Effective World Modeling for Humanoid Robots: Long-Horizon Prediction and Efficient State Compression

Technical Report (Team Duke)

Peter Liu Annabelle Chu Yiran Chen
Duke University

Abstract

World models enable robust policy evaluation of robots by simulating environment responses to agent actions. This paper details methodologies for two foundational components of such models, using datasets from the 1X World Model Challenge: (1) accurate, long-horizon visual prediction conditioned on robot poses, and (2) efficient compression of video data. For predictive sampling, we adapt Diffusion Forcing Transformer (DFoT). Our long horizion prediction model uses a UViT3D backbone, initially pretrained on RealEstate10K and then fine-tuned on the 1X humanoid raw video and state dataset. Accurate pose conditioning is achieved with a FiLM module in the ResBlocks and AdaLN-style mechanism in the TransformerBlocks. For state compression (cross-entropy evaluation), a conditional CNN is trained on 1X tokenized data, which predicts a future latent tokens. We achieve a 21.5578 PSNR for the frame generated two seconds into the future in sampling and a top-500 CE loss of 7.4976 in compression, ranking first in both categories.

1. Introduction

In general-purpose robotics, evaluating policies remains a significant hurdle. Real-world environments are time-consuming to set up and difficult to replicate with consistency across trials [1]. While traditional physics-based simulators offer control, they struggle with non-rigid objects and the overhead of asset creation. Learned world models offer a solution by learning to simulate an environment's response to an agent's actions directly from real-world video, absorbing its richness and complexity without extensive manual engineering [1].

Framed by the 1X World Model Challenge [2], this work addresses two foundational components: long-horizon predictive sampling and efficient data compression. Long-horizon predictive sampling tackles the challenge of generating coherent future sequences without the common failure modes of model collapse or hallucination. Efficient data compression is necessary to manage the significant storage and compute overhead of training on large, diverse datasets.

Figure 1. Diverse environments and tasks in the raw video and state dataset. Dataset available here [2].

2. Datasets

Our work utilizes two 1X Technologies datasets collected from EVE humanoid operations. The raw video and state dataset (for sampling) provides 512×512 raw videos (30 Hz) with synchronized robot states. This 25-dimensional state vector includes joint angles for all major limbs and the neck, alongside binary hand closure states and overall linear/angular velocities of the robot base. The tokenized video and state dataset (for compression) contains these videos processed by the NVIDIA Cosmos 8x8x8 tokenizer (64k vocabulary) into 3x32x32 tokens and the same synchronized states. Both datasets include training and validation splits, with 100 hours and 1 hour of video respectively.

3. Background Models and Techniques

Our work adapts established deep learning frameworks. For video prediction, we use the Diffusion Forcing Transformer (DFoT) [7], which iteratively refines predictions by enforcing consistency with historical context. This is built upon a UViT [4] backbone, a Vision Transformer architecture optimized for diffusion models. For action conditioning,

we adapt two mechanisms prominent in text-to-image synthesis: FiLM [6] modules modulate convolutional feature maps in ResBlocks, while AdaLN-style methods [5] adapt normalization layers in TransformerBlocks. For the compression task, our model predicts tokens generated by the NVIDIA Cosmos tokenizer [3], which quantizes spatio-temporal video patches into a finite vocabulary.

4. Long-Horizon Visual Prediction

4.1. Task and Evaluation

The 1X Sampling Challenge requires predicting a 512×512 video frame two seconds ahead (77th frame from a 17-frame, 30 fps context). This prediction is conditioned on past video and state trajectories (Sec. 2). We measure pixel-level fidelity via PSNR, defined as:

$$\mathrm{PSNR}(I, \hat{I}) = 10 \log_{10}\Big(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\Big), \tag{1}$$

$$\mathrm{MSE} = \frac{1}{H\,W} \sum_{i=1}^{H} \sum_{j=1}^{W} \big(I_{ij} - \hat{I}_{ij}\big)^2, \tag{2}$$

$$\mathrm{MAX}_I = \max_{1 \le i \le H,\ 1 \le j \le W} \{\, I_{ij} \}. \tag{3}$$

4.2. Approach: Action-Conditioned Diffusion Forcing Transformer

Our long horizon sampling model has 466M parameters and leverages the Diffusion Forcing Transformer (DFoT) framework [7]. We employ a UViT3D [4] backbone (Fig. 2; configuration in Sec. 4.3). This block structure leverages convolution for local feature processing in earlier ResBlock stages and the transformer's self-attention mechanism for global context modeling in deeper TransformerBlock layers. The denoising process is conditioned by robot pose vectors using adaptive modulation: FiLM for ResBlocks and AdaLN-style mechanisms for TransformerBlocks, detailed in Fig. 3.

Training and Guidance: The UViT3D backbone is initialized with RealEstate10K pretrained weights [7] at step 500k, then fine-tuned on 1X's raw train dataset for ≈ 1.2M steps. Pose-specific FiLM/AdaLN modules were trained from scratch, allowing them to specialize in mapping EVE robot kinematics to visual changes. Key parameters are in Table 1. Optimal PSNR was achieved using stabilized history guidance with a guidance scale of 3.0 and a stabilization level of 0.01.

Prediction and Post-processing: 17 frames were downsampled to 5 frames ($F_0, F_4, F_8, F_{12}, F_{16}$) serving as context. The model autoregressively generates 15 frames (at 4-frame intervals) to reach frame 77. Post-prediction, Gaussian blur ($\sigma = 2.0$) and histogram matching ($\alpha = 0.8$ with context frames) improved PSNR by approximately 1.6 dB.

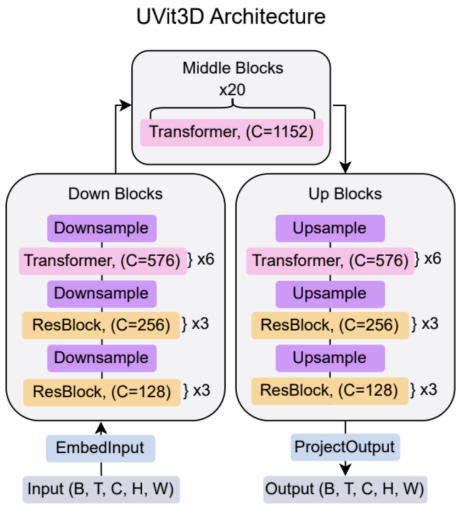

Figure 2. UViT3D backbone: U-Net with ResBlocks (C=128, 256) and TransformerBlocks (C=576, 1152). Adaptive conditioning (Fig. 3) uses pose signals.

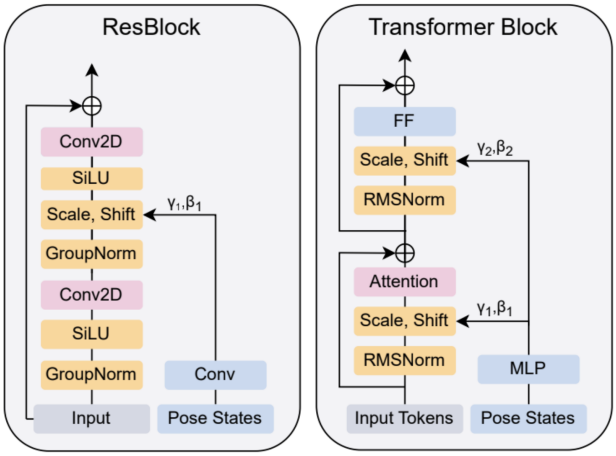

Figure 3. ResBlock (left) and TransformerBlock (right) internals, showing pose-conditioned FiLM and AdaLN-style modulation.

Table 1. Key Training Hyperparameters for Sampling Model

Parameter	Value
Trainable Parameters	466M
Base Model Pretraining	RealEstate10K (500k steps)
Fine-tuning Steps	1.16M (Total 1.66M)
Batch Size (per GPU)	1 (on 8 A5000s)
Init. LR	1e-5
LR Scheduler	Cosine decay to 1e-6 (w/ warmup)
Diffusion Beta Schedule	Cosine
Training Duration	≈ 1 month

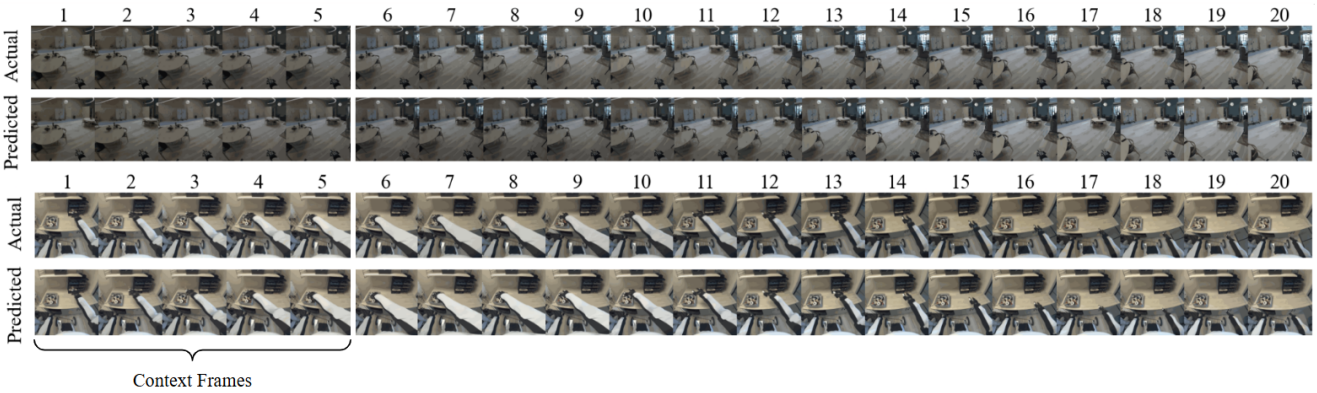

Figure 4. Examples of successful long-horizon prediction.
Top: Moving around in a kitchen; **Bottom:** Grabbing and placing an object.
Videos: https://huggingface.co/spaces/Ppffg/1xdemo.

4.3. UViT3D Configuration

Training and Guidance: The UViT3D backbone, along with blocks defined by Fig. 3, has 9 transformer heads, RoPE embeddings, and an EmbedInput patch size of 2. Timestep embeddings utilized Fourier features. Fine-tuning parameters are in Table 1.

4.4. Results: Predictive Quality and Examples

The model achieved a test PSNR of **21.5578**. Validations PSNR reached **25.6196 dB**, as shown in Fig. 5. Qualitative examples of successful predictions are presented in Fig. 4, and common failure modes illustrated in Fig. 6.

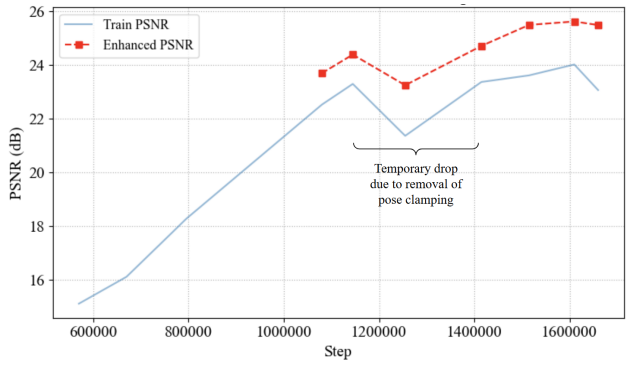

Figure 5. Validation PSNR curve. The normalized pose vectors were clamped at the start of training for stability and then unclamped (causing dip then rise in PSNR). "Post processed" includes blur ($\sigma = 2.0$) and histogram matching ($\alpha = 0.8$ with context frames only).

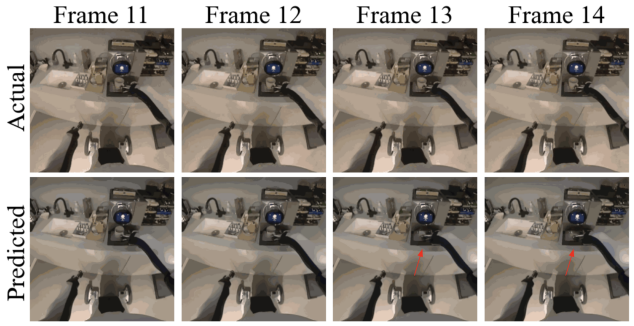

Figure 6. Common failure mode: object disappearance during long-horizon generation.

5. Efficient Compression

Training world models on large video datasets is computationally intensive, making efficient latent token compression crucial for scalability.

5.1. Task and Evaluation Metric

The 1X Compression Challenge evaluates this capability by tasking models to predict the next 17 tokenized frames (3x32x32) given 17 context frames and corresponding robot states. Performance is measured by the cross-entropy (CE) loss between predicted and ground-truth tokens, with the official leaderboard ranking based on the top-500 token CE.

5.2. Approach: Conditional CNN for Latent Token Prediction

To process the 3x32x32 token grid, we employ a 72M parameter FiLM-conditioned residual CNN (Fig. 7) trained from scratch. The CNN architecture is well-suited to capture local patterns in the latent space. It embeds video tokens and actions, processes them via ResNet-style blocks, and outputs token logits. Hyperparameters are in Table 2.

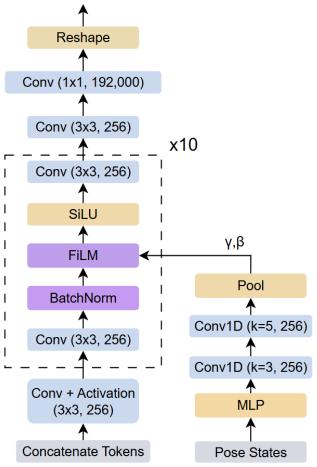

Figure 7. Conditional CNN architecture for latent token prediction.

Table 2. Key Hyperparameters for Compression Model

Parameter	**Value**
Trainable Parameters	72M
Initialization	From scratch on train set
Epochs	8
Batch Size	16
Init. LR	2×10^{-5}
LR Scheduler	Cosine anneal to 1×10^{-6}
Architecture (Selected)	
Embedding dim.	32
Channels	128
Residual blocks	**10**
Condition dim.	128
Loss Specific	
KL weight	1×10^{-4}
KL annealing	$0 \rightarrow 0.005$ (10k steps)

5.3. Results

The CNN achieved a top-500 CE loss of **7.4976** (1st place) and a full-vocabulary CE of **5.6759** (Table 3). Validation loss is shown in Fig. 8.

Table 3. Performance Summary on World Modeling Benchmarks

Benchmark Area	Metric	Our Score	Rank
Sampling	PSNR (Val)	25.3 dB	—
	PSNR (Test)	**21.5578 dB**	1st
Compression	CE (Full, Internal)	5.6001	—
	CE (Top-500, Test)	**7.4976**	1st

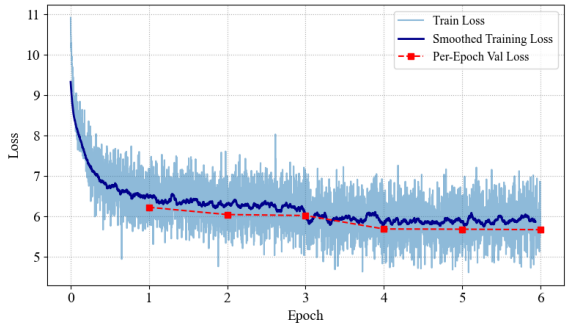

Figure 8. Validation full-vocab CE loss for compression model.

Table 4. Top 3 Performance on World Modeling Leaderboard

Benchmark Area	Submitter	Scores	Rank
Sampling	Duke	**21.5578 dB**	1st
	Micheal	18.5083 dB	2nd
	vjango	18.4823 dB	3rd
Compression	Duke	**7.4976**	1st
	a27sridh	7.9869	2nd
	WaterlooVipLab	7.9869	3rd

6. Discussion and Conclusion

We demonstrate effective strategies for long-horizon visual prediction and efficient state compression, achieving state-of-the-art performance on both 1X World Model Challenge benchmarks. Our adapted Diffusion Forcing Transformer and conditional CNN placed first in the Sampling (**21.5578** PSNR) and Compression (**7.4976** CE) tasks, respectively, outperforming the runners-up by significant margins (Table 4).

These results validate that adapting powerful generative architectures with precise, task-relevant conditioning is a highly effective strategy for building the core components of learned simulators. The model's success in coherent generation (Fig. 4) stems from effectively grounding visual changes in agent kinematics via the conditioning modules. Conversely, object disappearance (Fig. 6), highlights a remaining challenge in maintaining long-term state consistency for scene elements not directly manipulated by the agent. Addressing these coherence issues and integrating explicit physics priors remain critical open research areas. Future work building on these capabilities will be a key step toward realizing the "scaling laws" of robotics and, ultimately, more generalizable embodied intelligence.

References

[1] 1X Technologies. 1x world model. https://www.1x.tech/discover/1x-world-model, 2024. 1

[2] 1X Technologies and OpenDriveLab. 1X World Model Challenge Home. Hugging Face Spaces, 2025. [Online]. Available: https://huggingface.co/spaces/1x-technologies/1X_World_Model_Challenge_Home. 1

[3] Niket et al. Agarwal. Cosmos world foundation model platform for physical ai, 2025. 2

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models, 2023. 1, 2

[5] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 2

[6] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. 2

[7] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. 1, 2