# NaVLA$^2$: A Vision-Language-Audio-Action Model for Multimodal Instruction Navigation

**Jugang Fan**[1, 2*], **Peihao Chen**[3*], **Changhao Li**[1*], **Qing Du**[1†], **Jian Chen**[1†], **Mingkui Tan**[1, 2]

[1]South China University of Technology,
[2]Peng Cheng Laboratory,
[3]MindOn
{felixfjg123, phchencs, changhaoli65}@gmail.com;
{duqing, ellachen, mingkuitan}@scut.edu.cn

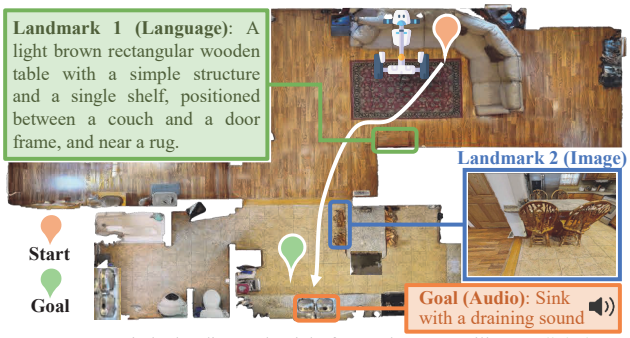
## Abstract

Embodied navigation is a fundamental capability for intelligent agents, yet remains challenging in partially observable environments where navigation instructions can be difficult to interpret. However, existing tasks only provide unimodal instructions, which are ambiguous in complex multimodal environments with multiple similar objects, and may result in misinterpretation and navigation failure. To overcome these limitations, we introduce MINav, a novel task where the navigation path is precisely described by a multimodal instruction. The instruction provides multimodal cues, including object categories, RGB images, language descriptions, and auditory descriptions, which help the agent to disambiguate and ground objects in the environment and navigate effectively. We further construct a large-scale dataset of 43.9K navigation episodes using a two-stage pipeline that first annotates multimodal references of objects and then synthesizes diverse multimodal instructions. We find that existing methods struggle on MINav task, indicating substantial room for improvement in agents' multimodal grounding. To address this, we propose NaVLA$^2$, a vision-language-audio-action model that additionally integrates spatial audio and employs a CoThinkAct module to jointly generate high-level reasoning and consistent low-level actions. Experimental results demonstrate that NaVLA$^2$ significantly outperforms competitive baselines on MINav benchmark. We hope that our proposed MINav and NaVLA$^2$ will facilitate future research toward agents with stronger multimodal understanding and grounding capabilities for navigation.

**Code** — https://github.com/felixfjg/NaVLA

## 1 Introduction

Embodied navigation (Zhu et al. 2017; Batra et al. 2020; Krantz et al. 2020) has received significant attention in recent years, as it serves as a fundamental capability for embodied agents interacting with humans and the environment. The task requires an agent to navigate to a specified target location in an unseen environment, solely based on its sensory inputs. Since robots often need to navigate to a target location before performing further interaction (Das et al. 2018; Majumdar et al. 2024; Zitkovich et al. 2023), navigation serves as a critical prerequisite and an essential skill in developing intelligent embodied agents. However, robust navigation remains highly challenging due to the partial observability of the environment and ambiguity in goal specification.




**Instruction:** Begin by heading to the right front, where you will spot a light brown rectangular wooden table with a simple structure and a single shelf, positioned between a couch and a door frame, and near a rug. Then, move to the right front to find <image>. Finally, go right front to reach the sink with a draining sound.

Figure 1: In Multimodal Instruction Navigation (**MINav**) task, the agent receives multimodal observations and follows the multimodal instructions to reach the goal location. The simulator renders RGB images and binaural spatial audio at each step. The multimodal instructions, composed of language descriptions, image references, and text-format auditory cues, guide the agent to move through a sequence of landmarks toward the goal object.

To tackle these challenges, prior works have formulated navigation problems into various tasks, which can be broadly categorized into goal-reaching and instruction-following navigation. In **goal-reaching navigation**, an agent must move to the vicinity of a given target specified as a point (Anderson et al. 2018a), an object category (Batra et al. 2020), an image (Zhu et al. 2017), or an audio cue (Chen et al. 2019). Although conceptually simple and intuitive, these tasks can lead to blind search when environments contain many similar objects or large open areas. **Instruction-following navigation** mitigates this issue by providing agents with natural

*Equal Contribution
†Corresponding Author

language descriptions (Anderson et al. 2018b; Krantz et al. 2020) that guide them step-by-step to the target. This helps reduce search space, but existing text-only instructions can still be ambiguous—for example, "go to the chair next to the table" is unclear in a room with multiple tables and chairs.

In contrast, humans naturally leverage multimodal information to localize themselves, recognize landmarks, and reason about the approximate direction of a target. For example, when told "Turn off the kitchen sink where the water is still running" in a scene with multiple sinks, a person probably first locates the kitchen using visual landmarks like cabinets, then combines the sound of running water with visual cues to locate the exact sink. However, this ability is not well reflected in existing agents and benchmarks. Inspired by this, we introduce a new **Multimodal Instruction Navigation (MINav) task**, designed to evaluate an agent's ability to integrate diverse sensory inputs and follow rich multimodal instructions. In this task, the agent is given an instruction that describes the passing landmark and final goal with multimodal information, and it is required to navigate in an unseen environment to reach the final goal object.

To support this task, we develop a simulator with visual and realistic acoustic rendering, and further design a two-stage dataset construction pipeline to generate a multimodal instruction navigation dataset. Firstly, we automatically annotate each object in the environment with aligned multimodal references, including object category, representative image, language description, and auditory description. Then, we synthesize diverse and natural instructions based on these references. As a result, we curate a large-scale dataset consisting of 43.9K multimodal instruction episodes.

While multimodal instruction navigation is clearer in terms of description-level, it requires the agent to have a higher multimodal understanding ability, making it more challenging. We found that existing navigation methods struggle on the MINav task, which also shows that there is still a lot of research space on the multimodal understanding ability of intelligent agents, further demonstrating the importance of our proposed task. Among existing methods, vision-language-action (VLA) models (Zhang et al. 2024b; Zheng et al. 2024a) have recently shown promising results, thanks to the powerful capabilities of the VLM backbone (Liu et al. 2023; Li, Wang, and Jia 2024). However, current VLA approaches face two major limitations: 1) they typically process only vision and language, neglecting the spatial sound, which is essential for accurately localizing sound-emitting objects in environments; 2) they operate in an end-to-end paradigm that maps inputs directly to actions, limiting interpretability and reasoning transparency. To address these issues, we propose **NaVLA$^2$**, the first model to integrate vision, language, and audio for navigation. Specifically, inspired by how humans locate sound-emitting objects, we design a spatial-semantic audio encoding module to extract features of spatial audio. To enhance transparency, we introduce a CoThinkAct module, which outputs a high-level chain-of-thought describing agent's intent and parallelly decodes the hidden state into intent-corresponding low-level action sequences. Experiments show that NaVLA$^2$ achieves state-of-the-art performance on MINav while providing interpretable decision making.

To summarize, our contributions are as follows:

- We introduce MINav, a new navigation benchmark that provides multimodal instruction to guide agents in better grounding and reaching target objects. By reflecting the multimodal nature of real-world navigation, MINav aims to advance the development of more perceptually grounded and capable multimodal intelligent agents.
- We develop an automated pipeline and a simulator to generate a richly annotated dataset at scale, enabling large-scale multimodal navigation episodes in 3D scenes. Our simulator dynamically renders spatial audio with acoustic properties, along with egocentric visual observations, enabling realistic multimodal perception during navigation.
- We propose NaVLA$^2$, a vision-language-audio-action model designed for multi-modal navigation. NaVLA$^2$ not only integrates spatial audio with vision and language inputs, but also features a CoThinkAct reasoning framework that produces interpretable decision making. Experiments show that NaVLA$^2$ significantly outperforms strong baselines on the MINav benchmark.

## 2 Related Work

**Goal-reaching Navigation.** Goal-reaching navigation refers to the task where an agent is given only high-level goals and must explore the environment to reach the specified object. Existing work can be roughly categorized into single-goal navigation and multi-goal navigation. In **single-goal navigation** tasks, the agent needs to reach the specified goal, including image-goal navigation (Zhu et al. 2017; Krantz et al. 2022), object-goal navigation (Batra et al. 2020; Chaplot et al. 2020; Zhang et al. 2023), and audio-goal navigation (Chen et al. 2019, 2020, 2022a). **Multi-goal navigation** (Wani et al. 2020; Marza et al. 2023; Chen et al. 2022b) involves finding multiple targets in a specific order, which is more challenging due to its increased task length. GOAT-Bench (Khanna et al. 2024) further extends this to open-vocabulary settings, where goals are flexibly described using categories, images, or language. Unlike previous tasks that only provide goal descriptions, our task requires the agent to navigate following fine-grained instruction descriptions. This setting is more challenging because the agent must ground each part of the instruction in the environment and take actions accordingly.

**Instruction-following Navigation.** Different from goal-reaching navigation, instruction-following navigation requires the agent to follow detailed language instructions that describe the path and landmarks, and move step by step to reach the goal. In early instruction-following navigation tasks (Anderson et al. 2018b), the agent navigates within a predefined navigation graph following language instructions. Subsequent works (Jain et al. 2019; Ku et al. 2020) expand on this by providing longer and more diverse instructions. VLN-CE (Krantz et al. 2020) introduces a more realistic but challenging setting, where agents can move freely in the environment instead of predefined nodes. LH-VLN (Song et al. 2025) introduces a new setting to handle long-term VLN tasks in multiple stages. Later tasks (Shridhar et al. 2020; Thomason et al. 2020; Gao et al. 2022) go a step further by asking the agent to complete both navigation and interaction

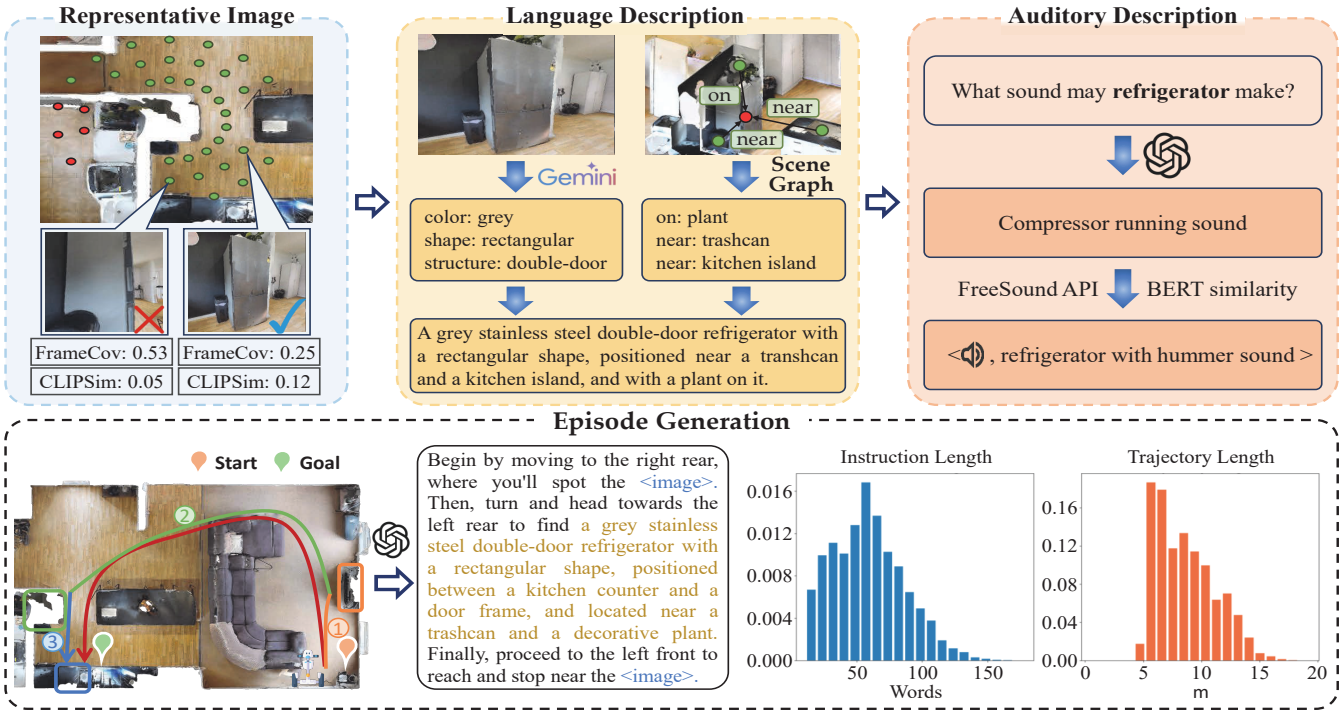



Figure 2: Two-stage dataset construction pipeline for MINav task. We first annotate each object with multimodal references (**Top**). These references are naturally aligned through the object itself (e.g., a refrigerator). Then, we sample trajectories in the scene and combine them with multimodal references to generate instructions, constructing diverse episodes (**Bottom**).

based on instruction or dialogue history. However, these tasks only use text as the instruction description, which may be ambiguous and lead to multiple plausible interpretations. In contrast, we propose a multimodal instruction navigation task that incorporates complementary visual and auditory cues to provide more informative guidance.

**Vision-Language-Action Models for Navigation.** Recent works (Zhang et al. 2024b; Zheng et al. 2024a; Xu et al. 2024) extend Vision-Language Models (Liu et al. 2023; Bai et al. 2023) to embodied settings, forming Vision-Language-Action (VLA) models that enable agents to act based on visual and language inputs. NavGPT-2 (Zhou et al. 2024) integrates a vision language model with a topological map-based navigation strategy to improve autonomous navigation in discrete environments. NaviLLM (Zheng et al. 2024a) introduces schema-based instructions to integrate data from different tasks into the training process, creating a generalist navigation agent. LEO (Huang et al. 2024) introduces a generalist agent that can understand the semantic information of 3D environments to perform multiple tasks, which can also be used for embodied navigation. Some studies (Zhang et al. 2024b; Xu et al. 2024) have applied VLA methods to navigation in real-world environments, demonstrating the generalizability and practicality of these methods. While prior VLA models focus solely on visual and textual inputs, ignoring audio and lacking explicit reasoning for navigation within continuous environments, we propose a NaVLA$^2$ model that integrates spatial audio perception and high-level thinking to enhance multimodal understanding and action planning.

## 3 Multimodal Instruction Navigation

Existing navigation tasks typically rely on either single-modal goal references or purely textual instructions. While such setups are intuitive and easy to design, they fall short in representing the multimodal cues in complex environments, which are important for robust object grounding and disambiguation during navigation. These limitations highlight the need for more expressive task formulations that better reflect how humans perceive and interpret multimodal surroundings. To this end, we propose MINav, a new multimodal instruction navigation task where instructions describe the path using rich multimodal references to guide agents' navigation.

In the following, we first define the MINav task (Section 3.1). We then describe our two-stage automatic dataset construction pipeline, composed of generating multimodal object references (Sections 3.2) and synthesizing navigation episodes (Sections 3.3) based on multimodal references. Finally, we introduce our simulation details (Sections 3.4).

### 3.1 Problem Formulation

In the MINav task, the agent is randomly initialized in an unseen indoor environment and receives a multimodal instruction. As shown in Figure 1, the instruction uses multimodal references as cues to guide the agent to move through landmarks toward the goal object. The multimodal references could be 1) object category, 2) representative image, 3) language description and 4) auditory description. We treat the last object mentioned in the instruction as the goal object, and the previous ones are landmarks.

At each timestep, the agent receives multimodal observations from the environment, including an egocentric RGB image and binaural spatial audio. Based on its understanding of the multimodal instruction and observations, the agent selects a discrete action from the action space $\mathcal{A} = \{\texttt{FORWARD}, \texttt{TURN\_LEFT}, \texttt{TURN\_RIGHT}, \texttt{STOP}\}$, where FORWARD moves the agent forward by 25cm and the turns rotate it by 30°. The episode ends when the agent predicts STOP or reaches the maximum number of steps, and success is determined by whether the agent stops within a 3-meter radius of the goal object.

### 3.2 Multimodal Reference of Objects

Each object in an indoor environment can be described and referenced through multiple modalities, conveying different aspects of its identity. As shown at the top of Figure 2, an object can be identified by its category (e.g., "refrigerator"), visually recognized through a representative image, described with natural language, and also text-format audio cues (e.g., "refrigerator with hummer sound"). However, these references are not well-prepared in existing datasets. To address this, we design an automatic annotation pipeline to generate a tuple of references $\langle r_c, r_i, r_l, r_a \rangle$ for each object in HM3DSem scenes (Yadav et al. 2023). These annotations enable flexible and expressive object references for navigation.

**Object Category $r_c$.** Although HM3DSem contains 1,660 object categories, many are noisy or redundant due to crowdsourcing. To reduce label sparsity and improve consistency, we follow Khanna et al. (2024) to map all objects into a cleaner taxonomy of 312 base categories, making it easier for agents to perform object grounding.

**Representative Image $r_i$.** A well-chosen image can convey more precise object information than a simple category label. Following Krantz et al. (2022), we collect candidate images for each object by rendering views from all valid agent positions within a 2-meter radius, oriented toward the object's center. From these views, we select representative images using a scoring function that considers three factors: 1) frame coverage, defined as the ratio of object pixels to total pixels based on semantic masks; 2) CLIP similarity between the full image and object category; and 3) CLIP similarity between the cropped object image and the category. The final image score is computed as the product of these three metrics, which avoids dominance from any single metric due to scale differences. The highest-scoring image is selected as the most representative image reference $r_i$.

**Language Description $r_l$.** To comprehensively describe an object's appearance and spatial context, it is essential to capture both intrinsic (e.g., color, material) and extrinsic (e.g., spatial relations) attributes. Prior works (Khanna et al. 2024) typically derive both of them from the same egocentric view, limiting their spatial expressiveness. To overcome this, we leverage scene graphs from VLA-3D (Zhang et al. 2024a) to extract rich object-to-object spatial relationships as extrinsic attributes (details in the Appendix). For intrinsic attributes, we feed the most representative image into a vision-language model to generate a caption. Finally, we prompt GPT to summarize both intrinsic and extrinsic cues into a concise and fluent description $r_l$ for the object.

**Auditory Description $r_a$.** In complex environments, sounds often carry distinctive semantic and spatial cues, which humans naturally use for navigation and disambiguation. To introduce this modality, we collect a set of representative audio clip samples for each object category through a three-step process: 1) prompt GPT to generate a description of sounds that an object might emit; 2) use this description to query the Freesound API and retrieve candidate audio clips; and 3) compute the BERT (Devlin et al. 2019) similarity between the GPT-generated description and each audio clip's metadata, discarding samples with similarity scores below 0.5. The GPT-generated textual description serves as the audio reference $r_a$ for this object, while the retrieved audio clip is used to insert into the simulator for spatial audio rendering.

### 3.3 Navigation Episode Generation

Based on annotated reference tuples $\langle r_c, r_i, r_l, r_a \rangle$ for each object, we construct the multimodal instruction navigation dataset. For each episode, we initialize the agent's starting and ending position following two rules from Khanna et al. (2024): 1) both positions must be on the same floor, and 2) their distance ranges between 5 and 30 meters. Then, we plan the shortest trajectory $\tau_{raw}$ between the start and end positions, shown as the red path in Figure 2. Along this trajectory, we select 1 to 5 nearby objects: the last is the goal object, and the rest serve as landmarks. For each object, we randomly choose one modality from its reference tuple to represent it. Next, we begin at the agent's initial position and set the first landmark as the target. A new trajectory is then planned toward this target, which we refer to as $\tau_{stage1}$. The next stage begins from the endpoint of current trajectory and takes the second landmark as the target. By repeating this process, we generate a sequence of trajectories and concatenate them to obtain the final navigation path, denoted as $\tau_{final} = [\tau_{stage1}, \tau_{stage2}, ..., \tau_{stageN}]$. To prevent large deviations, we discard episodes where the ratio between the geodesic distances length of $\tau_{final}$ and $\tau_{raw}$ exceeds 2.

For multimodal instruction generation, we compute the relative direction from the start of each sub-trajectory to its corresponding landmark. Then we use GPT-4o to generate a multimodal instruction, based on the object references and relative direction information (see details in the Appendix). In total, we generated 43,908 episodes across 145 training scenes and 2,628 episodes from 36 test scenes. We then further invite three human experts to select 10 high-quality episodes per test scene based on trajectory videos and instructions. The selected episodes ensure that all referenced objects are present in the environment and the instructions are clear, resulting in a final test-mini subset of 360 episodes.

### 3.4 MINav Simulator

We build our simulator based on the Habitat simulation platform (Savva et al. 2019) and SoundSpaces 2.0 (Chen et al. 2022a). To improve rendering efficiency, we dynamically enable the audio sensor only in episodes where an object is described via an auditory reference. In such episodes, we insert the corresponding audio clip into the center of the referenced object and add a pair of microphones at the agent's height to simulate binaural audio perception. The simulator

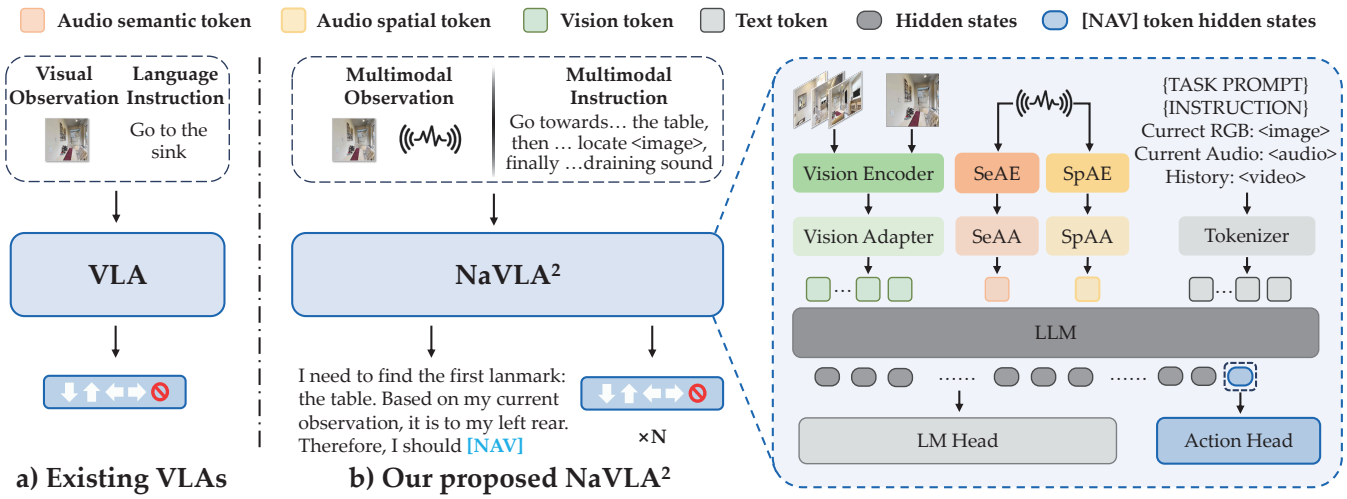



Figure 3: Overview of our **NaVLA²**. Different from existing VLA models, we enable the agent to perceive multimodal observation and understand multimodal instructions. Each modality is first processed by a modality-specific encoder to extract features, which are then concatenated and fed into the LLM in interleaved format. Featuring a two-branch design, NaVLA² simultaneously encodes both spatial and semantic aspects of binaural audio, and jointly produces high-level thinking and low-level action sequence. (Abbreviations: Se – Semantic, Sp – Spatial, AE – Audio Encoder, AA – Audio Adapter.)

renders room impulse responses (RIR), which are convolved with the raw audio clips to produce binaural spatial audio with acoustic properties. Visual observations are captured from the agent's egocentric RGB camera. Both audio and visual observations are updated at each time step, which is closer to real-time observation in real-world.

## 4 NaVLA²

Solving MINav task requires agents to perceive and reason over multimodal sensory inputs. However, existing VLA models typically lack audio grounding, and many treat decision-making as a black-box action generation process, limiting their effectiveness and interpretability for navigation in complex environments. To address these challenges, we further propose NaVLA², a **V**ision-**L**anguage-**A**udio-**A**ction model that integrates spatial audio perception and a CoThinkAct module. NaVLA² jointly reasons over multimodal inputs and produces interpretable multi-step action plans, enabling effective navigation in complex multimodal settings.

### 4.1 Model Architecture

As illustrated in Figure 3, the agent receives multimodal inputs, including multimodal observations (RGB images and binaural spatial audio) and multimodal instructions (image reference, language description, and text-format auditory description). Each input is first encoded using a modality-specific encoder, and then aligned into a shared representational space via lightweight adapters. These aligned embeddings are concatenated in the format of our designed prompt, and then fed into the LLM. The LLM outputs high-level reasoning about the agent's current objective and direction, with a special token that guides low-level action sequence prediction. We elaborate on our spatial-semantic audio encoding strategy and CoThinkAct module in more detail below.

**Spatial-Semantic Audio Encoder.** While prior MLLMs primarily focus on the semantic feature of audio, we argue that spatial cues embedded in stereo sound are equally critical for navigation—just as humans localize objects through directional hearing. To capture both the semantic and spatial features of sound, we introduce a dual-branch audio encoding module. Specifically, the binaural audio is downmixed to mono and encoded by the semantic audio encoder, followed by a learnable semantic audio adapter. In parallel, the original binaural signal is fed into a spatial audio encoder to extract directional cues, with its output passed through a spatial audio adapter. The resulting semantic and spatial embeddings are then concatenated to form the final audio representation. This dual-branch design ensures that the agent perceives not only what is sounding, but also where the sound is coming from, enhancing spatial awareness during navigation.

**CoThinkAct Module.** Most existing VLA models directly map fused representations to actions, making it difficult to interpret the agent's decision-making. To promote interpretable and coherent navigation, we propose the CoThinkAct module, which performs high-level reasoning and low-level action prediction in parallel, both conditioned on the final hidden states from the LLM. Specifically, the LLM last-layer embeddings are taken as shared contextual representation used in two branches: 1) One branch passes them through the `lm_head` to generate a chain-of-thought in natural language. This CoT explicitly outlines the agent's navigation intent, including identifying the current target (which landmark or goal) and estimating its relative direction. The output sequence is designed to end with a special token, `[NAV]`, which inherently summarizes the preceding reasoning process due to the autoregressive nature of LLMs. 2) The other branch extracts the last hidden embedding corresponding to `[NAV]` token and feeds it into the `action_head`, which decodes it into a

sequence of $N$ low-level intent-corresponding actions. This design allows the agent to think and act concurrently, yielding interpretable and consistent decision-making and efficient multi-step action prediction in a single forward pass of LLM.

### 4.2 The Training of the NaVLA$^2$

To enhance multimodal understanding and navigation ability, we adopt a three-stage training pipeline for NaVLA$^2$. The first stage aligns audio with text, the second stage enhances ability to understand each modality independently, and the third fine-tunes the model on navigation dataset with interleaved multimodal inputs. We use Vicuna-7B-v1.5 as the LLM (same as Zhang et al. (2024b) for fair comparison), CLIP ViT-L/14 as the vision encoder, CLAP and SpatialAST as semantic and spatial audio encoders, respectively. All adapters and the action head are implemented as two-layer MLPs.

**Stage 1: Audio-Text Alignment.** In this stage, we aim to align all modalities into a shared language space via modality adapters. Since LLaVA already provides a well-aligned vision adapter, we directly reuse it and focus solely on audio adapters. To perform spatial-semantic audio-text alignment, we first collect 271K semantic-audio-text pairs by merging and cleaning existing audio caption datasets (Kim et al. 2019; Mei et al. 2024; Drossos, Lipping, and Virtanen 2020). For each audio clip, we sample three reverberation-direction pairs from dataset provided by Zheng et al. (2024b). Then, we convolve each audio clip with the sampled reverberation RIR to obtain binaural audio, and append the direction information to the caption, generating 813K spatial–audio–text pairs. We train the audio adapters on this dataset for 2 epochs.

**Stage 2: Multimodal Instruction-tuning.** In this stage, we fully fine-tune the LLM to understand individual modalities, including image, video, and audio. We build a diverse multimodal QA dataset by sampling from existing datasets (Zhang et al. 2024c; Li et al. 2024; Zheng et al. 2024b), yielding a total of 1M samples. The model is trained for 1 epoch.

**Stage 3: MINav Fine-tuning.** We finally fine-tune NaVLA$^2$ on MINav dataset, where each input combines interleaved multimodal instructions and observations, presenting a more challenging setting. We generate 379K training samples from navigation episodes, each in the format: $\langle I,\ O^v_{0:t-1},\ O^v_t,\ O^a_t,\ A_{t:t+N-1}\rangle$ where $I$ is the multimodal instruction, $O^v_{0:t-1}$ is the historical visual observation, $O^v_t$ and $O^a_t$ denote the current visual and audio observations, and $A_{t:t+N-1}$ is the future $N$-step action sequence. Historical visual observations are uniformly sampled $K$ frames from all past RGB observations and encoded following the video representation method in Zhang et al. (2024c). Each sample is formatted into the prompt as below. We apply LoRA fine-tuning to the LLM with a rank of 16, and jointly train the token embeddings, lm head, and action head.

> The instruction guides you through landmarks to reach the final target object. Based on the current observation and historical observations, infer the direction of the landmark or target object relative to your position, and plan your next action.\n<instruction>{INSTRUCTION}</instruction>\n<visual_obs>Current RGB Observation: <image></visual_obs>\n<audio_obs>Current Audio Observation: <audio></audio_obs>\n<history_obs>History Observation: <video></history_obs>

## 5 Experiments

### 5.1 Comparison on MINav Benchmark

**Evaluation Metrics.** An episode is considered successful if the agent predicts the `STOP` action within a 3-meter radius of the goal object. We adopt metrics used in previous works (Krantz et al. 2020; Datta et al. 2021), including Success Rate (SR), Oracle Success Rate (OSR), Success weighted by Path Length (SPL), Navigation Error (NE), Trajectory Length (TL), and normalized Dynamic Time Warping (nDTW). All evaluations are conducted in previously unseen environments.

**Baselines.** To comprehensively evaluate the effectiveness of our proposed NaVLA$^2$, we compare it with a diverse set of baselines covering different paradigms, including map-based, reinforcement learning, imitation learning, MLLM, and VLA models. More details about baselines are in the Appendix.

- **Random** agent selects actions according to the action distribution from the training dataset, which is 54.1% `FORWARD`, 22.8% `TURN_LEFT`, 21.5% `TURN_RIGHT`, 1.6% `STOP`.
- **Qwen2.5-Omni** (Xu et al. 2025) supports a mix of vision, language and audio as input. We use a similar prompt as our method to let it make both reasoning and action planning.
- **Gemini-1.5** (Team et al. 2024) is another Omni-MLLMs. We use the same prompt as Qwen2.5-Omni.
- **CA-Nav** (Chen et al. 2024) parses instructions into subgoals using GPT-4 and plans via value mapping. As it is tailored for VLN, we use CLIP to replace image references with the most similar category label to convert instructions.
- **RL** baselines use DD-PPO to train two variants: providing either multimodal instructions (instr-nav) or the goal object only (goal-nav). The reward follows standard VLN, combining goal success, distance reduction, and time penalty.
- **IL** baseline is trained via behavior cloning with the same network as RL. It uses third-stage training data of NaVLA$^2$ and is supervised with cross-entropy loss on expert actions.
- **Navid** (Zhang et al. 2024b) is current open-source SoTA method trained on collected video-action dataset from VLN-CE. We use the same procedure as in CA-Nav to adapt it to our task. As Navid does not introduce its training details, we apply the same training setting as ours and train on the same amount of data for fair comparison.

**Results and Analysis.** As shown in Table 1, our proposed NaVLA$^2$ significantly outperforms all strong baselines, achieving a 27.2% SR (+11.6%) and 19.6% SPL (+4.4%). We also observe that existing VLN-CE SoTA methods, including CA-Nav and Navid, fall short on MINav. This highlights limitations of current methods when extended to more complex multimodal settings. These findings reveal there remains substantial room for future research on the multimodal understanding and navigation capabilities of agents, further demonstrating the importance of our proposed MINav task.

Notably, we find that RL(instr-nav) baseline significantly outperforms RL(goal-nav) on most metrics. This suggests that full-path instructions with intermediate landmarks provide richer supervision signals during training and facilitates more effective decision-making. In contrast, RL(goal-nav), which relies only on the final goal description, suffers from

| Method | Input | SR ↑ | OSR ↑ | SPL ↑ | SoftSPL ↑ | NE ↓ | nDTW ↑ | TL ↓ |
|---|---|---|---|---|---|---|---|---|
| **Zero-shot** | | | | | | | | |
| Random | – | 0.036 | 0.064 | 0.027 | 0.071 | 7.970 | 0.332 | 4.541 |
| Qwen2.5-Omni (Xu et al. 2025) | V, A, T | 0.039 | 0.039 | 0.024 | 0.069 | 7.722 | 0.359 | 3.854 |
| Gemini-1.5-flash (Team et al. 2024) | V, A, T | 0.056 | 0.086 | 0.049 | 0.105 | 7.388 | 0.380 | 4.365 |
| CA-Nav + GPT-4 (Chen et al. 2024) | V, T | 0.059 | 0.175 | 0.022 | 0.046 | 7.636 | 0.236 | 20.790 |
| **Finetuned** | | | | | | | | |
| IL(instr-nav) | V, A, T | 0.058 | 0.103 | 0.050 | 0.127 | 7.728 | 0.404 | 5.162 |
| RL(goal-nav) | V, A, T | 0.064 | 0.220 | 0.035 | 0.069 | 7.828 | 0.220 | 14.723 |
| RL(instr-nav) | V, A, T | 0.156 | 0.172 | 0.152 | **0.256** | 6.350 | **0.509** | **4.044** |
| Navid (Zhang et al. 2024b) | V, T | 0.131 | 0.336 | 0.085 | 0.121 | 8.020 | 0.313 | 11.506 |
| **NaVLA$^2$ (Ours)** | V, A, T | **0.272** | **0.433** | **0.196** | 0.232 | **6.242** | 0.474 | 12.078 |

Table 1: Comparison with different methods on MINav-bench. **Zero-shot** models are evaluated without fine-tuning. **Finetuned** models are trained using imitation, reinforcement, or supervised learning. Input modalities: V = Vision, A = Audio, T = Text.

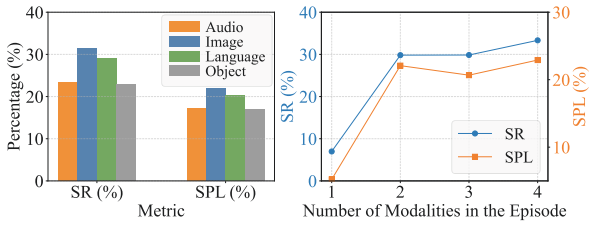



Figure 4: Impact of different modalities on performance.

| Spatial | Semantic | SR ↑ | SPL ↑ | NE ↓ |
|---|---|---|---|---|
| × | × | 0.206 | 0.162 | 6.677 |
| × | ✓ | 0.219 | 0.168 | 6.631 |
| ✓ | × | 0.258 | **0.201** | **6.147** |
| ✓ | ✓ | **0.272** | 0.196 | 6.242 |

Table 2: Ablation study of audio encoder design, showing how spatial and semantic branches contribute to performance.

sparser feedback and increased ambiguity in complex environments. These findings underscore the importance of instructions in guiding agents under multimodal settings.

We further analyze the impact of different modalities on NaVLA$^2$'s performance, as shown in Figure 4. On the left, we report metrics grouped by the modality of the final goal object; on the right, we show metrics grouped by the number of different modality references within episodes. Results indicate that relying on single modality leads to lowest navigation success rate. Providing more modalities helps the agent resolve ambiguity and reach the goal. Among all modalities, representative images contribute most to disambiguation, while using only object categories performs the worst.

## 5.2 Ablation Study

**Effect of Spatial Audio Perception.** Table 2 presents an ablation study on the design of our dual-branch audio encoder. Removing both branches leads to the lowest performance, highlighting the importance of audio perception. Adding only the semantic branch yields a modest improvement in SR and SPL, while adding only the spatial branch leads to a larger gain across all metrics, demonstrating its critical role in spatial reasoning. Combining both branches achieves the best success rate (27.2%), confirming the complementary effects of spatial and semantic information in audio for navigation.

**Effect of the CoThinkAct Module.** As shown in Table 3, removing the action head and instead using the LLM to predict actions in text form degrades performance, especially in SPL. This confirms that decoding actions from the `[NAV]` token embedding yields more effective trajectory planning consistent with high-level CoT. Disabling chain-of-thought and just outputting `[NAV]` token also leads to a performance drop, indicating that explicit reasoning supports better navigation decisions. These results demonstrate that the high-level thinking and low-level action sequence prediction work together to enhance both interpretability and navigation performance.

| Setting | SR ↑ | SPL ↑ | NE ↓ |
|---|---|---|---|
| No Action Head | 0.256 | 0.165 | 6.582 |
| No CoT | 0.261 | 0.192 | 6.304 |
| **CoThinkAct** | **0.272** | **0.196** | **6.242** |

Table 3: Ablation study on the CoThinkAct module.

# 6 Conclusion

We introduce MINav, a new multimodal instruction navigation task where the instruction provide detailed description about the path using multimodal references. We further propose NaVLA$^2$, a vision-language-audio-action model that integrates spatial audio and output both high-level thinking and low-level actions for effective navigation. Experiments show that existing state-of-the-art models struggle on MINav, while our approach significantly outperforms them. We hope our proposed MINav and NaVLA$^2$ will advance future research on developing agents with stronger multimodal understanding and grounding abilities.

## Acknowledgments

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U24A20327), the National Natural Science Foundation of China (Grant No. 62376099), and Natural Science Foundation of Guangdong Province (Grant No. 2024A1515010989).

## References

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018b. 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.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Batra, D.; Gokaslan, A.; Kembhavi, A.; Maksymets, O.; Mottaghi, R.; Savva, M.; Toshev, A.; and Wijmans, E. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.

Chaplot, D. S.; Gandhi, D. P.; Gupta, A.; and Salakhutdinov, R. R. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33: 4247–4258.

Chen, C.; Jain, U.; Schissler, C.; Gari, S. V. A.; Al-Halah, Z.; Ithapu, V. K.; Robinson, P.; and Grauman, K. 2019. Audio-visual embodied navigation. *environment*, 97: 103.

Chen, C.; Jain, U.; Schissler, C.; Gari, S. V. A.; Al-Halah, Z.; Ithapu, V. K.; Robinson, P.; and Grauman, K. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *The European Conference on Computer Vision*, 17–36.

Chen, C.; Schissler, C.; Garg, S.; Kobernik, P.; Clegg, A.; Calamia, P.; Batra, D.; Robinson, P.; and Grauman, K. 2022a. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35: 8896–8911.

Chen, K.; An, D.; Huang, Y.; Xu, R.; Su, Y.; Ling, Y.; Reid, I.; and Wang, L. 2024. Constraint-Aware Zero-Shot Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2412.10137*.

Chen, P.; Ji, D.; Lin, K.; Hu, W.; Huang, W.; Li, T.; Tan, M.; and Gan, C. 2022b. Learning active camera for multi-object navigation. *Advances in Neural Information Processing Systems*.

Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–10.

Datta, S.; Maksymets, O.; Hoffman, J.; Lee, S.; Batra, D.; and Parikh, D. 2021. Integrating egocentric localization for more realistic point-goal navigation agents. In *Conference on Robot Learning*, 313–328.

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: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 736–740.

Gao, X.; Gao, Q.; Gong, R.; Lin, K.; Thattai, G.; and Sukhatme, G. S. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4): 10049–10056.

Huang, J.; Yong, S.; Ma, X.; Linghu, X.; Li, P.; Wang, Y.; Li, Q.; Zhu, S.-C.; Jia, B.; and Huang, S. 2024. An Embodied Generalist Agent in 3D World. In *Proceedings of the International Conference on Machine Learning*.

Jain, V.; Magalhães, G.; Ku, A.; Vaswani, A.; Ie, E.; and Baldridge, J. 2019. Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1862–1872. Association for Computational Linguistics.

Khanna, M.; Ramrakhya, R.; Chhablani, G.; Yenamandra, S.; Gervet, T.; Chang, M.; Kira, Z.; Chaplot, D. S.; Batra, D.; and Mottaghi, R. 2024. GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16373–16383.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132.

Krantz, J.; Lee, S.; Malik, J.; Batra, D.; and Chaplot, D. S. 2022. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*.

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

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 Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4392–4412. Association for Computational Linguistics.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, Y.; Wang, C.; and Jia, J. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *The European Conference on Computer Vision*, 323–340.

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

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16488–16498.

Marza, P.; Matignon, L.; Simonin, O.; and Wolf, C. 2023. Multi-Object Navigation with dynamically learned neural implicit representations. In *The IEEE International Conference on Computer Vision*, 11004–11015.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 3339–3354.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; Parikh, D.; and Batra, D. 2019. Habitat: A Platform for Embodied AI Research. In *The IEEE International Conference on Computer Vision*.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10740–10749.

Song, X.; Chen, W.; Liu, Y.; Chen, W.; Li, G.; and Lin, L. 2025. Towards long-horizon vision-language navigation: Platform, benchmark and method.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, 394–406.

Wani, S.; Patel, S.; Jain, U.; Chang, A.; and Savva, M. 2020. Multion: Benchmarking semantic map memory using multi-object navigation. *Advances in Neural Information Processing Systems*, 33: 9700–9712.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Xu, Z.; Chiang, H. L.; Fu, Z.; Jacob, M. G.; Zhang, T.; Lee, T. E.; Yu, W.; Schenck, C.; Rendleman, D.; Shah, D.; Xia, F.; Hsu, J.; Hoech, J.; Florence, P.; Kirmani, S.; Singh, S.; Sindhwani, V.; Parada, C.; Finn, C.; Xu, P.; Levine, S.; and Tan, J. 2024. Mobility VLA: Multimodal Instruction Navigation with Long-Context VLMs and Topological Graphs. In *Conference on Robot Learning*, volume 270, 3866–3887.

Yadav, K.; Ramrakhya, R.; Ramakrishnan, S. K.; Gervet, T.; Turner, J.; Gokaslan, A.; Maestre, N.; Chang, A. X.; Batra, D.; Savva, M.; et al. 2023. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4927–4936.

Zhang, H.; Zantout, N.; Kachana, P.; Wu, Z.; Zhang, J.; and Wang, W. 2024a. VLA-3D: A dataset for 3D semantic scene understanding and navigation. *arXiv preprint arXiv:2411.03540*.

Zhang, J.; Dai, L.; Meng, F.; Fan, Q.; Chen, X.; Xu, K.; and Wang, H. 2023. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6672–6682.

Zhang, J.; Wang, K.; Xu, R.; Zhou, G.; Hong, Y.; Fang, X.; Wu, Q.; Zhang, Z.; and Wang, H. 2024b. NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation. In *Robotics: Science and Systems, 2024*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zheng, D.; Huang, S.; Zhao, L.; Zhong, Y.; and Wang, L. 2024a. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 13624–13634.

Zheng, Z.; Peng, P.; Ma, Z.; Chen, X.; Choi, E.; and Harwath, D. 2024b. BAT: Learning to Reason about Spatial Sounds with Large Language Models. In *Proceedings of the International Conference on Machine Learning*.

Zhou, G.; Hong, Y.; Wang, Z.; Wang, X. E.; and Wu, Q. 2024. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *The European Conference on Computer Vision*, 260–278.

Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J. J.; Gupta, A.; Fei-Fei, L.; and Farhadi, A. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE international conference on robotics and automation*, 3357–3364.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; Vuong, Q.; Vanhoucke, V.; Tran, H. T.; Soricut, R.; Singh, A.; Singh, J.; Sermanet, P.; Sanketi, P. R.; Salazar, G.; Ryoo, M. S.; Reymann, K.; Rao, K.; Pertsch, K.; Mordatch, I.; Michalewski, H.; Lu, Y.; Levine, S.; Lee, L.; Lee, T. E.; Leal, I.; Kuang, Y.; Kalashnikov, D.; Julian, R.; Joshi, N. J.; Irpan, A.; Ichter, B.; Hsu, J.; Herzog, A.; Hausman, K.; Gopalakrishnan, K.; Fu, C.; Florence, P.; Finn, C.; Dubey, K. A.; Driess, D.; Ding, T.; Choromanski, K. M.; Chen, X.; Chebotar, Y.; Carbajal, J.; Brown, N.; Brohan, A.; Arenas, M. G.; and Han, K. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning*, volume 229, 2165–2183.