EmbodiedRec: Preference-Grounded World Models for Spatially-Aware Recommendation

Abstract

Recommendation systems traditionally operate in abstract latent spaces, divorced from the physical contexts in which users make decisions. We propose ***EmbodiedRec****, a framework that equips an embodied agent with a* Preference World Model *(PWM): a structured representation that jointly captures physical scene dynamics and user preference distributions conditioned on spatial context. Rather than treating recommendation as a static lookup over interaction histories, EmbodiedRec frames it as an* ***active navigation problem****: the agent perceives its environment, maintains a spatially-evolving preference belief, and issues contextually grounded suggestions: what activity to pursue, what content to engage with, what next action to take: based on where it is and what it observes. As a preliminary validation, we augment MovieLens-100K with simulated spatial context and show that even a simple context-aware model outperforms context-free BPR by* ***+5.0% NDCG@10*** *and raises Spatial Preference Correlation from near-zero to* ***0.393****, motivating the full embodied framework.*

1. Introduction

Recommendation systems are increasingly embedded in agents that operate in the physical world: home robots, AR assistants, and interactive kiosks. Yet despite this deployment context, state-of-the-art recommenders remain fundamentally **disembodied**: they model user preferences from interaction histories alone, with no awareness of where a user is, what they can see, or how their physical surroundings shape momentary intent.

Human preferences are deeply **spatial**. A person in a kitchen is primed for cooking-related activities; the same person on a couch is oriented toward rest and entertainment; someone at a desk signals a desire for focused, information-rich content. These contextual signals are continuously visible to any co-located embodied agent, yet are invisible to a history-only recommender.

The Embodied AI community has made significant strides in world models for navigation, manipulation, and question answering [2, 4, 6]. These world models encode rich priors about physical dynamics, object relationships, and spatial structure. We argue that this physical grounding is the *missing ingredient* in recommendation.

Contributions.

- **EmbodiedRec framework**: the first formulation of recommendation as embodied world-model navigation, linking physical scene understanding with real-time preference inference.
- **Preference World Model (PWM)**: a spatially-indexed latent representation that jointly models scene dynamics and user preference evolution during agent traversal.
- **Preliminary validation**: experiments on MovieLens-100K with spatial context augmentation confirming that spatial grounding significantly improves both ranking quality and contextual coherence.
- **AI2-THOR-Rec (proposed)**: a fully embodied benchmark for spatially-aware recommendation as future work.

2. EmbodiedRec Framework

2.1. Problem Formulation

Let an agent operate in environment $\mathcal{E}$ with egocentric RGB-D observations $o_t \in \mathcal{O}$ and actions $\mathcal{A}$ (navigate, look, interact). A user u has a latent preference vector $\mathbf{p}_u \in \mathbb{R}^d$ over item space $\mathcal{I}$ (activities, content, or actions). At each timestep t, the agent outputs a ranked list $\hat{R}_t$ over $\mathcal{I}$, aiming to maximize:

$$\max_{\pi} \mathbb{E}\left[\sum_t \gamma^t \cdot \mathrm{Rel}(\hat{R}_t, u, o_t)\right], \tag{1}$$

where Rel rewards items simultaneously preferred by u *and* contextually coherent with the current observation.

2.2. Preference World Model (PWM)

Spatial Scene Encoder f_θ: A vision transformer [3] maps each egocentric frame o_t to a spatial feature map $S_t \in \mathbb{R}^{H \times W \times d}$, encoding object identity, functional affordance, and scene layout. Weights are initialized from a pretrained embodied navigation backbone.

Preference Dynamics Module g_ϕ: A GRU-based recurrent module evolves the preference estimate as the agent moves:

$$\hat{\mathbf{p}}_t = g_\phi(\hat{\mathbf{p}}_{t-1},\ \mathrm{Pool}(S_t),\ \mathbf{h}_{t-1}), \tag{2}$$

where $\mathbf{h}_{t-1}$ is the GRU hidden state. A kitchen scene upweights cooking activities; a living room at dusk upweights leisure content.

Grounded Recommendation Head h_ψ:

$$\mathrm{score}(i) = \hat{\mathbf{p}}_t^\top \mathbf{e}_i + \lambda \cdot \mathrm{CoSim}(S_t,\, \mathbf{v}_i), \tag{3}$$

where $\mathbf{e}_i$ is the item semantic embedding and $\mathbf{v}_i$ its visual representation via CLIP [7]. λ is learned jointly during fine-tuning.

2.3. Training

Stage 1 (Embodied Pre-training). f_θ and g_ϕ are pre-trained on embodied navigation trajectories (AI2-THOR [6]) with a next-frame prediction objective, without any recommendation signal.

Stage 2 (Recommendation Fine-tuning). The full PWM is fine-tuned end-to-end using a listwise LambdaRank [1] loss over spatially-conditioned user preference labels.

3. Proposed Benchmark: AI2-THOR-Rec

We propose **AI2-THOR-Rec** as a future benchmark for fully embodied spatially-aware recommendation. The plan is to augment AI2-THOR's 120 household scenes (kitchens, living rooms, bedrooms, offices) with a general-purpose activity and content catalog spanning five categories: Cooking, Exercise, Entertainment, Learning, and Relaxation. User profiles will be derived from behavioral interaction logs, encoding long-term preferences as latent vectors independent of any single domain. A key evaluation metric will be **Spatial Preference Correlation (SPC)**: the Spearman correlation between recommendation relevance scores and the cosine similarity of item semantic representations to the observed scene features: measuring whether recommendations are *semantically grounded* in the agent's surroundings, not merely history-matched.

4. Preliminary Validation

Setup. As a controlled preliminary test of the core hypothesis: that spatial context improves recommendation, we augment the MovieLens-100K dataset [5] with simulated spatial context. Movie genres are mapped to four room types (Living Room, Bedroom, Study, Kitchen) by semantic affinity (*e.g.*, Documentaries $\rightarrow$ Study, Romance $\rightarrow$ Bedroom). Each positive interaction is assigned to the semantically coherent room with probability 0.92, and a random room otherwise, simulating noisy real-world spatial signals. We apply a standard leave-one-out evaluation protocol with 99 sampled negatives, reporting NDCG@10, Hit@5, and SPC (Spearman ρ between relevance scores and spatial coherence across candidates).

Models. *BPR* [8]: standard matrix factorization with Bayesian Personalized Ranking, no context. *Context-BPR*: BPR extended with a learned room embedding projected into item space (a lightweight proxy for the PWM's context module).

Method	NDCG@10	Hit@5	SPC
BPR [8]	0.465	0.562	0.016
Context-BPR (spatial)	**0.489**	**0.610**	**0.393**

Table 1. Preliminary results on MovieLens-100K with simulated spatial context. Context-BPR improves NDCG@10 by +5.0%, Hit@5 by +8.7%, and raises SPC from near-zero (0.016) to 0.393, confirming that spatial context is a meaningful signal for personalized recommendation.

Results. Table 1 shows that even a simple room-embedding context model consistently outperforms context-free BPR. The most striking result is SPC: BPR produces recommendations with essentially no spatial coherence (0.016), while Context-BPR achieves 0.393—demonstrating that the model learns to align recommendations with the user's observed spatial context. This validates the core premise of EmbodiedRec, and motivates the full PWM framework where context is derived from live egocentric observations rather than a fixed room label.

5. Conclusion

We introduced **EmbodiedRec**, a framework that unifies general-purpose personalized recommendation with embodied world modeling. Preliminary results on MovieLens-100K confirm that spatial context is a meaningful inductive bias: a simple context-aware model improves NDCG@10 by 5.0% and raises spatial coherence dramatically over a context-free baseline. The full framework with the PWM deriving context dynamically from egocentric observations and the AI2-THOR-Rec benchmark are directions for ongoing work. We aim to release benchmark and code to support research at the intersection of embodied intelligence and context-aware recommendation, with applications to assistive home robots, museum guide agents, and personal AI companions.

References

[1] Chris Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010. 2

[2] Zhenyu Chen, Yanchao Liu, and Hao Wang. Embodied world models emerge from navigational task in open-ended environments. *arXiv preprint arXiv:2504.11419*, 2025. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Interna-*

tional Conference on Learning Representations (ICLR)*, 2021. 1

[4] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *Nature*, 620:757–762, 2023. 1

[5] F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015. 2

[6] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Ranjay Krishna, Ashish Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 1, 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 2

[8] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009. 2