VISTAv2: World Imagination for Indoor Vision-and-Language Navigation

Yanjia Huang Xianshun Jiang Xiangbo Gao Mingyang Wu Zhengzhong Tu*
Texas A&M University, TACO Group
https://taco-group.github.io/

Abstract

Vision-and-Language Navigation (VLN) requires agents to follow language instructions while acting in continuous real world spaces. Prior image imagination based VLN work shows benefits for discrete panoramas but lacks online, action-conditioned predictions and does not produce explicit planning values; moreover, many methods replace the planner with long-horizon objectives that are brittle and slow. To bridge this gap, we propose ***VISTAv2****, a generative world model that rolls out egocentric future views conditioned on past observations, candidate action sequences, and instructions, and projects them into an online value map for planning. Unlike prior approaches, VISTAv2 does not replace the planner. The online value map is fused at score level with the base objective, providing reachability and risk-aware guidance. Concretely, we employ a Conditional Diffusion Transformer video predictor aware of action to synthesize short-horizon futures, align them with the natural language instruction via a vision-language scorer, and fuse multiple rollouts in a differentiable Imagination-to-Value head to output an imagined egocentric value map. For efficiency, rollouts occur in VAE latent space with a distilled sampler and sparse decoding, enabling inference on a single consumer GPU. Evaluated on MP3D and RoboTHOR, VISTAv2 improves over strong baselines, and ablations show that action-conditioned imagination, instruction-guided value fusion, and the online value-map planner are all critical—suggesting that VISTAv2 offers a practical and interpretable route to robust VLN.*

1. Introduction

"The map is not the territory." - Alfred Korzybski

Enabling embodied agents to understand natural instructions and locate them both quickly and robustly in real-world environments without prior maps still remains a core challenge in Vision-and-Language Navigation tasks [1–5]. There are two different methods for approaching this, each with strengths and limitations. One relies on Large Vision-Language Models (VLMs) to match "currently observed objects" with "verbalized goals" through scoring the similarity, then employs frontier exploration or occupancy maps for "semantically driven mapping and search" (e.g., VLFM [6]). The other leverages world models to perform long-range rollouts in action space, evaluating path quality by "imagining future scenes" (e.g., Navigation World Model and VISTA [7, 8]). The former exhibits strong semantic generalization and simple deployment but lacks explicit evaluation of reachability and geometric information, often leading to looks right but turns out wrong detours. The latter can explicitly assess evidence of reachability but suffers from long-rollout fragility which makes it challenging to achieve stable gains.

In this paper, we argue that VLN benefits from *short-horizon, action-conditioned imagination that lives in map space*. We observe the followings. Pure language–vision matching tells what looks relevant but not whether it is reachable from the current pose. It cannot foresee occlusions or near-term collisions, and often over-scores visually salient but blocked targets (e.g., behind walls, glass, or outside the current room), leading to detours and backtracking [9–14]. Optimizing a full world model objective over long-horizons amplifies small pose errors and sampling noise; appearance drift across dozens of denoising steps causes the planner to chase artifacts rather than geometry. The compute budget required for long rollouts also reduces the number of candidates we can evaluate per step [15–19].

To address these challenges, we present **VISTAv2**, a language-conditioned, action-aware generative world model. VISTAv2 (i) rolls out short-horizon egocentric futures, (ii) turns them into an imagined value map by combining instruction–vision alignment with traversability and obstacle cues, and (iii) re-ranks planner candidates via score-level fusion with the base objective. Overall, VISTAv2 is a test-time plug-in to standard frontier-based exploration planners, runs online through latent-space diffusion with distillation, and shows improvements on VLN datasets.

Our contributions are summarized as follows:

*Corresponding author: tzz@tamu.edu

Figure 1. **VISTAv2 pipeline overview (§3.1).** From a language instruction and observations (RGB, depth, odometry), the agent: (1) builds a local map and proposes frontier-based candidate trajectories; (2) forms a language prior over the map (Value); (3) uses the world model to imagine short-horizon futures and converts them into an egocentric imagined value map; (4) fuses imagined value and the prior with the planner's native score to rank candidates (Eq. (2)) and executes the first control in a receding-horizon loop.

- A world model for VLN that performs *short-horizon, action-conditioned* rollouts and expresses their guidance as an *egocentric value map* in map space.
- An *Imagination-to-Value* head that converts predicted futures into a value map and guides planning via score fusion, without replacing the planner.
- Compared to VISTA, VISTAv2 improves Val-Unseen SR and SPL by +3.6 and +5.4 with shorter TL (13.26→10.73, -19%), and improves Test-Unseen SPL by +2.3 with shorter TL (14.20→12.44, -12%).

2. Related Work

Vision-and-Language Navigation (VLN). VLN links natural language instructions with embodied perception and control [20–29]. Early work posed the problem on panoramic navigation graphs with R2R and RxR [3, 30], later enabled at scale by the Habitat simulator [31] for photo-realistic, high speed training and evaluation. And beyond navigation graphs, VLN-CE lifts agents into continuous 3D control with egocentric sensing and realistic collisions, removing graph constraints and exposing geometric challenges central to real robots [32]. Recent trends leverage foundation or VLM models to strengthen language grounding and zero-shot generalization [8, 17, 18, 33–36], complementary work studies injecting visual imagination as an added modality to provide landmark cues, yielding measurable SR gains on VLN agents. However, current VLN systems still struggle to reason about reachability and collision risk, VLM only scoring often favors visually salient yet unreachable directions, and long-horizon rollouts can be brittle and slow, compounding errors in continuous settings.

Video Generation as World Models. Video generative models can act as controllable world models that roll out egocentric futures conditioned on past observations and actions. Recent systems show that such models can steer planning or augment downstream reasoning, either by playing out imagined trajectories or by ranking candidate plans using synthesized evidence [37–41]. Notable examples include large, action-conditioned controllable world models like Genie [42], which could generate interactive environments from both image or text prompts and allow agent control via input actions. Navigation World Models instantiate this idea for visual navigation with a controllable video generator that predicts future observations from past frames and navigation actions [7]. The model is trained on diverse egocentric videos and is used after training to simulate trajectories and verify whether a candidate plan reaches the

Figure 2. **World Model (§3.2).** Given the recent egocentric frames $I_{t-m+1:t}$, the instruction g, and a candidate trajectory A, we integrate poses to obtain $C(A) \in SE(2)^H$ and feed $(x_t, g, C(A))$ to the action-conditioned video diffusion model $\mathcal{W}_\theta$ (CDiT in VAE latent space). $\mathcal{W}_\theta$ produces a short-horizon egocentric rollout $\{\hat{I}_{t+\tau}\}_{\tau=1}^{H}$; only a stride-$\Delta t$ subset is decoded for downstream I2V scoring (§3.3).

goal. NWM is built on a Conditional Diffusion Transformer (CDiT), whose complexity is linear in the number of context frames, enabling scaling to large models while maintaining strong prediction quality. At deployment, NWM uses the generator in MPC and re-ranking for both insides and outsides of rollouts. This design reduces exposure to long-horizon drift without rewriting the planner.

Language Guided Mapping and Frontier Exploration. This line of work steers exploration by aligning egocentric observations with instruction text and projecting the scores onto maps or frontiers. A representative example is VLFM, which builds a frontier map and ranks candidate frontiers by vision–language similarity to the goal text, enabling zero-shot semantic navigation without task-specific training [6]. Follow up works extend the idea with panoramic cues and imagination from discrete viewpoints [17], further improving frontier selection under language guidance [34]. These methods provide strong semantic priors and simple deployment, yet they reason about reachability and collision risk only indirectly, high-scoring regions can be occluded or geometrically infeasible, and the scores are typically static [10, 43–46]. Our work complements this trend by injecting an *action conditioned, short-horizon* prior: we convert imagined futures into an egocentric value map and fuse it with the planner's native objective at score level, preserving the underlying map-based search while adding instruction consistent, geometry-aware guidance.

3. Method

We target *test-time enhancement* for Vision-and-Language Navigation in standard VLN environments: at inference, an agent augments its default policy with a language-conditioned, action aware generative world model that imagines future egocentric observations and converts them into an online value map for planning (Fig. 1). Our approach, VISTAv2, couples two components tightly: ❶ a video diffusion **world model** that rolls out short-horizon egocentric futures conditioned on the instruction and candidate actions; ❷ an **Imagination-to-Value (I2V)** head converts the imagined futures into an egocentric value map. At test-time, we do not replace the planner: the value map is fused at score level with the planner's native objective to rank candidate actions, keeping the search procedure intact.

Sec.3.1 overviews the VISTAv2 pipeline. Sec.3.2 formalizes the language-conditioned, action-aware world model. Sec.3.3 introduces the I2V head, which converts multi rollouts, instruction aligned signals into an egocentric value map and fuses it with the base planner's score for action selection. Finally, Sec.3.4 details the architecture and training setup used in our experiments.

3.1. Pipeline Overview

Fig. 1 illustrates our pipeline: given a language instruction g and egocentric observations at time t (RGB I_t, depth D_t, odometry P_t), VISTAv2 augments a standard frontier-based navigation stack with action-conditioned imagination and score-level fusion. Four components participate: a *World Model* $\mathcal{W}$, a *Language Reasoner* $\mathcal{R}$ (text–vision scorer), an *Imagination-to-Value* head $\mathcal{H}$, and a base *Planner* $\mathcal{P}$.

Given egocentric I_t, P_t, and an instruction goal g, we launch an n-step candidate–expansion loop instead of committing to an action directly. For each trajectory in the current candidate set such as frontiers with short-horizon paths, the world model rolls out H-step egocentric futures (we take H=4 here), yielding imagined observations. Conditioned on the instruction, a language–vision scorer (BLIP [47]) evaluates these imagined observations and (i) writes instructive evidence into a small buffer, and (ii) via an Imagination-to-Value (I2V) head converts them into an egocentric **imagined** value map. We fuse the imagined value with the planner's native score to rank candidates and keep the top-K to form the next layer. After the search, the planner executes the first control of the best candidate; mapping updates and the loop repeats at the next step. This *imagine* $\rightarrow$ *score* $\rightarrow$ *fuse* $\rightarrow$ *act* procedure augments a frozen navigation stack with world model priors and motion forecasts, improving VLN at test-time.

3.2. World Model Formulation

We treat the world model as an egocentric video generation simulator that rolls out a sequence of actions from previous observations. Given the last m RGB frames $I_{t-m+1:t}$, poses $P_{1:t}$, an instruction g, and a candidate control sequence $A_{t:t+H-1}$, the model predicts short-horizon egocentric futures that will later be converted into an imagined value map (Sec. 3.3).

Action Space. We use a continuous, egocentric control space

$$\mathcal{A} \subset \mathbb{R}^4, \quad a_\tau = (\Delta x_\tau,\, \Delta y_\tau,\, \Delta\theta_\tau,\, \kappa_\tau),$$

where $(\Delta x, \Delta y)$ are planar displacements in the local frame, $\Delta\theta$ is the yaw change, and κ scales step duration. Feasibility is enforced by platform limits $\|(\Delta x, \Delta y)\| \leq v_{\max}\kappa\Delta t$ and $|\Delta\theta| \leq \omega_{\max}\kappa\Delta t$, and a `STOP` primitive is available when appropriate.

Action Representation. A candidate $A_{t:t+H-1}$ from the base planner is resampled to a fixed horizon H. Each step is embedded as

$$\tilde{a}_\tau = \left[\tfrac{\Delta x_\tau}{\sigma_x},\, \tfrac{\Delta y_\tau}{\sigma_y},\, \sin\Delta\theta_\tau,\, \cos\Delta\theta_\tau,\, \tfrac{\kappa_\tau}{\sigma_\kappa}\right],$$

concatenated with a sinusoidal time code and a learned mode bit (turn vs. go-straight). The sequence $\tilde{A}_{t:t+H-1}$ is provided as per-timestep conditioning tokens.

Action-Driven Video Generation. Let $x_t \in \mathbb{R}^{H\times W\times 3}$ be the current egocentric frame and $C(A) = \big(c_1, \ldots, c_H\big) \in SE(3)^H$ the pose sequence induced by a candidate trajectory A starting from pose P_t (for ground robots $C(A) \subset SE(2)$). Our instruction- and pose-conditioned video diffusion model

$$\mathcal{W}_\theta :\ (x_t, g, C(A)) \mapsto\ \mathbf{X}(A) = \{\hat{x}_{t+\tau}\}_{\tau=1}^{H},\quad \hat{x}_{t+\tau} \in \mathbb{R}^{H\times W\times 3}. \tag{1}$$

maps the triplet $(x_t, g, C(A))$ to an *egocentric rollout* that follows the intended motion while emphasizing instruction-relevant visual evidence. In practice we operate in latent space, i.e., with $z_t = E(x_t)$ and $\hat{z}_{t+\tau}$ rolled out by a CDiT; only a sparse subset is decoded $\hat{x}_{t+\tau} = D(\hat{z}_{t+\tau})$ for downstream scoring.

Language Reasoner and Prior Map. We use a frozen vision–language encoder $\mathcal{R}$ (BLIP) to obtain a text embedding ϕ_g and per-pixel/patch image features $\phi(\cdot)$. Given the current observation set and the local map $\mathcal{M}_t$, we compute a *language prior* V_t^{prior} by projecting alignment scores to the egocentric grid:

$$s(u,v) = \mathrm{softmax}_\tau\big(\cos\langle \phi_g,\, \phi(I_t)[u,v]\rangle\big),$$
$$V_t^{\text{prior}} = \Pi\big(s(\cdot) \rightarrow \mathcal{M}_t\big).$$

where $\Pi(\cdot)$ denotes depth-aware splatting onto the map and τ is a temperature. V_t^{prior} captures instruction relevance but is action-agnostic; it is fused at score level in Eq. (2).

Score-level fusion. For a candidate $A_{t:t+H-1}$, we fuse imagined value and the planner's native score:

$$\begin{aligned} S_{\text{fused}}\big(A_{t:t+H-1}\big) = {} & S_{\text{base}}\big(A_{t:t+H-1}\big) \\ & + \lambda_1 \sum_{\tau=1}^{H} \gamma^\tau\, V_t^{\text{img}}\big(x_\tau(A), y_\tau(A)\big) \\ & + \lambda_2 \sum_{\tau=1}^{H} \gamma^\tau\, V_t^{\text{prior}}\big(x_\tau(A), y_\tau(A)\big). \end{aligned} \tag{2}$$

where $(x_\tau(A), y_\tau(A))$ are egocentric samples along A, $\gamma \in (0, 1]$ is a temporal discount, and $\lambda_1, \lambda_2 \geq 0$ are fusion weights. The highest-scoring candidate is chosen and only its first control is executed (receding horizon).

3.3. Imagination to Value (I2V)

Input. For a candidate $A_{t:t+H-1}$, the world model (Sec. 3.1) provides a sparse set of imagined egocentric frames $\{\hat{I}_{t+\tau}\}$, their poses $\{c_\tau\}$ and depth (predicted or monocular).

Frame scoring. Each $\hat{I}_{t+\tau}$ is converted to a single confidence map by linearly combining three dense cues: (i) instruction alignment (text–vision similarity),

Algorithm 1: $\mathcal{W}_\theta$.ROLLOUT — Action-Driven Video Generation

Input: $I_{t-m+1:t}$, P_t, instruction g, candidate $A_{t:t+H-1}$, horizon H, decode stride Δt, denoise steps K_d

Output: $C(A) = (c_1, \ldots, c_H)$, latent rollout $\hat{z}_{t+1:t+H}$, decoded frames $\mathbf{X}(A) = \{\hat{x}_{t+\tau}\}_{\tau \in \mathcal{S}}$

$C(A) \leftarrow$ INTEGRATEPOSES$(P_t, A_{t:t+H-1})$;
$\mathcal{S} \leftarrow \{\, \tau : \tau = 1, 1{+}\Delta t, \ldots, H \,\}$;
$z_{t-m+1:t} \leftarrow E(I_{t-m+1:t})$; $\hat{z}_{t+1:t+H} \leftarrow \emptyset$; $\mathbf{X}(A) \leftarrow \emptyset$;
for $\tau = 1$ **to** H **do**
  cond $\leftarrow \big(z_{t-m+1:t},\ \hat{z}_{t+1:t+\tau-1},\ \Phi(g),\ \tilde{A}_{t:t+\tau-1},\ \tau\big)$;
  $\hat{z}_{t+\tau} \leftarrow$ CDIT_DENOISE(cond; K_d);
  if $\tau \in \mathcal{S}$ **then**
    $\hat{x}_{t+\tau} \leftarrow D(\hat{z}_{t+\tau})$;
    $\mathbf{X}(A) \leftarrow \mathbf{X}(A) \cup \{\hat{x}_{t+\tau}\}$;
return $C(A)$, $\hat{z}_{t+1:t+H}$, $\mathbf{X}(A)$;

(ii) traversability/free-space, and (iii) obstacle/uncertainty penalty. The weights are learned, nonnegative, and the output is temperature-scaled and normalized to $[0, 1]$.

Egocentric projection. Using camera intrinsics and pose c_τ, the confidence map is softly projected onto a fixed egocentric grid around the agent (depth-aware splatting with bilinear accumulation). We apply a light morphological smoothing to account for the robot footprint and mask out regions outside the field of view.

Use. V_t^{img} is sampled along each candidate path and fused at score level with the planner's native score (Eq. (2)). For efficiency, we decode every Δt frames and use a compact grid (e.g., $80{\times}80$ within a 12 m window).

Multi-rollout aggregation. Given imagined frames at times $\tau \in \mathcal{S}$, we accumulate into the value map with discount γ:

$$V_t^{\text{img}}(u, v) = \text{LSE}_\beta\big\{\gamma^\tau\, \Pi(\tilde{v}_{t+\tau})(u, v)\big\}_{\tau \in \mathcal{S}},$$

where $\text{LSE}_\beta(\cdot)$ is a temperature-controlled log-sum-exp that smoothly approximates max-pooling and Π is the egocentric projection. Using LSE avoids brittleness of hard maxima while preserving peakiness at promising regions.

Uncertainty gating. We estimate the rollout uncertainty σ_A (from ensemble variance or diffusion variance) and disable imagination when it is high. We use a hard gate with a fixed threshold $\theta = 0.6$ across all experiments, see Figure 5.

Algorithm 2: ONESTEPPLAN: Imagine–Score–Fuse–Act (single step)

Require: current $I_{t-m+1:t}, D_{1:t}, P_t$, map $\mathcal{M}_t$, instruction g; world model $\mathcal{W}$; reasoner $\mathcal{R}$; I2V head $\mathcal{H}$; base planner $\mathcal{P}$; horizon H, #candidates K, discount γ, weights λ_1, λ_2, decode stride Δt

Ensure: next action a_t

$\mathcal{B} \leftarrow \mathcal{P}$.CANDIDATES$(\mathcal{M}_t, K)$; `// frontiers → paths A`
$V_t^{\text{prior}} \leftarrow \mathcal{R}$.PRIOR$(g, \mathcal{M}_t, P_t)$ **for *each*** $A \in \mathcal{B}$ **do**
  $C(A) \leftarrow$ INTEGRATEPOSES(P_t, A) ; `// poses along A`
  $\hat{\mathbf{X}}(A) \leftarrow \mathcal{W}$.ROLLOUT$(I_{t-m+1:t}, g, C(A); \Delta t)$
  $V_t^{\text{img}} \leftarrow \mathcal{H}$.I2V$(\hat{\mathbf{X}}(A), C(A))$
  $s_{\text{base}} \leftarrow \mathcal{P}$.SCORE$(A)$
  $s_{\text{img}} \leftarrow$ SAMPLEPATH$(V_t^{\text{img}}, A, \gamma)$
  $s_{\text{prior}} \leftarrow$ SAMPLEPATH$(V_t^{\text{prior}}, A, \gamma)$
  $S_{\text{fused}}(A) \leftarrow s_{\text{base}} + \lambda_1 s_{\text{img}} + \lambda_2 s_{\text{prior}}$; `// Eq. (2)`
$A^\star \leftarrow \arg\max_{A \in \mathcal{B}} S_{\text{fused}}(A)$;
return first control of $A^\star$;

3.4. World Model Details

Architecture. Our world model uses a CDiT-L backbone , with the same encoder and decoder used throughout the Navigation World Model (NWM) [7]. Conditioning is injected by per-timestep action tokens and an instruction embedding, as described in Sec. 3.1. All inputs are resized to 224×224 and processed in bf16 precision.

Training data. We curate **489k** egocentric RGB frames from HM3D and MP3D, organized as short trajectories with their pose sequences and actions. The train/val split follows our navigation splits; we use only indoor scenes and discard trajectories shorter than a fixed minimum length. Details and scene lists are provided in the Appendix.

Optimization and schedule. Unless otherwise noted, we train on 8× RTX 6000 Ada 48 GB for ∼8 days, totaling ≈200k optimizer steps. We use AdamW (lr $8{\times}10^{-5}$, weight decay 0.05, $\beta{=}(0.9, 0.999)$), cosine decay with 2k warmup, gradient clipping at 10.0, and an EMA with decay 0.9999. Batch size is 8 with *accumulation 2* (effective 16). Mixed precision is bf16. We evaluate every 2k steps and save short rollouts for visual inspection. Actions are z-score normalized over the training split.

Table 1. Performance on R2R. All methods use the 3 m success radius and identical stop rules.

Models	Validation Unseen				Test Unseen			
	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓
EnvDrop [48]	52.0	48.0	10.70	5.22	51.0	47.0	11.66	5.23
VLFM [6]	52.5	30.4	/	/	/	/	/	/
PREVALENT [49]	58.0	53.0	10.19	4.71	54.0	51.0	10.51	5.30
RecBERT [50]	63.0	57.0	12.01	3.93	63.0	57.0	12.35	4.09
HAMT [51]	66.2	61.5	11.46	3.62	65.0	60.0	12.27	3.93
HAMT-Imagine [52]	67.26	62.02	11.80	3.58	65.0	60.0	12.66	3.89
DUET [53]	71.52	60.41	13.94	3.31	69.0	59.0	14.73	3.65
DUET-Imagine [52]	72.12	60.48	14.35	3.19	71.0	60.0	15.35	3.52
PanoGen [54]	74.2	64.3	13.40	3.03	71.7	61.9	14.38	3.31
NavCoT [55]	40.23	36.64	9.95	6.26	/	/	/	/
OmniNav [56]	69.5	66.1	/	3.74	/	/	/	/
VISTA [8]	77.8	68.3	13.26	2.92	74.9	66.7	14.20	3.77
VISTAv2 (ours)	81.4	73.7	10.73	3.76	73.1	69.0	12.44	4.05

4. Experiments

4.1. Experiment Setup

Benchmarks. We evaluate in Habitat on continuous VLN-CE episodes derived from the Room-to-Room corpus in Matterport3D (R2R). For R2R we follow the standard splits: train 10,819 episodes over 61 scenes, val-seen 778 over 53, val-unseen 1,839 over 11, and test 3,408 over 18 scenes [3]. For HM3D, we adopt the public validation configuration with 2,000 episodes across 20 scenes and 6 goal categories to assess generalization in large-scale real-scan environments [57].

Metrics. We report standard VLN metrics [58]: ❶ *Success Rate (SR)* — fraction of episodes where the agent issues STOP within 3 meters of the goal; ❷ *Success weighted by Path Length (SPL)* — SR weighted by path efficiency relative to the shortest path; ❸ *Navigation Error (NE)* — shortest path distance (meters) from the final position to the goal; ❹ *Trajectory Length (TL)* — total distance traveled (meters). Higher SR/SPL and lower NE/TL indicate better navigation. All inference is on a single GeForce RTX 4090; for VISTAv2 this includes world model rollouts, I2V scoring, and planner evaluation.

4.2. Experiments Results

We evaluate VISTAv2 on both the R2R benchmark and RoboTHOR, and observe clear, consistent improvements over strong baselines across all splits in higher SR/SPL and with competitive NE/TL.

Baselines. We compare VISTAv2 with representative VLN/VLN-CE systems: EnvDrop, PREVALENT, RecBERT, HAMT, HAMT-Imagine, DUET, DUET-Imagine, VLFM, PanoGen(++), NavCoT, OmniNav, and the prior VISTA. When a method is defined on the panorama graph, we follow the standard R2R protocol and evaluate on the same splits; all methods share the same sensor suite and stopping rule for fair comparison.

Main Results. Table 1 reports results on R2R (Val-Unseen/Test-Unseen). VISTAv2 yields higher SPL and shorter trajectories on both splits, with modest SR/NE trade-offs. Val-Unseen: 81.4 SR / 73.7 SPL (+3.6/+5.4 vs. VISTA), TL 13.26→10.73 (–19%), NE 3.76 vs. 2.92. Test-Unseen: 73.1 SR / 69.0 SPL (–1.8/+2.3), TL 14.20→12.44, NE 4.05 vs. 3.77. Our fused policy tends to issue STOP once the imagined value concentrates near the goal and local confidence is high, which favors shorter, decisive paths yet can end just outside the 3 m success radius.

Imagination with value helps. Against semantic-only or frontier-style approaches (e.g., VLFM and PanoGen++ rows), VISTAv2 lifts SPL by $\approx 5-10$ points and SR by a similar margin, indicating that converting imagined futures into an *egocentric value map* improves action selection beyond language–vision matching.

Sequential imagination better than goal imagination. Compared to goal-only imagination baselines (e.g., HAMT-Imagine, DUET-Imagine), VISTAv2 yields large gains (Val: ↑SR by $\approx$ 9–14 points; Test: ↑SR by ≈ 8 points). We attribute this to our *action-conditioned, short-horizon* rollouts and score-level fusion, which respect near-term reachability and geometric risk during planning. Overall, these trends support our claim: *VLN agents perform better when imagination is expressed as a value in map space and used to re-rank candidates*, rather than relying on semantics alone or optimizing a long-horizon world-model objective.

RoboTHOR Results. Table 2 shows that VISTAv2 also brings sizable gains in the interactive RoboTHOR setting: it achieves 61.3 SR and 34.7 SWPL (Success-Weighted Path Length), outperforming VISTA by +18.2 SR and +5.9

Figure 3. **Effect of visual imagination on goal discovery.** Each panel shows one episode (same start/goal). **Left (No Imagination):** the base planner explores many frontiers (red circles) guided only by occupancy/prior; the value over the explored region is diffuse (green mask), the agent wanders (**132 steps**) and fails to localize the TV. **Right (With Imagination):** our world model rolls out egocentric futures and the I2V head produces a fan-shaped image value map (orange/yellow), which fused with the prior sharpens the score and steers the agent through the doorway to the TV room, succeeding in 30 steps. Top: current depth/RGB and occupancy with path; Right column: value maps; Bottom: frontier set and fused score along the chosen path.

Table 2. Comparison on RoboTHOR.

Model	SWPL↑	SR↑
CLIP-Ref. [59]	2.4	2.7
CLIP-Patch [59]	10.6	20.3
CLIP-Grad. [59]	9.7	15.2
MDETR [59]	8.4	9.3
OWL [59]	17.2	27.5
ESC [60]	22.2	38.1
VLTNet [29]	17.1	33.2
VISTA [8]	28.8	43.1
VISTAv2 (ours)	34.7	61.3

SWPL, and surpassing CLIP-based, OWL, ESC, and VLTNet baselines by a large margin. The concurrent SR and SWPL gains suggest that map-space value priors from short-horizon rollouts both reduce outright failures and improve path efficiency, consistent with our R2R results.

We further tried to use the NWM to directly generate and score long-horizon future frames, but this naive pixel-space planner performed significantly worse due to severe rollout drift and artifacts.

4.3. Ablations: From semantics to imagination to imagined value

Setup. We contrast three stages: (1) *VLFM* — frontier/map exploration scored only by vision–language similarity (no imagination); (2) *VISTA* — imagination for reasoning but *no map-space value fusion*; (3) *VISTAv2 (ours)* — short-horizon, action-conditioned imagination projected to an *egocentric value map* and fused at score level.

Table 3. **Ablation on R2R.** VLFM (semantics only) → VISTA (imagination w/o value) → VISTAv2 (imagination→value + fusion).

Val-Unseen				
Method	TL↓	NE↓	SR↑	SPL↑
VLFM [6]	–	–	52.5	30.4
VISTA [8]	13.26	2.92	77.8	68.3
VISTAv2 (ours)	10.73	3.76	81.4	73.7
Test-Unseen				
Method	TL↓	NE↓	SR↑	SPL↑
VLFM [6]	–	–	48.2	26.4
VISTA [8]	14.20	3.77	74.9	66.7
VISTAv2 (ours)	12.44	4.05	73.1	69.0

Impact of imagination (VLFM → VISTA). Adding imagination beyond language–vision scoring yields large gains on Val-Unseen: +25.3 SR and +37.9 SPL (77.8/68.3 vs. 52.5/30.4), fixing many visually plausible yet unreachable detours. The higher SPL indicates fewer backtracks and more direct paths, consistent with the model pruning dead-ends and avoiding "bait" views (mirrors, glass, long hallways). Qualitatively, short-horizon egocentric rollouts provide explicit *reachability evidence*. For example, whether a corridor bends behind the camera, a doorway is traversable, or a landmark is occluded by walls, frontiers that merely *look* semantically promising no longer dominate the ranking. Notably, this improvement comes *without* additional path supervision: simulated futures act as a geometry-aware prior that complements language matching and stabilizes decisions across scenes.

Figure 4. Qualitative visualization of the world-model rollout. For two trajectories in MP3D and HM3D. The rollouts capture room layout and semantics (doorways, arches, windows, tables and bookshelves), which are sufficient for planning even when textures appear stylized.

Impact of imagination-to-value fusion (VISTA → VISTAv2). Projecting short-horizon rollouts to a *map-space value* and fusing at score level improves both success and efficiency: on Val-Unseen, +1.6 SR, +5.4 SPL, and TL ↓19.1%; on Test-Unseen, SPL +2.3 and TL ↓12.4%, with a small SR trade-off (-1.8) and slightly higher NE (+0.28). Qualitatively, the value prior acts as a *reachability-aware tie-breaker* among candidates with similar base scores, suppressing imagined frontiers and zig-zag oscillations and yielding more direct paths (higher SPL, shorter TL). Because fusion *preserves* the planner, geometric constraints remain intact, while language alignment and traversability cues steer the search toward feasible, instruction-relevant corridors. This ablation isolates the effects: *frontier-only* → *imagination* → *imagination-to-value*, where the last step yields the largest SPL gains and clear TL reductions.

And we sweep the uncertainty–gating threshold θ (percentile of rollout variance σ_A) on R2R Val-Unseen. It exhibits a single-peak trend: too small θ degenerates to no fusion (VISTA), while too large θ admits noisy rollouts. We therefore set θ=0.6 consistent with our main results (Fig. 5).

Failure case study. In scenes with large mirrors or glass partitions, the imagined rollout can overestimate traversability and inflate language alignment behind reflective surfaces. In one episode, the instruction asks the agent to reach a visible landmark in the adjacent room. From the start pose, a full-wall mirror creates a strong false "continuation" of the corridor; monocular depth underestimates the barrier and the world model predicts a plausible forward corridor in a few steps. After projection, the I2V map places a high-value ridge straight ahead, and the fused score ranks this candidate above an alternative that first detours around the corner. As the agent advances, the base mapper eventually detects the blockage, triggering a local oscillation and an early stop just short of the correct doorway—yielding lower SR and slightly higher NE despite a short TL.

Figure 5. Uncertainty gating sweep on R2R (Val-Unseen). SR/SPL versus the gating threshold θ (right axis: fallback rate). Performance peaks at a moderate θ as 0.6.

5. Conclusion

We presented **VISTAv2**, a language-conditioned, action-aware world model for Vision-and-Language Navigation that rolls out short-horizon egocentric futures and expresses their guidance as an *egocentric value map* in map space. Instead of replacing the planner with a long-horizon objective, VISTAv2 fuses the imagined value *at score level* with a standard frontier based stack, injecting instruction-consistent and reachability-aware cues at test time. On R2R and RoboTHOR, VISTAv2 delivers consistent gains in SR and SPL with shorter paths, and ablations show that imagination helps beyond language priors and projecting imagination into map space value is key to efficiency and robustness. The value maps are interpretable and lightweight to deploy, suggesting a practical route for bringing generative world models into embodied navigation.

6. Limitations and Future Work

Our rollouts are short-horizon and rely on monocular depth and pose estimates; failures occur in low-texture, reflective, or cluttered scenes where uncertainty increases. Future work includes uncertainty-aware planning over longer horizons, richer action spaces, memory for multi-episode goals, and sim-to-real transfer on mobile robots.

References

[1] Wang Xin, Huang Qiuyuan, Celikyilmaz Asli, Gao Jianfeng, Shen Dinghan, Wang Yuan-Fang, Wang William, Yang, and Zhang Lei. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *arXiv preprint arXiv:1811.10092*, 2018. URL https://www.arxiv.org/abs/1811.10092. 1

[2] Chen Jinyu, Gao Chen, Meng Erli, Zhang Qiong, and Liu Si. Reinforced structured state-evolution for vision-language navigation. *arXiv preprint arXiv:2204.09280*, 2022. URL https://www.arxiv.org/abs/2204.09280.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 2, 6

[4] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model, 2024. URL https://arxiv.org/abs/2401.02695.

[5] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation, 2019. URL https://arxiv.org/abs/1811.10092. 1

[6] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation, 2023. URL https://arxiv.org/abs/2312.03275. 1, 3, 6, 7

[7] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2025. URL https://arxiv.org/abs/2412.03572. 1, 2, 5

[8] Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. Vista: Generative visual imagination for vision-and-language navigation, 2025. URL https://arxiv.org/abs/2505.07868. 1, 2, 6, 7

[9] Shah Dhruv, Osinski Blazej, Ichter Brian, and Levine Sergey. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *arXiv preprint arXiv:2207.04429*, 2022. URL https://www.arxiv.org/abs/2207.04429. 1

[10] Huang Chenguang, Mees Oier, Zeng Andy, and Burgard Wolfram. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2022. URL https://www.arxiv.org/abs/2210.05714. 3

[11] Vasudevan Arun, Balajee, Dai Dengxin, and Gool Luc, Van. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *arXiv preprint arXiv:1910.02029*, 2019. URL https://www.arxiv.org/abs/1910.02029.

[12] Zhang Jiazhao, Wang Kunyu, Xu Rongtao, Zhou Gengze, Hong Yicong, Fang Xiaomeng, Wu Qi, Zhang Zhizheng, and Wang He. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024. URL https://www.arxiv.org/abs/2402.15852.

[13] Wang Xin, Xiong Wenhan, Wang Hongmin, and Wang William, Yang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. *arXiv preprint arXiv:1803.07729*, 2018. URL https://www.arxiv.org/abs/1803.07729.

[14] Chen Shizhe, Guhur Pierre-Louis, Tapaswi Makarand, Schmid Cordelia, and Laptev Ivan. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. *arXiv preprint arXiv:2202.11742v1*, 2022. URL https://www.arxiv.org/abs/2202.11742v1. 1

[15] Chang Matthew, Gervet Theophile, Khanna Mukul, Yenamandra Sriram, Shah Dhruv, Min So, Yeon, Shah Kavit, Paxton Chris, Gupta Saurabh, Batra Dhruv, Mottaghi Roozbeh, Malik Jitendra, and Chaplot Devendra, Singh. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*, 2023. URL https://www.arxiv.org/abs/2311.06430. 1

[16] Nguyen Khanh, Dey Debadeepta, Brockett Chris, and Dolan and, Bill. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. *arXiv preprint arXiv:1812.04155*, 2018. URL https://www.arxiv.org/abs/1812.04155.

[17] Li Jialu and Bansal Mohit. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *arXiv preprint arXiv:2305.19195*, 2023. URL https://www.arxiv.org/abs/2305.19195. 2, 3

[18] Wei Meng, Wan Chenyang, Yu Xiqian, Wang Tai, Yang Yuqiang, Mao Xiaohan, Zhu Chenming, Cai Wenzhe, Wang Hanqing, Chen Yilun, Liu Xihui, and Pang Jiangmiao. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025. URL https://www.arxiv.org/abs/2507.05240. 2

[19] Zhiyuan Li, Yanfeng Lu, Yao Mu, and Hong Qiao. Cog-ga: A large language models-based generative agent for vision-language navigation in continuous environments. *arXiv preprint arXiv:2409.02522*, 2024. URL https://www.arxiv.org/abs/2409.02522. 1

[20] Zhao Ganlong, Li Guanbin, Chen Weikai, and Yu Yizhou. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. *arXiv preprint arXiv:2403.17334*, 2024. URL https://www.arxiv.org/abs/2403.17334. 2

[21] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments, 2021. URL https://arxiv.org/abs/2110.02207.

[22] Cui Yibo, Xie Liang, Zhang Yakun, Zhang Meishan, Yan Ye, and Yin Erwei. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. *arXiv preprint arXiv:2308.12587*, 2023. URL https://www.arxiv.org/abs/2308.12587.

[23] Kuo Chia-Wen, Ma Chih-Yao, Hoffman Judy, and Kira Zsolt. Structure-encoding auxiliary tasks for improved visual representation in vision-and-language navigation. *arXiv*

preprint arXiv:2211.11116, 2022. URL https://www.arxiv.org/abs/2211.11116.

[24] Raychaudhuri Sonia, Wani Saim, Patel Shivansh, Jain Unnat, and Chang Angel, X. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2109.15207*, 2021. URL https://www.arxiv.org/abs/2109.15207.

[25] Chi Ta-Chung, Eric Mihail, Kim Seokhwan, Shen Minmin, and Hakkani-tur Dilek. Just ask:an interactive learning framework for vision and language navigation. *arXiv preprint arXiv:1912.00915*, 2019. URL https://www.arxiv.org/abs/1912.00915.

[26] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. *arXiv preprint arXiv:1906.00347*, 2019. URL https://www.arxiv.org/abs/1906.00347.

[27] Loh Dillon, Bednarz Tomasz, Xia Xinxing, and Guan Frank. Adavln: Towards visual language navigation in continuous indoor environments with moving humans. *arXiv preprint arXiv:2411.18539*, 2024. URL https://www.arxiv.org/abs/2411.18539.

[28] Hong Yicong, Rodriguez-Opazo Cristian, Wu Qi, and Gould Stephen. Sub-instruction aware vision-and-language navigation. *arXiv preprint arXiv:2004.02707*, 2020. URL https://www.arxiv.org/abs/2004.02707.

[29] Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XVIII*, page 389–404, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-78455-2. doi: 10.1007/978-3-031-78456-9_25. URL https://doi.org/10.1007/978-3-031-78456-9_25. 2, 7

[30] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020. 2

[31] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[32] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments, 2020. URL https://arxiv.org/abs/2004.02857. 2

[33] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[34] Sen Wang, Dongliang Zhou, Liang Xie, Chao Xu, Ye Yan, and Erwei Yin. Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation. *Neural Networks*, 187:107320, July 2025. ISSN 0893-6080. doi: 10.1016/j.neunet.2025.107320. URL http://dx.doi.org/10.1016/j.neunet.2025.107320. 3

[35] Zhaohuan Zhan, Lisha Yu, Sijie Yu, and Guang Tan. Mc-gpt: Empowering vision-and-language navigation with memory map and reasoning chains, 2024. URL https://arxiv.org/abs/2405.10620.

[36] Shuhei Kurita and Kyunghyun Cho. Generative language-grounded policy in vision-and-language navigation with bayes' rule, 2020. URL https://arxiv.org/abs/2009.07783. 2

[37] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[38] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[39] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[40] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[41] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2

[42] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391. 2

[43] Sebastián Barbas Laina, Simon Boche, Sotiris Papatheodorou, Simon Schaefer, Jaehyung Jung, and Stefan Leutenegger. Findanything: Open-vocabulary and object-centric mapping for robot exploration in any environment, 2025. URL https://arxiv.org/abs/2504.08603. 3

[44] Mitsuaki Uno, Kanji Tanaka, Daiki Iwata, Yudai Noda, Shoya Miyazaki, and Kouki Terashima. Lgr: Llm-guided ranking of frontiers for object goal navigation, 2025. URL https://arxiv.org/abs/2503.20241.

[45] Jiajun Jiang, Yiming Zhu, Zirui Wu, and Jie Song. Dualmap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes, 2025. URL https://arxiv.org/abs/2506.01950.

[46] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 3554–3560. IEEE, October 2023. doi: 10.1109/iros55552.2023.10342512. URL http://dx.doi.org/10.1109/IROS55552.2023.10342512. 3

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 4

[48] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019. 6

[49] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13137–13146, 2020. 6

[50] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. *arXiv preprint arXiv:2011.13922*, 2020. 6

[51] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021. 6

[52] Akhil Perincherry, Jacob Krantz, and Stefan Lee. Do visual imaginations improve vision-and-language navigation agents?, 2025. URL https://arxiv.org/abs/2503.16394. 6

[53] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 6

[54] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation, 2023. URL https://arxiv.org/abs/2305.19195. 6

[55] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning, 2025. URL https://arxiv.org/abs/2403.07376. 6

[56] Xinda Xue, Junjun Hu, Minghua Luo, Xie Shichao, Jintao Chen, Zixun Xie, Quan Kuichen, Guo Wei, Mu Xu, and Zedong Chu. Omninav: A unified framework for prospective exploration and visual-language navigation, 2025. URL https://arxiv.org/abs/2509.25687. 6

[57] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. URL https://arxiv.org/abs/2109.08238. 6

[58] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping, 2019. URL https://arxiv.org/abs/1907.05446. 6

[59] Gadre Samir, Yitzhak, Wortsman Mitchell, Ilharco Gabriel, Schmidt Ludwig, and Song Shuran. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *arXiv preprint arXiv:2203.10421*, 2022. URL https://www.arxiv.org/abs/2203.10421. 7

[60] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation, 2023. URL https://arxiv.org/abs/2301.13166. 7