AR-Nav Benchmark: Augmented Reality Navigation with Vision and Language

Liqi Yan, Yihao Wu, Chenyi Xu, Chao Yang, Jianhui Zhang*, Pan Li*

Hangzhou Dianzi University
jh_zhang@hdu.edu.cn, lipan@ieee.org

Abstract

Augmented Reality (AR) navigation has emerged as a transformative tool for spatial intelligence, enabling users to interactively explore complex environments through wearable and mobile AR devices. However, current AR navigation systems struggle with low indoor localization accuracy, weak semantic understanding, and limited long-term memory, which severely limits their adaptability in dynamic, multi-floor, and large-scale real-world settings. To address these challenges, we present **AR-Nav benchmark**, a novel dataset with corresponding suite that leverages vision and language for AR navigation. First, to construct this benchmark, we proposed an Augmented Reality Visual-Language Memory Model (AR-VLM²), which generates structured, semantically rich, and temporally indexed representations for long-term AR navigation. Second, we design a lightweight navigation intent recommending module with hierarchical topological reasoning and language-grounded path planning, called ARN-Pilot, enabling low-latency and personalized route selection. Third, we introduce a closed-loop AR interaction module that supports real-time multi-modal feedback, dynamic memory updates, and human-in-the-loop query refinement. Extensive experiments in indoor multi-floor and outdoor parking scenarios show that AR-Nav suite significantly outperforms state-of-the-art AR navigation methods.

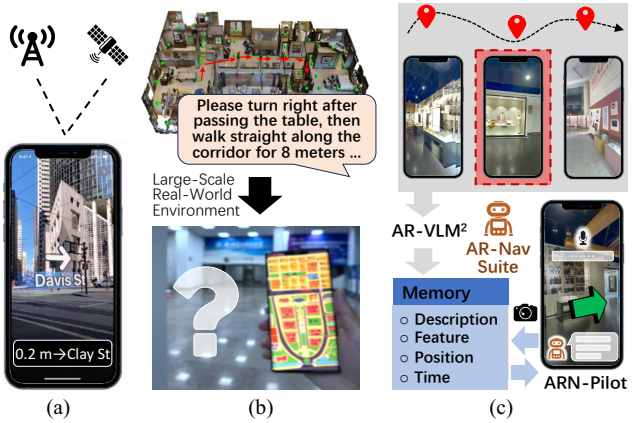


Figure 1: A comparison of different navigation methods. (a) GNSS- or signal-based navigation performs well outdoors but suffers from low indoor accuracy and requires labor-intensive annotations. (b) Vision-language navigation models trained in small simulated environments often struggle to generalize to large-scale real-world settings. (c) Our AR-Nav suite encodes both positional and descriptive information into a memory database, enabling scalable and flexible real-world navigation via real-time dialogue.

Introduction

Augmented Reality (AR) navigation plays a vital role in enabling intelligent spatial understanding and interaction in complex environments (Rehman and Cao 2016; Kim and Jun 2008; Chung, He, and Jung 2016; Katz et al. 2012). With the proliferation of AR-enabled devices such as smartphones, smart glasses and tablets, users increasingly expect seamless guidance that integrates real-time perception, intuitive display, and personalized assistance (Yan et al. 2022).

Existing AR navigation approaches span three main paradigms: marker-based systems that rely on QR codes or fiducials to anchor virtual content and guide users (Yeh et al. 2018; Jang 2012; Tadepalli, Ega, and Inugurthi 2021; Bopp et al. 2022), image-recognition or beacon-based approaches that utilize pre-annotated signage or Bluetooth/Wi-Fi/UWB beacons for localization (Kim and Jun 2008), and markerless methods built on Simultaneous Localization and Mapping (SLAM) or video-based positioning systems (VPS) that dynamically localize and map environments (Yan et al. 2024). While marker-based and beacon systems offer deployment simplicity and reasonable robustness in controlled indoor environments, they often require infrastructure installation and struggle in dynamic or cluttered spaces. On the other hand, markerless SLAM/VPS methods, such as ARCore or ARKit–based systems, enable flexible, infrastructure-free operation but face challenges in large-scale settings due to drift, occlusion, and fluctuating scene dynamics. Despite rapid advances, no existing approach fully achieves high accuracy, semantic awareness, real-time adaptability, and infrastructure independence, all essential for robust AR navigation in everyday environments. Nowadays, several researchers have begun exploring the use of Vision-Language Navigation (VLN) methods for AR navigation to meet the

*Corresponding Authors: Jianhui Zhang, Pan Li.

growing demand for systems that support high-level language queries (Zhou, Hong, and Wu 2024).

Despite progress in SLAM, VPS and VLN, existing AR navigation systems face several fundamental limitations. First, existing SLAM-based methods rely on GPS or manually pre-defined maps, which are often unavailable or inaccurate indoors. Indoor localization errors often exceed 5 meters, and systems typically cache only short-term perceptual data, making them incapable of supporting retrospective queries or adapting to real-world dynamics. Second, conventional VLN methods often suffer from poor generalization, especially when deployed in environments not seen during training. Their computational complexity increases quadratically with the number of candidate nodes, which makes them unsuitable for real-time operation in large-scale scenes. Finally, most LLMs are limited by a fixed-length context window and cannot retain structured memory across long time horizons, resulting in frequent forgetting of key spatial or temporal cues essential for AR-based guidance.

To overcome these challenges, we introduce AR-Nav, a new AR navigation benchmark and corresponding methods that leverages the seasoning ability of Large Language Models (LLMs), as shown in Fig. 2. **First**, to construct this benchmark, we introduce the Augmented Reality Visual-Language Memory Model (AR-VLM²), which converts ego-centric video into structured semantic memory entries with timestamps, camera poses, and semantic embeddings, with language-based spatial descriptions. Powered by a dual-stream spatio-semantic encoder and geometry-aware symbolic alignment, AR-VLM² builds language-grounded 3D semantic fields that enhance spatial recall, reasoning, and generalization in open-world navigation. **Second**, we present ARN-Pilot, a lightweight recommender and planner tailored for AR navigation, using a hierarchical, topology-aware reasoning pipeline. It models indoor multi-floor spaces as graph networks, performs multi-hop subgraph retrieval and dynamic path evaluation, and achieves sub-second inference with high retrieval precision by leveraging user preferences, temporal constraints, and environmental factors. **Third**, our AR Interaction Generation Module delivers dual-modality navigation guidance: visually, it overlays 3D arrows, spline paths, and floor-level indicators aligned with the environment; linguistically, it generates context-aware prompts like "*Walk past the red sofa*" or "*Turn right after the potted plant*", enabling semantic grounding and supporting human-in-the-loop goal refinements in real time.

Our main contributions are summarized as follows:

- We propose AR-Nav, a unified vision–language AR navigation benchmark that flexibly supports flexible navigation goals and quickly adapts to novel complex environments by fusing real-time visual perception and semantic language grounding.
- We design AR-VLM², a dual-stream visual-language memory model with symbolic-geometric alignment, enabling high-fidelity semantic memory construction and long-term spatio-temporal retrieval.
- We introduce ARN-Pilot, a lightweight yet expressive recommending and planning module that performs topology-aware goal retrieval and dynamic path generation in real-time, supporting personalized, low-latency navigation.
- We demonstrate that AR navigation models trained in AR-Nav benchmarks achieves state-of-the-art performance in real-world scenarios, including indoor multi-floor search, retrospective recall, and dynamic crowd-aware route planning.

We are releasing the AR-Nav benchmark, comprising a comprehensive suite of evaluation dimensions, evaluation protocols, curated prompts, and generated annotations. We are inviting the community to advance AR navigation by participating in the AR-Nav Challenge with novel models.

Related Work

AR Navigation

Augmented Reality navigation systems have evolved significantly with advances in wearable computing and spatial intelligence. Early systems like (Azuma et al. 2002) established fundamental AR tracking techniques, while more recent works (Lee et al. 2021) demonstrated mobile AR navigation prototypes. Commercial solutions such as LVV Live (Carballeira et al. 2021) leverage visual positioning systems for outdoor navigation, but struggle with indoor localization accuracy. Semantic-aware navigation systems (Liu, Qi, and Fu 2021; Yan et al. 2020) improved environment understanding but lack long-term memory capabilities. Learning-based approaches (Katragadda et al. 2024; Deng et al. 2022) introduced neural rendering for VR/AR navigation, yet suffer from high computational overhead. While existing AR navigation systems have made progress in wayfinding assistance (Juile, Chandrasekaran, and Surya 2024), they typically exhibit limited adaptability in dynamic, multi-floor environments due to their reliance on pre-built maps and weak semantic representations.

Vision-Based AR Navigation

Vision-based approaches have become predominant in AR navigation, with visual-inertial odometry (VIO) systems (Zuo et al. 2021) enabling robust tracking. Recent works (Radwan, Valada, and Burgard 2018) combine visual localization with learned features for improved accuracy. Language-grounded navigation methods (Anderson et al. 2018) incorporate natural language understanding, while multimodal systems (Hong et al. 2021) fuse vision and language for better human-computer interaction. However, these methods often process sensory data frame-by-frame without maintaining persistent environment memory (Zhou, Hong, and Wu 2024; Chen et al. 2025), leading to repetitive computation and poor scalability. Our AR-VLM^2 model addresses this by constructing structured, temporally-indexed representations that enable efficient long-term navigation while maintaining semantic richness.

Vision-Language Navigation for Agents

Vision-Language Navigation (VLN) (Anderson et al. 2018; Krantz et al. 2020; Majumdar et al. 2020) aims to teach

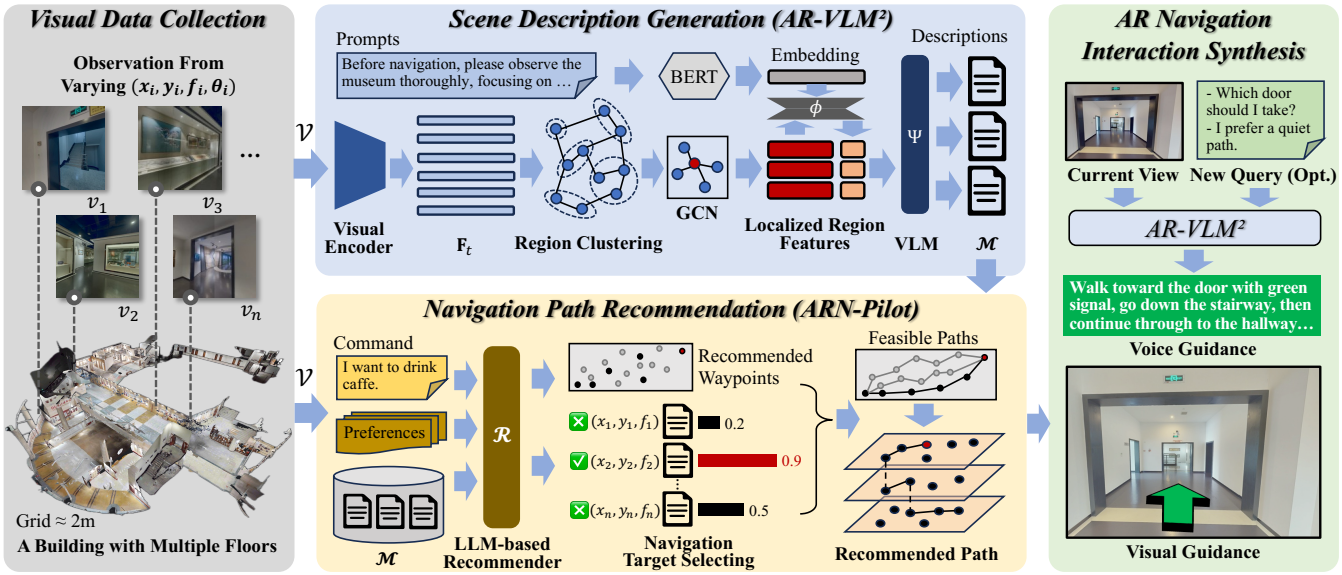


Figure 2: Overview of the AR-Nav framework, comprising four main stages: (1) data construction that collects RGB-D videos, 3D reconstructions, and natural language instructions across diverse AR navigation scenarios; (2) scene description via AR-VLM², which encodes egocentric visual-linguistic inputs into structured, spatially grounded memory representations; (3) target and path recommendation through ARN-Pilot, which aligns user intent with retrieved scene memory using a large language model; and (4) dual-channel interaction synthesis, which generates real-time AR overlays and speech instructions to enable intuitive, user-aligned navigation and human-in-the-loop feedback.

embodied agents (e.g., robots) to follow natural language instructions in 3D environments. Tasks require grounding commands like "Go to the kitchen and turn left at the table" into sequences of visual observations and navigation actions. Benchmarks such as R2R (Anderson et al. 2018), R4R (Jain et al. 2019), and RxR (Ku et al. 2020) use simulated indoor agents with egocentric views. Early works (Wang et al. 2019; Ma et al. 2019) adopted LSTM-based encoders, while recent ones (Chen et al. 2021; Zhou, Hong, and Wu 2024) employ transformers and LLMs. Yet, most VLN studies rely on reinforcement training in simulators and face transfer challenges due to domain gaps. In contrast, our work adapts VLN to Augmented Reality, tackling sim2real gap, dynamic queries, floor-level ambiguity, and natural scene diversity.

AR-Nav Suite

Overview of AR-Nav Benchmark Construction

Our goal is to establish a comprehensive evaluation platform for AR-based vision–language navigation in real-world environments. To this end, we collect a diverse set of AR navigation scenarios across three domains: indoor (offices, museums), outdoor (urban streets, parks), and semi-indoor (shopping malls, transit hubs). For each scenario, we record: ① RGB-D video sequences with typical AR device fields of view and motion paths. ② Reconstruction 2D/3D maps with language description as memory of sampled locations in the scene. ③ Ground-truth trajectories: the intended navigational route embedded in the scene. ④ Natural-language instructions, generated via crowd-sourcing and expert annotation, capturing both imperative (e.g., "Turn left at the red door") and descriptive cues (e.g., "You'll see a fountain on your right").

To capture rich visual context and viewpoint variations for each navigation check-point, we implement a structured image-sampling procedure: at regular intervals along each traversable path (e.g., every 2 meters), we sample the 6 DoF reachable region and collect six RGB-D photos spaced evenly at 60° intervals around the vertical axis. Each photo covers a 60° horizontal field-of-view, resulting in a 360° panorama around the agent. Depth maps and camera poses are recorded alongside each RGB image. This systematic multi-view sampling ensures comprehensive coverage of the navigable environment and supports both memory encoding and target recommendation modules in downstream tasks. We adopt a similar dense-view rendering methodology employed in outdoor AR/photogrammetry datasets

To generate other key components of AR-Nav dataset and support both static and incremental decision-making, we propose : (1) a Navigation Memory Database containing multi-modal representations of landmarks and prior context, created by AR-VLM², (2) a Path Recommendation Testbed called ARN-Pilot, where candidate goal location and customized path are predicted to be aligned with user intent, and (3) Dual-Channel Interaction as a user-facing layer, generating depth-aware visualized direction cues as well as contextual voice prompts. These components are produced during the construction pipeline to enable end-to-end evaluation, establishing a unified platform for advancing AR navigation using vision and language. Details of each components are provided in the following subsections.

Scene Description Generation (AR-VLM²)

To enable AR navigation that combines perception and user preference, we introduce AR-VLM², a vision-language modeling framework designed to generate structured scene descriptions from egocentric video streams and natural language instructions. The goal is to transform raw sensory input into semantically rich, retrievable representations that reflect human-style memory and support downstream tasks such as target recommendation and interaction generation.

AR-VLM² first encodes the visual stream $\mathcal{V} = \{v_1, v_2, \ldots, v_T\}$ by extracting spatially localized region features from each frame. Each frame v_t is processed through a vision encoder (e.g., Video Swin Transformer) to yield a dense feature map $\mathbf{F}_t$. Semantic-level regions $r_t^{(i)}$ are then identified using a region clustering method among frames to recognize the same region recorded in different frames, and each region is projected to a feature vector $\mathbf{v}_t^{(i)} \in \mathbb{R}^d$ via a Graph Convolution Neural Network (GCN). Simultaneously, the language stream $\mathcal{L} = \{l_1, \ldots, l_K\}$, including navigation prompts for global route intent, is tokenized and passed through a pretrained language encoder (e.g., BERT (Devlin et al. 2019) or T5 (Raffel et al. 2020)). The result is a sequence of token embeddings $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_K\}$, each in $\mathbb{R}^{d_l}$, capturing temporal and semantic dependencies. To build a cross-modal understanding of the scene, we introduce a multimodal fusion module based on cross-attention ϕ. Each visual region feature $\mathbf{v}_t^{(i)}$ attends to the textual context via:

$$\mathbf{h}_t^{(i)} = \phi(\mathbf{v}_t^{(i)}, \mathbf{L}) = \sum_{j=1}^{K} \alpha_{ij} \cdot \mathbf{l}_j, \tag{1}$$

$$\alpha_{ij} = \frac{\exp(\mathbf{v}_t^{(i)} \cdot \mathbf{l}_j)}{\sum_{k=1}^{K} \exp(\mathbf{v}_t^{(i)} \cdot \mathbf{l}_k)}. \tag{2}$$

This yields a set of fused scene representations $\mathbf{h}_t^{(i)} \in \mathbb{R}^d$, which encode local visual information contextualized by task-relevant language. Each fused representation $\mathbf{h}_t^{(i)}$ is then processed through a large Vision-Language Model (VLM) decoder Ψ_{VLM}, such as the language modeling head of BLIP-2 or LLaVA head, to generate the local semantic memory embedding:

$$\mathbf{m}_t^{(i)} = \Psi_{\text{VLM}}(\mathbf{h}_t^{(i)}), \tag{3}$$

where $\mathbf{m}_t^{(i)} \in \mathbb{R}^{d_m}$ captures object identity, spatial cues (e.g., relative orientation), and linguistic alignment. These memory entries are indexed with their corresponding 3D pose $(x_t, y_t, f_t, \theta_t)$, obtained from Structure-from-Motion (SfM), where x_t, y_t, f_t denote the coordinates and floor level. The full scene description memory is denoted as:

$$\mathcal{M} = \left\{ \left(\mathbf{m}_t^{(i)}, x_t, y_t, f_t, \theta_t \right) \right\}_{t,i}, \tag{4}$$

which constitutes the structured scene description set for the benchmark. To enrich this representation, we apply a set of augmentations. For robustness, we synthesize alternate views through viewpoint transformations $(\Delta x, \Delta y, \Delta\theta)$, and paraphrase instructions using pretrained generative language models. These techniques allow AR-VLM² to generalize across minor variations in appearance, phrasing, and trajectory deviation.

Unlike conventional captioning methods, AR-VLM² is designed to be compositional, spatially grounded, and retrieval-oriented. During benchmark construction, the outputs of AR-VLM² are used not only to populate the memory database but also to generate descriptive annotations that reflect the perception and linguistic alignment at each navigation step. These descriptions serve as a semantic backbone throughout the benchmark. In essence, AR-VLM² formulates scene description as a fusion-driven, egocentric encoding task that supports fine-grained retrieval, flexible target matching, and interpretable feedback generation. Its output bridges the raw visual-linguistic input with structured memory, enabling reliable navigation and human-compatible AR interaction.

Navigation Path Recommendation (ARN-Pilot)

To translate user command into actionable navigation decisions, ARN-Pilot first interprets the user's intent and scene-aware memory to select the most relevant target location. Specifically, a large language model $\mathcal{R}$ serves as the semantic backbone to align user command $\mathbf{l}_{1..t}$ with previously observed scene descriptions $\mathcal{M}$ generated by AR-VLM². ARN-Pilot retrieves the top-k memory embeddings $\{\mathbf{m}_{t,i}\}_{i=1}^{k}$ most relevant navigation goal aligned with the intent of the user based on the current visual context $\mathbf{v}_t$:

$$\{\mathbf{m}_{t,i}\}_{i=1}^{k} = \text{TopK}_{\mathbf{m} \in \mathcal{M}} \cos\big(\mathcal{E}_s(\hat{s}_t), \mathbf{m}\big) \tag{5}$$

where $\hat{s}_t$ is the generated scene caption from AR-VLM² and $\mathcal{E}_s$ its embedding. These retrieved scene memories serve as contextual candidates for recommendation.

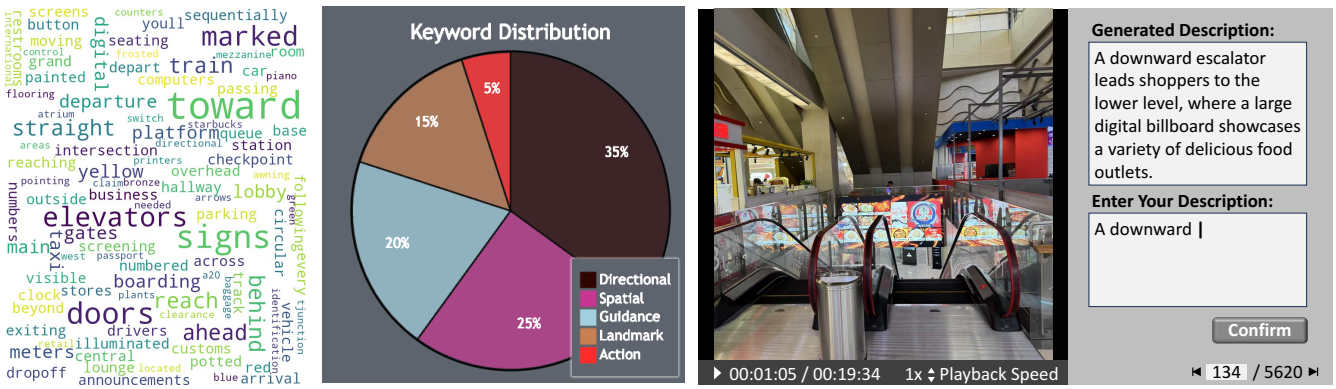
The retrieved set $\{\mathbf{m}_{t,i}\}$ is transformed into a personalized recommendation prompt, along with the recent instruction trajectory and user-specific preferences (metadata collected by history dialogues). Assuming that available paths have been computed from the 2D map constructed by SfM in AR-VLM², ARN-Pilot conditions a tuned LLaMA-based recommender $\mathcal{R}$ on this context to generate a ranked list of target, and recommend waypoint candidates $\{w_j\}$ from all feasible paths. This generation process is formulated as:

$$w_j = \arg\max_{w} P\big(w \mid \hat{s}_t, \{\mathbf{m}_{t,i}\}, \mathbf{l}_{1\ldots t}, \Theta_{\mathcal{R}}\big) \tag{6}$$

where $\Theta_{\mathcal{R}}$ denotes the model parameters and w represents candidate waypoints (described in natural language or 3D coordinates). Each waypoint candidate w_j is then scored by computing a matching score s_j that combines textual-semantic coherence and spatial feasibility:

$$\begin{aligned} s_j = \beta \cdot \cos\big(\mathcal{E}_s(w_j), \mathcal{E}_s(\hat{s}_t)\big) \\ + (1 - \beta) \cdot \text{ReLU}\big(d_{\max} - \|p_j - p_t\|\big) \end{aligned} \tag{7}$$

where p_j and p_t are the 3D positions of candidate w_j and current pose, $d_{\max}$ is a distance threshold, and β trades off semantic alignment and spatial proximity. The final ranking is determined by sorting s_j. The final path is chosen by maximizing the average waypoint score across feasible paths.



(a) (b) (c)

Figure 3: AR-Nav suit statistics and annotation checking interface. The first two graphs provide an overview of our AR-Nav suit. (a) The word cloud to visualize word distribution of our generated scene descriptions for navigation. (b) The ratio of descriptions across different content dimensions. (c) A human verification process through a web-based annotation checking interface to ensure high-quality descriptions for our dataset.

ARN-Pilot's LLM-based recommender is pretrained on our proposed navigation corpora and fine-tuned with prompts structured to mirror those in LLMRec (Wei et al. 2024). During benchmark construction, we execute ARN-Pilot over held-out trajectories to produce, at each checkpoint, a ranked list of recommended targets along with their scores. These recommendations constitute the testbed annotations, providing ground truth for downstream evaluation. By coupling scene description memory from AR-VLM² with a powerful LLM recommender, ARN-Pilot bridges perception, personalization, and action during navigation. This architecture supports interpretable, user-aligned waypoint predictions and establishes a unified interface between memory generation (AR-VLM²) and interaction guidance, positioning our benchmark at the forefront of AR navigation with language.

Dual-Channel Interaction Synthesis

To close the loop between memory-guided navigation and human-comprehensible instruction, we introduce a multimodal interaction synthesizer $\mathcal{I}$ that transforms navigation intent into real-time feedback across both visual and auditory channels. Built upon the structured memory $\mathcal{M}$, predicted waypoints $\{w_j\}$, and current visual context v_t, $\mathcal{I}$ synthesizes spatially grounded instructions delivered simultaneously through AR overlays (visable) and speech output (listenable). This dual-channel synthesis enables stepwise guidance that is perceptually coherent and easy for users to interpret during movement.

At each navigation step t, $\mathcal{I}$ conditions on the scene description $\hat{s}_t$ from AR-VLM², the selected navigation target w_{j^*} from ARN-Pilot, and the current pose (x_t, y_t, θ_t), generating instructions that integrate language-based direction with corresponding visual emphasis. For example, when the memory and current perception jointly identify a landmark such as a red sign above a glass door, $\mathcal{I}$ then output: "Walk toward the red sign above the glass door, then continue through to the hallway." Concurrently, the AR interface renders a bounding box highlighting the sign, projects a forward-pointing arrow toward the door, and speaks the instruction aloud. This synchronization of visable and listenable cues aligned with egocentric positioning ensures intuitive, timely, and context-aware feedback.

The interaction synthesizer also supports on-the-fly clarification, facilitating natural user inquiries that arise during navigation. For instance, if the user asks, "Do I need to pass the vase?" or "Which door should I take?" the system parses the query and conducts targeted memory retrieval from $\mathcal{M}$ based on both the current scene $\hat{s}_t$ and question context. Relevant memory entries are selected, and $\mathcal{I}$ refines its response by updating spatial relationships or visually emphasizing candidate paths. The revised guidance is then re-rendered with updated AR highlights and spoken confirmation, preserving dialogue coherence and enhancing user confidence. This human-in-the-loop refinement dynamically adapts to the user's cognitive state and navigation demands, showcasing the effectiveness of integrating vision-language memory with multi-modal AR guidance.

Experiments

Implementation Details

Dataset Construction and Evaluation Pipeline. We construct a large-scale navigation dataset consisting of 200 diverse scenes (120 indoor multi-floor and 80 outdoor/semi-indoor), sampled every ≈2m along traversable paths, yielding ≈50–70 locations per scene with 360° RGB-D panoramas captured at 6 views per point. Annotations span up to four floor levels. During evaluation, AR-VLM^2 adopts a dual-stream architecture, leveraging a Video Swin Transformer (Liu et al. 2022) and BERT-Base (Devlin et al. 2019) for visual and language processing, respectively. Region clustering employs a 2-layer GCN with 512-dimensional hidden features, while cross-modal fusion and semantic memory decoding utilize 512-dimensional embeddings via LLaVa (Liu et al. 2023). Each memory entry is indexed to SfM-based poses $(x_i, y_i, f_i, \theta_i)$ and enriched with

Method	Descriptive Question Accuracy ↑			Positional Error (m) ↓			Temporal Error (s) ↓		
	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
Zero-Shot Methods Without LLM/VLM									
ZSON	0.17±0.2	0.11±0.3	0.05±0.3	13.6±34.9	98.2±152.0	108.5±167.1	9.7±8.9	31.2±29.8	36.7±35.4
CoW	0.23±0.3	0.15±0.3	0.09±0.4	9.8±22.5	72.1±126.8	96.2±134.6	7.5±6.7	20.3±19.1	25.4±24.1
V3MVN	0.41±0.3	0.29±0.4	0.16±0.4	7.5±14.2	54.3±85.4	83.6±117.5	5.3±4.9	13.6±12.8	19.4±18.2
VLFM	0.55±0.4	0.37±0.4	0.21±0.3	6.1±11.5	48.6±79.7	71.4±105.9	4.2±4.0	8.7±8.1	15.2±14.3
LLM-Based Methods									
LLaMa-7B	0.52±0.5	0.42±0.5	0.38±0.5	6.5±12.3	33.7±68.2	62.5±93.4	4.7±4.3	7.2±6.9	12.7±12.0
Qwen3	0.60±0.4	0.59±0.4	0.53±0.5	5.1±10.2	28.4±52.6	50.2±77.2	3.4±2.9	6.0±5.7	9.6±8.7
DeepSeek-R1	0.57±0.4	0.55±0.4	0.52±0.5	5.3±10.5	29.6±56.7	52.9±85.8	3.9±3.3	6.1±5.8	10.2±9.4
GPT-4	0.68±0.3	0.64±0.4	0.61±0.4	4.9±9.7	27.3±51.1	48.6±70.5	2.3±2.1	5.9±5.6	8.4±7.9
VLM-Based Methods									
LLaVa-1.5	0.48±0.5	×	×	6.2±11.3	×	×	3.5±3.0	×	×
Qwen-VL	0.62±0.4	×	×	4.7±9.6	×	×	2.1±1.8	×	×
DeepSeek-VL2	0.59±0.4	×	×	5.1±10.0	×	×	2.4±2.2	×	×
GPT-4o	0.71±0.3	×	×	4.0±8.6	×	×	1.8±1.7	×	×
AR-Nav Suite (Ours)	**0.83**±0.2	**0.79**±0.2	**0.72**±0.3	**3.4**±7.2	**13.7**±27.8	**25.2**±50.1	**1.5**±1.4	**3.3**±3.1	**6.6**±6.2

Table 1: Results on AR-Nav. We evaluate our proposed AR-VLM² and ARN-Pilot against two baselines: one that processes all captions simultaneously (LLM-based), and another that ingests all video frames at once (VLM-based). All methods operate at a uniform frame sampling rate of 2 FPS. In the LLM-based setup, the entire textual caption is forwarded to the LLM in a single pass. In contrast, the VLM-based method receives all video frames as input to the VLM at an effective rate of 0.07 FPS due to computational constraints. Our results indicate that GPT-4o-based methods achieve the highest overall performance. Notably, our approach consistently surpasses the LLM-based baseline and remains competitive with the VLM-based strategy on Short videos. However, the VLM-based method fails to scale to Medium and Long videos due to input length limitations and is thus marked with an ×.

±0.5m/±15° perturbations and two paraphrases per instruction. A human verification process ensures annotation quality through a web-based interface (Fig. 3(c)). The ARN-Pilot module integrates a LLaMA-7B (Touvron et al. 2023) recommender, retrieving top-5 memory candidates with cosine similarity ≥0.7, computing waypoint scores with $\beta = 0.6$, and selecting paths within 500ms latency. Its interaction module renders real-world aligned AR overlays and present-tense natural-language instructions, supporting human-in-the-loop refinement.

Baselines. We compare our AR-Nav suite with several previous SOTA methods, which can be divided into three types: (1) Zero-shot methods without LLM or VLM, including ZSON (Majumdar et al. 2022), CoW (Gadre et al. 2023), ESC (Zhou et al. 2023), L3MVN (Yu, Kasaei, and Cao 2023), and VLFM (Yokoyama et al. 2024). (2) LLM-based methods, including LLaMa-7B (Touvron et al. 2023), Qwen3 (Yang et al. 2025), DeepSeek-R1 (Guo et al. 2025), and GPT-4 (Achiam et al. 2023). (3) VLM-based methods, including LLaVa-1.5 (Liu et al. 2023), Qwen-VL (Wang et al. 2024), DeepSeek-VL2 (Wu et al. 2024), and GPT-4o (Hurst et al. 2024). LLM-based methods first generate visual captions and then forward all of them to the LLM to directly output the suggested direction, while VLM-based methods directly input image sequences. In contrast, our method stores the scene description and corresponding information into memory for the LLM to refer to.

Evaluation Metrics. The AR-Nav dataset comprises four distinct answer types, each evaluated using tailored metrics. (1) **Descriptive Question Accuracy.** For spatial questions, which yield (x, y, z) coordinates, we compute the L2 distance to the ground-truth location. A prediction is considered correct if it falls within 15 meters of the target. (2) **Positional Error.** Temporal point-in-time and duration questions result in scalar predictions (e.g., "15 minutes"), for which we report the L1 error. A temporal prediction is deemed correct if it is within 2 minutes of the reference. (3) **Temporal Er-**

LLMs	Overall Correctness ↑		
	Short	Medium	Long
AR-Nav Suite	**0.78**±0.3	**0.62**±0.4	**0.69**±0.3
- w/o memory augmentation	0.73±0.3	0.59±0.4	0.63±0.4
- w/o ϕ	0.65±0.4	0.46±0.5	0.54±0.4
- w/o LLM-based Rec	0.57 ±0.4	0.42 ±0.5	0.48±0.5
- w/o preference condition	0.67±0.3	0.51±0.5	0.56±0.4

Table 2: Ablation Study. We perform ablations to assess the impact of key components in the AR-Nav suite. Memory augmentation and symbolic-geometric fusion in AR-VLM^2 significantly enhance contextual reasoning and spatial grounding. In ARN-Pilot, the LLM-based recommendation and user preference conditioning are essential for generating adaptive and personalized navigation instructions.

Top-k	Overall Correctness ↑	Positional Error (m) ↓	Temporal Error (s) ↓	Latency (ms) ↓
1	0.63	5.9	4.8	**280**
3	0.72	4.1	3.7	360
5	**0.78**	**3.4**	**3.3**	490
7	0.77	3.6	3.5	610
10	0.75	3.8	3.6	770

Table 3: Effect of Memory Size (*Top-k*). Optimal correctness is achieved at $k = 5$.

ror. Descriptive questions produce binary or free-form textual responses, where correctness is assessed via binary accuracy. To accelerate evaluation, textual answers are judged by an LLM for correctness, following prior work (Majumdar et al. 2024). Notably, structured outputs are enforced for spatial, temporal, and binary questions to ensure consistent evaluation. All experiments are conducted using three random seeds, while baseline results are reported with a single seed due to computational constraints. To mitigate the variability introduced by non-deterministic seeds, we report micro-averaged performance across seeds. Owing to the inherent variance in question difficulty, we introduce an **Overall Correctness** metric for ablation studies by thresholding spatial (e.g. <20m) and temporal metrics (e.g. <10s) to derive a unified binary indicator of correctness, thereby reducing the influence of outliers.

Main Results

We evaluate our proposed AR-VLM^2 and ARN-Pilot modules on the AR-Nav benchmark across a range of complex, real-world AR navigation tasks. As shown in Table 1, our approach significantly outperforms state-of-the-art methods across three primary metrics. Specifically, our method achieves a Descriptive Question Accuracy of 0.83, 0.79, and 0.72 on short, medium, and long trajectories, respectively, outperforming the best-performing GPT-4o baseline by over 12% absolute on average. This performance gap is attributed to the explicit long-term spatio-temporal memory encoding enabled by AR-VLM^2, which grounds language to structured visual regions. Notably, our method maintains high robustness across navigation horizons, while VLM-based models suffer degradation due to input length limitations and computational bottlenecks. In terms of positional precision, our method reports the lowest Positional Error across all lengths, achieving 3.4m, 13.7m, and 25.2m for short, medium, and long trajectories, respectively. These improvements validate the efficacy of ARN-Pilot's topological reasoning and personalized recommendation mechanisms. Similarly, our approach yields the lowest Temporal Error of 1.5s, 3.3s, and 6.6s, suggesting enhanced temporal alignment between predicted and ground-truth navigation steps. Collectively, these results demonstrate that AR-VLM^2 and ARN-Pilot enable accurate, robust, and personalized AR navigation in challenging scenarios.

Ablation Study

To understand the contribution of each component in our AR-Nav pipeline, we conduct a series of ablation studies summarized in Table 2. We consider the Overall Correctness across the three main factors above.

Effect of Memory Augmentation. We disable the synthetic viewpoint augmentation and paraphrased instructions during AR-VLM^2 training. The overall correctness drops by 6.4% (from 0.78 to 0.73 on short sequences), indicating that viewpoint and language diversity are critical for robust retrieval under visual and linguistic variations.

Effect of Symbolic-Geometric Fusion. We replace the cross-attention fusion ϕ in Eq. (1) with a naive concatenation of vision and language embeddings. This leads to a 16.7% (from 0.78 to 0.65 on short sequences) decrease in overall correctness, confirming that symbolic-geometric alignment substantially improves spatial grounding and generalization across complex layouts.

Effect of LLM-Based Recommendation. We replace LLM-Based recommendation as a naive similarity comparator. This results in a 26.9% drop in overall accuracy (from 0.78 to 0.57 on short sequences). The result emphasizes the importance of dynamic user intent modeling in achieving coherent and efficient AR guidance.

Effect of User Preference Condition. We remove the user preference conditioning from ARN-Pilot and use uniform path priors. This results in a 14.1% drop in overall accuracy (from 0.78 to 0.67 on short sequences). The result emphasizes the importance of dynamic user intent modeling in achieving coherent and efficient AR guidance.

Memory Size and Retrieval Precision. We also vary the size of the top-k memory entries retrieved by ARN-Pilot from 1 to 10, as shown in Table. 3. We observe performance saturation beyond $k = 5$, balancing recall and noise. Our default configuration ($k = 5$) provides the best trade-off between retrieval precision and computational efficiency.

Conclusion

We introduced AR-Nav, a comprehensive vision-and-language benchmark tailored for AR navigation in realistic, multi-floor, and dynamic environments. To enable structured long-term memory and personalized navigation planning, we proposed AR-VLM^2 and ARN-Pilot, two complementary modules in AR-Nav suite that together bridge egocentric visual understanding, semantic memory construction, and user-aligned interaction generation. Extensive experiments on over 200 environments demonstrate that our approach achieves state-of-the-art performance across diverse metrics and significantly outperforms leading LLM and VLM baselines. Ablation studies further validate the necessity of symbolic-geometric fusion, personalized planning, and memory augmentation for robust navigation under real-world constraints. We believe AR-Nav will facilitate future research in human-centric navigation, spatial memory modeling, and multimodal AR systems. We plan to open-source the benchmark, codebase, and trained models to support reproducibility and accelerate community progress.

Acknowledgments

This project is supported by the Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020019), the Fundamental Research Funds for Provincial Universities of Zhejiang (No. GK249909299001-033), and the Key Laboratory of Data Science and Intelligence Education (Hainan Normal University), Ministry of Education (No. DSIE202403).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Azuma, R.; Baillot, Y.; Behringer, R.; Feiner, S.; Julier, S.; and MacIntyre, B. 2002. Recent advances in augmented reality. *IEEE computer graphics and applications*, 21(6): 34–47.

Bopp, M. H.; Corr, F.; Saß, B.; Pojskic, M.; Kemmling, A.; and Nimsky, C. 2022. Augmented reality to compensate for navigation inaccuracies. *Sensors*, 22(24): 9591.

Carballeira, P.; Carmona, C.; Díaz, C.; Berjón, D.; Corregidor, D.; Cabrera, J.; Morán, F.; Doblado, C.; Arnaldo, S.; del Mar Martín, M.; et al. 2021. FVV live: A real-time free-viewpoint video system with consumer electronics hardware. *IEEE Transactions on Multimedia*, 24: 2378–2391.

Chen, J.; Lin, B.; Liu, X.; Ma, L.; Liang, X.; and Wong, K.-Y. K. 2025. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23568–23576.

Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34: 5834–5847.

Chung, C. O.; He, Y.; and Jung, H. K. 2016. Augmented reality navigation system on android. *International Journal of Electrical and Computer Engineering*, 6(1): 406.

Deng, N.; He, Z.; Ye, J.; Duinkharjav, B.; Chakravarthula, P.; Yang, X.; and Sun, Q. 2022. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11): 3854–3864.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*, 4171–4186.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2023. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 23171–23181. IEEE.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; and Gould, S. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 1643–1653.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jain, V.; Magalhaes, G.; Ku, A.; Vaswani, A.; Ie, E.; and Baldridge, J. 2019. Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1862–1872.

Jang, S. H. 2012. A qr code-based indoor navigation system using augmented reality. In *GIScience–Seventh International Conference on Geographic Information Science*.

Juile, J.; Chandrasekaran, S.; and Surya, P. 2024. ARGuide Pro: An AR-based Indoor Navigation. In *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–4. IEEE.

Katragadda, S.; Lee, W.; Peng, Y.; Geneva, P.; Chen, C.; Guo, C.; Li, M.; and Huang, G. 2024. Nerf-vins: A real-time neural radiance field map-based visual-inertial navigation system. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 10230–10237. IEEE.

Katz, B. F.; Kammoun, S.; Parseihian, G.; Gutierrez, O.; Brilhault, A.; Auvray, M.; Truillet, P.; Denis, M.; Thorpe, S.; and Jouffrais, C. 2012. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality*, 16(4): 253–269.

Kim, J.; and Jun, H. 2008. Vision-based location positioning using augmented reality for indoor navigation. *IEEE Transactions on Consumer Electronics*, 54(3): 954–962.

Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; and Lee, S. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 104–120. Springer.

Ku, A.; Anderson, P.; Patel, R.; Ie, E.; and Baldridge, J. 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4392–4412.

Lee, L.-H.; Braud, T.; Hosio, S.; and Hui, P. 2021. Towards augmented reality driven human-city interaction: Current research on mobile headsets and future challenges. *ACM Computing Surveys (CSUR)*, 54(8): 1–38.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Liu, Z.; Qi, X.; and Fu, C.-W. 2021. 3d-to-2d distillation for indoor scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4464–4474.

Ma, C.-Y.; Lu, J.; Wu, Z.; AlRegib, G.; Kira, Z.; Socher, R.; and Xiong, C. 2019. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations (ICLR)*.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Majumdar, A.; Ajay, A.; Zhang, X.; Putta, P.; Yenamandra, S.; Henaff, M.; Silwal, S.; Mcvay, P.; Maksymets, O.; Arnaud, S.; et al. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Majumdar, A.; Shrivastava, A.; Lee, S.; Anderson, P.; Parikh, D.; and Batra, D. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, 259–274. Springer.

Radwan, N.; Valada, A.; and Burgard, W. 2018. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4): 4407–4414.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rehman, U.; and Cao, S. 2016. Augmented-reality-based indoor navigation: A comparative analysis of handheld devices versus google glass. *IEEE Transactions on Human-Machine Systems*, 47(1): 140–151.

Tadepalli, S. K.; Ega, P. A.; and Inugurthi, P. K. 2021. Indoor navigation using augmented reality. *International journal of scientific research in science and technology*, 7(4): 588–592.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.-F.; Wang, W. Y.; and Zhang, L. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6629–6638.

Wei, W.; Ren, X.; Tang, J.; Wang, Q.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, 806–815.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yan, L.; Liu, D.; Song, Y.; and Yu, C. 2020. Multimodal aggregation approach for memory vision-voice indoor navigation with meta-learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5847–5854. IEEE.

Yan, L.; Ma, S.; Wang, Q.; Chen, Y.; Zhang, X.; Savakis, A.; and Liu, D. 2022. Video captioning using global-local representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6642–6656.

Yan, L.; Wang, Q.; Zhao, J.; Guan, Q.; Tang, Z.; Zhang, J.; and Liu, D. 2024. Radiance field learners as uav first-person viewers. In *European Conference on Computer Vision*, 88–107. Springer.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yeh, H.-T.; Chen, B.-C.; Yang, C.-T.; and Weng, P.-L. 2018. New navigation system combining QR-Code and augmented reality. *Journal of Internet Technology*, 19(2): 565–571.

Yokoyama, N.; Ha, S.; Batra, D.; Wang, J.; and Bucher, B. 2024. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, 42–48. IEEE.

Yu, B.; Kasaei, H.; and Cao, M. 2023. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3554–3560. IEEE.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7641–7649.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 42829–42842. PMLR.

Zuo, X.; Merrill, N.; Li, W.; Liu, Y.; Pollefeys, M.; and Huang, G. 2021. CodeVIO: Visual-inertial odometry with learned optimizable dense depth. In *2021 ieee international conference on robotics and automation (icra)*, 14382–14388. IEEE.