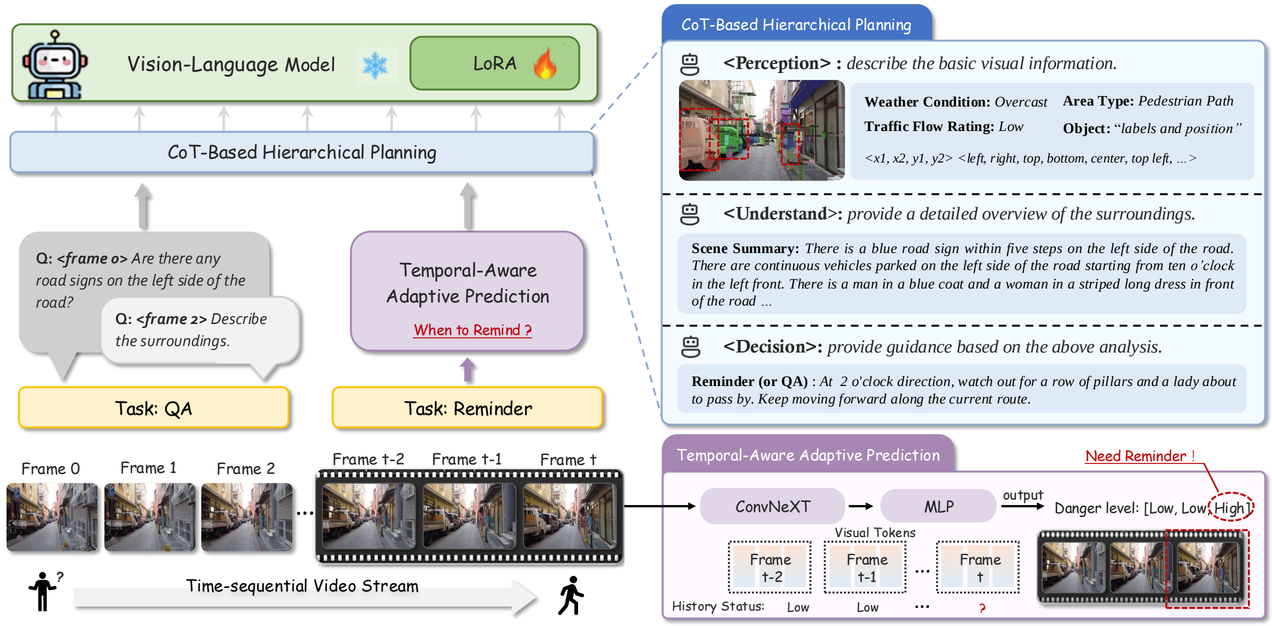


Figure 7. An overview of the proposed WalkVLM framework. WalkVLM employs CoT-based hierarchical planning to summarize the static attributes and understanding of scenes, thereby facilitating the subsequent reminder and QA tasks. Furthermore, temporal-aware adaptive prediction has been proposed to calculate the trigger state of VLM, thereby reducing the temporal redundancy of outputs.

reminders frame by frame or at regular intervals, it will produce a substantial amount of temporal redundancy for the user, resulting in a suboptimal user experience. Secondly, continuous VLM inference also brings computational pressure to hardware devices. Identifying and implementing solutions to this challenge is a key component in the effective utilization of VLM for video streaming processing.

To address the aforementioned issues, we come up with a temporal-aware adaptive prediction (TAP) module that incorporates historical information to pre-calculate whether to trigger the VLM currently, thereby reducing the inference pressure on hardware. Specifically, as shown in the right of Figure 7, we utilize a lightweight model to analyze historical N frames and determine whether to trigger the VLM at the current moment based on the historical output states. Given the frames $[I_{t_{-N}}, ..., I_{t_{-1}}, I_{t_0}]$, we utilize a 3D convolutional model to extract the features f_v from the sequence. Simultaneously, the predicted trigger states from the previous N moments are independently embedded, concatenated, and then passed through multiple layers of perceptrons to generate the state feature f_s. Furthermore, f_v and f_s are integrated by a multi-layer MLP to generate the current trigger probability $\mathcal{P}_t$. Three levels of triggers are defined, which correspond to the degrees of danger in the WAD dataset.

The TAP model is used to trigger the reminder of VLM, and subsequent experiments have verified that this module can effectively reduce the temporal redundancy when generating walking guidance.

5. Experiments

5.1. Settings

Models & Details. WalkVLM is implemented with the MiniCPM-V2.6 model [36], which is an 8B multimodal model built upon Qwen2-7B [46]. We add LoRA to all the linear layers of MiniCPM-V2.6 with a rank of 64, while maintaining the video stream sampling rate of 2 FPS. The number of historical frames N is set to 3, and the visual extraction backbone in the TAP module is ConvNext3D [47]. We compared WalkVLM with multiple popular multimodal models, including GPT-4o [44], Qwen2-VL(7B) [45], MiniCPM-V2.6(8B) [36], DeepSeek(1.3B&7B) [42], Yi-VL(6B) [43]. All the prompts of the large models used in this paper can be found in Appendix B.

Metrics. We use the following metrics to evaluate the models: **(a) ROUGE.** This metric measures the similarity between the generated text and the reference text by comparing overlapping words or phrases, including ROUGE-1, ROUGE-2, and ROUGE-L [48]. **(b) TF-IDF Similarity (TF-IDF).** Combine term frequency and inverse document frequency to evaluate the weight of words, represent the text as a TF-IDF vector, and then measure the semantic similarity between texts [49]. **(c) GPT Score.** GPT4 is used to evaluate the superiority ratio between the generation results of different multimodal models and the ground truth (GT) [50, 51]. **(d) Temporal Redundancy F1-Score (TRF).** Given the historical model state and historical frames, let the model predict the danger level of the current moment,

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
DeepSeek (1.3B) [42]	0.073	0.098	0.015	0.090	0.060	0.182	0.103	0.020	0.095	0.042
DeepSeek (7B) [42]	0.132	0.073	0.009	0.068	0.006	0.189	0.088	0.021	0.081	0.125
Yi-VL (6B) [43]	0.112	0.093	0.009	0.085	0.054	0.113	0.091	0.012	0.082	0.021
MiniCPM-V2.6 (8B) [36]	0.111	0.071	0.007	0.064	0.010	0.192	0.139	0.025	0.120	0.104
GPT-4o [44]	0.116	0.078	0.008	0.072	0.405	**0.242**	0.163	0.034	0.145	0.125
Qwen2-VL (7B) [45]	0.106	0.107	0.010	0.097	0.018	0.232	0.182	0.037	0.162	0.063
WalkVLM	**0.166**	**0.191**	**0.062**	**0.173**	**0.447**	0.189	**0.202**	**0.051**	**0.174**	**0.521**

Table 2. Quantitative comparison of different methods on reminder and QA tasks. WalkVLM leads in almost all the TF-IDF, ROUGE, and GPT Score metrics. The higher the metric, the better the result. **Bold** and underline indicate the best and the second-best, respectively.

and calculate the F1-Score between the prediction and the GT. TF-IDF and ROUGE evaluate similarity from semantic similarity and word granularity, respectively, while the GPT Score determines the optimal result by comparing results with GT. TRF measures the temporal redundancy of the model's output; the higher it is, the less temporal redundancy is generated.

5.2. Quantitative Results

Table 2 presents the quantitative metrics of different models on the reminder and QA task. On the ROUGE metric, WalkVLM has achieved the best results in both tasks, verifying that the model's output is closest to the GT at the word granularity. On the TF-IDF metric for measuring semantic similarity, WalkVLM performs the best in reminder tasks, indicating that the model can generate more concise and accurate results like GT. While in QA tasks, WalkVLM's performance on TF-IDF scores does not stand out significantly. This could be attributed to the fact that during training, the model is encouraged to generate concise answers, which may inadvertently diminish its capacity to offer elaborate explanations of the questions. The GPT score represents the overall evaluation of the LLM on the generated results and the GT. WalkVLM outperforms other models such as GPT-4o in terms of GPT scores for reminder and QA tasks, validating that the model's output has the most consistent distribution with the GT.

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

Table 3. Temporal redundancy assessment of the reminder task, our method achieved the highest TRF score.

We use TRF to evaluate the temporal redundancy of the output from various VLMs. Specifically, we utilize multiple frames of images along with historical dangerous states as inputs, letting the model to generate a dangerous level discrimination identifier, thereby determining whether a reminder should be triggered currently. As shown in Table 3, compared to other models, WalkVLM has achieved the highest TRF indicator, which indicates that this model can better reduce the redundancy of reminders in temporal.

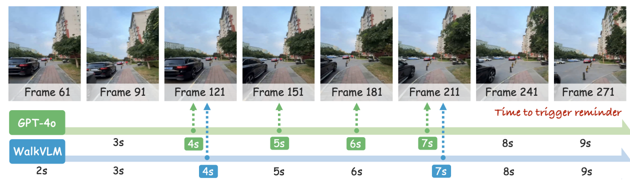


Figure 8. Visualization of triggering moments of GPT-4o and WalkVLM. WalkVLM triggers with less redundancy, providing information to users in a more timely manner.

Model	Reminder Task		QA Task	
	Concise.	Semantic.	Concise.	Semantic.
DeepSeek(1.3B)	0.026	0.080	0.091	0.114
DeepSeek(7B)	0.002	0.197	0.061	0.114
Yi-VL	0.085	0.023	0.121	0.022
MiniCPM-V2.6	0.026	0.122	0.061	**0.205**
GPT-4o	0.056	0.195	0.030	**0.205**
Qwen2-VL	0.121	0.168	0.061	0.170
WalkVLM	**0.683**	**0.216**	**0.576**	0.170

Table 4. User study results on conciseness and semantic similarity across different tasks. Higher score indicates better performance.

5.3. Qualitative Results

Figure 9 presents the visual comparison in reminder task between different VLM models. Compared to other methods such as GPT-4o, WalkVLM can generate more concise and informative responses, thus providing a better experience for users. In the left case, whereas other models offer highly detailed responses, WalkVLM simply provides a concise prompt to the user, effectively highlighting the crucial aspect. As in the right case, WalkVLM perceives the car coming from the one o'clock direction and conveys the focus to the user, which other models have not accomplished.

Figure 8 shows a qualitative comparison of GPT-4o and WalkVLM in terms of temporal redundancy. Our model triggers VLM with lower temporal redundancy and can provide information to users in a more timely manner. Appendix C presents more qualitative results, including the comparison with other VLMs on actual video streams.

5.4. Subjective Results

As illustrated in Table 4, we requested nine annotators to perform a subjective evaluation of various VLM models with respect to language conciseness and semantic similarity to the GT. Participants are required to rank the results individually, and we use the top-1 superiority ratio to eval-

CVPR
#12248

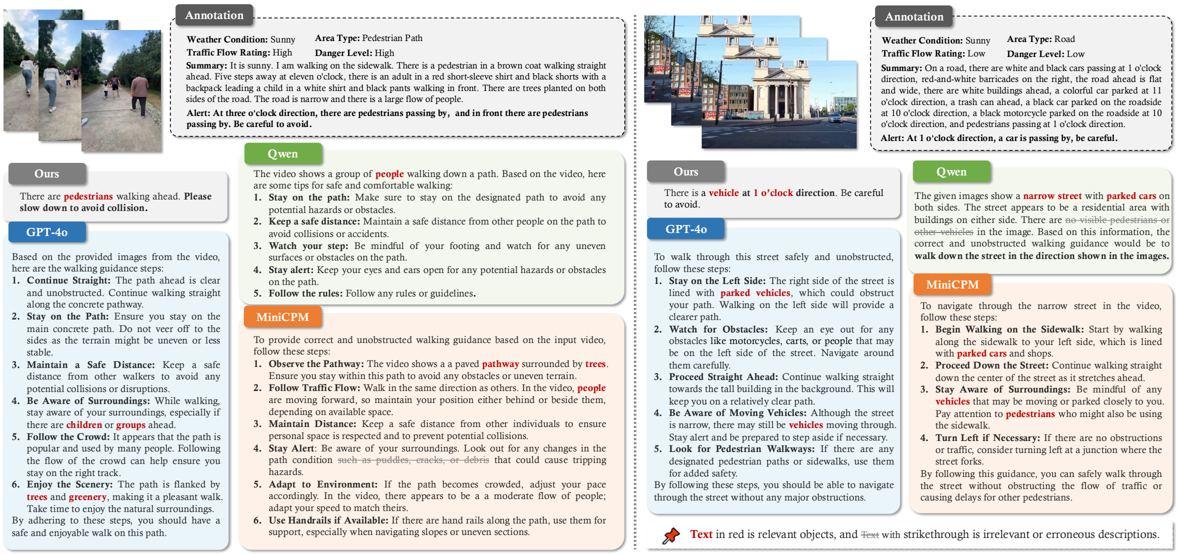


Figure 9. Visualization comparison of different VLM models. Compared to other models, WalkVLM is able to generate concise and informative answers, providing users with a good experience in blind walking. Refer to Appendix C for more visualization comparisons.

uate the performance of the method, where a higher score indicates better performance. Compared with other methods, WalkVLM has far surpassed other models in terms of conciseness, both in reminder and QA tasks. In the semantic similarity evaluation against the GT, WalkVLM performs marginally better than GPT-4o in the reminder task but slightly worse in the QA task. The suboptimal performance of WalkVLM in the semantic evaluation of QA tasks, can be attributed to the conciseness of its output, which means that a small amount of output information is difficult to cover all the semantics of the GT.

5.5. Ablative Study

The ablation study of WalkVLM is shown in Table 5 to verify the effectiveness of CoT-based hierarchical planning (CHP) and POLM prior. We conducted three sets of ablation experiments: **(a) w/o CHP.** Remove the CHP mechanism and generate reminder directly based on the input visual information. **(b) w/o Pos Prior.** Remove the approximate position of significant obstacles in POLM. **(c) w/o POLM Prior.** Remove the input filtered target exact location and category. In these experiments, when the CHP mechanism was removed, the model's degradation was significant, which may be due to the model's inability to fully perceive the scene, resulting in the inconsistency between the distribution of generated reminder and the GT distribution. While CHP, enables the model to conduct more detailed analysis from static attributes and scene summaries, thereby obtaining more concise results. For the case of lacking POLM prior, the model's ROUGE performance is worse compared to lacking position prior, indicating that the model relies more on the visual details.

Configuration	TF-IDF	ROUGE-1	ROUGE-2	ROUGE-L
w/o CHP	0.094	0.073	0.007	0.066
w/o Pos Prior	0.151	0.189	**0.062**	0.171
w/o POLM Prior	0.152	0.178	0.056	0.164
Full	**0.166**	**0.191**	**0.062**	**0.173**

Table 5. Ablation study on reminder task. CHP stands for CoT-based hierarchical planning, Pos Prior stands for the general area where obstacles are located in POLM, and POLM Prior stands for the pixel point where the filtered target is exactly located.

6. Conclusion

To fulfill the mission of technology for good, this paper presents WalkVLM, a vision-language model for blind walking task, which employs chain of thought for hierarchical planning to generate concise but focused reminders, and utilizes temporal-aware adaptive prediction to reduce the redundancy of reminders in the time series. Additionally, we have constructed a diverse, extensive, and unbiased walking awareness dataset, aimed at providing a more robust data foundation for this field. Comprehensive experiments show that, compared to other VLM models, WalkVLM can generate more concise reminder and better temporal adaptability when handling video streaming in blind walking task.

7. Limitations

This paper proposes a WAD dataset and systematically establishes the blind walking task based on the vision-language model, thereby setting up an extensive benchmark and offering valuable data support to this field. Although the WAD dataset covers dozens of cities, its generalization capability is still relatively limited in practical applications, making the collection of additional data an essential endeavor. Moreover, we devised the WalkVLM to make the reminders concise and opportune, but still leave considerable room in inference efficiency.

References

[1] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2117–2126, 2013. 1

[2] Santiago Real and Alvaro Araujo. Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. *Sensors*, 19(15):3404, 2019. 1

[3] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. Viassist: Adapting multi-modal large language models for users with visual impairments. *arXiv preprint arXiv:2404.02508*, 2024. 1, 2

[4] Askat Kuzdeuov, Olzhas Mukayev, Shakhizat Nurgaliyev, Alisher Kunbolsyn, and Huseyin Atakan Varol. Chatgpt for visually impaired and blind. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 722–727. IEEE, 2024. 1

[5] Nicholas A Giudice, Benjamin A Guenther, Toni M Kaplan, Shane M Anderson, Robert J Knuesel, and Joseph F Cioffi. Use of an indoor navigation system by sighted and blind travelers: Performance similarities across visual status and age. *ACM Transactions on Accessible Computing (TACCESS)*, 13(3):1–27, 2020. 1

[6] Gaurav Jain, Yuanyang Teng, Dong Heon Cho, Yunhao Xing, Maryam Aziz, and Brian A Smith. " i want to figure things out": Supporting exploration in navigation for people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28, 2023.

[7] Marion A Hersh and Michael A Johnson. *Assistive technology for visually impaired and blind people*, volume 1. Springer, 2008. 1

[8] Haiying Xia, Cong Yao, Yumei Tan, and Shuxiang Song. A dataset for the visually impaired walk on the road. *Displays*, 79:102486, 2023. 1, 2, 4

[9] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2, 4

[10] Zongming Yang, Liang Yang, Liren Kong, Ailin Wei, Jesse Leaman, Johnell Brooks, and Bing Li. Seeway: Vision-language assistive navigation for the visually impaired. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 52–58. IEEE, 2022. 1

[11] Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. *arXiv preprint arXiv:2403.12415*, 2024. 1, 2, 4

[12] Ruiping Liu, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ke Cao, Yufan Chen, Kailun Yang, and Rainer Stiefelhagen. Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1857–1867, 2023. 1, 2

[13] Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. Vialm: A survey and benchmark of visually impaired assistance with large models. *arXiv preprint arXiv:2402.01735*, 2024. 1, 2

[14] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. Emerging practices for large multimodal model (lmm) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv:2407.08882*, 2024. 1

[15] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024. 1

[16] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 1

[17] Zain Merchant, Abrar Anwar, Emily Wang, Souti Chattopadhyay, and Jesse Thomason. Generating contextually-relevant navigation instructions for blind and low vision people. *arXiv preprint arXiv:2407.08219*, 2024. 1, 2, 4

[18] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[19] Maria De Marsico, Chiara Giacanelli, Clizia Giorgia Manganaro, Alessio Palma, and Davide Santoro. Vqask: a multimodal android gpt-based application to help blind users visualize pictures. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, pages 1–5, 2024. 1

[20] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2266–2274, 2024. 2

[21] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 2

[22] Xingli Zhang, Lei Liang, Shenglu Zhao, and Zhihui Wang. Grfb-unet: A new multi-scale attention network with group receptive field block for tactile paving segmentation. *Expert Systems with Applications*, 238:122109, 2024. 2

[23] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. A dataset for crucial object recognition in blind and low-vision individuals' navigation. *arXiv preprint arXiv:2407.16777*, 2024. 2

[24] Wu Tang, De-er Liu, Xiaoli Zhao, Zenghui Chen, and Chen Zhao. A dataset for the recognition of obstacles on blind sidewalk. *Universal Access in the Information Society*, 22(1):69–82, 2023. 2, 4

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[26] Shishun Tian, Minghuo Zheng, Wenbin Zou, Xia Li, and Lu Zhang. Dynamic crosswalk scene understanding for the visually impaired. *IEEE transactions on neural systems and rehabilitation engineering*, 29:1478–1486, 2021. 2

[27] Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885, 2024. 2

[28] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[30] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-ϕ: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024. 2

[31] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024.

[32] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[33] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 2

[34] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[35] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[36] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 6, 7

[37] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2

[38] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. Identifying crucial objects in blind and low-vision individuals' navigation. *arXiv preprint arXiv:2408.13175*, 2024. 4

[39] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3, 5

[40] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5

[42] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 6, 7

[43] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 6, 7

[44] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[46] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[47] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[48] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[49] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003. 6

[50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 6

[51] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6