This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

WalkGPT: Grounded Vision–Language Conversation with Depth-Aware Segmentation for Pedestrian Navigation

Rafi Ibn Sultan[1] Hui Zhu[1] Xiangyu Zhou[1]
Chengyin Li[2] Prashant Khanduri[1] Marco Brocanelli[3] Dongxiao Zhu[1,4]
[1]Department of Computer Science, Wayne State University [2]Department of Radiation Oncology, Henry Ford Health
[3]Department of Electrical and Computer Engineering, The Ohio State University [4]Institute for AI and Data Science, Wayne State University

Abstract

*Ensuring accessible pedestrian navigation requires reasoning about both semantic and spatial aspects of complex urban scenes, a challenge that existing Large Vision–Language Models (LVLMs) struggle to meet. Although these models can describe visual content, their lack of explicit grounding leads to object hallucinations and unreliable depth reasoning, limiting their usefulness for accessibility guidance. We introduce **WalkGPT**, a pixel-grounded LVLM for the new task of Grounded Navigation Guide, unifying language reasoning and segmentation within a single architecture for depth-aware accessibility guidance. Given a pedestrian-view image and a navigation query, WalkGPT generates a conversational response with segmentation masks that delineate accessible and harmful features, along with relative depth estimation. The model incorporates a Multi-Scale Query Projector (MSQP) that shapes the final image tokens by aggregating them along text tokens across spatial hierarchies, and a Calibrated Text Projector (CTP), guided by a proposed Region Alignment Loss, that maps language embeddings into segmentation-aware representations. These components enable fine-grained grounding and depth inference without user-provided cues or anchor points, allowing the model to generate complete and realistic navigation guidance. We also introduce **PAVE**, a large-scale benchmark of 41k pedestrian-view images paired with accessibility-aware questions and depth-grounded answers. Experiments show that WalkGPT achieves strong grounded reasoning and segmentation performance. The source code and dataset are available on the project website*[1].

1. Introduction

Safe and accessible pedestrian navigation is essential for independent mobility in diverse environments. However, pedestrian routes often contain static and dynamic barriers such as stairs, uneven terrain, parked vehicles, and temporary obstructions [15, 28, 43], posing significant risks for individuals with mobility challenges [51]. While most automated navigation systems are designed for vehicles [20, 61], pedestrian-level navigation remains largely underexplored. Addressing this gap requires systems that understand the environment from a first-person view and reason about geometry, depth, and accessibility to navigate complex urban scenes.

Figure 1. Overview of WalkGPT for *accessibility-aware* grounded navigation guide. The model grounds language on segmentation masks enriched with depth information, providing holistic spatial understanding that captures both object shapes and depth cues for interpretable *accessibility analysis*.

Recent advances in Large Vision–Language Models (LVLMs) [12, 27, 30, 31] demonstrate strong visual understanding and language reasoning capabilities, suggesting potential for conversational guidance in pedestrian navigation. However, existing LVLMs lack explicit spatial reasoning [3, 6, 9] needed to infer geometric structure and depth

[1] https://sites.google.com/view/walkgpt-26/

relationships in real-world scenes. Recent spatial-aware approaches [5, 6, 11] attempt to address this but often rely on *user-provided visual cues* to estimate object depth, which is impractical for pedestrian navigation because users cannot manually supply such inputs. LVLMs also tend to hallucinate, describing objects not present in the scene [58], which can lead to hazardous guidance.

Grounded LVLMs [25, 38, 39, 56, 57] extend conventional LVLMs by aligning textual references with corresponding image regions, improving spatial consistency and reducing hallucinations [58]. Their segmentation masks provide explicit shape cues, making them promising for pedestrian navigation. However, these masks are 2D and lack the depth information required for spatial reasoning. Consequently, grounded LVLMs remain limited in understanding relative distances and spatial hierarchy, which are critical for safe, accessibility-aware navigation. Thus, existing grounded LVLMs have not been adapted for pedestrian navigation, a domain challenged by complex urban scenes and the absence of large-scale pedestrian-view datasets with joint question–answering and grounding annotations. These limitations motivate the research question: *Can we develop a grounded LVLM capable of providing depth-aware navigation guidance for pedestrians?*

To address these limitations, we propose **WalkGPT**, a grounded LVLM with spatial reasoning for the task of *Grounded Navigation Guide*. As illustrated in Figure 1, given a pedestrian-view image and a user query about the path ahead, WalkGPT generates a grounded response integrating free-form reasoning with depth-aware segmentation masks for fine-grained accessibility guidance. Unlike prior spatial reasoning approaches [5, 11], it operates without requiring user-provided visual cues or anchor points. By grounding accessible and harmful elements along with their segmentation masks and relative distances, WalkGPT provides comprehensive navigation feedback that supports safe movement, including for individuals with disabilities.

WalkGPT, a unified architecture (illustrated in Figure 2), enables grounded conversations by reasoning over spatial cues from text and segmentation for navigation understanding. It introduces two novel modules: the Multi-Scale Query Projector (MSQP), which aggregates multi-scale visual context, and the Calibrated Text Projector (CTP), guided by a proposed Region Alignment Loss to refine language-to-vision grounding. Using structured tokens such as `<p>`, `<assessment>`, `<SEG>`, and `<distance>`, WalkGPT structures its outputs to explicitly link language with segmentation and distance cues, enabling automated depth-aware navigation guidance without user-provided inputs. As no existing dataset addressed this problem, we introduce **PAVE**, a large-scale VQA dataset of 41k pedestrian-view images with accessibility-aware questions and answers containing embedded depth information.

Our main contributions are summarized as follows:

- We introduce **WalkGPT**, the first-of-its-kind LVLM for pedestrian accessibility via grounded spatial reasoning.
- We propose a streamlined architecture aligning visual and language representations to enhance pixel-level grounding using the novel MSQP and CTP with structured token supervision.
- We curate **PAVE**, a large-scale VQA dataset with depth annotations for accessibility and spatial understanding.
- WalkGPT achieves state-of-the-art performance on grounded navigation guidance and sets a benchmark for accessibility-aware AI in pedestrian environments.

2. Related Works

Grounded Large Vision–Language Models (LVLMs). LVLMs [2, 4, 12, 27, 30] have advanced multimodal understanding by integrating large language models (LLMs) with visual encoders through large-scale pretraining and instruction tuning. In particular, pixel-grounded LVLMs enable region-level understanding where textual descriptions are explicitly linked to image pixels. Early works performed coarse grounding via bounding boxes [7, 8, 24, 33, 37, 49, 52, 55]. More fine-grained grounding using integrated segmentation encoder–decoder architectures has followed two directions: reasoning segmentation [25, 39, 45, 53, 57] and visually grounded conversation [38, 47, 48, 56], both aiming to unify semantic reasoning with pixel-level grounding. Beyond these general-domain efforts, domain-specific studies have adapted visually grounded models to medical [10] and remote-sensing imagery [42]. However, grounded LVLMs have not yet been explored for accessibility-aware pedestrian navigation, which requires spatial reasoning in complex real-world environments.

Spatial Reasoning LVLMs. LVLMs [5, 6, 11, 16] designed for spatial reasoning typically use depth maps or visual anchors to capture spatial relations between objects, whereas others [34] handle spatial reasoning without performing depth estimation. Although effective for general spatial understanding, these models often depend on user-provided visual cues or anchor points and, in some cases, omit depth prediction, which limits their suitability for navigation-oriented scenarios.

Accessibility-Aware Pedestrian Navigation. Basic pedestrian navigation studies [1, 18, 21, 41, 60] focus on static object detection or scene labeling with limited insight into route accessibility. More recent LVLM-based or multimodal systems [15, 28, 51, 54] address this gap, but rely on synthetic or metadata-heavy inputs and still lack the fine-grained grounding needed for real-world guidance. Consequently, grounded LVLMs have yet to be explored for accessibility-aware navigation, which requires both pixel-level grounding and depth reasoning in complex urban environments.

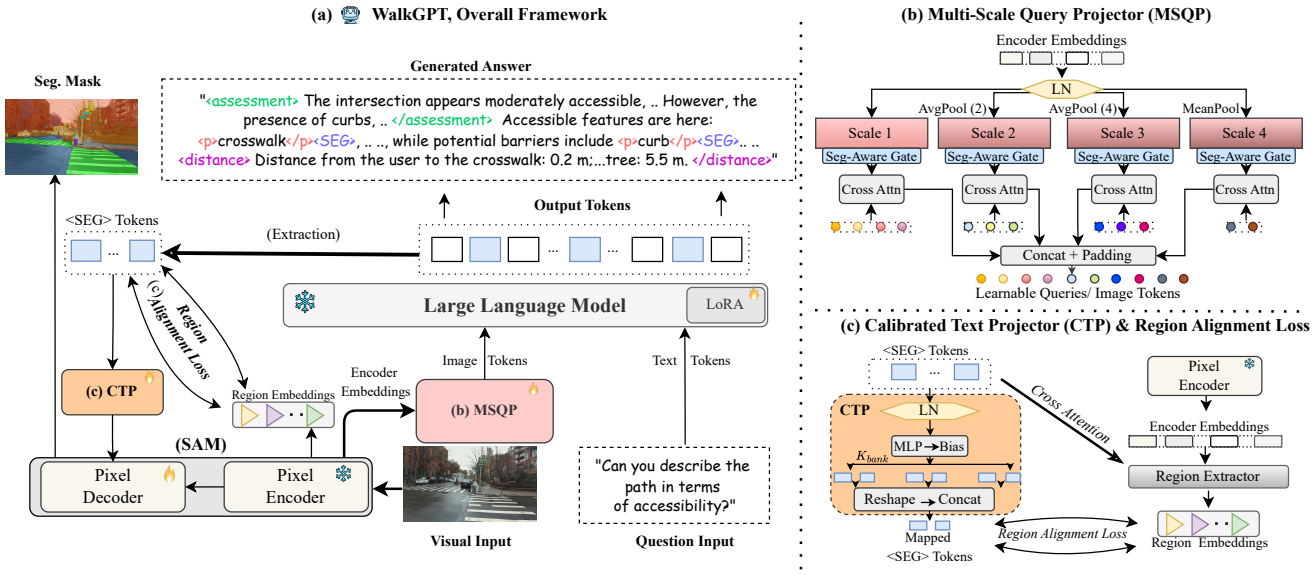

Figure 2. Overview of **WalkGPT** for grounded navigation guidance. (a) Overall framework. (b) The **Multi-Scale Query Projector (MSQP)**, which aggregates multi-level visual features into spatially aligned image tokens for language reasoning. (c) The **Calibrated Text Projector (CTP)**, guided by the proposed **Region Alignment Loss**, maps <SEG> tokens into the visual space. Structured tokens (<SEG>, <distance>, <assessment>, <p>) link language generation with segmentation and depth reasoning.

3. Methods

3.1. WalkGPT: The Architecture

Figure 2 illustrates WalkGPT, a grounded conversational model that generates navigation-aware responses linked to segmentation masks. Text tokens are concatenated with image tokens from a shared SAM-based pixel encoder [23] and processed by the LLM to produce the output sequence. Any <SEG> tokens in the response are passed to the SAM pixel decoder to obtain spatially aligned masks. WalkGPT introduces two architectural components, MSQP and CTP, enabling a shared encoder to support both conversation generation and segmentation.

Multi-Scale Query Projector (MSQP). The MSQP maps pixel encoder embeddings into semantically aligned image tokens in the language space for LLM input (Figure 2b). Unlike standard MLP projectors, MSQP aggregates visual features across multiple spatial levels, preserving both local detail and global structure and producing more spatially informative representations.

Given encoder embeddings $\mathbf{Z} \in \mathbb{R}^{B\times L\times C}$ from the pixel encoder, where B is the batch size, $L = H \times W$ flattened image tokens, and C the feature dimension, we apply a linear projection $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{C\times d_{\text{proj}}}$ to obtain $\mathbf{F} = \mathbf{Z}\mathbf{W}_{\text{proj}} \in \mathbb{R}^{B\times L\times d_{\text{proj}}}$, with $d_{\text{proj}} = 1024$ as the working dimension. Then $\mathbf{F}$ is reshaped into a grid and average-pooled at multiple scales, generating token banks $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^4, \mathbf{x}^g\}$ corresponding to native, pooled-by-2, pooled-by-4, and global-mean resolutions. Each scale-specific feature bank is modulated by a segmentation-aware gating function $g(\cdot)$ that highlights structure- and edge-rich regions before attention. For each scale s, a small set of learnable query embeddings $\mathbf{Q}^s \in \mathbb{R}^{Q_s\times d_{\text{proj}}}$ interacts with the gated tokens $\mathbf{x}^s$ through two layers of cross-attention, producing refined outputs $\mathbf{O}^s \in \mathbb{R}^{B\times Q_s\times d_{\text{proj}}}$. Each output token $\mathbf{o}_i^s$ is computed as a content-weighted mixture $\mathbf{o}_i^s = \sum_j \alpha_{ij}\mathbf{x}_j^s$, where attention weights α_{ij} are obtained by softmax normalization over query-key similarities. The outputs from all scales are first concatenated and padded to a fixed length $Q = \sum_s Q_s + 4 = 36$ to ensure consistency across scales. The padded sequence is then linearly projected to the LLM hidden dimension H, producing the final image tokens $\mathbf{V}_{\text{proj}} \in \mathbb{R}^{B\times Q\times H}$. By attending across multiple spatial hierarchies, MSQP condenses fine-grained details and global scene context into a compact set of tokens $\mathbf{V}_{\text{proj}}$.

Calibrated Text Projector (CTP). In our framework, the grounded <SEG> tokens in the generated response (see Figure 2a) act as textual prompts that guide the pixel decoder for mask prediction, linking language reasoning with pixel-level segmentation. Unlike prior methods that use a linear projection into the segmentation space, CTP expands each token into structured sub-embeddings, preserving fine-grained semantics. These calibrated embeddings improve spatial correspondence, supporting segmentation-aware mask generation.

Given the hidden states of the <SEG> tokens $\mathbf{T} \in \mathbb{R}^{B\times M\times H}$ from the LLM, where M is the number of <SEG> tokens, we apply a linear projection $\mathbf{W}_{\text{vis}} \in$

$\mathbb{R}^{H\times d_{\text{vis}}}$ to transform them, yielding $\mathbf{U} = \mathbf{T}\mathbf{W}_{\text{vis}} \in \mathbb{R}^{B\times M\times d_{\text{vis}}}$, where d_{vis}=256 matches the visual backbone dimension. As illustrated in Figure 2c, to prevent loss of token diversity, each reduced vector $\mathbf{u}_i$ is expanded into a small set of K_{bank} calibrated embeddings through a bias-augmented transformation $\mathbf{E}_i = \text{MLP}(\mathbf{u}_i) + \mathbf{B}$, where $\mathbf{E}_i \in \mathbb{R}^{K_{\text{bank}}\times d_{\text{vis}}}$ and $\mathbf{B} \in \mathbb{R}^{K_{\text{bank}}\times d_{\text{vis}}}$ contains learnable biases that encode modality-specific priors such as objectness and boundary structure. This expansion allows each token to generate multiple complementary sub-embeddings that capture different aspects of its semantics. The resulting calibrated bank $\mathbf{E} \in \mathbb{R}^{B\times M\times K_{\text{bank}}\times d_{\text{vis}}}$ is reshaped to $\mathbb{R}^{B\times(MK_{\text{bank}})\times d_{\text{vis}}}$ and concatenated along the token dimension to serve as the text prompt for the pixel decoder.

Region Alignment Loss. Since CTP maps the rich H-dimensional LLM embeddings (4096) into the lower d_{vis}-dimensional visual space (256), substantial information loss may occur. To preserve semantic detail, we introduce *Region Alignment Loss*, a contrastive regularization that enforces consistency between text embeddings and their corresponding visual regions. For each <SEG> token, the method aligns its projected embedding with the visual features of the target region while pushing it away from unrelated areas, encouraging a semantically faithful $H\to d_{\text{vis}}$ mapping. With the pixel encoder frozen, this regularization improves projection fidelity and language–region alignment.

For each sample $b \in \{1,\dots,B\}$ in the batch, we take the pre-projection <SEG> token embeddings $\mathbf{T}_b \in \mathbb{R}^{M\times H}$ (from the LLM hidden states) and the flattened pixel-encoder embeddings $\mathbf{Z}_b \in \mathbb{R}^{L\times C}$. Since the <SEG> tokens interact with specific image regions during generation, they are used to cross-attend and retrieve the most relevant spatial areas. The attended regions act as pseudo-targets to supervise their projected counterparts from CTP, establishing fine-grained region–token correspondence through cross-attention between the original <SEG> tokens and the visual embeddings. Specifically, we compute

$$\begin{aligned}&\mathbf{q} = \mathbf{t}_{b,m}\mathbf{W}_q, \quad \mathbf{K}_b = \mathbf{Z}_b\mathbf{W}_k, \quad \mathbf{V}_b = \mathbf{Z}_b\mathbf{W}_v,\\ &\boldsymbol{\pi} = \text{softmax}\left(\frac{\mathbf{K}_b\mathbf{q}^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^L,\end{aligned} \tag{1}$$

where $\mathbf{t}_{b,m}$ is the m-th <SEG> token in batch b and d_k=d_{vis} is the shared query–key projection dimension.

To focus alignment on salient object regions and suppress background noise, we emphasize the top-K image tokens with the highest attention weights. The selected indices are $\mathcal{I}_K = \text{TopK}(\boldsymbol{\pi}, K)$, and the normalized weights are $\alpha_i = \pi_i / \sum_{j\in\mathcal{I}_K}\pi_j$. The positive region embedding is then computed as

$$\mathbf{z}^+_{b,m} = \mathbf{W}_o\left(\sum_{i\in\mathcal{I}_K}\alpha_i\,\mathbf{v}_{b,i}\right) \in \mathbb{R}^{d_{\text{vis}}}, \tag{2}$$

where $\mathbf{v}_{b,i}$ denotes the i-th value vector from $\mathbf{V}_b$ and full attention is used when $K = L$, i.e., $\mathcal{I}_K$ contains all tokens.

The mapped tokens from CTP, $\mathbf{E}_b = [\mathbf{e}_{b,1},\dots,\mathbf{e}_{b,M}]^\top \in \mathbb{R}^{M\times d_{\text{vis}}}$, are aligned with their corresponding positive regions using an InfoNCE loss [35]. For each <SEG> token, the paired region embedding $\mathbf{z}^+_{b,m}$ serves as a positive example, while image tokens from other images (and non-attended regions) act as negatives. With L_2-normalized vectors $\hat{\mathbf{e}}_{b,m}$ and $\hat{\mathbf{z}}^+_{b,m}$, the positive logit is $s^+_{b,m} = \langle\hat{\mathbf{e}}_{b,m}, \hat{\mathbf{z}}^+_{b,m}\rangle$. The loss is defined as

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{BM}\sum_{b,m}\log\frac{\exp(a_{b,m})}{\exp(a_{b,m}) + \sum_{k\in\mathcal{K}^-}\exp(r_{b,mk})}, \tag{3}$$

where $a_{b,m} = s^+_{b,m}/\tau$, $r_{b,mk} = \langle\hat{\mathbf{e}}_{b,m}, \hat{\mathbf{z}}^-_k\rangle/\tau$, and τ is the temperature parameter.

3.2. PAVE: The VQA Dataset

Dataset Summary. We introduce **PAVE** (**P**edestrian **A**ccessibility and **V**isual-grounded **E**valuation), a spatially grounded VQA dataset for accessibility reasoning in complex pedestrian environments. It captures diverse real-world scenes with occlusions, reflections, motion blur, and dense object layouts typical of urban navigation. PAVE is built from the real-image subset of SANPO [44], which provides head-mounted pedestrian-view frames with human-annotated semantic and instance masks and corresponding depth maps. We focus exclusively on real images to preserve natural visual artifacts present in pedestrian scenes and therefore exclude synthetic data [44, 51]. SANPO's panoptic labels are converted into unified semantic maps, while depth information is used separately to compute object-level distances for each feature. The dataset contains 41k image–question–answer triplets, each consisting of an RGB frame, a question about path accessibility, and a free-form answer describing accessible features, harmful features, their distances from the camera, and an overall accessibility assessment. Additional details are provided in the Appendix.

Dataset Curation Pipeline. Each question–answer pair in PAVE is generated from a SANPO frame containing (1) the RGB image, (2) segmentation masks identifying visible objects/features, (3) depth information for each feature, and (4) accessibility labels (e.g., *sidewalk* as accessible, *vehicle* or *stair* as harmful). For each feature mask, we compute the minimum pixel depth to represent its closest visible distance from the pedestrian viewpoint, providing object-level distance information for the final annotations. As shown in Figure 3, these structured scene attributes are then provided to GPT-5-nano [36] to generate accessibility-aware VQA pairs. A system prompt instructs the model to act as a navigation assistant, with a one-shot example guiding the output format. GPT-5-nano produces both the natural-

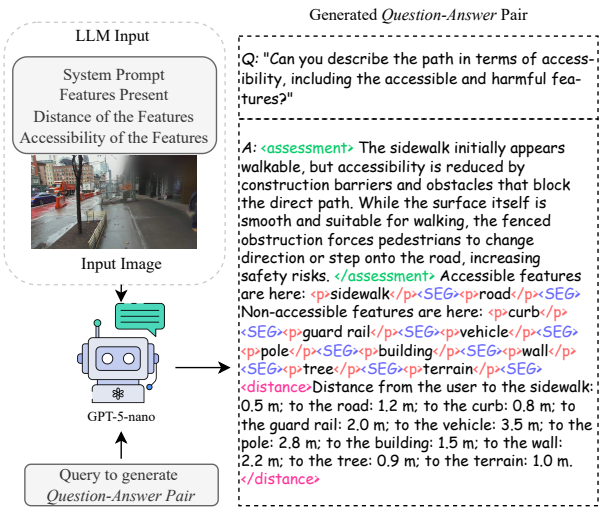

Figure 3. Pipeline for generating accessibility-aware VQA pairs in the PAVE dataset. The LLM receives the system prompt, detected features, their distance values, and the accessibility of the features, and generates structured outputs containing <assessment>, <distance>, <SEG>, and <p> tokens.

language question and a structured answer containing four elements: a concise <assessment> of overall accessibility, a <distance> tag listing object-level distances in text form, and <p> and <SEG> tokens that spatially ground the conversation. Additional details on prompt design and formatting are provided in the Appendix.

Verification Pipeline. To ensure annotation reliability at scale, we combine automated validation with selective human review. Programmatic checks verify object references, <SEG> token associations, and consistency with sensor-derived depth across all samples. Manual inspection is then applied to the experimental subset to confirm structural correctness, including formatting and token usage.

3.3. The Training Recipe

Conversation Generation. The causal language model is trained autoregressively with teacher forcing, predicting each output token conditioned on preceding tokens. Let x denote the concatenated textual input (system prompt, question, and prior outputs), and let $\mathbf{V}_{\text{proj}}$ denote the projected image tokens produced by the MSQP from the frozen pixel encoder (see Section 3.1). The model generates the answer sequence $y_{1:S}$ by jointly attending to textual and visual embeddings through cross-attention. The training objective is the standard cross-entropy loss,

$$\mathcal{L}_{\text{CE}} = -\frac{1}{S}\sum_{s=1}^{S} \log P(y_s \mid y_{<s},\, x,\, \mathbf{V}_{\text{proj}}), \qquad (4)$$

where S denotes the number of answer tokens. Tokens from the prompt and question contribute to the conditioning context but are excluded from loss computation. Structured tokens provide task-specific control over the generated output: <assessment> produces an accessibility judgment, <p> grounds referenced objects within the dialogue, <SEG> provides the segmentation prompt for the pixel decoder, and <distance> expresses object-level depth in text form. Representing these outputs as language tokens allows WalkGPT to unify conversational reasoning, spatial grounding, segmentation prompting, and depth description within a single next-token prediction process.

Segmentation Mask Prediction. As illustrated in Figure 2a, <SEG> tokens generated within the textual response serve as spatial prompts for the pixel decoder, enabling segmentation grounded in the language output. Each <SEG> token corresponds to a semantic entity mentioned in the answer and is associated with an image region. The hidden state of each <SEG> token is mapped by the CTP into the visual space and used by the decoder together with the projected visual embeddings $\mathbf{V}_{\text{proj}}$ to predict one mask per token, following [38]. Segmentation supervision combines Dice and cross-entropy losses, $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}_{seg}}$, encouraging accurate mask prediction and alignment between linguistic entities and visual regions.

Depth Estimation. WalkGPT predicts object-level depth as part of the same autoregressive sequence used for grounded conversation. Distance information is expressed through the <distance> ...</distance> span of the generated answer, where objects referenced in the response are associated with textual distance descriptions. During training, sensor-derived depth maps are used to compute the minimum visible distance for each segmented object, which is inserted into the ground-truth answers as discretized natural-language distance expressions (Figure 3). These tokens are learned through the same cross-entropy objective used for conversation generation. Because the referenced objects are already grounded through <SEG> tokens, the model learns to associate segmentation-aligned visual regions with their corresponding distance descriptions. Repeated supervision of these distance statements encourages attention to visual cues correlated with relative depth, allowing depth estimation to emerge through language prediction without dense depth supervision or dedicated depth heads.

Overall Training Objective. The final objective integrates three complementary components: (1) cross-entropy loss $\mathcal{L}_{\text{CE}}$ for grounded conversation generation, (2) segmentation loss $\mathcal{L}_{\text{seg}}$ for mask prediction, and (3) contrastive alignment loss $\mathcal{L}_{\text{NCE}}$ for visual–textual correspondence. The total loss is expressed as

$$\mathcal{L}_{\text{total}} = \alpha_1\, \mathcal{L}_{\text{CE}} + \alpha_2\, \mathcal{L}_{\text{seg}} + \alpha_3\, \mathcal{L}_{\text{NCE}}, \qquad (5)$$

where α_1, α_2, and α_3 are scalar weights controlling each component's contribution to jointly optimize conversation generation, spatial grounding, and cross-modal alignment.

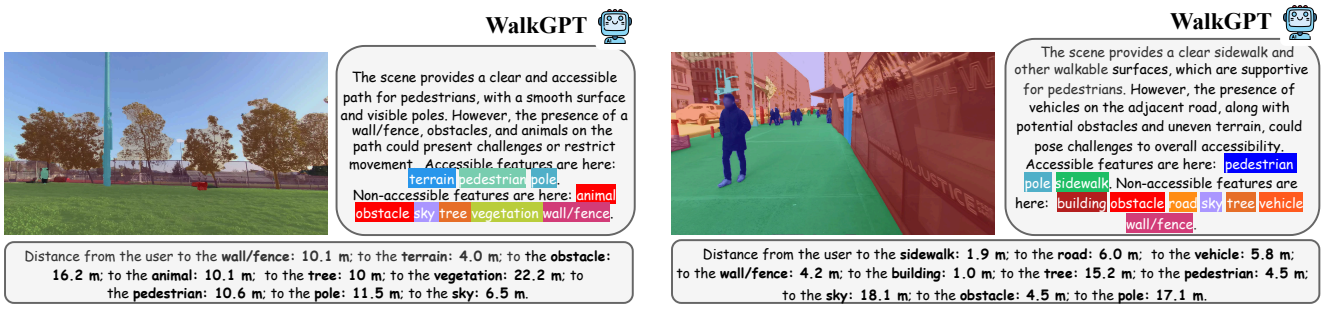

Figure 4. Qualitative results of WalkGPT on the PAVE validation set. Given a scene image, WalkGPT generates grounded conversations together with segmentation masks and depth-aware distance estimates, reflecting its understanding of accessibility and spatial context. Additional examples are provided in the Appendix.

4. Experiments

4.1. Implementation Details

WalkGPT uses a single SAM ViT-H pixel encoder [23] shared across all components, providing consistent visual grounding for both text generation and mask prediction. The language model is initialized with pretrained checkpoints from [39] for the 13B version and [38] for the 7B version. We apply LoRA [14] to fine-tune the language model parameters in a lightweight and efficient manner. All experiments are run on a single NVIDIA H100 GPU (80 GB) with Python 3.10.8 and CUDA 12.1. The PAVE dataset contains 91 pedestrian-view video sessions recorded at high frame rates. Because frames are sequential, consecutive images often change very little, which can create strong temporal redundancy that can encourage memorization during training. To obtain a more balanced representation of each scene, we uniformly sample 100 frames per session. This reduces redundancy while preserving scene-level diversity. We use 85 sessions for training (about 8.5k frames) and hold out 6 sessions (around 600 frames) for evaluation.

Training proceeds in two stages. In the pretraining stage, only the MSQP module is optimized while all other components remain frozen, allowing MSQP to learn stable visual tokenization across heterogeneous datasets. We use ADE20K [59] and the RefCOCO family [22] for this stage. The fine-tuning stage jointly trains MSQP, CTP, the pixel decoder, and the LoRA parameters on the PAVE dataset. All models share the same optimization settings for fair comparison. Additional hyperparameter details and ablations are provided in the Appendix.

4.2. Baselines

We compare WalkGPT with leading grounded vision-language models, including GLAMM [38], LISA [25], PixelLM [39], GSVA [48], OMG-LLaVA [56], and Sa2VA [53], using the validation split of the PAVE dataset. Each model is evaluated in both zero-shot and fine-tuned settings and adapted to our task by extending its tokenizer with our four structured token types (`<SEG>`, `<distance>`, `<p>`, `<assessment>`), ensuring all baselines operate under the same multimodal interface. We evaluate three aspects of grounded navigation: text generation, segmentation, and depth estimation. CIDEr and METEOR measure text quality, while AP50, mIoU, and Recall assess segmentation of objects of interest. Depth Accuracy and Absolute Relative Error (AbsRel) quantify depth estimation. Depth Accuracy measures the proportion of estimations within $[0.5\times, 2\times]$ of ground truth, and AbsRel computes the mean absolute difference normalized by ground-truth depth. Additional metric details are provided in the Appendix. To assess generalization beyond navigation, we also evaluate Referring Expression Segmentation (RES) on RefCOCO, RefCOCO+, and RefCOCOg datasets. Following standard protocols [25, 39], we report precision at multiple IoU thresholds and mean IoU, and compare against baselines including MCN [32], VLT [13], CRIS [46], LAVT [50], ReLA [29], LISA [25], and PixelLM [39].

4.3. Results

Grounded Navigation Conversation Generation. Figure 4 shows qualitative examples of WalkGPT generating grounded navigation conversations from single pedestrian-view images. The model provides an overall assessment and segments *accessible* and *harmful* features, including their distances from the user, helping pedestrians understand what to avoid and how far potential obstacles are. These examples span visually distinct scenes and illustrate that WalkGPT maintains consistent spatial grounding across different layouts. Table 1 summarizes performance on text generation, segmentation, and depth estimation. Zero-shot baselines fail across all metrics and cannot predict depth even with added `<distance>` tokens, highlighting the difficulty of this real-world task. Fine-tuning improves results but remains inadequate. WalkGPT achieves substan-

Table 1. Performance comparison on grounded navigation conversation generation. Models marked with † are zero-shot; "-FT" indicates fine-tuned on PAVE. Depth metrics for zero-shot models are listed as N/A because they fail to produce any depth estimations. Best results are **bold-faced.**

Model	Text Generation		Segmentation Performance			Depth Estimation	
	CIDEr↑	METEOR↑	AP50↑	mIoU↑	Recall↑	Depth Acc.↑	AbsRel↓
GLAMM† [38]	1.32	21.98	1.23	2.01	2.24	N/A	N/A
GLAMM-FT [38]	37.96	39.12	15.21	18.23	25.01	38.95	77.05
LISA† [25]	0.97	17.75	1.02	1.50	1.84	N/A	N/A
LISA-FT [25]	35.14	36.17	13.71	15.07	24.11	35.46	81.22
PixelLM† [39]	1.08	21.87	1.22	1.59	2.39	N/A	N/A
PixelLM-FT [39]	37.49	38.02	15.97	18.10	28.92	39.00	74.61
GSVA† [48]	0.78	20.74	1.78	1.87	2.54	N/A	N/A
GSVA-FT [48]	35.78	38.15	14.67	17.34	29.71	36.55	78.19
OMG-LLaVA† [56]	0.97	19.99	2.05	3.21	2.02	N/A	N/A
OMG-LLaVA-FT [56]	38.01	38.96	15.74	18.02	28.05	39.02	75.01
Sa2VA† [53]	1.28	21.02	2.71	3.14	1.78	N/A	N/A
Sa2VA-FT [53]	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

Table 2. Performance comparison on the referring expression segmentation (RES) benchmark. Best results are **bold-faced.**

Method	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
MCN [32]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [13]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS [46]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT [50]	72.7	75.8	68.8	62.8	68.4	55.1	61.2	62.1
ReLA [29]	73.8	76.5	70.2	66.0	71.0	57.7	65.5	66.0
LISA [25]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
PixelLM [39]	73.0	76.5	68.2	66.3	**71.7**	58.3	69.3	70.5
WalkGPT (Ours)	**76.2**	**78.5**	**68.3**	**70.0**	71.1	**60.5**	**72.6**	**71.6**

tial gains over strong fine-tuned baselines such as PixelLM-FT and OMG-LLaVA-FT: the 13B model improves mIoU by more than 10% (20.16 vs. 18.10) and raises depth accuracy by over 25% (48.95 vs. 39.00), while the 7B variant surpasses prior best text-generation scores by over 10% in CIDEr. These improvements stem from WalkGPT's unified pixel encoder together with MSQP and CTP, where MSQP provides fine-grained multi-scale visual tokens and CTP, guided by region alignment, maintains consistent grounding between visual regions and structured outputs. Together, they support robust multimodal grounding across model sizes.

Referring Expression Segmentation (RES). RES requires segmenting target regions in an image given a natural-language referring expression. Standard RES models [13, 25, 29, 32, 39, 46, 50, 62, 63] take the expression directly as input. For WalkGPT and the grounded LVLMs in Table 2, we place the expression into a short instruction template and prompt the model to generate a response containing a <SEG> token, whose embedding is decoded by the pixel decoder to produce the mask. Following the setup in Section 4.2, we evaluate WalkGPT on public RES benchmarks. Although not designed specifically for RES, WalkGPT shows strong generalization as a grounded vision–language model. As shown in Table 2, it achieves 76.2% on refCOCO-val and 72.6% on refCOCOg-val, outperforming LISA and PixelLM by up to about 3–4%. These gains reflect stronger visual–language grounding, enabling more precise localization of the referred region without RES-specific training.

Table 3. Segmentation performance on PAVE compared with representative vision-only segmentation benchmarks.

Model	mIoU↑	Recall↑
U-Net [40]	16.85	28.34
nnU-Net [19]	18.55	28.41
Swin-UNETR [17]	**20.60**	30.65
WalkGPT (Ours)	20.16	**32.71**

Challenges and Failure Analysis. To contextualize segmentation performance on PAVE, we compare WalkGPT with strong vision-only benchmarks, including U-Net [40], nnU-Net [19], and Swin-UNETR [17]. Although specialized for segmentation, these models achieve under 21% mIoU, reflecting the difficulty of PAVE's dense pedestrian-view scenes characterized by heavy occlusions, small regions of interest, and severe class imbalance. WalkGPT attains a comparable mIoU (20.16) without segmentation-specific fine-tuning, indicating that even expert models struggle under these conditions. These results underscore the inherent difficulty of PAVE's real-world scenes for grounded segmentation in navigation settings. Figure 5 provides a representative example: strong road reflections on the building façade resemble physical obstacles, caus-

ing WalkGPT to segment them as real objects and generate flawed accessibility guidance. This misinterpretation propagates to the associated segmentation output, further degrading mask quality. Such failures are particularly common in single-view images, where reflective surfaces distort visual depth cues and make it difficult to distinguish true geometry from appearance. Similar issues arise from motion blur, noisy surfaces, and significant class imbalance (see Appendix), all of which blur the boundary between true structures and incidental visual artifacts.

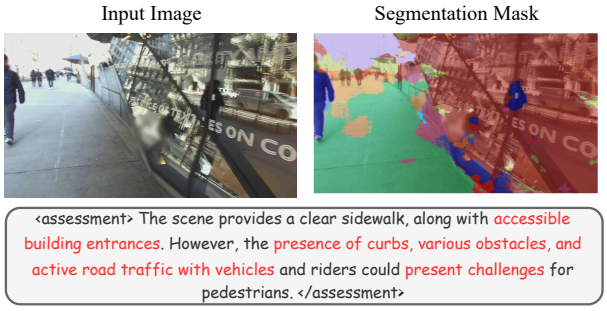

Figure 5. Failure case study on PAVE. WalkGPT misinterprets strong road reflections on the building façade as physical obstacles, producing incorrect guidance even though the path itself is fully accessible. Part of the image is blurred for privacy.

Grounded LVLM for Hallucination Mitigation. Table 4 reports hallucination performance against non-grounded LVLMs. LLaVA-1.5 [31], LLaVA 1.6 Mistral [26], and Qwen-VL-Chat [4] show high hallucination rates and limited coverage of visible objects, reflecting their lack of explicit grounding mechanisms. WalkGPT incorporates pixel-level grounding by linking text generation to segmentation-informed visual features. This connection reduces unsupported object mentions and improves recognition of image regions that are actually present, leading to substantially lower $CHAIR_i$ and higher Cover scores. Constraining language to visual evidence allows WalkGPT to produce scene descriptions that remain faithful to the underlying image.

Ablation Study. We evaluate several design variations of WalkGPT to assess their impact on grounded navigation. Table 5 shows that replacing MSQP with a simple MLP leads to a clear drop in performance (43.01→39.50 METEOR, 20.16→17.40 mIoU, 48.95→43.39 Depth Acc.), highlighting MSQP as a key component. Removing its multi-scale aggregation also reduces performance (41.60 METEOR, 19.30 mIoU, 44.70 Depth Acc.), confirming that multi-resolution features are important for capturing spatial structure and depth cues. Reducing the number of MSQP queries to Q=8 (from Q=32) yields a further decline (38.10 METEOR, 16.20 mIoU, 45.33 Depth Acc.), indicating that query diversity supports robust grounding and distance reasoning. Substituting CTP with a linear mapping (40.70 METEOR, 18.60 mIoU, 47.98 Depth Acc.) or removing the contrastive loss L_{NCE} (41.00 METEOR, 18.90 mIoU, 47.00 Depth Acc.) mainly affects language and segmentation while leaving depth largely unchanged. Freezing the LLM reduces performance across metrics (35.20 METEOR, 17.80 mIoU, 40.21 Depth Acc.), suggesting that lightweight language adaptation also supports distance-aware guidance. Finally, removing the `<distance>` token primarily harms depth prediction (48.95→38.77) while leaving segmentation nearly unchanged (20.16→20.01 mIoU), confirming the importance of explicit distance supervision. Overall, these results confirm that WalkGPT's performance arises from the joint contribution of MSQP, CTP-based alignment, and structured distance tokens.

Table 4. Comparison of hallucination ($CHAIR_i$) and object coverage (Cover) scores across LVLMs on the PAVE dataset.

Model	$CHAIR_i$↓	Cover↑
LLaVA-1.5 [31]	22.16	33.04
LLaVA 1.6 Mistral [26]	23.56	38.83
Qwen-VL-Chat [4]	26.78	31.42
WalkGPT (Ours)	**18.49**	**83.66**

Table 5. Ablation study examining the impact of different design choices in WalkGPT.

Variant	METEOR↑	mIoU↑	Depth Acc.↑
WalkGPT (Full)	**43.01**	**20.16**	**48.95**
w/o MSQP → MLP	39.50	17.40	43.39
w/o MSQP multi-scale	41.60	19.30	44.70
MSQP queries Q=8	38.10	16.20	45.33
CTP → Linear	40.70	18.60	47.98
w/o L_{NCE}	41.00	18.90	47.00
w/o LoRA (LLM frozen)	35.20	17.80	40.21
w/o `<distance>`	41.22	20.01	38.77

5. Conclusion

WalkGPT advances grounded multimodal reasoning by reframing pedestrian navigation as an interpretable, accessibility-aware dialogue grounded in pixels and depth. By jointly modeling conversational guidance, segmentation, and distance estimation, and introducing the PAVE dataset, it establishes a benchmark for accessibility-aware reasoning and opens new directions for trustworthy assistive navigation systems. Beyond this task, our framework highlights the importance of tightly coupling language reasoning with spatial grounding for real-world multimodal AI.

Limitations and Future Work. WalkGPT may still be affected by dataset artifacts that introduce ambiguity. Future work will explore improved depth estimation and evaluate cross-domain generalization on additional navigation and grounding datasets.

Acknowledgments

This work was supported by the National Institutes of Health (NIH), National Eye Institute (NEI), under Grant R61EY037504.

References

[1] Andrew Abbott, Alex Deshowitz, Dennis Murray, and Eric C Larson. Walknet: A deep learning approach to improving sidewalk quality and accessibility. *SMU Data Science Review*, 1(1):7, 2018. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] Anonymous. Spatialab: Can vision–language models perform spatial reasoning in the wild? In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review. 1

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 8

[5] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P. Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 2

[7] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024. 2

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[9] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning*, 2025. 1

[10] Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24732–24741, 2025. 2

[11] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 1, 2

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16321–16330, 2021. 6, 7

[14] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023. 6

[15] Jon E Froehlich, Alexander J Fiannaca, Nimer M Jaber, Victor Tsaran, and Shaun K Kane. Streetviewai: Making street view accessible using context-aware multimodal ai. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2025. 1, 2

[16] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[17] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 7

[18] Hochul Hwang, Sunjae Kwon, Yekyung Kim, and Donghyun Kim. Is it safe to cross? interpretable risk assessment with gpt-4v for safety-aware street crossing. In *2024 21st International Conference on Ubiquitous Robots (UR)*, pages 281–288. IEEE, 2024. 2

[19] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 7

[20] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4524–4536, 2025. 1

[21] Yuhan Jiang, Sisi Han, Dapeng Li, Yong Bai, and Mingzhu Wang. Automatic concrete sidewalk deficiency detection and mapping with deep learning. *Expert Systems with Applications*, 207:117980, 2022. 2

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[23] 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. 3, 6

[24] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 2

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 6, 7

[26] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[28] Zongrong Li, Junhao Xu, Siqin Wang, Yifan Wu, and Haiyang Li. Streetviewllm: Extracting geographic information using a chain-of-thought multimodal large language model. *arXiv preprint arXiv:2411.14476*, 2024. 1, 2

[29] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 6, 7

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1, 8

[32] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 6, 7

[33] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 2

[34] Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7851–7860, 2025. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[36] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. Model family including GPT-5, GPT-5-mini, and GPT-5-nano variants. 4

[37] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[38] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2, 5, 6, 7

[39] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 2, 6, 7

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015. 7

[41] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. 2

[42] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. In *Forty-second International Conference on Machine Learning*, 2025. 2

[43] Rafi Ibn Sultan, Chengyin Li, Hui Zhu, Prashant Khanduri, Marco Brocanelli, and Dongxiao Zhu. Geosam: Fine-tuning sam with multi-modal prompts for mobility infrastructure segmentation. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, pages 501–508. IOS Press, 2025. 1

[44] Sagar M Waghmare, Kimberly Wilber, Dave Hawkey, Xuan Yang, Matthew Wilson, Stephanie Debats, Cattalyya Nuengsigkapian, Astuti Sharma, Lars Pandikow, Huisheng Wang, et al. Sanpo: A scene understanding, accessibility and human navigation dataset. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7866–7875. IEEE, 2025. 4

[45] Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774, 2024. 2

[46] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 6, 7

[47] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24710–24721, 2025. 2

[48] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 2, 6, 7

[49] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13838–13848, 2024. 2

[50] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022. 6, 7

[51] Kedi Ying, Ruiping Liu, Chongyan Chen, Mingzhe Tao, Hao Shi, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. mmwalk: Towards multi-modal multi-view walking assistance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 1, 2, 4

[52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[53] Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 2, 6, 7

[54] Zhiqiang Yuan, Ting Zhang, Yeshuang Zhu, Jiapei Zhang, Ying Deng, Zexi Jia, Peixiang Luo, Xiaoyue Duan, Jie Zhou, and Jinchao Zhang. Walkvlm: Aid visually impaired people walking by vision language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9845–9854, 2025. 2

[55] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2

[56] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024. 2, 6, 7

[57] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2

[58] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Forty-second International Conference on Machine Learning*, 2025. 2

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6

[60] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7641–7649, 2024. 2

[61] Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[62] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 7

[63] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 7