



Introducing WalkGPT

Background

- Pedestrian navigation requires understanding walkable areas, obstacles, and depth.
- Existing Large Vision Language Models (LVLMs) often lack reliable grounding and spatial reasoning.
- WalkGPT provides depth-aware grounded guidance from pedestrian-view images.

What WalkGPT Does

- Generates navigation guidance from an image and question.
- Grounds accessible and harmful features with segmentation masks.
- Estimates object-level distances for safer accessibility guidance.

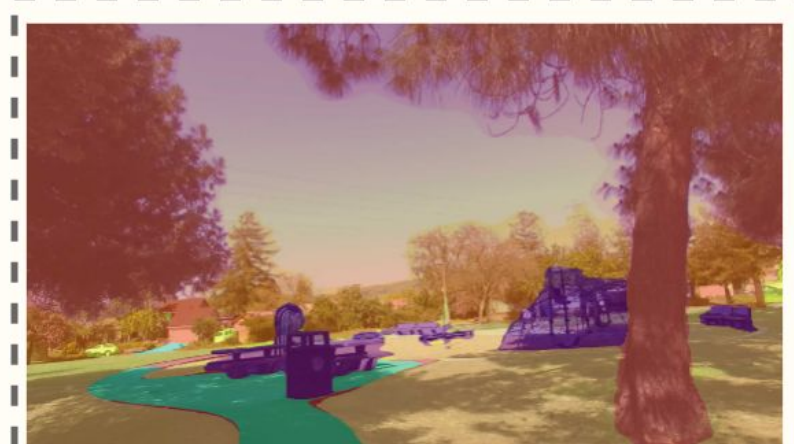
PAVE Dataset

- A new 41k real pedestrian-view image-QA pairs.
- Includes accessibility labels, masks, and depth-grounded answers.



Can you describe the path in terms of **accessibility** for a pedestrian, including segmentation masks for accessible and harmful features?

Pedestrian



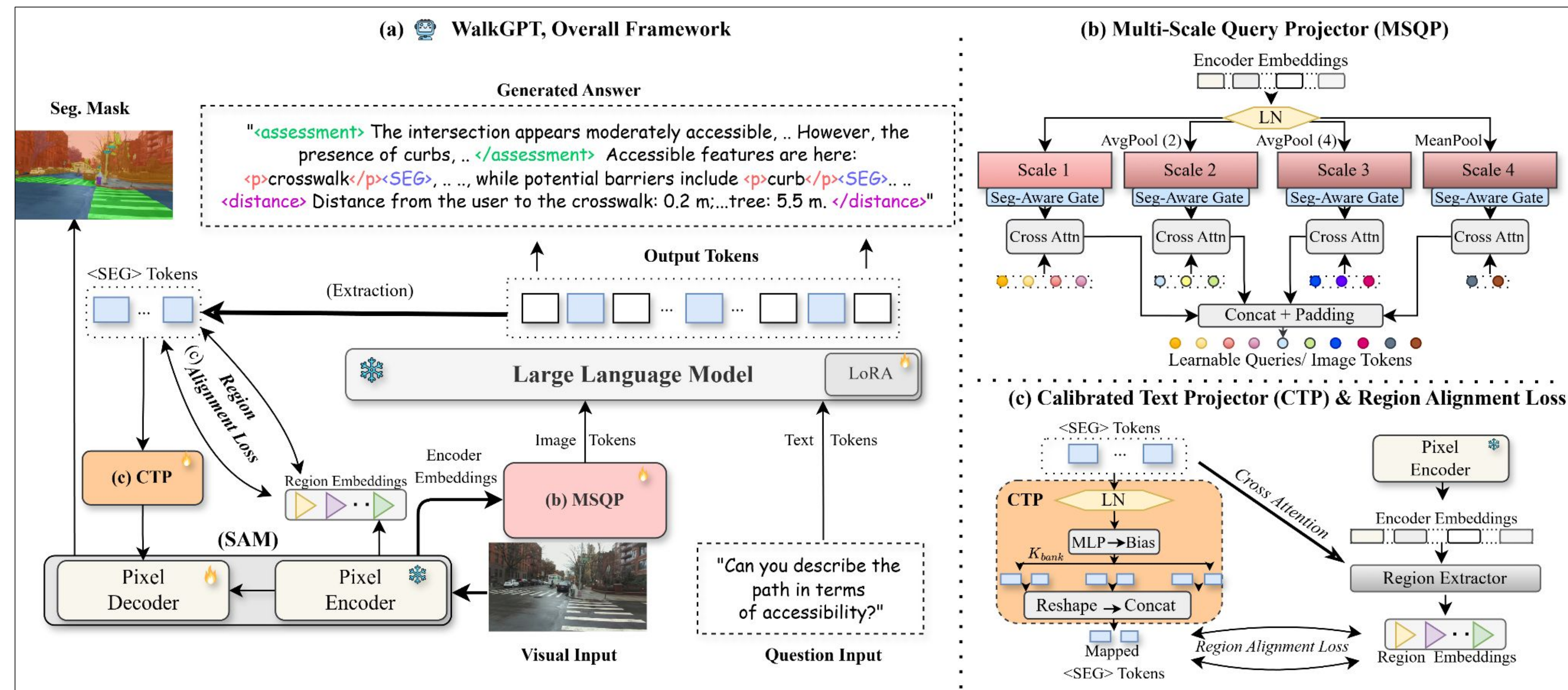
The park appears largely walkable, with a smooth paved path and open visibility that makes it comfortable for pedestrians. Natural elements like trees are around the path but don't block the main walking area.

Accessible features are here: **sidewalk** and **paved trail**. Non-accessible features are here: **tree**, **vegetation**, **building**, **vehicle**, and **obstacle**. Distance measured to the **sidewalk**: 0.5 m; to the **paved-trail**: 2.0 m; to the **tree**: 1.2 m; to the **vegetation**: 2.5 m; to the **building**: 50.6 m; to the **obstacle**: 18.5 m.

WalkGPT



Methodology



Framework

- Takes a pedestrian-view image and question, then generates accessibility guidance with structured tokens.
- Links object mentions to segmentation masks and distance estimates for pedestrian guidance.

Multi-Scale Query Projector (MSQP)

- Views the scene at multiple zoom levels.
- Captures small obstacles, curb edges, and the full path layout

Calibrated Text Projector (CTP)

- Expands each <SEG> token into a label-specific mask prompt.
- Preserves richer object cues for segmentation.

Region Alignment Loss

- Matches calibrated <SEG> prompts to their visual regions.
- Keeps the language-to-mask mapping semantically faithful.

Key Takeaway

WalkGPT moves pedestrian assistance from generic scene description toward pixel-grounded, depth-aware navigation dialogue.

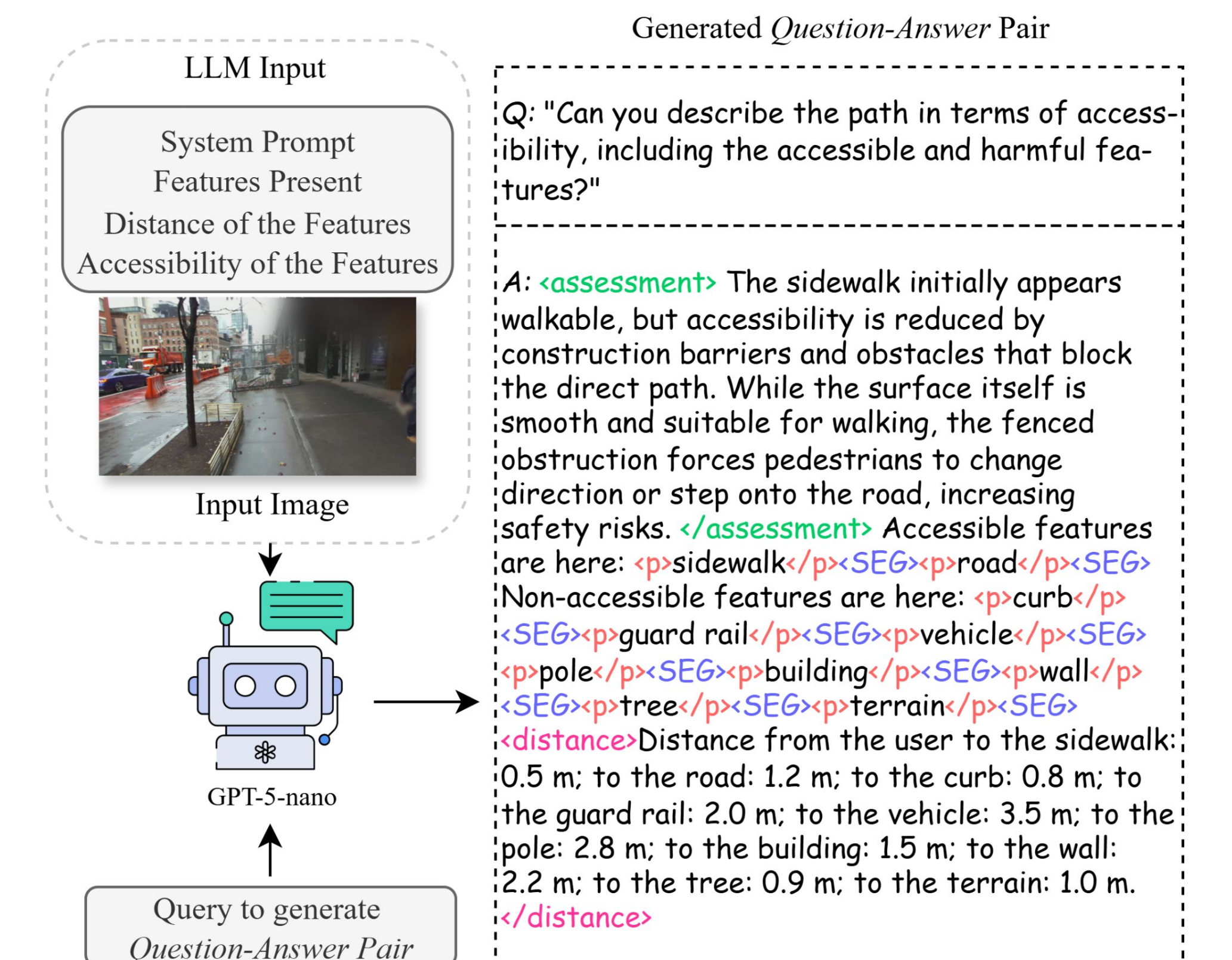


Project Website

Our Lab

R61EY037504

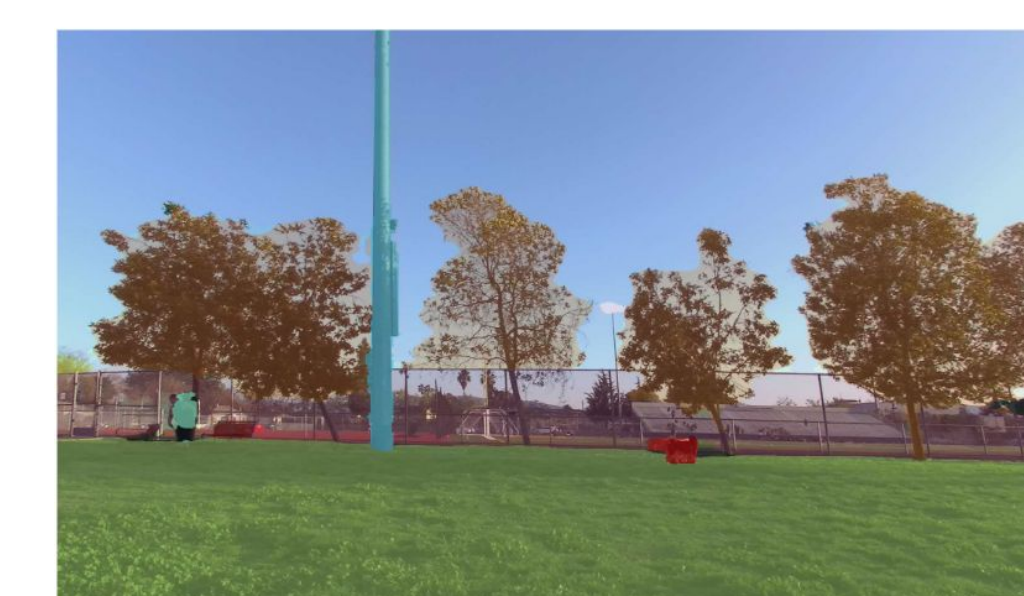
PAVE Dataset Construction



Results

Model	Text Generation		Segmentation Performance			Depth Estimation	
	CIDER↑	METEOR↑	AP50↑	mIoU↑	Recall↑	Depth Acc.↑	AbsRel↓
GLAMM	37.96	39.12	15.21	18.23	25.01	38.95	77.05
LISA	35.14	36.17	13.71	15.07	24.11	35.46	81.22
PixelLM	37.49	38.02	15.97	18.10	28.92	39.00	74.61
GSVA	35.78	38.15	14.67	17.34	29.71	36.55	78.19
OMG-LLaVA	38.01	38.96	15.74	18.02	28.05	39.02	75.01
Sa2VA	38.82	39.66	18.72	16.10	29.20	40.54	73.82
WalkGPT (7B)	41.97	42.36	16.66	19.95	31.55	41.97	67.88
WalkGPT (13B)	41.17	43.01	17.26	20.16	32.71	48.95	70.66

Qualitative Results



The scene provides a clear and accessible path for pedestrians, with a smooth surface and visible poles. However, the presence of a wall/fence, obstacles, and animals on the path could present challenges or restrict movement. Accessible features are here: **terrain**, **pedestrian**, **pole**. Non-accessible features are here: **animal**, **obstacle**, **sky**, **tree**, **vegetation**, **wall/fence**.

Distance from the user to the wall/fence: 10.1 m; to the terrain: 4.0 m; to the obstacle: 16.2 m; to the animal: 10.1 m; to the tree: 10 m; to the vegetation: 22.2 m; to the pedestrian: 10.6 m; to the pole: 11.5 m; to the sky: 6.5 m.