CVPR
#14729

WalkVLM : Aid Visually Impaired People Walking by Vision Language Model

Supplementary Material

Contents

A. Walking Awareness Dataset

A.1. Data Regional Distribution

Table A1 shows the data distribution and corresponding duration in the WAD dataset. The WAD dataset covers ten cities and contains a wide range of data sources. Figure A1 illustrates the relevant regional distribution. As illustrated, our dataset is spread across Asia and Europe, showing a relatively balanced distribution between different regions. Furthermore, the sampling across different regions is relatively uniform, with a large number of samples at various locations to avoid bias, which has good generalization characteristics.

A.2. Dataset Category Definition

As shown in Table A2, WAD dataset contains multiple predefined data categories. For weather conditions, we have selected the most common types, avoiding scenarios such as rainy or snowy days that make visually impaired people (VIPs) difficult to go outside. For location types, we have selected the types where VIPs are likely to appear, avoiding rare locations. For the traffic flow rating, we instructed annotators to count the number of people in each video segment and used this count as the basis for classification. For scene summarization, during annotation, we required annotators to summarize static attributes such as road conditions, pedestrian flow, and vehicle flow, providing a comprehensive description of the current environment. Currently, the granularity of our dataset is still relatively coarse. In the future, we will continue to refine different fine-grained categories and gradually expand the size of the dataset.

City	Country	Hours
Amsterdam	Netherlands	1:21h
Bangkok	Thailand	2:55h
Chiang Mai	Thailand	1:07h
Istanbul	Turkey	1:08h
Kuala Lampur	Malaysia	1:12h
Singapore	Singapore	1:36h
Stockholm	Sweden	1:06h
Venice	Italy	1:50h
Zurich	Switzerland	1:05h
Beijing	China	2:33h

Table A1. The source region and duration of the WAD dataset. Refer to Fig. A1 for visualization results.

A.3. Annotation Process

We use the page shown in Figure A2 to request annotators to make marks. For static tags, we have provided relevant options for the annotators. For scene summary, we require annotators to describe aspects such as the scene, road conditions, pedestrian flow, and vehicle flow. For reminder and QA, we require annotators to expand on different situations, as described in Section 3.2 of the main paper. Since descriptive tags carry a temporal dimension, we have adopted the annotation method in Table A3 for labeling. After the text categorization is completed, we perform a quality inspection on it and use LLama3.1[1] to normalize the samples that pass the inspection to debias.

A.4. Detection Model

The Detic model [1] has achieved excellent results on the LVIS benchmark [2] in open-world detection tasks by training the detector classifier on image classification data. In view of the model's good generalization ability, we use it to perform preliminary target extraction on the WAD dataset. Figure A3 presents some example of the detection results

[1] https://ai.meta.com/blog/meta-llama-3-1/

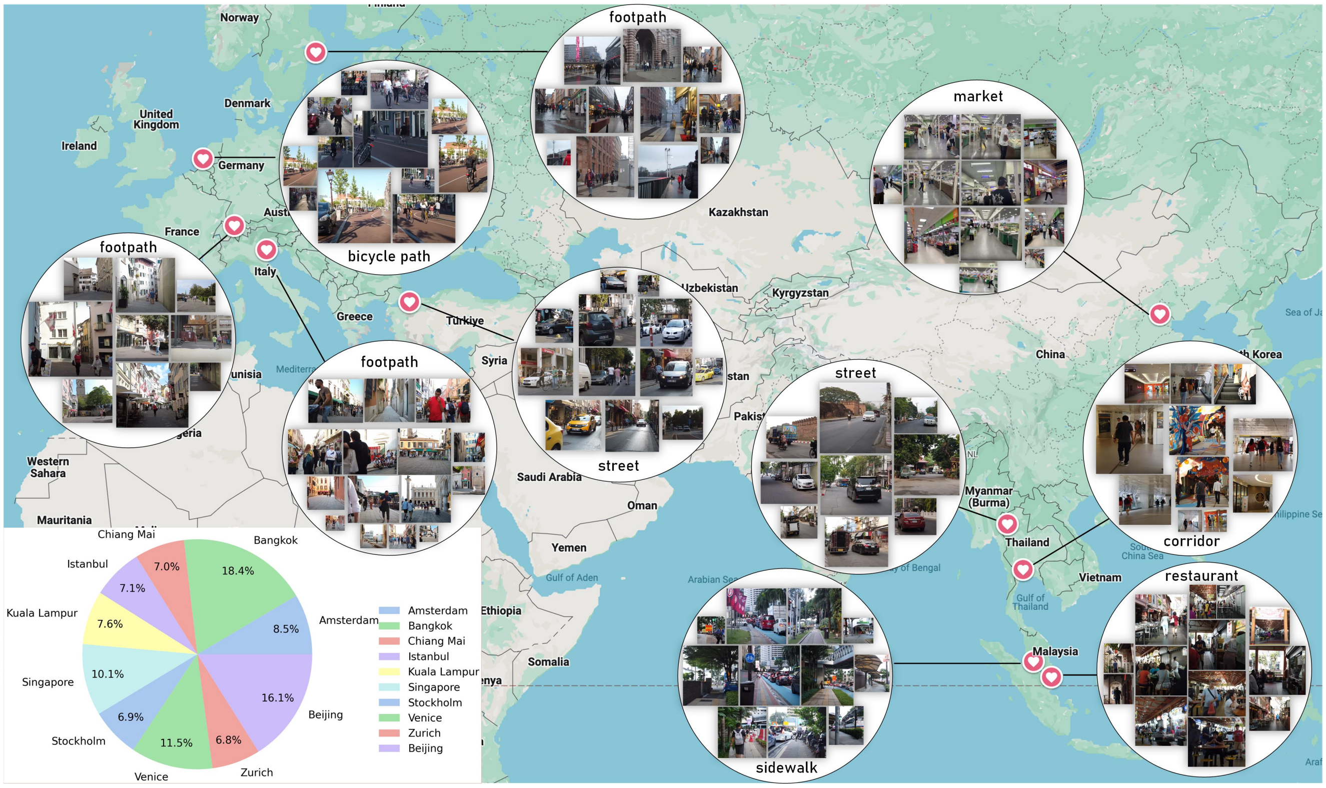


Figure A1. Visualization results of the WAD dataset sorted by region. The WAD dataset has a wide range of sources, and the samples and categories shown are randomly obtained from the dataset. The pie chart in the lower left corner shows the proportion of video length from different regions.

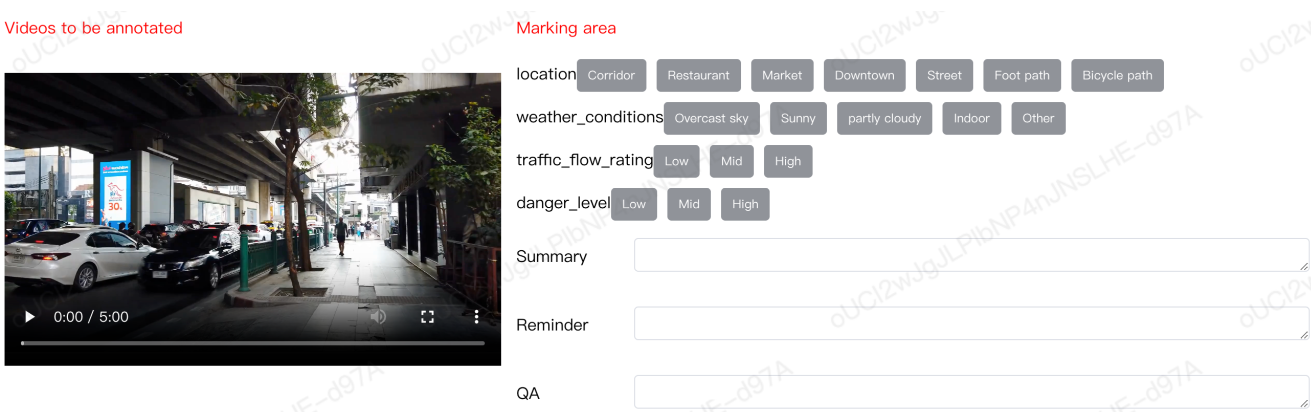


Figure A2. Annotation tool interface. Annotators mark the static attributes of the video in the video, record the time points of reminders and QA, and enter corresponding text descriptions.

of the Detic model in the WAD dataset, demonstrating that the model has a strong ability to extract small and complex targets. After using the model for detection, we conducted manual confirmation and deleted some false positive boxes, thus obtaining the final detection results.

A.5. Sample Visualization

Figure A4 and Figure A5 show more sample visualization results in the WAD dataset. Our dataset has wide coverage, diverse types, and possesses ideal reminder attributes to train VLM to have guiding capabilities in blind walking tasks.

Tag Type	Category	Note
Weather Conditions	Sunny	-
	Night	Not make fine-grained distinctions
	Overcast	-
	Cloudy	-
	Indoor	Not make fine-grained distinctions
	Other	Severe weather conditions such as rain and fog for walking
Location Type	Busy Street	Open-air commercial streets
	Road	Roads where vehicles can travel normally
	Restaurant	Food stalls gathered together, inside large canteens
	Padestrain Path	Walking paths in parks and other places for healthy walking
	Corridor	Indoor walking paths
	Bicycle Lane	Bicycle roads with bicycle signs
	Shopping Mall	Large shopping supermarkets
	Other	Niche scenarios
Traffic Flow Rating	Low	Fewer than 2 people appear in the sliced video
	Mid	Between 2 and 10 people appear in the sliced video
	High	More than 10 people appear in the sliced video
Danger Level	Low	The road is clear, the pedestrian flow is low, and no dangers within 15 steps
	Mid	Other scenarios that do not belong to low or high
	High	Potential collision factors, such as narrow roads, bumpy roads, vehicle warnings
Scene Description	-	Detailed description of the current environment, level of danger, and pedestrian flow
QA	-	The three types of inquiries mentioned in the paper and concise responses
Reminder	-	Brief walking directions to provide to the user based on the current scenario

Table A2. The interpretation of label categories contained in the WAD dataset.

Figure A3. The detection results provided in the WAD dataset, which were pre-detected by the Detic model [1], and then manually reviewed to ensure the correctness of the results. See here for more detection samples.

A.6. Data Analysis

Figure A7 shows the distribution of the top 100 categories contained in the WAD dataset, while Table A4 shows all the categories included. Figure A6 presents a word cloud distribution with annotated descriptions, where the most frequently used words include *oclock*, *pedestrain*, *direction*. We have counted the word count distribution in different annotated texts in Figure A8. For reminder and QA scenarios, the data contained in WAD is shorter in length, while for summary scenario descriptions are more detailed.

A.7. Benchmark Data Splits

To ensure the diversity of test data, we adopted a category-based combined clustering method. Through this method, we carefully selected a certain number of samples from the

Weather condition: Sunny **Area type:** Street **Danger level:** Low **Traffic flow rating:** Mid

Summary: The current road is flat. There is a parked vehicle five steps to the left. There is a huge building on the right. There are trees at twelve o'clock. There is a huge building behind the trees. The current weather is sunny and there are many pedestrians on the current road. There is a yellow billboard above the position ten steps to the left.

Q: How long to the intersection?

A: Go straight about twenty steps to reach, there are a row of iron piers at 11 o'clock direction, and there are two iron pillars on the ground right in front.

Weather condition: Night **Area type:** Street **Danger level:** Mid **Traffic flow rating:** Mid

Summary: In the evening on a street, there are trees and street lamps on the right. The sky is dim and the street lamps are on. About 20 steps at one o'clock, there are several passersby. About 30 steps in front, there are several passersby and a car. There are some shops on both sides of the road. About 30 steps at one o'clock, there is a building with colored lights. The road surface is flat but slippery.

Q: How to go?

A: Go towards 11 o'clock, avoiding the street lights and trees in the middle of the walkway.

Weather condition: Sunny **Area type:** Pedestrian Path **Danger level:** Low **Traffic flow rating:** Low

Summary: On a walking path, there are rows of cars parked in front and on the right. On the silver-grey car on the right, there is a black-and-white cow cat squatting. About 15 steps ahead, there is a pedestrian in white. On the left, there is a row of metal railings, and inside the railings are residential buildings. The road is narrow, there are few pedestrians, and no vehicles passing.

Q: Current road situation.

A: The current road is a small road with cars parked on both sides. There are few pedestrians on the road. Now, walk to the left side of the road, which is relatively narrow. Be careful to avoid the cars parked on the right side and pay attention to safety.

Figure A4. Visual examples of QA samples in WAD dataset. See here for dynamic samples.

clustering results to form our test set. Ultimately, we selected 1007 reminders and 134 QA pairs as our testset. Furthermore, we conducted a thorough analysis of the distribution of the test set to confirm that they are accurate and that the same type of data is represented in the training set.

B. Model & Details

B.1. All Prompts Used in Paper

Table A5 displays all the prompts utilized in this paper under various circumstances such as normalizing annotation

CVPR
#14729

Weather condition: Sunny **Area type:** Other **Danger level:** Mid **Traffic flow rating:** Mid

Summary: Right side is glass door, 10 steps away on the right side there is a cart, and there are pedestrians ahead.

Reminder: There is a glass wall in front, be careful to avoid it.

Weather condition: Overcast **Area type:** Pedestrian Path **Danger level:** Mid **Traffic flow rating:** Mid

Summary: Walking on a stone bridge. The left side is red, and the right side is paved with stone bricks. There are stone railings on both sides of the bridge. There are many trees below and on both sides of the bridge. There is a pedestrian in a black coat about five steps ahead. The large flow of people is mainly concentrated about fifteen steps ahead. There is no road nearby, and the traffic flow is zero.

Reminder: at 10 o'clock direction, there are pedestrians passing by. please move slowly towards 11 o'clock direction.

Weather condition: Sunny **Area type:** Pedestrian Path **Danger level:** Mid **Traffic flow rating:** Low

Summary: On the sidewalk on the right side of the road, there is a downward step on the left. A yellow car passes on the left side of the road. There is a row of green plants on the right. There is a row of trees at the one o'clock direction. There is a sign at the two o'clock direction. There are cars parked on the roadside at the eleven o'clock direction. The current road is narrow, and there are few pedestrians.

Reminder: At 11 o'clock direction there is a car, at 1 o'clock direction there is a sign, be careful to avoid.

Weather condition: Sunny **Area type:** Pedestrian Path **Danger level:** High **Traffic flow rating:** Mid

Summary: On a sidewalk, there is a telegraph pole at 10 o'clock, and a billboard in front, about to be hit. At 10 o'clock on the left, two pedestrians are pushing items forward. A takeaway motorcycle is parked on the roadside at 1 o'clock. As the lens moves forward, a couple are walking hand in hand on the sidewalk at 10 o'clock. At 1 o'clock, about five steps away, there is an electric box and a telegraph pole. There is a stall at 11 o'clock. Many cars are parked on the right side of the road waiting to pass. The road is narrow with many roadblocks.

Reminder: At 11 o'clock direction, there are pedestrians passing by. about five steps in front, there are telegraph poles and electric boxes. be careful to avoid.

Figure A5. Visual examples of reminder samples in WAD dataset. See here for dynamic samples.

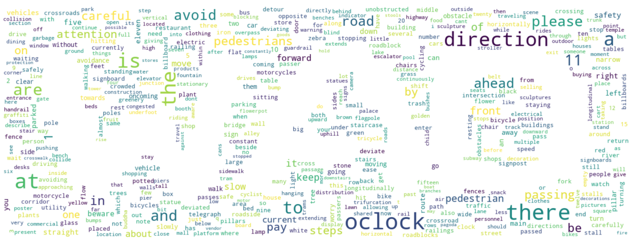


Figure A6. Word cloud distribution of the description in Walking Awareness Dataset.

...
$\langle 2m30s - AE \rangle$
almost hit the wall, go forward in the 11 o'clock direction to return to the main route.
$\langle 2m43s - B \rangle$
five steps ahead is the fork in the road, go forward in the 10 - o'clock direction to return to the main route.
...
$\langle 3m39s - O \rangle$
Q: describe the current scene
A: at a crossroads with many vehicles, keep still to avoid, there are some obstacles ahead, be careful to avoid
...

Table A3. Example of reminder and QA result annotation with a temporal dimension. We required annotators to mark the time when events occurred in the video, the question and reminder categories, as well as concise responses.

results, reasoning with VLM, and conducting evaluations. Normalize the annotation results are crucial for ensuring the consistency and uniformity of annotation results, and this prompt are used in the preprocessing stage to correct bias in the data. For the inference prompt of other models, we input historical multi-frame images and historical states to enable it to generate trigger states and reminders for the user. In the prompt of WalkVLM, we make the model predict different levels of labels step by step and gradually output the results. The evaluation prompt based on GPT4 compares different results with the ground truth to obtain the proportion statistics of the optimal model.

B.2. Evaluation of Temporal Redundancy F1-Score

This section systematically evaluates the redundancy of temporal outputs of different models. Temporal redundancy refers to the excessive frequency of output information in this paper. In order to evaluate the temporal redundancy of different models, we decompose the test video to ensure that each sample contains historical N frames and N states, thereby predicting the trigger state under the current situation. We collected 834 such samples as a test set. The predicted labels are divided into three levels, corresponding to the degree of danger. When the degree of danger is high, we regard it as triggering VLM. By comparing the predicted different states with the ground truth, the distribution gap between the two sets of data can be calculated, thereby calculating the F1-score.

C. Experiment

C.1. Visualization of Hierarchical Reasoning

We have demonstrated the results of hierarchical reasoning using WalkVLM in Figure A9. WalkVLM can effectively extract static attributes from video streams and generate a comprehensive summary of the current scene. After integrating fragmented attributes, the model produces concise and informative walking instructions.

C.2. Visual Comparison of Different Models

Figure A10 and A11 presents a comparison of additional visualization results between WalkVLM and other models. Our approach yields more streamlined results, enabling a superior human-machine interaction experience during blind walking task.

C.3. Comparison of Video Streaming Inference

In this section, we deployed WalkVLM and MiniCPM-V2.6 [3] on cloud devices to verify the differences in performance between the two models in real-world scenarios. The visualization results of the two models on the video stream can be viewed here, where WalkVLM is capable of generating less temporal redundancy. As shown in A12, on real-time video streams, for two models with the same size parameters, WalkVLM can generate more concise and accurate walking guidance.

However, the current model still has certain **limitations** in practical applications. **Firstly**, the model has a weak ability to prioritize events, making it difficult to identify the most urgent actions that need reminders in the scene. Facing this issue, our next attempt is to establish an event priority model that enables the model to propose necessary events and obtain priority results through ranking. **Second**, the model still has certain misjudgments in obstacle recognition and direction. Going forward, we will attempt to inject more prior knowledge about obstacles into the model and try to design some rule-based methods to verify the output of WalkVLM, so as to enhance its usability. **Thirdly**, there is still significant room for improvement in the model's recognition of fine-grained obstacles. We believe that this can be compensated for by collecting more available data.

Although there are the aforementioned shortcomings, compared to other models, WalkVLM has made a solid advancement in the blind walking task. We will continue

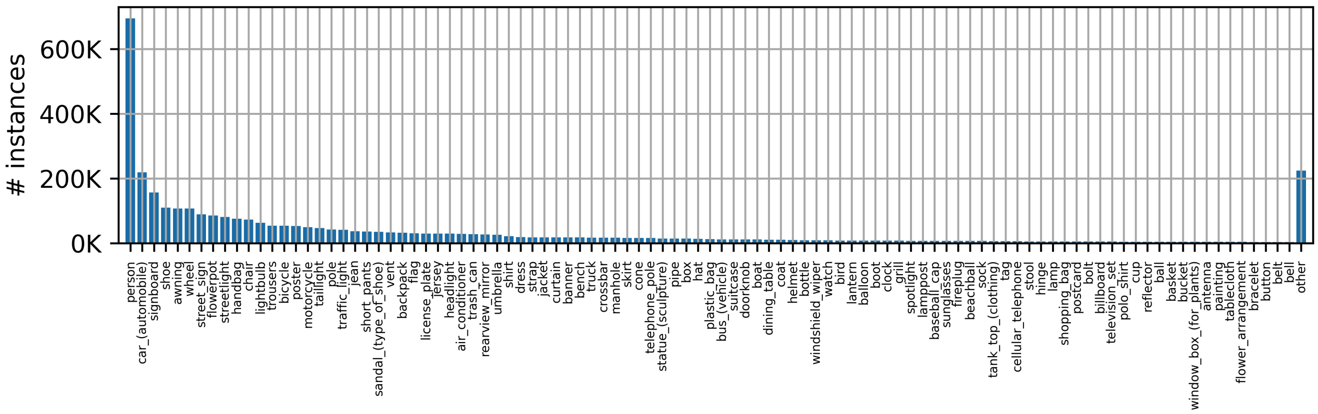
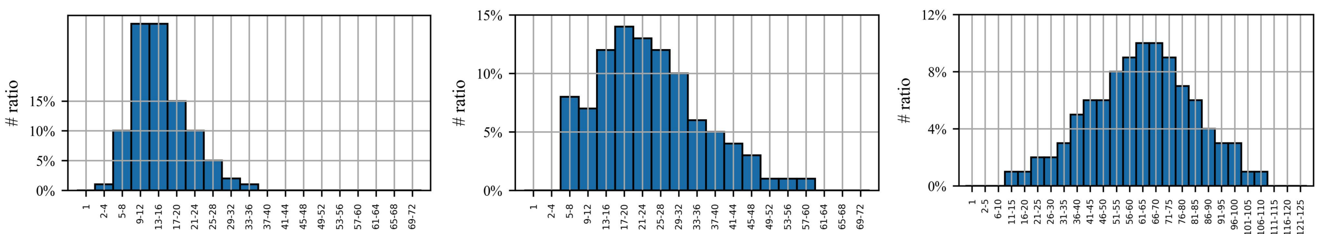


Figure A7. Detect target distribution. For clarity, display the top 100 with the highest frequency of occurrence.



(a) Reminder (b) QA (c) Scene summary

Figure A8. Data length distribution in different text annotation types.

to iterate on this model to further enhance its usability in real-world scenarios!

D. Discussion

In the context of the increasingly popular vision-language model field, it is crucial to explore how to use it to address the daily challenges faced by visually impaired patients. Our work on the WalkVLM model and the walking awareness dataset represents a significant step in this direction, aiming to empower individuals with visual impairments through advanced technological solutions.

One of the most rewarding aspects of this research has been the opportunity to apply cutting-edge AI research to a problem that has profound real-world implications. We are deeply committed to leveraging technology to enhance the quality of life for everyone, and our work on WalkVLM exemplifies this mission. By providing a tool that can offer more accurate and context-aware guidance, we hope to make a tangible difference in the lives of blind individuals, enabling them to navigate their environments with greater independence and confidence.

However, we also recognize that our current approach has several limitations that need to be addressed to fully realize its potential. One major limitation is the geographical scope of our dataset, which currently covers only Europe and Asia. To develop a truly global solution, we need to expand our data collection efforts to include a wider range of regions and environments. This will ensure that our model can adapt to the diverse conditions and challenges faced by blind individuals around the world.

Another important consideration is the need for more real-time capabilities in our model. While WalkVLM offers significant advancements in understanding and interpreting walking-related data, achieving rapid inference is essential for practical applications. Real-time processing allows for immediate feedback and adjustments, which are critical for ensuring the safety and effectiveness of assistive technologies.

Additionally, integrating Retrieval-Augmented Generation (RAG) techniques [4, 5] could further enhance the information provided by our model. By combining WalkVLM with RAG, we can incorporate a broader range of perspectives and data sources, leading to more informative and contextually relevant responses. This approach not only improves the accuracy and utility of our model but also fosters a more dynamic and interactive user experience.

In conclusion, while our work on WalkVLM achieves a significant advancement in the field of assistive technologies for the visually impaired, there is still much to be done. By addressing the limitations mentioned above and continuing to innovate, we hope to build on our current achievements and contribute to a future where technology empowers indi-

person, car-(automobile), signboard, shoe, awning, wheel, street-sign, flowerpot, streetlight, handbag, chair, lightbulb, trousers, bicycle, poster, motorcycle, taillight, pole, traffic-light, jean, short-pants, sandal-(type-of-shoe), vent, backpack, flag, license-plate, jersey, headlight, air-conditioner, trash-can, rearview-mirror, umbrella, shirt, dress, strap, jacket, curtain, banner, bench, truck, crossbar, manhole, skirt, cone, telephone-pole, statue-(sculpture), pipe, box, hat, plastic-bag, bus-(vehicle), suitcase, doorknob, boat, dining-table, coat, helmet, bottle, windshield-wiper, watch, bird, lantern, balloon, boot, clock, grill, spotlight, lamppost, baseball-cap, sunglasses, fireplug, beachball, sock, tank-top-(clothing), tag, cellular-telephone, stool, hinge, lamp, shopping-bag, postcard, bolt, billboard, television-set, polo-shirt, cup, reflector, ball, basket, bucket, window-box-(for-plants), antenna, painting, tablecloth, flower-arrangement, bracelet, button, belt, bell, baby-buggy, flagpole, ladder, bowl, spectacles, vase, clock-tower, blouse, book, stop-sign, handle, banana, refrigerator, toy, sunhat, beanie, doughnut, necklace, train-(railroad-vehicle), bottle-cap, fan, tarp, vest, crate, orange-(fruit), magazine, apple, skateboard, parking-meter, postbox-(public), necktie, dog, earring, vending-machine, sweatshirt, barrel, lampshade, chandelier, cowboy-hat, minivan, newsstand, choker, hook, dish-antenna, scarf, camera, pizza, mask, drawer, weathervane, figurine, motor-scooter, magnet, pigeon, speaker-(stero-equipment), cart, cooler-(for-food), blackboard, roller-skate, hot-air-balloon, flip-flop-(sandal), unicycle, headscarf, cabinet, hatbox, mirror, legging-(clothing), candle, satchel, teddy-bear, lanyard, log, glove, pennant, wall-socket, shower-cap, blinker, canister, pottery, robe, gargoyle, steering-wheel, newspaper, suspenders, dumpster, water-bottle, easel, kite, cushion, apron, horse, wreath, pew-(church-bench), dispenser, tomato, towel, melon, pumpkin, doormat, fire-extinguisher, sombrero, walking-cane, can, telephone-booth, thermostat, wineglass, heart, bandanna, tambourine, cat, jar, peach, carton, ring, frisbee, pot, carrot, watering-can, surfboard, mailbox-(at-home), headband, buoy, coconut, hose, card, sweater, lemon, remote-control, butterfly, grape, plate, knob, gravestone, knocker-(on-a-door), elephant, globe, mast, paper-plate, raincoat, wristlet, projector, watermelon, tote-bag, pirate-flag, mail-slot, tray, bulletproof-vest, brass-plaque, handcart, table, tricycle, towel-rack, laptop-computer, belt-buckle, fire-alarm, bow-(decorative-ribbons), slipper-(footwear), sink, papaya, sawhorse, briefcase, glass-(drink-container), cake, latch, coat-hanger, step-stool, fish-(food), napkin, pastry, motor, shopping-cart, sofa, silo, doll, toilet, tank-(storage-vessel), cookie, crucifix, oven, bamboo, tassel, hairnet, golfcart, fish, bread, cow, monitor-(computer-equipment) computer-monitor, lion, seashell, microwave-oven, earphone, Christmas-tree, water-jug, wagon-wheel, airplane, locker, broom, calendar, pop-(soda), barrette, mammoth, Rollerblade, avocado, blazer, scoreboard, hippopotamus, birdbath, shield, rubber-band, paper-towel, music-stool, straw-(for-drinking), poncho, neckerchief, pinwheel, houseboat, crutch, green-bean, birthday-card, sunflower, pickup-truck, grocery-bag, wine-bottle, faucet, halter-top, wine-bucket, sandwich, life-buoy, basketball-backboard, bullhorn, aerosol-can, tapestry, toilet-tissue, bathtub, tripod, goldfish, gourd, fireplace, stepladder, orange-juice, edible-corn, oil-lamp, garden-hose, potato, shower-curtain, water-tower, knife, onion, apricot, tennis-racket, piggy-bank, ashtray, puppet, sculpture, pretzel, fedora, brassiere, milk-can, cantaloup, blimp, blanket, guitar, kiwi-fruit, brake-light, armor, shawl, scissors, table-tennis-table, toothbrush, birdcage, lettuce, cylinder, radiator, turban, kimono, birdhouse, slide, envelope, Dixie-cup, Ferris-wheel, microphone, swimsuit, lime, beer-bottle, shaving-cream, fishbowl, ice-skate, camper-(vehicle), hairpin, pillow, underwear, oar, bonnet, chinaware, cymbal, penguin, sausage, strawberry, costume, dishtowel, gull, sword, bagel, spoon, crown, harmonium, duffel-bag, candle-holder, camcorder, horse-buggy, jumpsuit, clothes-hamper, knee-pad, bathrobe, comic-book, beer-can, giant-panda, map, phonograph-record, bell-pepper, toolbox, solar-array, rhinoceros, booklet, cupcake, shower-head, binoculars, monkey, matchbox, hand-towel, deer, pan-(for-cooking), dove, wheelchair, armoire, camel, goose, hair-dryer, dress-hat, tiger, tennis-ball, place-mat, bridal-gown, ottoman, cornice, mug, pear, sail, boxing-glove, passenger-car-(part-of-a-train), cap-(headwear), horse-carriage, urn, wig, wind-chime, thermos-bottle, fume-hood, crock-pot, bubble-gum, cherry, drum-(musical-instrument), wagon, bed, clarinet, eyepatch, tissue-paper, padlock, cigarette, parasol, baseball-bat, teacup, mandarin-orange, aquarium, bun, bowling-ball, telephone, lemonade, dog-collar, windmill, saltshaker, tartan, zucchini, lab-coat, tinsel, radar, pitcher-(vessel-for-liquid), pug-dog, sheep, coffee-maker, folding-chair, pinecone, visor, octopus-(animal), medicine, cassette, yogurt, saddlebag, wardrobe, basketball, persimmon, tape-(sticky-cloth-or-paper), tights-(clothing), baseball-glove, water-heater, cauliflower, cover, garbage-truck, forklift, bath-mat, chopping-board, computer-keyboard, propeller, wristband, gift-wrap, duck, railcar-(part-of-a-train), violin, football-helmet, blueberry, chopstick, piano, starfish, lawn-mower, fork, diaper, frying-pan, shark, wallet, duct-tape, pineapple, elk, toaster, earplug, wall-clock, cab-(taxi), zebra, bow-tie, hog, mallet, boiled-egg, knitting-needle, keycard, condiment, dragonfly, garlic, pepper-mill, drumstick, snowman, thumbtack, gasmask, pouch, teapot, sling-(bandage), barrow, bulldozer, spear, bookmark, mat-(gym-equipment), coffee-table, sleeping-bag, bat-(animal), runner-(carpet), iron-(for-clothing), bath-towel, coatrack, musical-instrument, bulletin-board, pie, tinfoil, overalls-(clothing), bib, pelican, egg, mascot, cistern, bookcase, giraffe, pad, trench-coat, bandage, chalice, flannel, clipboard, dustpan, celery, sweet-potato, headset, bread-bin, bowler-hat, walking-stick, saddle-blanket, phonebook, seahorse, clasp, lollipop, desk, broccoli, nailfile, anklet, dress-suit, rag-doll, beanbag, gondola-(boat), bear, mushroom, cider, dishwasher, alcohol, clementine, flap, rifle, icecream, ski, snowboard, vacuum-cleaner, automatic-washer, trailer-truck, hamper, television-camera, cigar-box, tobacco-pipe, bouquet, candy-bar, ferry, bead, banjo, ladybug, pacifier, shovel, control, fishing-rod, cruise-ship, washbasin, whipped-cream, pen, goggles, pan-(metal-container), flipper-(footwear), cucumber, nightshirt, dolphin, water-cooler, cloak, mop, pendulum, canoe, artichoke, heater, hammock, water-gun, almond, paintbrush, shredder-(for-paper), pita-(bread), liquor, eggbeater, scale-(measuring-instrument), dresser, ski-boot, cigarette-case, teakettle, armband, frog, file-cabinet, tow-truck, squid-(food), mouse-(computer-equipment), keg, tongs, deadbolt, quesadilla, hair-curler, koala, asparagus, platter, bobbin, coaster, milk, inhaler, salami, flamingo, life-jacket, coffeepot, urinal, eggplant, business-card, mattress, fig-(fruit), corkboard, raft, cash-register, cabana, suit-(clothing), kitchen-table, corset, gorilla, cocoa-(beverage), yacht, salmon-(fish), spice-rack, parachute, coil, squirrel, ironing-board, projectile-(weapon), coverall, trophy-cup, thread, measuring-stick, dinghy, crowbar, ski-pole, trunk, salad, dartboard, bedpan, award, rabbit, cincture, parka, colander, windsock, home-plate-(baseball), baboon, green-onion, eclair, toothpaste, saucer, highchair, handkerchief, pajamas, saxophone, potholder, ladle, spatula, first-aid-kit, veil, parakeet, scrubbing-brush, clip, blender, stapler-(stapling-machine), parrot, measuring-cup, owl, ice-maker, sweat-pants, videotape, corkscrew, marker, muffin, tiara, cast, beret, gun, tape-measure, generator, cowbell, sushi, hookah, seabird, crow, tachometer, cream-pitcher, battery, alligator, spider, Band-Aid, lightning-rod, hamburger, elevator-car, checkbook, hockey-stick, syringe, beeper, gelatin, wrench, water-scooter, hornet, fire-hose, Lego, stove, key, palette, chicken-(animal), deck-chair, chaise-longue, hairbrush, flashlight, smoothie, mitten, flute-glass, crab-(animal), bagpipe, clothespin, soap, lizard, river-boat, boom-microphone, radish, paperweight, fire-engine, candy-cane, bow-(weapon), sponge, wedding-cake, hourglass, ice-pack, tea-bag, cappuccino, eagle, machine-gun, salmon-(food), wet-suit, clutch-bag, cube, brussels-sprouts, wolf, toothpick, kennel, soccer-ball, prawn, hamster, identity-card, egg-yolk, pegboard, honey, duckling, pencil, ham, saddle-(on-an-animal), gameboard, hot-sauce, amplifier, alarm-clock, tortilla, manatee, brownie, nutcracker, popsicle, funnel, hotplate, trampoline, crib, heron, shampoo, butter, army-tank, date-(fruit), bottle-opener, cornet, camera-lens, jelly-bean, griddle, atomizer, armchair, bass-horn, hummingbird, salsa, baguet, sweatband, arctic-(type-of-shoe), footstool, power-shovel, drone, tractor-(farm-equipment), bunk-bed, food-processor, radio-receiver, cufflink, scarecrow, cock, cougar, chocolate-cake, wok, raspberry, ping-pong-ball, blackberry, dollhouse, space-shuttle, skewer, bobby-pin, school-bus, puffin, car-battery, razorblade, stirrup, drill, truffle-(chocolate), fighter-jet, thermometer, cupboard, screwdriver, sled, eel, pipe-bowl, broach, plume, sofa-bed, ferret, turtle, escargot, crescent-roll, printer, quilt, chocolate-bar, paddle, toaster-oven, motor-vehicle, puffer-(fish), soya-milk, cork-(bottle-plug), cabin-car, walrus, patty-(food), police-cruiser, skullcap, baseball, handsaw, Sharpie, stagecoach, cape, receipt, notebook, rib-(food), paperback-book, perfume, ballet-skirt, stirrer, steak-(food), telephoto-lens, barbell, record-player, mound-(baseball), dental-floss, sparkler-(fireworks), microscope, strainer, wooden-leg, dish, peeler-(tool-for-fruit-and-vegetables), hammer, milkshake, detergent, octopus-(food), limousine, chessboard, Tabasco-sauce, curling-iron, convertible-(automobile), underdrawers, freight-car, dalmatian, notepad, seaplane, burrito, dishrag, packet, birthday-cake, binder, wooden-spoon, pool-table, sewing-machine, pitchfork, cardigan, crayon, manger, kettle, CD-player, barge, flash, rolling-pin, cleansing-agent, dagger, waffle, hardback-book, toast-(food), puppy, egg-roll, chili-(vegetable), kitchen-sink, chocolate-mousse, router-(computer-equipment), pencil-sharpener, pin-(non-jewelry), kayak, sharpener, grater, nut, shoulder-bag, pantyhose, plow-(farm-equipment), mint-candy, crisp-(potato-chip), needle, pea-(food), beef-(food), sherbert, pepper, iPod, bullet-train, polar-bear, headboard, volleyball, bulldog, crape, reamer-(juicer), birdfeeder, table-lamp, pocketknife, jewelry, meatball, pudding, hand-glass, Bible, money, stylus, sugarcane-(plant), cayenne-(spice), shepherd-dog, lip-balm, soup-bowl, cornbread

Table A4. Full list of the target categories present in the walking awareness dataset, sorted by the number of occurrences in the dataset.

viduals with visual impairments to lead more independent and fulfilling lives. Our commitment to this cause remains unwavering, and we look forward to the next steps in this journey!

E. Societal Impact

Our contribution extends beyond the realm of technological advancement, offering significant societal benefits that can greatly improve the quality of life for visually impaired individuals. By introducing the WalkVLM model and the accompanying walking awareness dataset, we are taking a substantial step towards enhancing the independence and safety of blind individuals as they navigate through their daily environments.

Firstly, the WalkVLM model and dataset address a critical need for more accessible and effective assistive technologies

Application Scenario	**Models**	**Input Prompt**
Normalize the annotation results	Llama 3.1	*Please normalize the following manually annotated output to reduce information redundancy and maintain as standardized an output as possible. During the processing, please follow these guidelines:* *a. Convert all labels to lowercase.* *b. Remove any superfluous spaces or special characters.* *c. Retain the accurate position of objects in the sentence, such as what time or how many steps away.* *d. For similar or repetitive semantic annotations, reduce the redundancy of semantics.* *e. Output the result directly.* *The annotated text is as follows:* ...
Instruct VLM to provide guidance on blind walking based on the provided video	DeepSeek (1.3B&7B) Yi-VL (6B) MiniCPM-V2.6 (8B) GPT-4o Qwen2-VL (7B)	”request”: ”format the sentence below into the format, given in English”, ”restriction”: ”answer in json like the format given below without code block”, ”sentence”: ”Analyze the following video frames and determine the danger level for a blind person if they were to walk straight ahead. The danger levels are categorized as follows:: Low (open areas, with few people or obstacles): Mid (moderate danger, such as areas with some obstacles or moderate activity): High (high danger, such as narrow pathways, crowded areas, or busy roads)are the danger levels for the first two frames:1:{history_states[0]}2:history_states[1]}on this information and the provided image below, please provide the danger level for Frame 3.that danger level are indicated by single letters only. (A, B, or C)., provide walking instructions based on the provided image to ensure safe navigation.”, ”format”: { ”data”: { ”Frame 3 Danger Level”: ”string (A, B, or C)”, ”Walking Guidance”: ”string” } } ...
Training and inference of WalkVLM	WalkVLM	”request”: ”format the sentence below into the format, given in English”, ”restriction”: ”answer in json like the format given below without code block”, ”sentence”: ”You are now a guide. I can’t see the path and will be walking solely based on your instructions. Each input frame displays the road information ahead. The main objects in each image are { json_str }. Please provide clear and unobstructed walking directions. Describe in order: 1. Location (e.g., corridor, restaurant, market, downtown, street, foot path, bicycle path), 2. Weather conditions (e.g., overcast sky, sunny, partly cloudy, indoor), 3. Traffic flow rating (e.g., low: 0-4 people/minute, medium: 4-10 people/minute, high: 10+ people/minute), 4. Describe the overall scene based on the input images and all the information from the above three points, 5. Please guide me on how to proceed based on the input images and all previous descriptions.”, ”format”: { ”data”: { ”1. Location”: ”string”, ”2. Weather conditions”: ”string”, ”3. Traffic flow rating”: ”string”, ”4. Describe the overall scene in the image”: ”string”, ”5. Instructions on how I should proceed”: ”string” } } ...
Use LMM to evaluate the similarity between generated results and ground truth	GPT4	*Please act as an impartial judge and evaluate the quality of the responses provided by multiple assistants displayed below. You should choose the assistant that matches the GT answer. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not favor certain names of the assistants. Be as objective as possible. The answer should be the most closest to the semantics of the GT result and have the most concise answer. After providing your explanation, strictly follow the following format to output your final verdict: if assistant A is better, output ”[[A]]”, if assistant B is better, output ”[[B]]”, and similar formats for other answers. Request you select a relatively optimal result and directly output the option.* {*GT*} {} *The Start of Assistant A’s Answer* {} *The End of Assistant A’s Answer* *The Start of Assistant B’s Answer* {} *The End of Assistant B’s Answer* ... *The Start of Assistant G’s Answer* {} *The End of Assistant G’s Answer*

Table A5. All prompts utilized in this paper.

for the visually impaired. Traditional navigation aids often fall short in providing the necessary real-time information and adaptability required for complex environments. Our model, with its advanced capabilities in understanding and interpreting walking-related data, can offer more precise and context-aware guidance, thereby reducing the risks associated with independent travel.

Moreover, the dataset we have compiled is a valuable resource that can foster further research and development in the field of assistive technologies. By making this dataset publicly available, we encourage collaboration and innovation among researchers, leading to the creation of even more

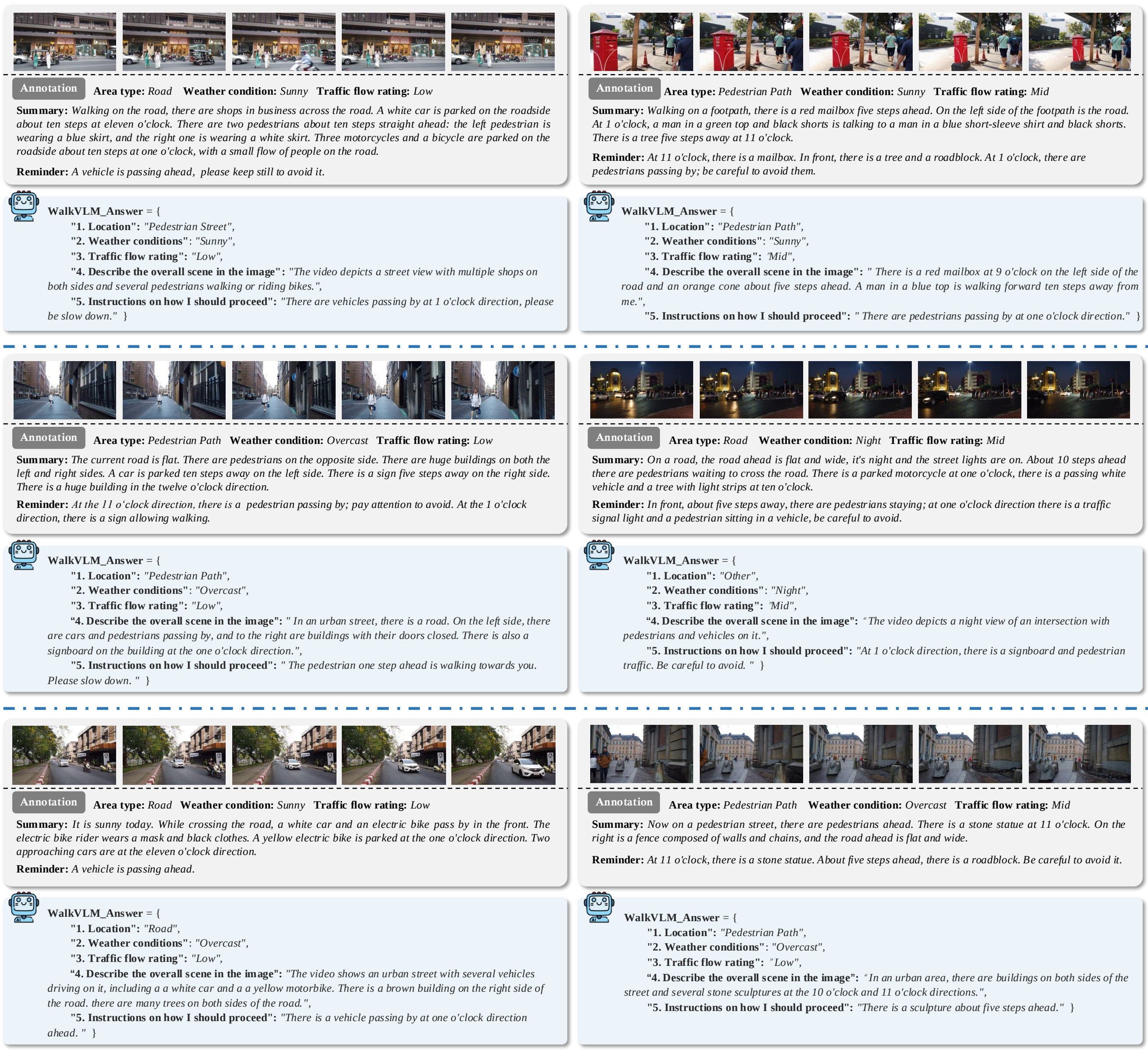


Figure A9. Hierarchical inference results visualization of WalkVLM.

sophisticated solutions that can cater to the diverse needs of blind individuals. This collaborative effort can ultimately result in technologies that are not only more effective but also more widely adopted and accessible.

From an educational standpoint, our work can also play a pivotal role in raising awareness about the challenges faced by the visually impaired community. By showcasing the potential of AI and machine learning in addressing these challenges, we hope to inspire more individuals and organizations to contribute towards creating a more inclusive society. This increased awareness can lead to more supportive policies and initiatives that focus on improving the quality of life for the visually impaired.

Additionally, the WalkVLM model and dataset have the potential to impact various industries beyond assistive technologies. For instance, they can be adapted for use in smart city planning, where understanding pedestrian behavior and safety is crucial. This broader application can lead to safer and more accessible urban environments for everyone, not just the visually impaired.

In summary, our contribution not only advances the state of the art in AI and machine learning but also has far-reaching societal implications. By providing a robust benchmark and a rich dataset, we are paving the way for innovative solutions

Annotation

Weather Condition: Overcast **Area Type:** Street **Traffic Flow Rating:** High **Danger Level:** High

Summary: The current road is flat. There are pedestrians in the same and opposite directions, and there are many pedestrians on the current road. Five steps to the left is a clothing store. At twelve o'clock is an orange guardrail. In front of the guardrail is a highway with cars running on it. On the left side of the highway is a huge building.

Q: Which direction should I walk?

A: Please keep moving in the one o'clock direction.

Ours

Walking towards 1 o'clock direction.

MiniCPM

You should **walk forward**, as the people in the video are moving in that direction.

GPT-4o

To continue walking, head straight along the path in the last image. This will take you past the people and towards the area with the **orange barriers** and the **building** in the background.

Qwen

Based on the image, it appears that you are already walking in the direction you need to go. The people in the background are moving in the same direction as you, and the storefronts and signposts are also facing in the same direction. Therefore, you should **continue walking in the same direction.**

Annotation

Weather Condition: Sunny **Area Type:** Other **Traffic Flow Rating:** High **Danger Level:** High

Summary: On the sidewalk, there are pedestrians in black tops and white pants at 11 o'clock, and in mint-green tops and grey pants in front walking. A child in a black short-sleeve and jeans is standing in front of a toy. At 1 o'clock, there is a grocery store. Five steps away, there is an ice cream cabinet at 2 o'clock. A man wearing gray short-sleeves and white pants and a woman wearing a white dress are picking out ice cream. There are also many pedestrians on the current road.

Q: Describe the scene ahead for me.

A: There are pedestrians passing straight ahead. At 1 o'clock direction, a pedestrian is buying ice cream on the roadside. On both sides of the road, there are rows of trees except for ice cream vendors.

Ours

In front there is an **outdoor food stall with some pedestrians buying food around it.** On the left side of the road there is a also passing by.

GPT-4o

Based on the provided images from the video, The scene depicts **a pathway lined with trees** on the left side and **small stalls or shops** on the right. The pathway is relatively narrow and appears to be part of an outdoor market or fair. Various items are displayed for sale at the stalls, and there are a few **people** walking along the pathway, some stopping to look at the items. The weather seems pleasant with a clear sky, and the overall atmosphere is casual and relaxed.

MiniCPM

The scene ahead depicts a a bustling market or street fair environment. There are several **stalls** with various items on display, such as toys and possibly food-related products. The presence of **people** browsing the stalls indicates active commerce and engagement with the offerings. The setting appears to be **outdoors**, with **trees** providing shade and contributing to a relaxed atmosphere typical of such venues.

Qwen

The video shows a group of **people** walking down **a path outdoors**. The setting appears to be a park or a similar recreational area, with **trees and greenery** visible in the background. There is a building or structure on the right side of the image, and ~~a vehicle parked nearby~~. The people are carrying bags and seem to be engaged in some activity, possibly related to outdoor recreation or a group outing. The overall atmosphere is casual and relaxed.

Text in red is relevant objects, and ~~Text~~ with strikethrough is irrelevant or erroneous descriptions.

Figure A10. More QA visualization comparisons of different models.

that can significantly enhance the lives of blind individuals and promote a more inclusive society.

References

[1] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[3] 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. 6

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 7

[5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 7

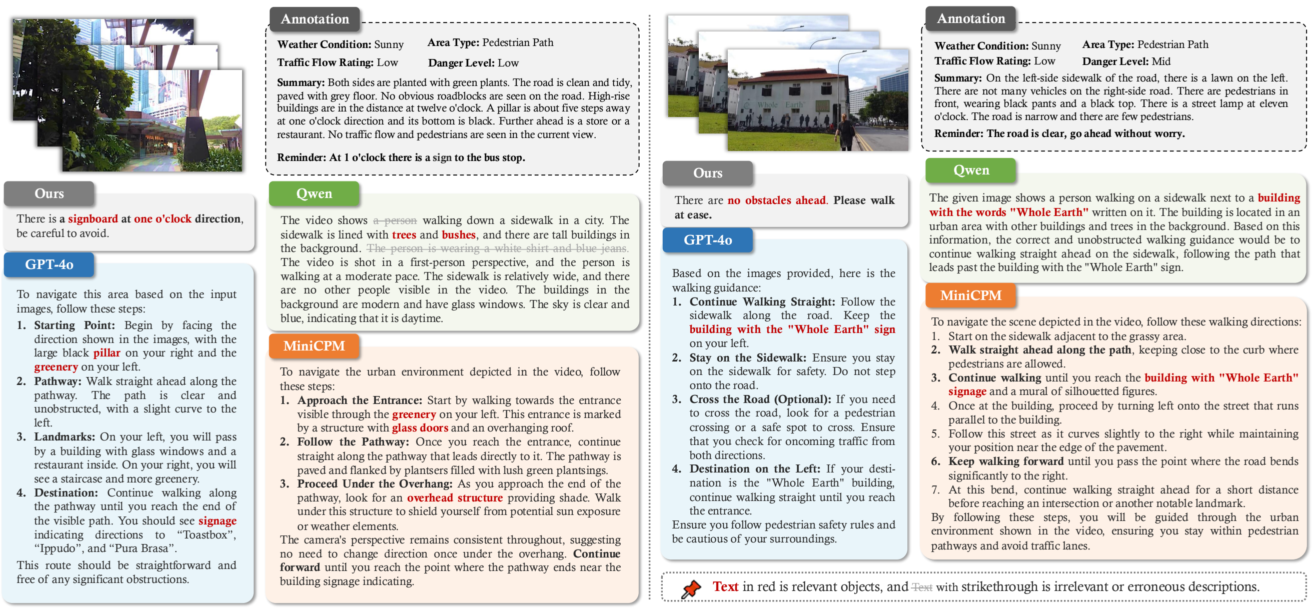


Figure A11. More reminder visualization comparisons of different models.

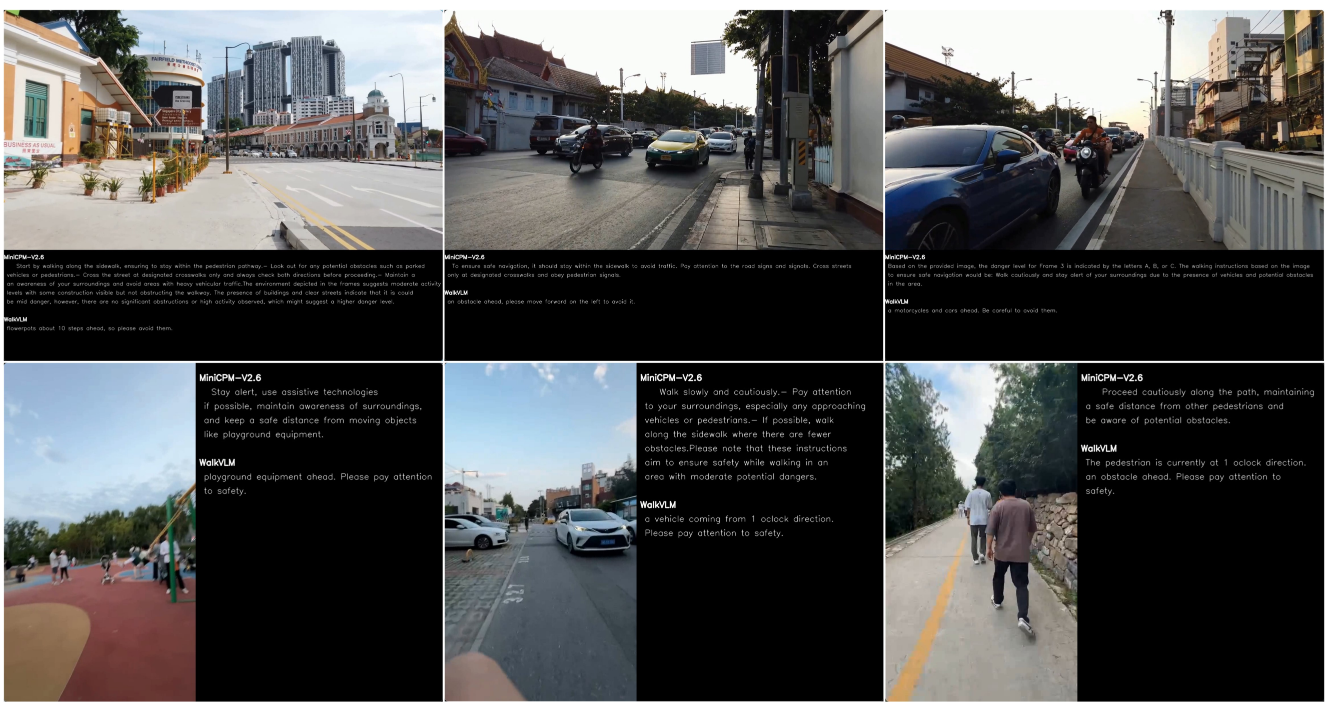


Figure A12. Sampling results of video stream inference in the blind walking task. Zoom in to view the generated results. See here for the video inference results. WalkVLM is capable of generating less temporal redundancy and providing more concise and informative responses.