World2Minecraft: Occupancy-Driven Simulated Scenes Construction

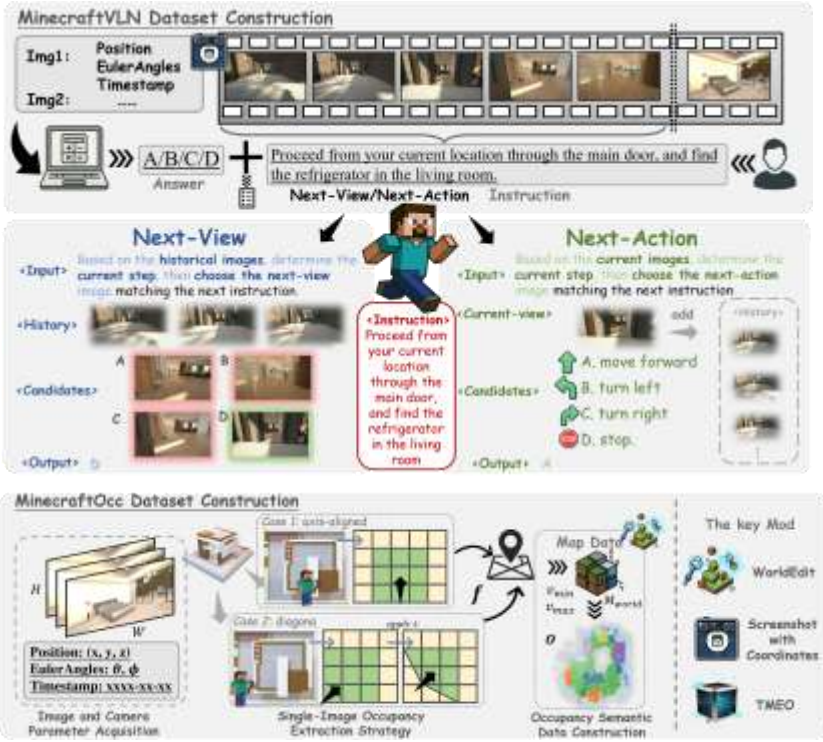
Lechao Zhang[1], Haoran Xu[1], Jingyu Gong[1], Xuhong Wang[2], Yuan Xie[1,3], Xin Tan[1,2,*]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Artificial Intelligence Laboratory
[3]Shanghai Innovation Institute

ICLR

Motivation

Embodied AI needs worlds that are both **realistic and editable**. Existing methods usually offer **only one side**: real-scene simulators are hard to edit, while Minecraft is editable but far from reality. Moreover, reliable transfer relies on 3D semantic occupancy prediction, which is limited by costly and scarce data. This motivates us to **bridge reality and Minecraft**.

Dataset Construction



Method

Framework of World2Minecraft: real-world scenes are reconstructed into Minecraft via occupancy prediction, enabling VLN tasks such as Next-View and Next-Action in the generated environments.

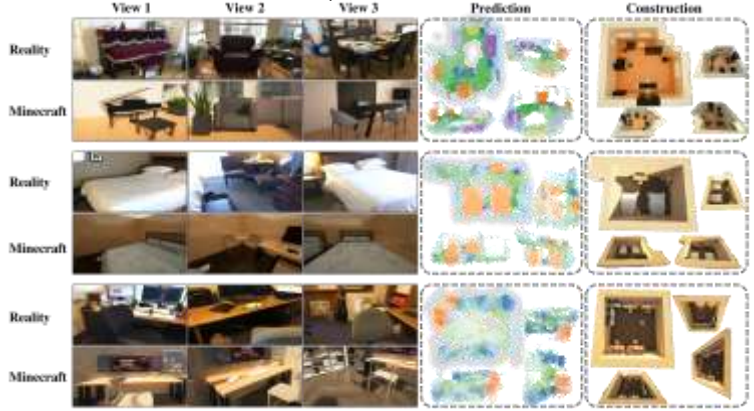
Overview of World2Minecraft: multi-view observations are converted into fused 3D occupancy and then translated into executable Minecraft commands for scene reconstruction.

Algorithm 1 World2Minecraft: Reality-to-Virtual Transfer

Require: **Input:** Image set $\mathcal{I} = \{I_1, \ldots, I_N\}$
Camera intrinsic parameters $\mathcal{K}$
Camera extrinsic parameters $\mathcal{E}$
Pretrained models $\mathcal{F}_{\text{mono}}$, $\mathcal{F}_{\text{emb}}$
Ensure: **Output:** Reconstructed Minecraft scene
procedure RECONSTRUCTSCENE($\mathcal{I}, \mathcal{K}, \mathcal{E}, \mathcal{F}_{\text{mono}}, \mathcal{F}_{\text{emb}}$)
  $\mathcal{O}_{\text{mono}} \leftarrow \emptyset$ ▷ Initialize monocular predictions set
  for each image $I_i \in \mathcal{I}$ **do** ▷ Process each view
    $\mathcal{O}^i_{\text{mono}} \leftarrow \mathcal{F}_{\text{mono}}(I_i, \mathcal{K})$ ▷ Generate per-view occupancy
    $\mathcal{O}_{\text{mono}} \leftarrow \mathcal{O}_{\text{mono}} \cup \{\mathcal{O}^i_{\text{mono}}\}$
  $\hat{\mathcal{O}}_{\text{scene}} \leftarrow \mathcal{F}_{\text{embodied}}(\mathcal{O}_{\text{mono}}, \mathcal{K}, \mathcal{E})$ ▷ Fuse multi-view predictions
  $\mathcal{D} \leftarrow \mathbf{K} * \hat{\mathcal{O}}_{\text{scene}}$ ▷ Compute density map via 3D convolution
  $\mathcal{C} \leftarrow \{\mathbf{v} \in \hat{\mathcal{O}}_{\text{scene}} \mid \mathcal{D}(\mathbf{v}) \geq \tau\}$ ▷ Extract centers above threshold τ
  $\mathcal{C}' \leftarrow \text{Cluster}(\mathcal{C}, \eta)$ ▷ Merge centers within distance η
  $\mathcal{M} \leftarrow \text{TranslateToMinecraft}(\mathcal{C}')$ ▷ Generate Minecraft building commands
  ExecuteCommands($\mathcal{M}$) ▷ Render scene in Minecraft
  return MinecraftScene ▷ Return reconstructed virtual scene

Result

Reality-to-Minecraft reconstruction results. Across different viewpoints, the reconstructed scenes stay consistent with real-world observations



VLN results in reconstructed scene. Guided by Gemini-2.5-Pro, the agent follows instruction "Go to the piano" and successfully reaches the target.

Comparative Experimental Results

Method	OOB ↓	Collision ↓	Semantic ↑	Visual ↑	Complete ↑	Aesthetic ↑
LayoutGPT	0.279	4.5	0.689	5.000	3.856	4.582
I-Design	0.423	**0**	0.884	6.001	4.734	5.352
LayoutVLM	**0**	0.9	0.348	3.625	2.270	2.708
World2Minecraft (Ours)	0.024	0.2	**0.913**	**6.145**	**5.186**	**6.022**

Dataset Composition	Task	Qwen2.5-VL-3B No Train	Qwen2.5-VL-3B SFT	Qwen2.5-VL-3B RFT	Qwen2.5-VL-7B No Train	Qwen2.5-VL-7B SFT	Qwen2.5-VL-7B RFT
Base	Next-View	0.2195	0.5610	0.2927	0.3905	**0.5854**	0.4390
	Next-Action	0.1943	0.7200	0.6343	0.3829	**0.8000**	0.6343
Extend	Next-View	0.2261	**0.7087**	0.3043	0.2913	0.6826	0.6043
	Next-Action	0.3657	0.5437	**0.6667**	0.3786	0.6019	0.6343
Combined	Next-View	0.2288	0.5609	0.3137	0.2878	0.6642	**0.6753**
	Next-Action	0.3037	0.4835	**0.6570**	0.3760	0.6281	0.6219