LEARNING COMPOSABLE TRANSITION REPRESENTATIONS FOR VISUAL NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

What latent representation should an image-goal navigation policy use? Direct visual policies are efficient, but their representations are often learned primarily through action supervision and may fail to capture the transition structure underlying viewpoint changes. We propose NavCompose, a self-supervised pre-training paradigm that learns structured navigation latents from unlabeled visual trajectories. NavCompose encodes relative visual transitions and explicitly regularizes them to be composable across consecutive observations and continuous along local trajectories. The learned representation is renderable and supports interpolation between observed viewpoints as well as short-horizon extrapolation beyond them, providing a compact model of local visual transitions. For downstream image-goal navigation, we freeze the pre-trained backbone and train only a lightweight waypoint prediction head. Across multiple navigation datasets, NavCompose achieves state-of-the-art waypoint prediction performance while maintaining strong accuracy with substantially fewer labeled trajectories. Its lightweight downstream architecture further enables fast inference, outperforming existing navigation models in runtime efficiency. Code will be released.

1 INTRODUCTION

Image-goal navigation requires an embodied agent to reach a location specified by a target image using only egocentric visual observations, without access to the goal pose or a pre-built global map. The central challenge is cross-view spatial reasoning: given current and goal observations, the agent must infer how the viewpoint should change and translate this change into navigation decisions. This suggests that effective navigation representations should capture not only what a scene looks like, but also how observations evolve across viewpoints.

Most image-goal navigation methods learn visual representations jointly with navigation policies from trajectories annotated with actions or waypoints. ViNT (Shah et al., 2023), for example, directly maps context and goal images to waypoint predictions, while recent methods improve navigation through diffusion (Cai et al., 2025; Zeng et al., 2025; Sridhar et al., 2024), flow-based learning (Gode et al., 2025), and geometry-aware planning (Peng et al., 2025; Meng et al., 2025). These approaches rely on navigation-specific trajectory supervision, which is substantially more expensive to collect than unlabeled visual sequences. This motivates a fundamental question:

Can navigation-relevant visual representations be pre-trained from unlabeled sequences and transferred to navigation tasks with limited labeled trajectories?

Recent advances in self-supervised visual learning demonstrate the potential of large-scale unlabeled data for transferable representation learning. Methods such as DINO (Oquab et al., 2023; Siméoni et al., 2025), JEPA (Assran et al., 2023), and RAE (Singh et al., 2026) learn powerful visual features without task-specific labels, while view-synthesis methods (Jiang et al., 2025; Zhao et al., 2026; Mitchel et al., 2025) capture aspects of viewpoint variation. However, navigation requires more than representing individual observations or isolated pairwise changes. A useful representation of viewpoint evolution should satisfy two structural properties. First, *compositionality*: consecutive local viewpoint changes should compose consistently to represent longer-range transformations. Second, *continuity*: the representation should vary smoothly as the viewpoint evolves, enabling meaningful transitions between observed states and beyond them.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

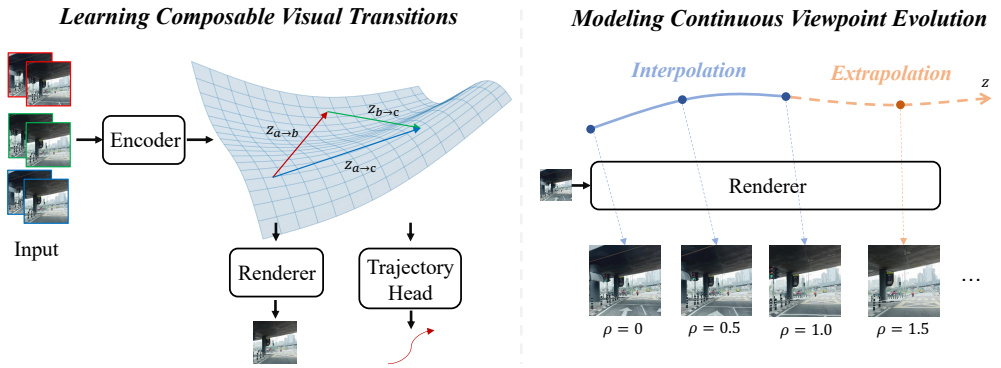

Figure 1: **Learning a structured latent space for visual transitions.** NavCompose learns relative viewpoint changes from unlabeled videos as *composable and continuous* latent transitions. Consecutive transitions can be composed and decoded into future observations, while a lightweight trajectory head directly predicts navigation waypoints. The continuous latent structure further supports viewpoint interpolation ($\rho \in [0, 1]$) and extrapolation ($\rho > 1$) beyond the observed trajectory.

Existing transition-based representation learning provides important foundations but does not explicitly address these requirements for observation-driven navigation. XFactor (Mitchel et al., 2025) learns relative-view representations through view reconstruction, but models individual transitions without imposing trajectory-level compositional structure. Latent action models such as ALAM (Tang et al., 2026) and AC-LAM (Wei et al., 2026) investigate compositional latent dynamics, but primarily study action-conditioned manipulation. In contrast, we study *observation-grounded visual transitions*, learning relative viewpoint changes directly from unlabeled visual trajectories without action or pose supervision.

To bridge this gap, we propose NavCompose, a self-supervised framework that learns structured visual transition representations from unlabeled video. Given two observations, NavCompose encodes their relative viewpoint change into a latent transition, which is grounded through future-view reconstruction. This objective encourages the latent space to capture meaningful viewpoint transformations, where transitions compose consistently along trajectories rather than behaving as arbitrary latent codes. We further regularize the transition space to evolve smoothly over trajectories, allowing it to represent both observed and future viewpoint changes. After pre-training, we freeze the transition encoder and train only a lightweight waypoint head on labeled navigation trajectories, enabling large-scale unlabeled videos to provide transferable representations with limited navigation supervision. The learned transition space provides a lightweight local visual transition model, where latent transitions can be composed and decoded to synthesize intermediate and short-horizon future observations. This capability relates to navigation world models (Bar et al., 2025; Zhang et al., 2026; Zhou et al., 2026), but differs in purpose: view prediction serves as a self-supervised grounding signal for representation learning rather than an action-conditioned rollout mechanism. During navigation, NavCompose directly predicts waypoints from the learned transition representation without requiring visual generation. Our contributions are summarized as follows:

- We formulate observation-grounded visual transition learning for image-goal navigation, shifting visual pre-training from representing individual observations to learning relative viewpoint transformations directly from unlabeled trajectories.
- We propose NavCompose, a self-supervised framework that learns a composable and continuous latent transition space through view reconstruction, multi-step transition composition, and trajectory-level continuity regularization.
- We demonstrate that the learned transition representation transfers effectively to image-goal navigation, achieving state-of-the-art waypoint prediction across multiple public benchmarks while substantially reducing the amount of navigation-specific trajectory supervision required.

2 Related Work

2.1 Image-Goal Conditioned Visual Navigation

Image-goal conditioned visual navigation aims to learn policies that drive an agent from the current visual observation to a target image. Early generalist navigation policies, such as ViNT (Shah et al., 2023), formulate navigation as a goal-reaching problem and learn transferable visual-motor affordances from diverse robot trajectories. NoMaD (Sridhar et al., 2024) further introduces a goal-masked diffusion policy, unifying goal-directed navigation and goal-agnostic exploration within a single policy. More recent methods continue to improve the expressiveness and robustness of navigation policies through diffusion-based action generation, stronger visual perception, and planning-oriented policy designs, including NavDP (Cai et al., 2025), NaviDiffuser (Zeng et al., 2025), FlowNav (Gode et al., 2025), LoGoPlanner (Peng et al., 2025), and BEINGS (Meng et al., 2025). These methods demonstrate strong progress in learning navigation behaviors from visual trajectories. However, their representations are often learned implicitly through action prediction or policy optimization. In contrast, our work explicitly studies the structure of the latent space itself and asks what kind of representation is more suitable for downstream navigation policy learning.

2.2 Visual Representation Spaces for Navigation

Another line of work investigates latent representation spaces for navigation world models and planning. Navigation World Models (Bar et al., 2025) learn action-conditioned visual dynamics by predicting future observations, showing that learned predictive models can support navigation planning in visual space. However, many world-model-based approaches rely on compressed VAE latents, which are efficient but may discard fine-grained spatial information that is important for precise navigation. Recent work therefore explores alternative representation spaces. For example, Kim et al. (2026) studies compact VAE latent spaces for efficient planning, while ReL-NWM (Zhang et al., 2025) leverages a pre-trained DINOv3 (Siméoni et al., 2025) representation encoder and introduces mechanisms for integrating action signals and historical context. RAE-NWM (Zhang et al., 2026) instead models navigation dynamics in a dense DINOv2 (Oquab et al., 2023) representation space, motivated by the stronger structural and predictive properties of dense visual features. NavDreamer (Huang et al., 2026b) further explores generative visual models as a navigation interface. Beyond navigation-specific models, self-supervised representation learning methods such as DINOv2 (Oquab et al., 2023) and JEPA (Assran et al., 2023) learn strong visual features without action labels, while geometry-oriented models such as VGGT (Wang et al., 2025) expose 3D structure from image collections. Value-shaped latent models (Destrade et al., 2025) further suggest that representation geometry can be explicitly organized for planning. These works indicate that the geometry of the latent space is crucial for decision making. Our work follows this perspective but focuses on visual navigation: rather than using generic semantic or predictive features directly, we learn a pose-consistent latent space whose relative-motion embeddings are explicitly constrained to be compositionally additive.

2.3 Learning Navigation from Large-Scale Web Data

Large-scale data has become increasingly important for building generalizable navigation agents (Kumar et al., 2020; Ramrakhya et al., 2022; Kuno et al., 2026; Lin et al., 2026; Dalal et al., 2025). Instead of relying only on robot-collected trajectories, recent works explore the use of web-scale human videos and automatically generated annotations. UrbanNav (Mei et al., 2026), for example, constructs large-scale language-guided urban navigation data from web videos and uses it to train navigation policies in diverse real-world city environments. Such efforts show the potential of scalable data sources for improving robustness and generalization in open-world navigation. Our work is complementary to this direction. While large-scale web data improves the diversity of navigation experience, we focus on how such visual trajectory data should be represented. Specifically, we study how to pre-train a latent space that preserves relative motion structure and supports downstream trajectory prediction with a lightweight policy head.

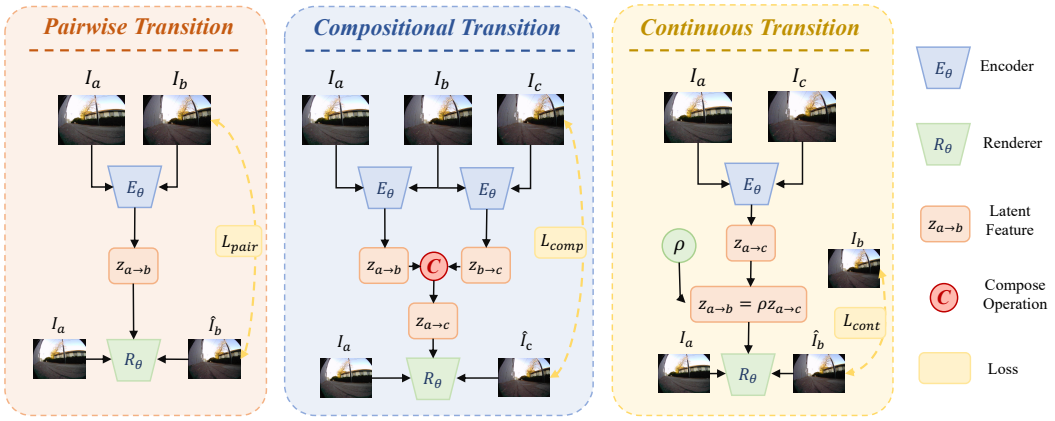

Figure 2: **Overview of NavCompose.** We learn a structured visual transition space from unlabeled trajectories using three complementary objectives. *Pairwise transition grounding* learns renderable relative-viewpoint latents that capture the visual and geometric changes between observations. *Compositional transition structuring* encourages local transitions to accumulate consistently along trajectories, while *continuous transition structuring* regularizes smooth latent evolution with viewpoint progress. Together, these objectives establish a structured representation of viewpoint transformations, where transition latents encode trajectory-related motion patterns and can be effectively transferred to downstream waypoint prediction.

3 Method

3.1 Overview

NavCompose adopts a two-stage framework that decouples visual transition learning from navigation policy learning, as illustrated in Figure 2. *In the first stage*, we pre-train a transition encoder E_θ and a source-conditioned renderer R_θ on large-scale unlabeled visual trajectories. Given two observations, the encoder maps them to a latent representation of their relative viewpoint change, while the renderer grounds this transition by reconstructing the target observation. We further impose *compositionality* and *continuity* constraints, encouraging transitions to accumulate consistently across multiple steps and vary smoothly with trajectory progress. *In the second stage*, we transfer the pre-trained representation to image-goal navigation using labeled navigation trajectories. We discard the renderer and freeze the pre-trained transition encoder. Given recent observations and a goal image, the frozen encoder produces goal-relative transition representations, which are mapped to future waypoints by a lightweight trajectory head trained with labeled navigation trajectories. Thus, large-scale unlabeled videos provide the visual transition representation, while navigation supervision is required only for the final waypoint prediction module.

3.2 Pairwise Transition Grounding

We first learn a visual representation that explicitly captures the relative transformation between observations. Following the view-transfer formulation of XFactor (Mitchel et al., 2025), we represent the viewpoint change from a source observation to a target observation as a latent transition rather than encoding the two observations independently. This transition representation is learned through a transition encoder and a source-conditioned view renderer.

Given a source observation I_a and a target observation I_b, the transition encoder predicts $z_{a\to b} = E_\theta(I_a, I_b)$, where $z_{a\to b}$ encodes the relative visual transition from I_a to I_b. A renderer then reconstructs the target observation conditioned on the source image and the transition latent: $\hat{I}_b = R_\theta(I_a, z_{a\to b})$. We optimize the reconstruction objective

$$\mathcal{L}_{\text{pair}} = \|I_b - \hat{I}_b\|_1 + \lambda_{\text{lpips}}\text{LPIPS}(I_b, \hat{I}_b). \tag{1}$$

This objective encourages the transition latent to capture viewpoint-related information that is sufficient for reconstructing the target observation. Building on this visually grounded representation, we further introduce geometric constraints to structure the latent space for visual navigation.

3.3 COMPOSITIONAL TRANSITION STRUCTURING

The pairwise view-prediction objective introduced above encourages the latent representation to encode the relative transformation between two observations. However, it does not explicitly constrain the geometry of the latent space. In particular, independently learned pairwise pose embeddings are not guaranteed to maintain consistent relationships across observations along a trajectory, nor do simple latent-space operations necessarily correspond to meaningful sequential motion. Such structure is especially important for visual navigation, where trajectories naturally arise from sequences of local viewpoint changes.

We therefore introduce a compositional constraint to encourage a structured transition space. Given three temporally ordered observations (I_a, I_b, I_c), we first encode two consecutive transitions:

$$z_{a\to b} = E_\theta(I_a, I_b), \qquad z_{b\to c} = E_\theta(I_b, I_c). \tag{2}$$

We then approximate the longer transition from I_a to I_c by adding the two local transitions together with a learned residual correction:

$$\tilde{z}_{a\to c} = z_{a\to b} + z_{b\to c} + s f_\phi\left([z_{a\to b}; z_{b\to c}]\right), \tag{3}$$

where f_ϕ is a lightweight MLP and s controls the magnitude of the residual term. The residual branch is initialized to zero, so the composition initially reduces to simple latent addition and gradually learns nonlinear corrections during training. Rather than directly enforcing $\tilde{z}_{a\to c}$ to match another latent vector, we supervise the composed transition through its visual consequence. Specifically, the composed latent is used with the initial observation to predict the final view:

$$\hat{I}_c = R_\theta(I_a, \tilde{z}_{a\to c}), \tag{4}$$

with the objective

$$\mathcal{L}_{\text{comp}} = \|I_c - \hat{I}_c\|_1 + \lambda_{\text{lpips}}\text{LPIPS}(I_c, \hat{I}_c). \tag{5}$$

This objective encourages latent composition to reflect the accumulated viewpoint change along the trajectory, imposing a locally consistent structure on the transition space. Importantly, compositionality is grounded through future-view prediction rather than an abstract latent-space constraint, ensuring that the composed representation remains meaningful in observation space.

Compositional latent structure has also been explored in concurrent latent-action models such as ALAM (Tang et al., 2026) and AC-LAM (Wei et al., 2026), which organize action-like latents to support downstream manipulation policies. In contrast, we study compositionality as a representation-learning principle for image-goal navigation, where the learned transition latent encodes relative viewpoint changes and is directly transferred as a visual representation for waypoint prediction.

3.4 CONTINUOUS TRANSITION STRUCTURING

Geometric motion is inherently continuous: along a local trajectory, intermediate observations correspond to partial progress toward a farther viewpoint. We therefore encourage the transition latent to vary consistently with trajectory progress, such that changes in the latent representation correspond to continuous changes in viewpoint. We consider both interpolation between observed viewpoints and short-horizon extrapolation beyond them.

For interpolation, consider a source observation I_a, an endpoint observation I_b, and an intermediate observation I_q sampled between them along the same local trajectory. We first encode the full transition $z_{a\to b} = E_\theta(I_a, I_b)$. Let $\rho \in (0, 1)$ denote the relative temporal progress of I_q within the source–endpoint interval. We construct the intermediate transition by directly scaling the full displacement:

$$\tilde{z}_{a\to q} = \rho z_{a\to b}. \tag{6}$$

The scaled latent is then decoded from the source observation:

$$\hat{I}_q = R_\theta(I_a, \tilde{z}_{a\to q}). \tag{7}$$

This encourages proportional scaling of a transition latent to correspond to partial progress along the associated viewpoint change.

For extrapolation beyond the encoded transition, rather than directly scaling the latent by a factor larger than one, we extend the transition through repeated composition. Given a local displacement z, an integer number of steps is accumulated using the composition operation introduced above, while a fractional remainder is represented by a proportionally scaled displacement. For example, a $1.5\times$ transition is formed by composing z with $0.5z$, whereas a $2\times$ transition is obtained by composing z with itself. This allows short-horizon extrapolation to reuse the same compositional structure learned from consecutive viewpoint changes.

For both interpolation and extrapolation, the predicted query view is supervised by

$$\mathcal{L}_{\text{cont}} = \|I_q - \hat{I}_q\|_1 + \lambda_{\text{lpips}}\text{LPIPS}(I_q, \hat{I}_q). \tag{8}$$

Together, these objectives encourage the latent transition space to support both proportional progress within a local motion and consistent accumulation beyond it, introducing a continuous structure into the learned representation.

3.5 Pre-Training Objective and Task Transfer

The overall pre-training objective is defined as:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{pair}} + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}. \tag{9}$$

After pre-training, we transfer the learned transition representation to image-goal navigation. We discard the renderer, freeze the transition encoder E_θ, and train only a lightweight trajectory head A_ϕ using labeled navigation trajectories. Given recent context observations and a goal image, the frozen transition encoder extracts a sequence of goal-relative transition latents, which are fed into a lightweight trajectory head to predict future waypoints parameterized by $(\Delta x, \Delta y, \Delta\psi)$. Only the trajectory head is optimized with waypoint supervision, while the pre-trained transition encoder remains fixed.

4 Experiments

4.1 Datasets

We evaluate image-goal navigation on four real-world trajectory benchmarks: Go Stanford (Hirose et al., 2019), RECON (Shah et al., 2021), SACSoN (Hirose et al., 2023), and SCAND (Karnan et al., 2022). Each evaluation sample consists of a sequence of egocentric context observations, a goal image sampled from the same trajectory, and corresponding local waypoint annotations. For representation pre-training, we augment the navigation datasets with unlabeled videos from DL3DV (Ling et al., 2024), ScanNet (Dai et al., 2017), LAVN (Johnson et al., 2024), TartanGround (Patel et al., 2025), and CityWalker (Liu et al., 2025). For all downstream benchmarks, representation pre-training uses only the training split, with no validation or test trajectories observed.

4.2 Experimental Details

As shown in Table 1, we evaluate the transferability of unsupervised visual representations to downstream image-goal navigation. We first include existing *navigation models* as task-specific reference methods for direct comparison on downstream navigation. We then compare two families of unsupervised pre-trained representations. *Unsupervised foundation models* learn general-purpose visual features from large-scale unlabeled images or videos. *Unsupervised geometry models*, in contrast, focus on learning geometric and viewpoint-dependent representations from unlabeled visual observations, typically through novel-view synthesis.

To isolate representation quality, we use the same frozen-backbone evaluation protocol for all representation-based methods. Each pre-trained visual backbone is frozen, and only a trajectory prediction head is trained with navigation supervision. This ensures that performance differences mainly reflect the quality of the learned representations. Representation pre-training uses no action, waypoint, pose, or other explicit geometric annotations; these labels are used only for downstream trajectory-head training. We evaluate trajectory prediction using Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) (Sturm et al., 2012), where lower is better.

Table 1: Image-goal navigation results. †We train XFactor using the same training data as NavCompose. Lower is better for ATE and RPE.

Method	Go Stanford		RECON		SACSoN		SCAND		Average	
	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓
Navigation Models										
ViNT (Shah et al., 2023)	0.854	0.110	0.659	0.215	0.616	0.169	0.507	0.286	0.659	0.195
NoMaD (Sridhar et al., 2024)	1.158	0.133	1.220	0.262	1.274	0.286	1.025	0.345	1.169	0.257
NavDP (Cai et al., 2025)	0.840	0.242	0.952	0.259	0.738	0.439	0.649	0.413	0.795	0.338
FlowNav (Gode et al., 2025)	1.058	0.141	1.021	0.351	1.091	0.435	0.950	0.527	1.030	0.364
NWM (Bar et al., 2025)	0.895	0.120	0.326	0.115	0.497	0.258	0.463	0.231	0.545	0.181
Unsupervised Foundation Models										
VAE (Rombach et al., 2022)	0.803	0.105	0.379	0.108	0.603	0.168	0.677	0.276	0.616	0.164
DINOv2 (Oquab et al., 2023)	0.781	0.104	0.347	0.095	0.485	0.141	0.662	0.273	0.569	0.153
DINOv3 (Siméoni et al., 2025)	0.772	0.103	0.306	0.083	0.471	0.137	0.621	0.258	0.543	0.145
RAE (Zheng et al., 2025)	0.771	0.102	0.299	0.091	0.459	0.136	0.603	0.251	0.533	0.145
RAEv2 (Singh et al., 2026)	0.742	0.100	0.276	0.075	0.465	0.152	0.607	0.259	0.523	0.147
JEPA (Assran et al., 2023)	0.775	0.103	0.305	0.087	0.532	0.156	0.673	0.277	0.571	0.156
V-JEPA2 (Assran et al., 2025)	0.793	0.103	0.322	0.094	0.573	0.163	0.612	0.281	0.575	0.160
Unsupervised Geometry Models										
RayZer (Jiang et al., 2025)	0.783	0.104	0.327	0.094	0.616	0.172	0.645	0.281	0.593	0.163
ERayZer (Zhao et al., 2026)	0.769	0.101	0.253	0.070	0.462	0.138	0.587	0.253	0.518	0.141
NAS3R (Huang et al., 2026a)	0.773	0.102	0.302	0.083	0.524	0.148	0.649	0.258	0.562	0.148
XFactor† (Mitchel et al., 2025)	0.782	0.103	0.289	0.078	0.586	0.166	0.641	0.263	0.575	0.153
NavCompose	0.745	0.099	0.250	0.067	0.457	0.135	0.543	0.229	0.499	0.133

Our training consists of two stages. For *representation pre-training*, we train NavCompose on four NVIDIA RTX 4090 GPUs at a resolution of 256×256, using a 256-dimensional transition latent. The model is optimized for 150k steps with AdamW (Loshchilov & Hutter, 2017), using a learning rate of 4×10^{-4}, weight decay of 5×10^{-3}, and a per-GPU batch size of 32. For *trajectory-head training*, all pre-trained representation backbones are frozen, and only a lightweight trajectory prediction head is optimized with navigation supervision. The head is implemented as a 4-layer transformer decoder and is trained for 50k steps using AdamW with a batch size of 256.

4.3 Main Results

As shown in Table 1, NavCompose transfers effectively to image-goal navigation despite being pre-trained solely on unlabeled visual observations. Among methods leveraging unsupervised visual pre-training, NavCompose achieves state-of-the-art performance across multiple navigation benchmarks, demonstrating that structured viewpoint transitions provide a strong and transferable representation for downstream navigation. Compared with existing unsupervised geometry-based representations, NavCompose consistently improves navigation performance. For example, compared with ERayZer, NavCompose reduces the average ATE/RPE from $0.518/0.141$ to $0.499/0.133$, with improvements across individual benchmarks. NavCompose also outperforms XFactor, which uses the same visual data, enabling a direct comparison under matched pre-training conditions. These results indicate that explicitly modeling compositional and continuous viewpoint transitions enables visual-only pre-training to better capture navigation-relevant spatial dynamics. Notably, NavCompose achieves competitive performance with task-specific navigation models while requiring only lightweight waypoint supervision after pre-training. Overall, these results demonstrate that large-scale unlabeled video pre-training with structured transition modeling can effectively bridge visual representation learning and downstream image-goal navigation.

4.4 Ablation Studies

Unless otherwise stated, all ablations are conducted on RECON (Shah et al., 2021) using the same training settings. We first analyze the effectiveness of our learned transition representation through navigation data-efficiency and model-profile studies. We further provide detailed ablations on the latent transition structure, composition function, and latent dimension in the Appendix.

Navigation data efficiency. Finally, we examine how effectively the pre-trained representation transfers to navigation under limited waypoint supervision. We freeze the pre-trained transition encoder and vary the fraction of labeled navigation trajectories used to train only the trajectory head.

As shown in Table 2, NavCompose already achieves an ATE/RPE of $0.288/0.079$ using only 10% of the navigation data, compared with $0.250/0.067$ when using the full training set. Increasing the supervision to 30% further reduces the error to $0.266/0.071$, while using 50% of the data reaches $0.258/0.069$, approaching the performance obtained with full supervision. The gradual improvement and early saturation indicate that the pre-trained representation captures navigation-relevant structure that can be efficiently adapted with substantially reduced task-specific supervision.

Model profile. We further compare NavCompose with existing navigation models in terms of both navigation accuracy and inference efficiency. As shown in Table 3, NavCompose achieves the best performance on RECON, with an ATE/RPE of $0.250/0.067$, while requiring only 7.152 ms per inference under standard eager execution, corresponding to 139.81 FPS. Compared with ViNT, NavCompose reduces ATE from 0.659 to 0.250 and RPE from 0.215 to 0.067, while running $1.61\times$ faster. The efficiency advantage becomes even more pronounced compared with recent navigation models: NavCompose is over $6\times$ faster than NoMaD and FlowNav and over $10\times$ faster than NavDP. Overall, NavCompose offers a favorable accuracy-efficiency trade-off, achieving strong navigation performance while maintaining efficient real-time inference.

Table 2: Navigation performance with different fractions of labeled data.

Training data	ATE ↓	RPE ↓
10%	0.288	0.079
30%	0.266	0.071
50%	0.258	0.069
100%	0.250	0.067

Table 3: **Model profile comparison.** Inference latency and FPS are measured on a single NVIDIA RTX 4090 GPU, averaged over 500 inference runs.

Method	ATE ↓	RPE ↓	Latency (ms) ↓	FPS ↑
ViNT	0.659	0.215	11.547	86.60
NoMaD	1.220	0.262	43.212	23.14
NavDP	0.952	0.259	82.238	12.16
FlowNav	1.021	0.351	45.961	21.76
NavCompose	**0.250**	**0.067**	**7.152**	**139.81**

4.5 Qualitative Results

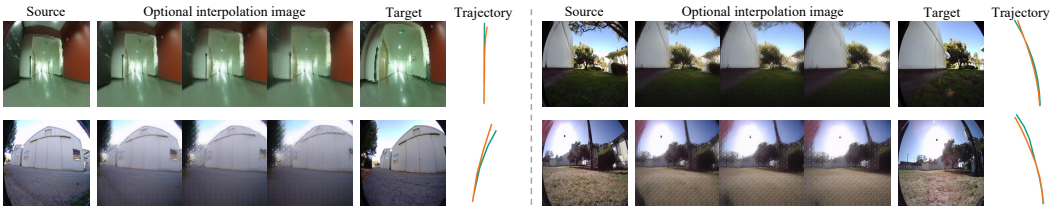

Figure 3: **Qualitative results.** Each example contains a source image, several intermediate views rendered by interpolating in the learned latent transition space, and a target image. The right side shows the corresponding ground-truth (GT) trajectory and predicted (Pred) trajectory.

Trajectory prediction. We further visualize trajectory prediction results in Figure 3. The interpolated views exhibit smooth visual transitions between the source and target observations, suggesting that the learned latent space captures continuous viewpoint changes. Moreover, the predicted trajectories remain close to the demonstrated paths across both indoor and outdoor scenes, indicating that the learned transition representations preserve navigation-relevant spatial structure.

Latent interpolation and extrapolation. Beyond navigation trajectories, we further evaluate whether the learned transition latent captures general visual dynamics. As shown in Figure 4, the model can generate plausible intermediate views through latent interpolation and predict future observations through extrapolation. These results demonstrate that the learned transition representation models coherent visual evolution between observations.

4.6 Real-World Evaluation

Experimental setup. To evaluate transfer to closed-loop physical navigation, we deploy NavCompose on a mobile robot. We compare NavCompose with ViNT and NoMaD across four routes: Corridor, Garden, Pantry, and Office. All methods use the same route-specific topological maps and downstream navigation and control stack. We conduct three trials per method on each route, yield-

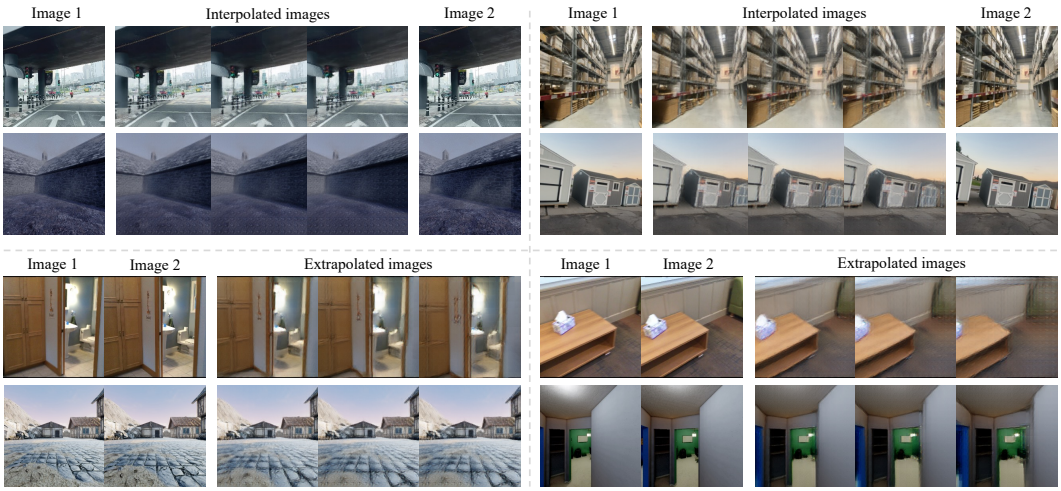

Figure 4: **Latent interpolation and extrapolation.** Top: given two observations, we interpolate the learned transition latent to render intermediate viewpoints. Bottom: given two observations, we extrapolate the transition latent to predict future viewpoints beyond the observed pair. Across diverse indoor and outdoor scenes, generated views evolve smoothly and coherently, indicating that the transition space captures continuous visual dynamics and serves as a lightweight world model.

ing 12 trials per method and 36 trials in total. A trial is successful if the robot reaches the predefined physical goal region and stops autonomously within it, without collision or human intervention. Additional deployment details and qualitative visualizations are provided in the Appendix.

Table 4: **Real-world topological navigation results.** Route-specific entries report successful trials out of three. The overall success rate (SR) is computed across 12 trials per method.

Method	Corridor	Garden	Pantry	Office	Overall SR
ViNT	3/3	0/3	0/3	2/3	41.7%
NoMaD	1/3	0/3	0/3	0/3	8.3%
NavCompose	3/3	1/3	2/3	3/3	75.0%

Results. As shown in Table 4, NavCompose achieves the highest overall success rate of 75.0% (9/12), compared with 41.7% (5/12) for ViNT and 8.3% (1/12) for NoMaD. NavCompose successfully completes all trials on Corridor and Office, and also achieves higher success rates on Garden and Pantry. These results demonstrate that NavCompose transfers effectively from offline training to closed-loop real-world navigation across diverse routes.

5 Conclusion

We presented NavCompose, a self-supervised framework for learning structured visual transition representations from unlabeled videos for image-goal navigation. By grounding relative viewpoint changes through image reconstruction and modeling their compositional and continuous structure, NavCompose learns how observations evolve along trajectories. The resulting representation supports transition composition, viewpoint interpolation, and short-horizon extrapolation, while transferring effectively to navigation through a lightweight waypoint prediction head without visual rollout at inference. Experiments across multiple benchmarks demonstrate state-of-the-art waypoint prediction, label efficiency, and efficient inference. These results highlight the potential of unlabeled sequences for learning structured and transferable representations for embodied navigation.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, LaTeX formatting, and system environment preparation, such as configuring experimental environments and assisting with deployment setup. The tools were not used to formulate research hypotheses, design the research methodology or experiments, or interpret the experimental results. All AI-assisted text and configurations were reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, and experimental artifacts produced with the aid of generative AI.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Wenzhe Cai, Jiaqi Peng, Yuqiang Yang, Yujian Zhang, Meng Wei, Hanqing Wang, Yilun Chen, Tai Wang, and Jiangmiao Pang. Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Dwip Dalal, Utkarsh Mishra, Narendra Ahuja, and Nebojsa Jojic. City navigation in the wild: Exploring emergent navigation from web-scale knowledge in mllms. *arXiv preprint arXiv:2512.15933*, 2025.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models. *arXiv preprint arXiv:2601.00844*, 2025.

Samiran Gode, Abhijeet Nayak, Débora NP Oliveira, Michael Krawez, Cordelia Schmid, and Wolfram Burgard. Flownav: Combining flow matching and depth priors for efficient navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17762–17768. IEEE, 2025.

Noriaki Hirose, Fei Xia, Roberto Martín-Martín, Amir Sadeghian, and Silvio Savarese. Deep visual mpc-policy learning for navigation. *IEEE Robotics and Automation Letters*, 4(4):3184–3191, 2019.

Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

Ranran Huang, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. From none to all: Self-supervised 3d reconstruction via novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37358–37369, 2026a.

Xijie Huang, Weiqi Gai, Tianyue Wu, Congyu Wang, Zhiyang Liu, Xin Zhou, Yuze Wu, and Fei Gao. Navdreamer: Video models as zero-shot 3d navigators. *arXiv preprint arXiv:2602.09765*, 2026b.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4929, 2025.

Faith Johnson, Bryan Bo Cao, Kristin Dana, Shubham Jain, and Ashwin Ashok. A landmark-aware visual navigation dataset. *arXiv preprint arXiv:2402.14281*, 2024.

Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7 (4):11807–11814, 2022.

Dongwon Kim, Gawon Seo, Jinsung Lee, Minsu Cho, and Suha Kwak. Planning in 8 tokens: A compact discrete tokenizer for latent world model. *arXiv preprint arXiv:2603.05438*, 2026.

Ashish Kumar, Saurabh Gupta, and Jitendra Malik. Learning navigation subroutines from egocentric videos. In *Conference on Robot Learning*, pp. 617–626. PMLR, 2020.

Shoya Kuno, Yumo Ouchi, and Kanata Suzuki. Co-training with ego-centric video and demonstration for robot navigation task. *arXiv preprint arXiv:2606.01951*, 2026.

Xingyao Lin, Guojin Zhong, Tianyi Lu, Ziyi Ye, Yichen Zhu, Zuxuan Wu, and Yu-Gang Jiang. Activemimic: Egocentric video pretraining with active perception. *arXiv preprint arXiv:2606.06194*, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6875–6885, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yanghong Mei, Yirong Yang, Longteng Guo, Qunbo Wang, Ming-Ming Yu, Xingjian He, Wenjun Wu, and Jing Liu. Urbannav: Learning language-guided embodied urban navigation from web-scale human trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18505–18513, 2026.

Wugang Meng, Tianfu Wu, Huan Yin, and Fumin Zhang. Beings: Bayesian embodied image-goal navigation with gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5252–5258. IEEE, 2025.

Thomas W Mitchel, Hyunwoo Ryu, and Vincent Sitzmann. True self-supervised novel view synthesis is transferable. *arXiv preprint arXiv:2510.13063*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20524–20531. IEEE, 2025.

Jiaqi Peng, Wenzhe Cai, Yuqiang Yang, Tai Wang, Yuan Shen, and Jiangmiao Pang. Logoplanner: Localization grounded navigation policy with metric-aware visual geometry. *arXiv preprint arXiv:2512.19629*, 2025.

Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5173–5183, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70. IEEE, 2024.

Jürgen Sturm, Wolfram Burgard, and Daniel Cremers. Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark. In *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, volume 13, pp. 6, 2012.

Zuojin Tang, Haoyun Liu, Xinyuan Chang, Changjie Wu, Dongjie Huo, Yandan Yang, Bin Liu, Zhejia Cai, Feng Xiong, Mu Xu, et al. Alam: Algebraically consistent latent transitions for vision-language-action models. *arXiv preprint arXiv:2605.10819*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Hangxing Wei, Xiaoyu Chen, Chuheng Zhang, Tim Pearce, Jianyu Chen, Alex Lamb, Li Zhao, and Jiang Bian. Learning additively compositional latent actions for embodied ai. *arXiv preprint arXiv:2604.03340*, 2026.

Yiming Zeng, Hao Ren, Shuhang Wang, Junlong Huang, and Hui Cheng. Navidiffusor: Cost-guided diffusion model for visual navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11994–12001. IEEE, 2025.

Mingkun Zhang, Wangtian Shen, Fan Zhang, Haijian Qin, Zihao Pei, and Ziyang Meng. Rae-nwm: Navigation world model in dense visual representation space. *arXiv preprint arXiv:2603.09241*, 2026.

Zhiwei Zhang, Hui Zhang, Kaihong Huang, Chenghao Shi, and Huimin Lu. Efficient image-goal navigation with representative latent world model. *arXiv preprint arXiv:2511.11011*, 2025.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7525–7535, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Changqing Zhou, Yueru Luo, Zeyu Jiang, and Changhao Chen. Uninav: A unified world-action diffusion model for visual navigation, 2026. URL https://arxiv.org/abs/2608.03244.