Walk through Paintings: Egocentric World Models from Internet Priors

Anurag Bagchi[1] Zhipeng Bao[1] Homanga Bharadhwaj[1]
Yu-Xiong Wang[2] Pavel Tokmakov[3*] Martial Hebert[1*]
[1]CMU [2]UIUC [3]TRI

egowm.github.io

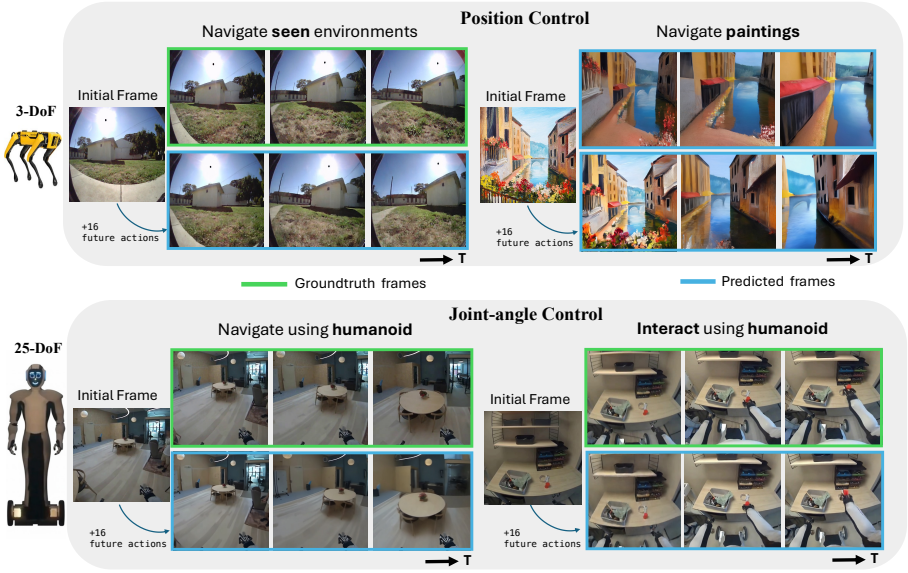

Fig. 1: Our framework generates future frame predictions (shown in blue) that accurately follow the provided robot actions (ground-truth frames shown in green). Notably, it effortlessly generalizes to vastly different embodiments, tasks, and domains, including non-realistic ones, such as navigating within paintings (top right). This generalization is enabled by our simple, universal architecture.

Abstract. What if a video generation model could not only imagine a plausible future, but the correct one – accurately reflecting how the world changes with each action? We answer this by presenting the Egocentric World Model (EgoWM), a simple, architecture-agnostic method that transforms any pre-trained video diffusion model into an action-conditioned world model, enabling precisely controllable future predic-

* Equal advising.

tion. Rather than training from scratch, we repurpose the rich world priors of Internet-scale video models by injecting appropriately compressed motor commands through lightweight conditioning layers. This allows our model to follow actions faithfully while preserving generalization and realism. Our approach scales naturally across embodiments and action spaces – from 3-DoF mobile robots to 25-DoF humanoids, where predicting egocentric joint-angle driven dynamics is substantially more challenging. The model produces coherent rollouts for both navigation and manipulation, requiring only modest fine-tuning. To evaluate physical correctness independent of appearance, we introduce the Structural Consistency Score (SCS), which measures whether stable scene elements evolve consistently with the provided actions. Our method improves SCS by up to 65% over the prior state of the art, Navigation World Models; applies seamlessly to three different video diffusion model architectures; and effectively utilizes Internet priors to generalize to unseen environments, including navigation and manipulation inside paintings. Finally, we demonstrate the applicability of EgoWM to robotic planning.

Keywords: Generative Models · World Models · Video Diffusion

1 Introduction

Modeling how the future visual state of the world evolves in response to an agent's actions – often referred to as *world modeling* [19] – is a fundamental capability for organisms with vision. It enables planning in navigation and manipulation tasks and even creative problem-solving [33, 39]. Animals acquire this ability through a lifetime of interaction and observation of others, allowing them to anticipate outcomes even in novel or physically impossible scenarios.

For artificial agents, however, especially for the ones with complex embodiments (like humanoids), action-conditioned videos in diverse real-world settings are expensive to acquire – an agent must interact with the world to get this data. As a result, many existing works train world models on narrow, simulator-defined datasets [3, 47] or with bespoke designs [7, 43, 51, 60], which constrains their scalability. This leads to a crucial question: can we re-purpose off-the-shelf video models trained on large-scale *passive* data, and convert them into world models using only a small amount of paired *action-observation* data?

Several prior works have explored learning action-conditioned video prediction models [7, 18, 21, 47, 61]. For example, the recent Navigation World Models (NWM) approach [7] trains a 1-billion-parameter diffusion model on egocentric videos of robots and humans to model navigation behavior, while Ctrl-World [18] fine-tunes StableVideoDiffusion [8] on the large-scale DROID dataset [26] for table top manipulation.

However, these methods remain domain-specific, learned for a single robot setup in limited set of environments. In contrast, modern video generation models [2, 49] capitalize on the more scalable Diffuion Transformer (DiT) architecture [38] and Internet-scale video-language datasets [13] to enable emulating

real-world interactions for any embodiment and environment when given suitable prompts [55]. Crucially, these models continue to improve at an astonishing pace [9, 35].

The key challenge is to introduce a mechanism for precise controllability, which is universally applicable to *any existing or future* video diffusion model architecture, without compromising generalization acquired during pre-training. We build on the observation that *every* video diffusion architecture uses a denoising timestep embedding to control the generation process, and inject the action control signal directly into this universal pathway. A central question is aligning actions with the *latent temporal resolution* of each model. We therefore match the model's latent compression factor and downsample the actions accordingly. The resulting action embeddings are combined with the diffusion timestep embedding, replicated along the temporal dimension of the latent video.

This design is *architecture-agnostic*: it does not alter the original model's architecture and applies across different backbones. We validate this with both UNet-based [42] and DiT-based [38] models with varying temporal compression factors (Section 5) and systematically ablate our design choices (Section 5.3). The results show improved action controllability and stronger generalization compared to prior conditioning approaches. Qualitatively, we demonstrate never-before-seen levels of generalization to novel environments, including navigation and manipulation inside paintings (Figure 6; see supplementary for discussion).

Our method is also *embodiment-agnostic*. It scales seamlessly from a 3-DoF setting to the 25-DoF joint space of the EVE 1X humanoid, even though large parts of the robot's body are not visible in the egoview. As illustrated in Figure 1, this enables high-dimensional humanoid navigation and manipulation – a capability not demonstrated by prior open-source world models.

To quantitatively evaluate our approach, we propose a new metric: **Structural Consistency Score (SCS)** in Section 4. Prior metrics such as LPIPS [57] or FVD [46] focus on perceptual similarity or realism [31, 54], which can conflate visual fidelity with physical correctness. A model could produce sharp, realistic frames that nevertheless depict physically inconsistent outcomes. SCS complements these appearance-centric measures by capturing whether the *structure* of the generated environment evolves correctly in response to the agent's actions, independent of texture. Concretely, we automatically identify *passive* (*e.g.*, walls, furniture, objects manipulated by the robot) scene elements, as well as the visible parts of the robot itself, in the first frame and track them across both real and generated videos using a segmentation tracker. Comparing the resulting trajectories allows us to assess whether the model's predicted world evolves coherently with the true environment. A higher SCS indicates stronger structural and physical consistency with the provided action sequence.

In summary, our contributions are as follows:

- We propose a simple yet powerful recipe, **EgoWM**, to convert *any* image-to-video diffusion model into an **action-conditioned world model**, regardless of base *architecture*, temporal *compression* and *action space* complexity.

- We demonstrate that our method works across various video diffusion backbones and extends seamlessly to **25-DoF** humanoids, enabling high-resolution video prediction for navigation and manipulation, from an **egocentric** view, generalizing to highly **Out-of-Distribution paintings**.
- We introduce the **Structural Consistency Score (SCS)**, a metric to evaluate the structural and physical plausibility of generated videos, measuring action alignment independently, for both *navigation* and *manipulation*.
- Finally, we demonstrate that EgoWM is effective for downstream robot trajectory **planning** in both 3-DoF navigation and 25-DoF manipulation.

By learning from large-scale passive datasets and a small amount of action-labeled data, our framework offers a path toward scalable, general-purpose visual dynamics models.

2 Related Work

World Models and Video Prediction have long been studied in reinforcement learning and robotics. The classic work of Ha *et al*. [19] trained a variational RNN to simulate simple games, allowing an agent to plan within its own latent dream. Follow-up works demonstrated video prediction for control in robotics – *e.g*., visual foresight model [14], which learned to predict future camera frames given robot actions and used model-predictive control to push objects to the target location, or SV2P [4], which used stochastic latent variables for multi-modal future predictions. These pioneering approaches, however, were limited to relatively constrained environments (toy simulations or single-task robot setups).

In the gaming domain, several works have proposed to visually emulate a game engine from screen pixels and keyboard actions [27, 34, 52]. More recently, there is a trend of utilizing high-capacity generative models for world modeling. The DIAMOND agent [3] uses a diffusion-based world model to better preserve visual detail in Atari games. GameNGen [47] takes a similar diffusion-based approach to simulate an entire 3D game in real time. These results underscore the potential of diffusion models to capture complex dynamics, but they have been confined to narrow domains, trained from scratch on domain-specific data.

In another line of work, world models have been trained for table-top manipulation scenarios [18, 37, 60, 61], but focus on narrow domains with static, exo-centric cameras. However, in the wild agents rarely have access to exo-centric views and need to move inside their environment. Very recently, Navigation World Model (NWM) [7] broadened the scope by training a single custom designed autoregressive diffusion model (cDiT) across embodiments for 3-DoF planar egocentric navigation. However, their design prevents them from leveraging internet-scale video diffusion pretraining, limiting generalisation. In this work, we push further on generality: rather than collecting more data or designing a new model, we adapt existing pre-trained models to a wide variety of ego-centric tasks, from 3-DoF navigation to 25-DoF humanoid loco-manipulation.

GrndCtrl [21], IRASim [61], and Ctrl-World [18] are recent works that leverage video generation pretraining for navigation and table-top manipulation. GrndCtrl re-uses the global action-conditioning mechanism of Cosmos [2], which we show to provide limited control precision. IRASim and Ctrl-World build on earlier video diffusion architectures [8, 59] and do not account for temporal compression — a key component of modern DiT-based models. Our approach, instead, provides a universal conditioning framework compatible with arbitrary video diffusion backbones and enables precise control in complex, high-dimensional action spaces.

Video Diffusion Models have emerged as a powerful paradigm for video generation. [22] demonstrated that a straightforward extension of image diffusion models can produce high-quality short videos. This was followed by text-to-video models leveraging massive paired visual-language samples [8,12,50]. These models are trained on observation-only video data (*e.g.*, crawled web videos) and are not inherently action-conditioned, but they learn strong priors about physics and appearances. There have also been advances in architectural design: Diffusion Transformers (DiT) [38] replaced the U-Net [42] backbone, achieving excellent video generation performance [2, 49, 56].

More recently, diffusion-based generative models have been adapted beyond pure synthesis to support downstream vision and control tasks — leveraging their rich priors to enable generalization to unseen domains. For example, in the image domain, several works [25, 30, 36, 58] have shown that a pretrained image diffusion backbone can be used to bring several classic vision tasks into the open world. Video diffusion representations have been successfully adapted for video segmentation [5,62], dynamic views synthesis [6,48], as well as for robot policy learning [23, 29], demonstrating never-before-seen levels of generalization for these tasks. In contrast, in world model learning, most of the approaches – while adopting the video diffusion objective – utilize custom model design, limiting the benefits of pre-trained representations.

Evaluation Metrics used in video generation were originally designed to capture image fidelity or distributional similarity. Frame-level measures such as SSIM [53] and PSNR [32] evaluate brightness, contrast, and structure per frame, while LPIPS [57] compares neural network features. Sequence-level metrics like Fréchet Video Distance (FVD) [46], which measures feature distribution differences using a 3D ConvNet [11], were later introduced to capture temporal quality. World modeling approaches typically adopt these metrics to evaluate predicted futures, yet high scores often fail to reflect action alignment. We address this gap with the Structural Consistency Score (SCS), which explicitly disentangles action-following accuracy from visual fidelity.

3 Method

3.1 Preliminaries

Modern video diffusion models [8, 12] synthesize future frames conditioned on an observed frame x_0 (and optionally texts) by iteratively denoising a Gaussian

latent. Let f_{vdm} denote a pre-trained image-to-video model parameterized for T_s diffusion steps. Sampling is performed by drawing $\epsilon \sim \mathcal{N}(0, I)$ and applying the reverse denoising process:

$$\hat{X} = f_{\text{vdm}}(c_t, \epsilon, T_s), \tag{1}$$

where c_t encodes the frame-level conditioning signal.

To reduce computation, these models operate in a low-dimensional latent space. A video $X \in \mathbb{R}^{T \times H \times W \times 3}$ is mapped by a spatio-temporal VAE [28, 41] into a latent tensor $z = \mathcal{E}(X) \in \mathbb{R}^{\frac{T}{k} \times h \times w \times c}$, where k denotes the temporal downsampling factor, carefully accounting for which is key for precise action conditioning.

Generation proceeds by decoding the predicted trajectory:

$$\hat{X} = \mathcal{D}(f_{\text{vdm}}(c_t, \epsilon, T_s)) \,. \tag{2}$$

During training, a noisy latent z_{t_s} is produced by the forward diffusion process, and the model learns to predict the injected noise:

$$\min_{\theta} \mathbb{E}_{z, t_s, \epsilon} \left[\| \epsilon - \epsilon_\theta(z_{t_s}, e_c, t_s) \|_2^2 \right] . \tag{3}$$

Here e_c denotes the conditioning embedding and $\epsilon_\theta(\cdot)$ is implemented by a U-Net [42] or a DiT-based [38] denoiser. Although architectures vary widely, all models share a common mechanism: latent features are modulated by the denoising timestep t_s, which forms the basis of our action-conditioning strategy.

3.2 Action Conditioning

Our objective is to convert passive, image-conditioned video diffusion models into action-conditioned world models. Formally, given an action sequence $\mathbf{A} \in \mathbb{R}^{D \times T}$, where D denotes the action-space dimensionality and T the prediction horizon, and an initial observation x_0, the model is fine-tuned to generate future frames:

$$\hat{\mathbf{X}}_{1:T} = f_{\text{vdm}}(x_0, \mathbf{A}, \epsilon, T_s). \tag{4}$$

To enable this, we design a general framework, that is architecture-agnostic, allowing a single conditioning strategy to generalize across multiple diffusion models and embodiments. Next, we describe the key components of our approach in more detail.

Action Projection. To condition a video diffusion model on a control trajectory $\mathbf{A}$, we first embed each D-dimensional action vector into a latent representation $\mathbf{Z_A} \in \mathbb{R}^{d \times T}$, where d is the embedding dimension. In its general form, the module, shown in Figure 2 (top), consists of a sequence of lightweight multi-layer perceptrons (MLPs) that map each action vector to the embedding space. For video diffusion models that use temporal latent compression (*i.e.*, VAEs with factor $k > 1$), action embeddings are further downsampled by 1D convolutions, producing $\mathbf{Z} \in \mathbb{R}^{d \times T/k}$ to ensure alignment with the latent video frame rate.

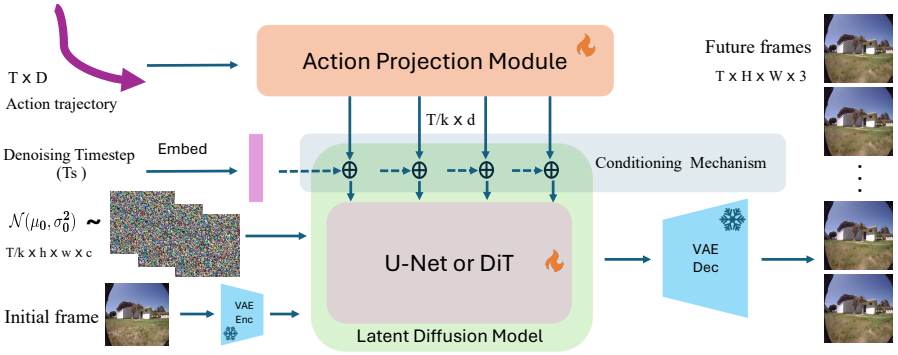

Fig. 2: EgoWM embeds high dimentional action sequences into a universal feature space and injects these embeddings into a pre-trained video diffusion model by reusing its timestep-conditioning pathway (shown in the top center). This enables turning any passive video generation model into a world model without compromising the pre-trained representation.

Different embodiments use different parameterizations: for *3-DoF Navigation*, we adopt the standard control representation $(\Delta x, \Delta y, \Delta\phi)$; for *25-DoF Humanoid Control*, the standard representation is adopted from [1], but otherwise our design is left unchanged. Because the projection module operates independently of the base architecture and only interfaces through the shared latent space, it can adapt to different temporal resolutions, action dimensionalities, and embodiment types without architectural changes.

Injecting Actions via Timestep Modulation. Rather than introducing model-specific action conditioning layers, as in prior work [18, 61], we leverage the universal diffusion timestep pathway and adapt it to the model's latent temporal resolution. Concretely, let the action embeddings be $Z_a \in \mathbb{R}^{d\times T/k}$, where k denotes the temporal compression factor. We replicate the global embedding of denoising timestep t_s along the the compressed temporal dimension to obtain $Z_{t_s} \in \mathbb{R}^{d\times T/k}$. At each location where the base model applies timestep-dependent modulation to the video latent, we add the corresponding action embedding (Figure 2, middle). This enables each temporally aligned action embedding to directly modulate the representation of its associated latent video frames. Formally, for each modulation block i, we redefine:

$$P_i^{\text{scale}}, P_i^{\text{shift}}, P_i^{\text{gate}} = F_i(Z_{t_s} + Z_a)\,, \tag{5}$$

where $F_i(\cdot)$ is the block-specific projection used by the base model.

Initial State Ambiguity in Humanoids. For humanoid control, where many body parts are not visible in the camera view, we additionally include the embedding of the initial agent state Z_s:

$$P_i^{\text{scale}}, P_i^{\text{shift}}, P_i^{\text{gate}} = F_i(Z_{t_s} + Z_a + Z_s)\,. \tag{6}$$

Fig. 3: Each column shows two generated rollouts compared to the corresponding ground-truth frame, with their LPIPS, DreamSim, and SCS scores. Perceptual metrics incorrectly favor the visually sharper but physically inconsistent sample, while SCS correctly identifies the sequence that follows the true action trajectory, accurately capturing structural consistency.

This simple formulation makes the modulation parameters frame-specific and action-aligned, while preserving architectural compatibility with both U-Net–based and DiT-based video diffusion models. Despite its simplicity, this strategy yields fine-grained motion control in both low-DoF and high-DoF settings, while allowing extreme generalization.

4 Structural Consistency Score

Existing action-conditioned video generation models are typically evaluated using perceptual similarity metrics such as LPIPS [57] or DreamSim [16], computed between the generated and ground-truth frames [7]. While effective for measuring visual fidelity, these metrics were designed to capture semantic similarity between object-centric images, not structural alignment in complex scenes. As shown in Figure 3, two model predictions can achieve similar perceptual similarity scores despite very different action-following behavior. Moreover, applying per-frame similarity metrics to long-horizon navigation ignores the inherent stochasticity of the problem: early in a trajectory, a model can meaningfully reconstruct the ground truth, but as the agent moves further, multiple plausible futures emerge.

To address these limitations and accurately evaluate action following, we *augment* the existing appearance-centric metrics with **S**tructural **C**onsistency **S**core (**SCS**). To compute SCS, we first trim each evaluation sequence to exclude frames containing only novel regions unseen in the initial observation, since structural correspondence is ill-defined there. We detect this transition automatically using the dense point tracking method of Harley et al. [20], to identify when all points visible in the first frame have left the field of view.

For the remaining frames, we select key scene structures — passive objects whose apparent motion stems solely from the agent's actions (*e.g.*, buildings, trees, or objects manipulated by the robot). Restricting evaluation to passive

Table 1: Comparison with state-of-the-art methods on the RECON validation set after training on three navigation datasets. Each column group reports LPIPS, DreamSim, and SCS scores for different prediction horizons. All of our variants, even the temporally compressed (k=4) models, outperform Navigation World Model, especially at longer horizons, highlighting the effectiveness of our approach.

Method	Frame 2			Frame 4			Frame 8			Frame 16		
	LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑
NWM ($k = 1$)	0.26	0.10	58.4	0.30	0.11	56.2	0.35	0.13	46.8	0.45	0.19	33.4
SVD ($k = 1$)	0.25	0.10	62.0	0.27	0.10	**61.7**	0.31	0.12	**57.2**	0.39	0.17	**55.2**
Cosmos-2B ($k = 4$)	**0.20**	**0.07**	**63.1**	**0.22**	**0.08**	60.4	**0.26**	**0.09**	56.4	**0.33**	**0.10**	47.5
Wan-14B ($k = 4$)	0.22	0.08	62.7	0.26	**0.08**	59.0	0.29	0.10	53.9	0.34	0.11	49.7

objects reduces ambiguity and allows us to leverage recent foundational models for video object segmentation [10, 40], which, as we show in the supplementary, are robust to the artifacts in generated videos. Each selected object is annotated either via sparse point clicks, or via natural language queries (see supplmentary for details), and tracked through both ground-truth and predicted sequences.

The final Structural Consistency Score is obtained by averaging mask IoU [15] across all T frames and N tracked objects:

$$\mathrm{SCS} = \frac{1}{NT} \sum_{j=1}^{N} \sum_{t=1}^{T} \frac{|\mathcal{M}_{\mathrm{pred}}^{(t,j)} \cap \mathcal{M}_{\mathrm{gt}}^{(t,j)}|}{|\mathcal{M}_{\mathrm{pred}}^{(t,j)} \cup \mathcal{M}_{\mathrm{gt}}^{(t,j)}|}, \tag{7}$$

where $\mathcal{M}_{\mathrm{pred}}^{(t,j)}$ and $\mathcal{M}_{\mathrm{gt}}^{(t,j)}$ are the binary masks of object j at frame t in the predicted and ground-truth videos, respectively. Higher SCS values indicate stronger structural alignment between the predicted and ground-truth sequences.

Figure 3 illustrates that SCS correlates strongly with action following, in contrast to perceptual metrics that emphasize visual realism. By focusing on structural evolution rather than appearance, SCS provides a quantitative measure of how faithfully a model predicts the causal consequences of agent's actions.

5 Experiments

Datasets and Evaluation. We evaluate our approach across three progressively challenging settings — (i) 3-DoF ego-centric navigation in real-world robot environments, (ii) humanoid navigation, and (iii) humanoid manipulation with a 25-DoF action space — using four open-source datasets. RECON [44] contains diverse real-world indoor navigation trajectories. SCAND [24] features long, continuous paths through structured indoor environments, emphasizing smooth motion and spatial consistency. TartanDrive [45] includes challenging outdoor trajectories with uneven terrain and dynamic backgrounds, testing robustness to visual and proprioceptive noise. Finally, the 1X Humanoid Dataset [1] includes both navigation and manipulation episodes paired with 25-DoF joint-angle states, enabling evaluation of high-dimensional embodied control.

All datasets are used for training, unless stated otherwise, and we report results on the validation splits of RECON and 1X Humanoid. Following prior

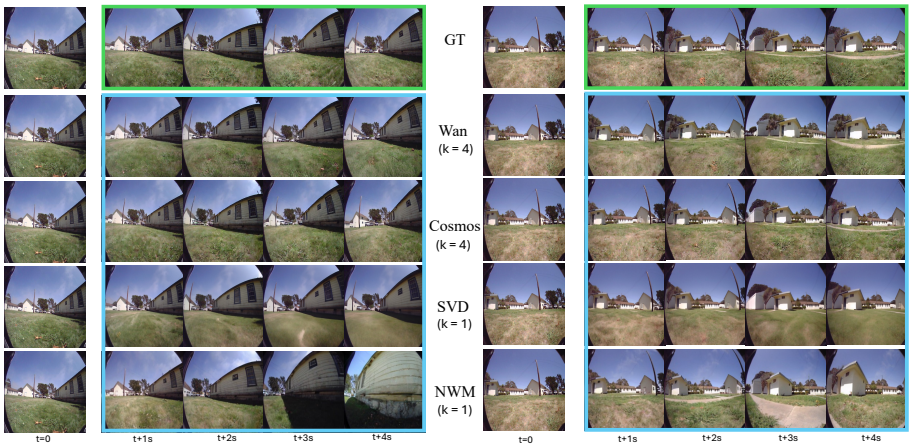

Fig. 4: Qualitative results on egocentric navigation. All variants of our model generate realistic, temporally coherent sequences that accurately follow the provided action trajectories regardless of compression, while NWM exhibits noticeable drift (*e.g.*, veering right in the left column). Results best viewed on the project page.

work [7], we report the standard perceptual similarity metrics LPIPS [57] and DreamSim [16] to assess visual fidelity between predicted and ground-truth frames. To better capture differences in action following accuracy between the models, we additionally report our SCS metric introduced in Section 4. We report results up to 4 seconds into the future at 4 FPS, with additional results over 4× longer horizons provided in Section C of the supplementary.

Implementation Details. Across all experiments, we evaluate three variants of our method built on top of the SVD [8], Cosmos [2] (Predict2-2B) and Wan2.1-14B [49] models. We use two 1D convolutional layers to downsample the action embeddings to match the temporally compressed latent resolution of Cosmos and Wan ($k = 4$). Fine-tuning is performed using the original pretraining objective of the base diffusion model (denoising or flow matching; see Section 3.1). The learning rate for the newly introduced action projection module is set to 10× higher than that of the base model weights.

5.1 Navigation Results

We fine-tune all three (SVD, Cosmos and Wan) variants of our method on the training sets of RECON, SCAND, and Tartan-Drive for a fair comparison with NWM [7] and report results on the validation set of RECON. Table 1 reports LPIPS, DreamSim, and our proposed SCS metrics for prediction horizons of 2 to 16 frames. All variants of our method outperform NWM across all evaluation horizons and metrics, with particularly large improvements in SCS and at greater distances from the initial frame (up to +65%), indicating substantially stronger action alignment. Cosmos and Wan achieve higher perceptual fidelity, reflecting the stronger visual priors induced by large-scale pretraining, while SVD

Table 2: Results of SVD, Cosmos, Wan and ablation variants on humanoid navigation and manipulation. Our approach allows seamless generalization to this complex 25-DoF action space. The pre-trained SVD variant outperforms the baseline trained from scratch by significant margins, demonstrating the value of Internet-scale pre-training in this challenging scenario.

Task	Variant	Frame 2			Frame 4			Frame 8			Frame 16		
		LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑	LPIPS↓	DreamSim↓	SCS↑
Navigation	**SVD**	0.11	0.07	**75.6**	0.16	0.09	**66.4**	**0.26**	0.13	**50.3**	**0.35**	0.18	**34.4**
	SVD (scratch)	0.17	0.15	69.7	0.25	0.20	56.0	0.37	0.30	40.0	0.47	0.40	21.6
	Cosmos-2B	0.11	0.04	65.2	0.19	0.06	54.0	0.27	0.09	42.3	0.40	0.19	27.0
	Wan-14B	**0.09**	**0.04**	70.2	**0.15**	**0.06**	61.3	0.27	**0.10**	45.2	0.39	**0.17**	30.2
Manipulation	**SVD**	0.04	0.04	**86.8**	**0.06**	0.05	81.8	0.10	0.07	75.0	0.13	0.10	**76.6**
	Cosmos-2B	**0.04**	**0.03**	86.2	0.07	**0.05**	**82.0**	**0.10**	**0.06**	**78.2**	**0.12**	**0.07**	67.7
	Wan-14B	0.05	0.04	81.1	0.07	0.06	78.5	0.11	0.07	76.0	0.14	0.09	65.6

exhibits more stable structural consistency over long trajectories due to the absence of temporal downsampling. Nevertheless, our action conditioning design enables modern, temporally compressed DiT-based models to attain strong SCS performance despite operating at reduced latent temporal resolution.

Figure 4 visualizes predicted rollouts on the RECON validation set. All variants of our model produce realistic sequences that closely follow the provided trajectories, while NWM exhibits noticeable drift. Cosmos and Wan generate sharper and more detailed predictions, whereas SVD exhibits mild fidelity loss due to its chunk-wise autoregressive inference. Overall, our results demonstrate that leveraging pre-trained video diffusion models for world modeling yields substantial improvements over specialized NWM architecture, combining high perceptual quality with robust action following.

In terms of efficiency, NWM is trained on 64 H100 GPUs and predicts at 224×224, while our SVD and Cosmos variants use only 8 A100 GPUs and generate higher-resolution outputs (512×512 and 480×640, respectively). On the same 20 RECON samples using a single A100 GPU, EgoWM achieves up to $6\times$ faster 64-frame inference than NWM, while using less action data and approximately $8\times$ less training compute; see Supplementary Sec. D for the detailed compute, data, resolution, and latency comparison.

5.2 Humanoid Results

We next evaluate our approach on the 1X Humanoid Dataset, a significantly more challenging setting involving a 25-DoF control space spanning all major joints, neck motion, and gripper actions. As shown in the 25-DoF navigation results in Table 2, all variants exhibit somewhat lower SCS scores toward the end of the trajectory compared to the 3-DoF navigation experiments. This reflects the increased difficulty of high-dimensional humanoid control, where maintaining coherent motion requires reasoning over many coupled joints and partial egocentric visibility. Nevertheless, our universal conditioning approach applies to both embodiments without any architectural modifications, demonstrating its scalability across action spaces that differ by nearly an order of magnitude in

complexity. The SVD model also outperforms its variant trained from scratch, with the gap being especially significant on perceptual similarity, confirming the value of Internet-scale, passive pretraining. We further compare to concurrently propsed humanoid WM architectures [17] in the ablation analysis in Section 5.3.

In the lower part of Table 2, we quantitatively evaluate our models on 25-DoF manipulation. Since the scene changes very little during manipulation compared to navigation, the LPIPS and DreamSim scores are much lower even at frame 16, compared to navigation. We also see high SCS scores, showing that all variants learn to accurately follow the manipulation commands. SVD is better at action following further into the future, compared to Cosmos and Wan, a common trend across all results, as a consequence of the temporal compression.

Figure 5 illustrates prediction of our model on humanoid sequences in the validation set of 1X. Despite the dramatic increase in embodiment complexity to 25 DoF, EgoWM produces stable trajectories that remain consistent with the provided actions for navigation, reaching, and grasping. Our model maintains coherent global scene structure while capturing detailed arm, hand, and gripper articulation throughout. Crucially, the predicted videos remain visually consistent and physically plausible, enabling planning applications in Section 6.

5.3 Ablation Analysis

In Table 3 we compare the core design choices of EgoWM against alternative conditioning strategies used in prior work. We first assess the effectiveness of our universal action-conditioning mechanism in the high-dimensional humanoid action space by comparing it to two common alternatives: global conditioning used in [21] and the Cosmos-Predict-2B World Model [huggingface], and the chunking strategy proposed in concurrent work [17]. While perceptual metrics remain largely unchanged across designs, controllability, measured by SCS, drops by 44.8% under global conditioning and by 12% with chunking. These results highlight the importance of temporally aligned action injection for precise control. Next, we evaluate how the action-conditioning pathway affects generalization by comparing our method to the cross-attention conditioning used in the recent Ctrl-World model [18]. As out-of-distribution data is not available for the 1X humanoid embodiment, we conduct this study in the 3-DoF navigation setting using SVD as the base model to ensure a faithful reproduction of [18]. Both variants are trained on SCAND (*urban campus*) and evaluated on the unseen RECON environments (*residential suburbs*). We observe that our denoising timestep-based conditioning is better at preserving priors learned during Internet-scale pre-training of video diffusion models, compared to the cross-attention approach of Ctrl-World, as indicated by the appearance-centric metrics. Moreover, it also provides more precise action controllability, yielding a 23.6% gain in SCS.

5.4 Extreme Generalization

Figure 6 shows EgoWM generalization results in highly unrealistic domains. In the top part of the figure, we evaluate our 3-DoF Wan variant by conditioning on

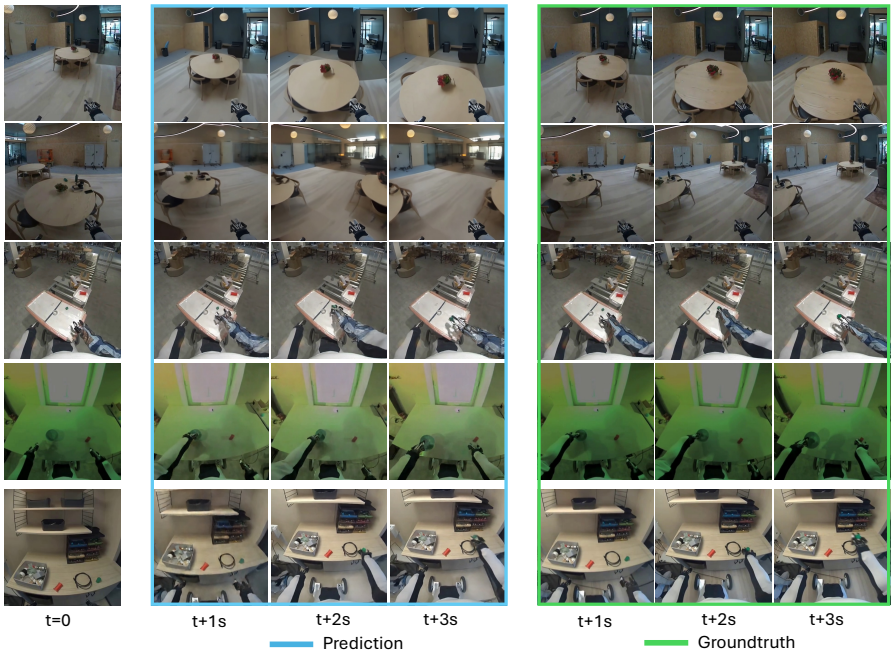

Fig. 5: Our action-conditioned video world model enables a humanoid agent to navigate and execute long-horizon, contact-rich manipulation skills by conditioning on desired actions. Across diverse scenes and object geometries, the model produces physically consistent body motions, stable grasps, and smooth end-effector trajectories that closely follow the provided action sequence. Results best viewed on the project page.

the same initial painting and generating distinct, unseen navigation trajectories from it, highlighting its ability to produce different motion-consistent futures in OOD scenarios. We further show similar generalization in 25-DoF manipulation, where the Wan variant rolls out an action sequence from the 1X validation set to simulate a humanoid looking down, reaching with the right arm, and picking up an apple inside a painting. More results are available on the project page.

6 Planning with EgoWM

In this section we evaluate whether EgoWM can support downstream decision-making with its simulated futures, allowing model-based action selection.

Experiment Setup. We use goal-conditioned planning with the Cross-Entropy Method (CEM), following NWM [7]. Candidate action sequences are sampled, rolled out with our world model, and selected based on perceptual similarity between their predicted final frames and a goal image.

We evaluate this setup in two settings: 3-DoF navigation on RECON and 25-DoF humanoid manipulation on held-out 1X validation tasks. For navigation, we

Table 3: Ablation studies at the 16-frame horizon. We first validate the accuracy of our action-injection mechanism on humanoid navigation, and then report out-of-distribution evaluation on RECON to assess generalization. Our design achieves both more accurate controllability and stronger generalization compared to the alternatives.

Dataset / Task	Variant	LPIPS↓	DreamSim↓	SCS↑
Complex action space controllability at $k = 4$				
Humanoid Navigation	Cosmos-2B (Ours)	**0.40**	0.19	**27.0**
	Cosmos-2B (Global)	0.46	0.20	14.9
	Cosmos-2B (Chunked)	0.42	**0.17**	23.7
Out-of-distribution generalization				
RECON (OOD)	SVD (Ours)	**0.59**	**0.43**	**29.7**
	SVD (X-attn)	0.64	0.46	22.7

Table 4: 3-DoF navigation planning evaluation on RECON over 16 frame trajectories (4 seconds). Both EgoWM variants achieve lower average trajectory error (ATE).

Method	ATE ↓
NWM	3.25
EgoWM (SVD)	3.18
EgoWM (Cosmos)	**3.13**

Table 5: 25-DoF manipulation results on 1X validation tasks. We report success rate, with planned trajectories selected by EgoWM and GT demonstrations evaluated in EgoWM.

Task	SR (Planned) ↑	SR (GT) ↑
P&P towel in basket	100.0%	100.0%
P&P block on board	80.0%	94.0%
Pick up obj from pile	60.0%	75.0%
Pick up cup	66.7%	100.0%

follow the NWM protocol and use the same trajectory prior: candidate trajectories are straight-line motions with equidistant intermediate steps. We sample 120 trajectories radiating from the robot's current pose, roll each candidate out in EgoWM, compare the predicted final frame to the goal image, and average the top-scoring actions to obtain the final planned trajectory. It is then evalauted against the ground-truth trajectory using Average Trajectory Error (ATE) [7].

For manipulation, as there are many possible way to sucesfully complete a task, we measure the success rate of the generated trajectories (SR) instead of ATE. A simulator is not available for 1X, so we use human assessment to compute SR, following Ctrl-World [18]. To sample canidate trajectories we split episodes into validation and test subsets and use the val set to estimate CEM sampling parameters. We then apply the inference procedure described above on the test set to obtain the final planned trajectories. Only held-out tasks that were not used to train EgoWM are included in the evalaution to test generalziation.

Results. In 3-DoF navigation, as shown in Table 4, both EgoWM variants outperform NWM. The gains are moderate because this short-horizon setting is strongly constrained by the straight-line trajectory prior and dense coverage of sampled directions. Still, the improvement indicates that EgoWM predictions can effectively support navigation planning.

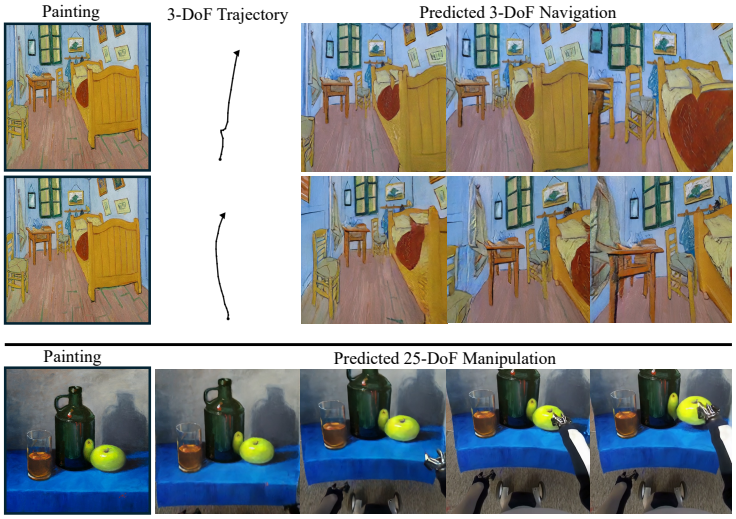

Fig. 6: EgoWM generalizes to paintings, including navigating inside a room and picking up an apple from a table. Despite the drastic shift in appearance, EgoWM preserves coherent motion, effectively "walking into" and "interacting with" painted worlds by leveraging Internet-scale pre-training. Results best viewed on the project page.

In 25-DoF humanoid manipulation, we report results with the Cosmos backbone in Table 5 (middle) and observe that CEM-selected trajectories achieve high success across several tasks, despite the higher-dimensional action space. We further report the success rate of ground-truht demonstrations rolled out in EgoWM in the third column. For most tasks the target success rate of 100% is recovered, demonstrating the potential of our model for policy evalaution.

7 Conclusion

We presented a simple and general framework for transforming pre-trained video diffusion models into world models. By introducing lightweight action conditioning into existing architectures, our method leverages the priors learned from Internet-scale data to produce physically and semantically consistent futures. Through extensive experiments, we demonstrated strong generalization across diverse embodiments and environments, including humanoids, and effectivness in downstream decision making. Finally, we proposed the Structural Consistency Score (SCS) to faithfully evaluate world models' action following. These contributions open a path toward scalable, controllable, and generalizable world models, bridging the gap between passive and active visual prediction.

Acknowledgment

This project was supported by Toyota Research Institute.

References

1. 1X Technologies: 1X World Model Challenge (jun 2024)
2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical AI. arXiv:2501.03575 (2025)
3. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in Atari. NeurIPS (2024)
4. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. ICLR (2018)
5. Bagchi, A., Bao, Z., Wang, Y.X., Tokmakov, P., Hebert, M.: ReferEverything: Towards segmenting everything we can speak of in videos. In: ICCV (2025)
6. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: ReCamMaster: Camera-controlled generative rendering from a single video. In: ICCV (2025)
7. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: CVPR (2025)
8. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv:2311.15127 (2023)
9. ByteDance: Seedance 2.0. `https://seed.bytedance.com/en/seedance2_0` (2026), accessed: 2026-03-02
10. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. In: ICLR (2026)
11. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
12. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR (2024)
13. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70M: Captioning 70M videos with multiple cross-modality teachers. In: CVPR (2024)
14. Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., Levine, S.: Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. arXiv:1812.00568 (2018)
15. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010). `https://doi.org/10.1007/s11263-009-0275-4`
16. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning new dimensions of human visual similarity using synthetic data. In: NeurIPS (2023)
17. Gao, S., Liang, W., Zheng, K., Malik, A., Ye, S., Yu, S., Tseng, W.C., Dong, Y., Mo, K., Lin, C.H., et al.: DreamDojo: A generalist robot world model from large-scale human videos. arXiv:2602.06949 (2026)
18. Guo, Y., Shi, L.X., Chen, J., Finn, C.: Ctrl-World: A controllable generative world model for robot manipulation. In: ICLR (2026)
19. Ha, D., Schmidhuber, J.: World models. In: NeurIPS (2018)
20. Harley, A.W., You, Y., Sun, X., Zheng, Y., Raghuraman, N., Gu, Y., Liang, S., Chu, W.H., Dave, A., Tokmakov, P., You, S., Ambrus, R., Fragkiadaki, K., Guibas, L.J.: AllTracker: Efficient dense point tracking at high resolution. In: ICCV (2025)

21. He, H., Patrikar, J., Kim, D.K., Smith, M., McGann, D., akbar Agha-mohammadi, A., Omidshafiei, S., Scherer, S.: Grndctrl: Grounding world models via self-supervised reward alignment. arXiv:2512.01952 (2025)
22. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: ImaGen Video: High definition video generation with diffusion models. arXiv:2210.02303 (2022)
23. Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. In: ICML (2025)
24. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. RAL (2022)
25. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
26. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: DROID: A large-scale in-the-wild robot manipulation dataset. In: RSS Wrokshops (2024)
27. Kim, S.W., Zhou, Y., Philion, J., Torralba, A., Fidler, S.: Learning to simulate dynamic environments with GameGAN. In: CVPR (2020)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
29. Liang, J., Tokmakov, P., Liu, R., Sudhakar, S., Shah, P., Ambrus, R., Vondrick, C.: Video generators are robot policies. arXiv:2508.00795 (2025)
30. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3D object. In: ICCV (2023)
31. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: EvalCrafter: Benchmarking and evaluating large video generation models. In: CVPR (2024)
32. Mannos, J.L., Sakrison, D.J.: The effects of a visual fidelity criterion on the encoding of images. IEEE Trans. Information Theory **20**(4), 525–536 (1974). https://doi.org/10.1109/TIT.1974.1055250
33. McNamee, D., Wolpert, D.M.: Internal models in biological control. Annual review of control, robotics, and autonomous systems **2**(1), 339–364 (2019)
34. Oh, J., Guo, X., Lee, H., Lewis, R.L., Singh, S.: Action-conditional video prediction using deep networks in Atari games. NeurIPS (2015)
35. OpenAI: Sora 2. https://sora.chatgpt.com/ (2025), accessed: 2026-03-02
36. Ozguroglu, E., Liu, R., Surıs, D., Chen, D., Dave, A., Tokmakov, P., Vondrick, C.: pix2gestalt: Amodal segmentation by synthesizing wholes. In: CVPR (2024)
37. Pang, J.C., Tang, N., Li, K., Tang, Y., Cai, X.Q., Zhang, Z.Y., Niu, G., Sugiyama, M., Yu, Y.: Learning view-invariant world models for visual robotic manipulation. In: ICLR (2025)
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
39. Pezzulo, G., Parr, T., Friston, K.: Active inference as a theory of sentient behavior. Biological Psychology **186**, 108741 (2024)
40. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. In: ICLR (2025)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
42. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)

43. Seo, Y., Hafner, D., Liu, H., Liu, F., James, S., Lee, K., Abbeel, P.: Masked world models for visual control. In: CoRL (2023)
44. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. In: CoRL (2021)
45. Triest, S., Sivaprakasam, M., Wang, S.J., Wang, W., Johnson, A.M., Scherer, S.: TartanDrive: A large-scale dataset for learning off-road dynamics models. In: ICRA (2022)
46. Unterthiner, A., van Steenkiste, S., Keysers, D., Kipf, T., D'Amour, A., Sorrenson, P., Bousquet, O.: Towards accurate generative models of video: A new metric and challenges. In: NeurIPS (2018)
47. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. In: ICLR (2025)
48. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV (2024)
49. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv:2503.20314 (2025)
50. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: ModelScope text-to-video technical report. arXiv:2308.06571 (2023)
51. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
52. Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P.S., Long, M.: PredRNN: A recurrent neural network for spatiotemporal predictive learning. PAMI (2022)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
54. Wickrema, C., Leary, S., Sarkar, S., Giglio, M., Bianchi, E., Mace, E., Twardowski, M.: Benchmarking image similarity metrics for novel view synthesis applications. arXiv:2506.12563 (2025)
55. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv:2509.20328 (2025)
56. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
57. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
58. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV (2023)
59. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv:2412.20404 (2024)
60. Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., Gupta, A.: Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In: RSS (2025)
61. Zhu, F., Wu, H., Guo, S., Liu, Y., Cheang, C., Kong, T.: IRASim: A fine-grained world model for robot manipulation. In: ICCV (2025)
62. Zhu, Z., Feng, X., Chen, D., Yuan, J., Qiao, C., Hua, G.: Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In: ECCV (2024)