Supplementary Material

A Training Details

Detailed training configurations, including hyperparameters and implementation details, are available in the accompanying code repository: `https://github.com/XiaFire/Clone_Deterministic_Environment`.

B Additional Evaluation Metrics

In deterministic environments, there exists a single ground-truth trajectory for a given initial condition and action sequence. As a result, pixel-wise MSE provides a direct and reliable measure of fidelity, and is used as our primary metric.

To complement this evaluation, we additionally report perceptual metrics on Maze 9×9. These metrics capture structural similarity and distributional alignment that are not fully reflected by MSE. All metrics are computed in pixel space over rollout sequences. As shown in Table 1, the improvements are consistent with those observed under MSE.

Table 1: Supplementary perceptual metrics on Maze 9×9.

	SSIM ↑		rFID ↓	
Model	Baseline	GR	Baseline	GR
DF	0.8448	0.8516	4.4345	2.8729
VD	0.5979	0.7537	18.1453	7.2813
SD	0.8367	0.8369	6.4686	4.5200

C Additional Atari Experiments

We include additional experiments on Atari. The goal of this experiment is to extend our evaluation beyond static 3D environments to dynamic settings. We report PSNR and SSIM computed on rollout frames in pixel space. As shown in Table 2, GRWM consistently improves over the VAE-based world model across the evaluated environments. We note that this experiment is limited in scale and is not intended as a full Atari benchmark, but rather as a supplementary validation.

Table 2: Additional Atari results.

Env.	PSNR ↑		SSIM ↑	
	VAE-WM	GRWM	VAE-WM	GRWM
Asterix	28.57	29.04	0.9479	0.9518
Breakout	34.23	37.76	0.9848	0.9872

D Dataset Details

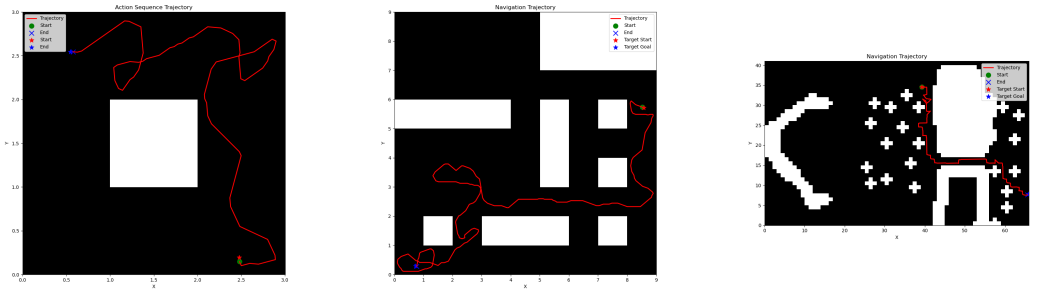
We evaluate our models on two environments: a memory Maze environment and a Minecraft environment.

Maze. For the Maze environment, we fix the random seed to generate a consistent set of maps. We use Memory-Maze Environment (Pasukonis et al., 2022). The rendered images are obtained using the MuJoCo engine. The agent has a discrete action space consisting of {move forward, turn left, turn right}. Trajectories are collected with a noisy A* algorithm to ensure sufficient coverage of the maze.

Minecraft. For the Minecraft environment, we adopt the map from Gornet and Thomson (2024) and enclose the area with wooden fences to restrict exploration. Trajectories are generated using a noisy A* policy under the same action space as in the Maze environment, providing diverse yet structured coverage. We first record the underlying deterministic state trajectories, and then render the corresponding pixel observations using Blender.

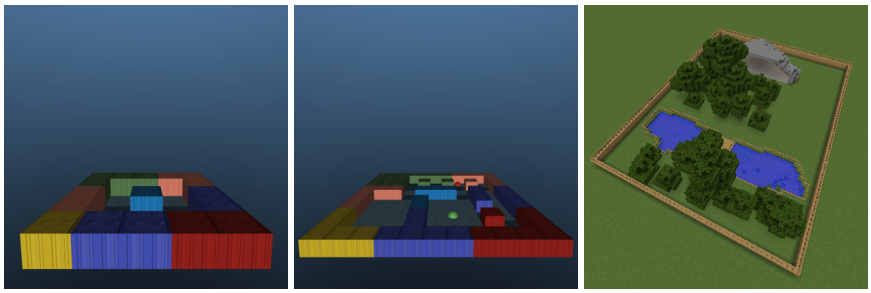
Statistics. Each trajectory contains up to 1000 frames, though most consist of several hundred frames. Each dataset contains 5000 trajectories in total.

Trajectory Visualization. To provide an intuitive understanding of the datasets, we visualize several representative trajectories. Figure 1 shows examples from three settings: M3x3-DET, M9x9-DET, and MC-DET.



M3x3-DET M9x9-DET MC-DET

Figure 1: Representative trajectories from the three datasets. Each plot shows a sample trajectory overlaid on the environment layout.



(a) M3×3-DET (b) M9×9-DET (c) MC-DET

Figure 2: High-angle perspective views of the three evaluation environments. These renderings provide an intuitive, three-dimensional understanding of the maze layouts that complements the 2D top-down maps in the main text.

E Ablation Studies

We conduct ablation studies to validate the contribution of our core components and design choices. Specifically, we examine four aspects: (1) the necessity of the two core loss terms, (2) the role of the projection head, (3) the impact of latent dimension, and (4) the effect of critical design choices on model performance.

E.1 Importance of Core Regularization Terms

Both slowness and uniformity losses are essential and complementary. We evaluate four model variants: a vanilla VAE, the full model, and two partial variants without $\mathcal{L}_{\text{uniform}}$ or $\mathcal{L}_{\text{slow}}$. For rollout evaluation, both partial variants diverged and produced NaN values, so we only report their loss statistics from autoencoder training. For completeness, we report the values of all loss terms during autoencoder training, even when they are not directly optimized. As shown in Table 3, removing either regularization leads to a substantial drop in rollout performance.

When optimizing for slowness alone (w/o $\mathcal{L}_{\text{uniform}}$), we observe a classic case of representation collapse. The model aggressively minimizes the slowness loss by mapping all representations to a tiny region of the latent space, evidenced by a very high uniformity loss and low slowness loss. This confirms that $\mathcal{L}_{\text{uniform}}$ is indispensable for preventing trivial solutions.

Table 3: Ablation study on the effect of regularization terms. We report the reconstruction loss $\mathcal{L}_{\text{recon}}$, the monitored slowness loss $\mathcal{L}_{\text{slow}}$, and the monitored uniformity loss $\mathcal{L}_{\text{uniform}}$.

Model	$\mathcal{L}_{\text{recon}}$	$\mathcal{L}_{\text{slow}}$	$\mathcal{L}_{\text{uniform}}$
VAE-WM	0.00042	0.88	-3.04
GRWM	0.00067	0.11	-3.13
GRWM w/o $\mathcal{L}_{\text{uniform}}$	0.00052	0.00015	0
GRWM w/o $\mathcal{L}_{\text{slow}}$	0.00100	0.46	-2.47

When optimizing for uniformity alone (w/o $\mathcal{L}_{\text{slow}}$), we interestingly observe that the slowness metric naturally decreases. We attribute this effect to pushing different trajectories apart, which implicitly encourages representations from the same trajectory to cluster. However, explicitly including $\mathcal{L}_{\text{slow}}$ accelerates and reinforces this trend.

E.2 Role of the Projection Head

Table 4: Effect of the projection head. The projection head reduces reconstruction loss while maintaining latent probing performance.

Model	$\mathcal{L}_{\text{recon}}$	Probing MSE
VAE-WM	0.00039	0.136
GRWM (w/ proj)	0.00061	0.058
GRWM (w/o proj)	0.00291	0.054

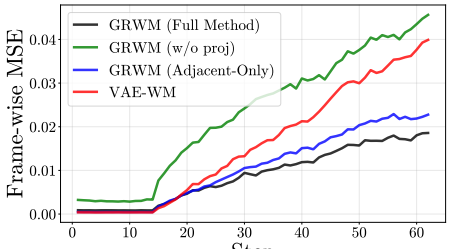


Figure 3: Frame-wise prediction MSE across ablation variants.

The projection head disentangles representation structuring from pixel-level reconstruction. While the model without a projection head can still learn a reasonably well-aligned latent space (as measured by latent probing MSE), its reconstruction loss is higher, as shown in table 4. By introducing the projection head, we allow the regularization losses to act in a separate subspace, freeing the primary latent space z to focus on accurate reconstruction. This decoupling ultimately leads to better overall predictive performance and lower frame-wise MSE, as shown in figure 3.

E.3 Analysis on Latent Dimension

We conducted an ablation study on the dimensionality of the latent space, testing dimensions of 16, 32, 64, and 128. The results is presented in Figure 4.

The benefits of our regularization are independent of the latent space size. GRWM consistently outperforms the vanilla VAE-WM across all tested dimensions. For every capacity, the rollout error of GRWM (solid lines) is significantly lower than that of the corresponding baseline (dashed lines).

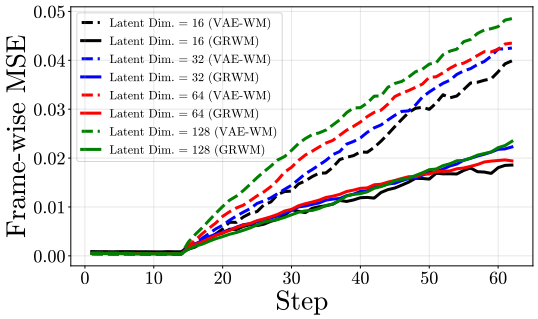


Figure 4: Ablation study on the impact of latent dimension. GRWM (solid lines) consistently and significantly outperforms the vanilla VAE baseline (dashed lines) across all tested latent dimensions (16, 32, 64, and 128). Notably, our method's performance is remarkably robust to the choice of latent dimension, while the baseline's performance is highly sensitive.

More importantly, our method demonstrates remarkable robustness to this hyperparameter. The performance curves for our model with latent dimensions 16, 32, 64, and 128 are nearly indistinguishable, indicating that our regularization technique successfully structures the latent space and learns a compact representation of the true state manifold, regardless of the available capacity. In contrast, the baseline's performance is highly sensitive to the latent dimension. For the vanilla VAE-WM, a larger latent space appears to be detrimental, leading to faster error accumulation. Without proper regularization, a higher capacity latent space may overfit to irrelevant visual details or capture noise, which harms long-term prediction. Our method effectively mitigates this issue, ensuring stable and predictable performance, which is a highly desirable property for practical applications.

E.4 Design of the Slowness Loss

All-pairs temporal consistency is crucial for preventing degenerate solutions. A critical design choice in our $\mathcal{L}_{slow}$ formulation is to pull all pairs of frames within a trajectory's context window closer, rather than only adjacent pairs. We compare our "All-Pairs" approach with an "Adjacent-Only" baseline (Figure 3). The results show that the "All-Pairs" strategy is superior. We attribute this to the causal nature of our encoder.

Since the encoder's output for frame t is already conditioned on frames $t-k, \ldots, t-1$, simply minimizing the distance between z_t and z_{t-1} presents a "lazy" optimization problem due to their overlapping inputs. In contrast, our "All-Pairs" strategy enforces smoothness across distant frames with non-overlapping inputs (e.g., z_t and z_{t-k}), leading to globally coherent latent trajectories and stronger long-horizon prediction.

F Additional Rollout Visualizations

We provide additional rollout visualizations for both the M9x9-DET environment in Figure 5 and the MC-DET environment in Figure 6. As shown in the M9x9-DET results (Figure 5), our method significantly outperforms the VAE baseline: while the VAE predictions are already inaccurate at 100 steps, our model maintains high fidelity at this horizon. In some cases, our method can occasionally produce accurate predictions even at 400 steps, demonstrating the improved consistency of the latent-space trajectories with the true environment. The MC-DET results (Figure 6) further confirm this robustness on complex trajectories. These results are not cherry-picked; the figure shows samples from randomly selected starting points, illustrating the typical performance of both methods over long-horizon rollouts.

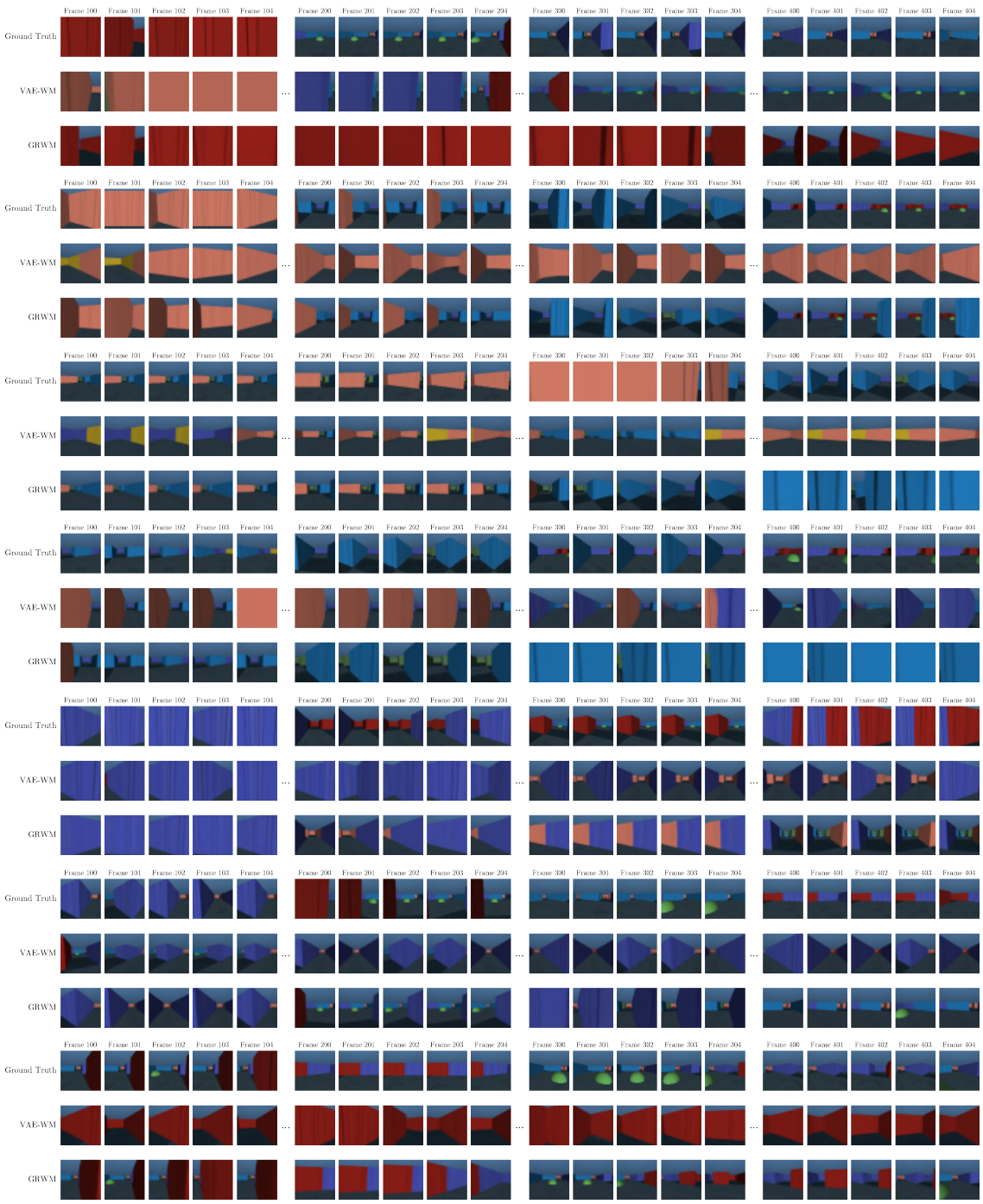


Figure 5: Visualization of generated frames at multiple time points from a single starting state. We show frames near steps 100, 200, 300, 400, sampled randomly — no cherry-picking. Our method significantly outperforms the VAE baseline: while the VAE predictions are already inaccurate at 100 steps, our model maintains high fidelity at this horizon. In some case, our method can occasionally produce accurate predictions even at 400 steps, demonstrating the improved consistency of the latent-space trajectories with the true environment.

Figure 6: Visualization of generated frames from the MC-DET sequence.

REFERENCES

James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6(7):820–833, 2024.

Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3d mazes. *arXiv preprint arXiv:2210.13383*, 2022.