DVG-WM: Disentangled Video Generation Enables Efficient Embodied World Model for Robotic Manipulation

Ziyu Shan[1], Zhenyu Wu[2], Xiaofeng Wang[3], Zheng Zhu[3], and Ziwei Wang[1]*

[1] Nanyang Technological University, Singapore
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] GigaAI
http://zyshan0929.github.io/DVGWM

Abstract. Video-based embodied world models provide an appealing substrate for robotic manipulation by predicting future states, yet current approaches remain limited by a fundamental entanglement: accurately modeling dynamics typically requires low-level temporal reasoning, while producing high-resolution frames demands expansive visual synthesis according to high-level semantics. This entanglement results in slow inference speed for iterative planning or too coarse predictions to retain contact-rich details. To solve this dilemma, we present Disentangled Video Generation World Model (DVG-WM), an efficient framework that explicitly **decomposes world modeling into dynamics learning and visual synthesis**. Conditioned on an initial observation and a language instruction, our model first generates a plausible sequence of intermediate visual states to preview the physical interaction and refines them to obtain high-fidelity videos. Furthermore, an efficient cascading mechanism is proposed, where DVG-WM leverages flow matching to directly map the dynamics to video latents, and introduces a latent degradation mechanism to enable the capability of regenerating contact-rich details. Experiments on LIBERO and real-world platforms demonstrate improved video quality with up to 3.97 × acceleration, validating that disentangled video generation can be an efficient embodied world model for robotic manipulation.

Keywords: Robotic Manipulation · Embodied World Model · Imitation Learning

1 Introduction

Embodied world models have gained increasing attention in the robotics community for their capability to model and predict the physical dynamics based on an initial observation and a conditioning input [12, 47]. The action-conditioned

* Corresponding author.

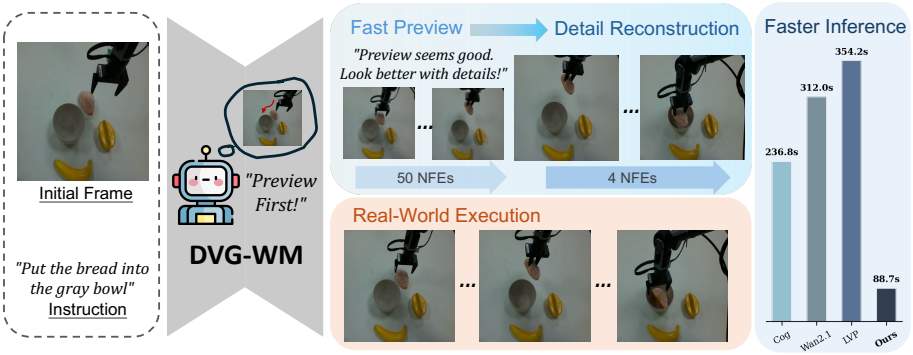

Fig. 1: Illustration of our DVG-WM. Given the initial observation and a language instruction, DVG-WM first generates a sequence of low-resolution imagined future trajectory to preview the physical interaction, then reconstructs the high-fidelity video with faster inference speed (right). In addition to the improved visual quality, DVG-WM also achieves impressive performance on real-world manipulation tasks.

world models support evaluating policies within imagination by simulating fine-grained action outcomes, while another significant role of embodied world models is to support goal-conditioned action planning by forecasting future states towards the given goal images or initial language instructions [9, 23, 51].

The mainstream of embodied world models generally employs video prediction [36, 43, 49] as the paradigm of world modeling. The typical pipeline involves fine-tuning a pretrained video generation backbone with conditioning input (*e.g.*, language, action, and goal images) to produce latent dynamics or pixel-level future predictions [8, 51]. Despite the impressive visual quality and temporal coherence, the generated videos fail to satisfy the requirement of efficient inference for iterative planning in real-world scenarios. This inefficiency of these world models stems from a fundamental limitation: *dynamics modeling is tightly entangled with high-fidelity visual synthesis*. Specifically, accurate dynamics modeling requires low-level temporal reasoning to produce smooth transitions such as contacts, slips, and occlusions, which concentrates on physical interaction patterns and is not necessarily learned in high-resolution settings with redundant textures [28, 40]. Conversely, producing high-resolution frames demands expansive visual synthesis based on high-level semantics and rich texture information of the surroundings [41]. Existing approaches [24, 51, 53] often attempt to solve both requirements within a single monolithic video generator, resulting in slow inference speed for high-resolution observations or coarse predictions to retain contact-rich details. This becomes particularly problematic in real-world iterative planning where the world model must be queried repeatedly to predict the interaction process.

A straightforward solution to this dilemma is to learn the dynamics and visual synthesis separately. For the typical diffusion-based video world models [23, 32, 51], learning dynamics and visual synthesis simultaneously requires a large number of function evaluations (NFEs, *i.e.* denoising steps for diffusion

models [15]) under high-resolution settings. In contrast, learning dynamics in a low-resolution latent space can significantly reduce the computation load for each NFE, allowing for low-cost preview to ensure the physical interaction is reasonable and consistent with the language instruction. Furthermore, the predicted latent dynamics can be refined into high-fidelity videos with fine-grained details with a minimal number of NFEs. This disentangled design allows the world model to efficiently learn the dynamics and visual synthesis separately in their optimal resolution.

Following this philosophy, we propose Disentangled Video Generation World Model (DVG-WM), an efficient video-based world model for robotic manipulation that explicitly decomposes dynamics learning and visual synthesis, as shown in Fig. 1. Conditioned on an initial observation and a language instruction, DVG-WM first generates a sequence of low-resolution intermediate visual states that captures the plausible evolution of the interaction, providing a fast dynamics preview suitable for iterative planning. Subsequently, the refinement stage produces high-fidelity videos with minimal NFEs (*i.e.* 4 steps) to restore high-frequency details and contact-sensitive cues. Furthermore, a novel cascading mechanism is developed to efficiently connect the two stages, where flow matching is leveraged to directly map the low-resolution dynamics to high-resolution video latents, and a latent degradation mechanism is also introduced to enhance the model's capability to regenerate accurate structures of objects and the end-effector at high resolutions, which is crucial for contact-rich manipulation.

We evaluate DVG-WM on LIBERO [26] and real-world platforms, showing that disentangled generation yields high-quality visual predictions with up to 3.97× speed up compared to the baselines. For the real-world tasks, the DVG-WM achieves better performance equipped with an action expert [10,45], demonstrating its capability of serving as an action planner. Furthermore, the ablation studies demonstrate the disentangled design successfully learns the expected patterns separately. These results reveal that disentangled video generation can be a practical foundation for embodied world models. The main contributions of this paper are summarized as follows:

- We propose DVG-WM, a disentangled video generation world model that explicitly decomposes dynamics learning and visual synthesis into a two-stage pipeline, enabling efficient preview and refinement for robotic manipulation.
- We propose a novel cascading mechanism to connect stages. Specifically, we leverage flow matching to directly map the dynamics to video latents, and introduce a latent degradation mechanism to enhance the model's capability to regenerate accurate structures for contact-rich manipulation.
- We evaluate DVG-WM on both simulation and real-world platforms, demonstrating improved visual prediction quality and accuracy with limited inference cost, validating disentangled video generation as an efficient embodied world model for robotic manipulation.

2 Related Works

2.1 Embodied World Models

According to applied condition types, the embodied world models can be categorized into two main types: action-conditioned and goal-conditioned models. The action-conditioned world models typically learn the outcomes of actions by predicting future latent states or videos, improving the policy generalization capability [1, 2, 11, 12, 17, 22, 33, 53]. Simultaneously, many works [9, 18, 28, 29, 46] focus on multi-modal generation conditioned on action, such as 3D point clouds, gaussians, depth, and mask. These action-conditioned world models can be incorporated into robotic manipulation as synthetic data generators that produce diverse successful rollouts for imitation learning and self-improvement [10,21,44,45] in unseen scenarios, and serve as imagination-based evaluators that rank candidate policies [35] or simulators for model-based reinforcement learning that reduce the sample complexity of real-world interaction [20, 39, 40, 42, 54].

Another type of embodied world models is goal-conditioned, which learns to predict the future states towards a goal image or language instruction [5, 50]. These models can be used for planning by forecasting the future states and optionally output the action using an inverse dynamic model [8]. Among these, KeyWorld [23] and LongScape [32] utilize language instruction as the goal and uses the world model as video planner. GE-Act [24] adopts a bi-model architecture with a world model predicting future visual features based on language instruction. WorldVLA [6] aligns vision and action learning within a unified latent space. Act2Goal [51] introduces a visual goal image with multi-scale temporal sampling for closed-loop control. Despite the impressive visual quality of generated videos, the inference efficiency of these world models remains a bottleneck for real-world iterative planning. Our work addresses this issue by disentangling dynamics learning for the physical interactions and visual synthesis, enabling efficient preview and refinement for robotic manipulation.

2.2 Video Diffusion

Diffusion models [10, 15, 25] have emerged as the dominant framework for high-fidelity video synthesis [3, 4, 14]. Video diffusion methods generate a sequence via iterative denoising of spatiotemporal representations, often in a compressed latent space to reduce memory and compute, while introducing explicit temporal modules to encourage coherence across frames. Beyond vanilla full-sequence denoising, recent work improves temporal modeling and long-horizon rollouts by imposing structured uncertainty over time, enabling causal or rolling generation beyond the training horizon [7, 31], and developing diffusion-based pipelines for long-horizon generation and super-resolution [13, 37, 41, 49, 52]. These advances substantially improve temporal consistency and visual fidelity, but generally retain the high inference cost inherent to iterative denoising. This limitation is particularly problematic for embodied manipulation, where iterative planning is needed. Our model addresses this mismatch by allocating computation asymmetrically through learning dynamics and visual synthesis at optimal resolutions.

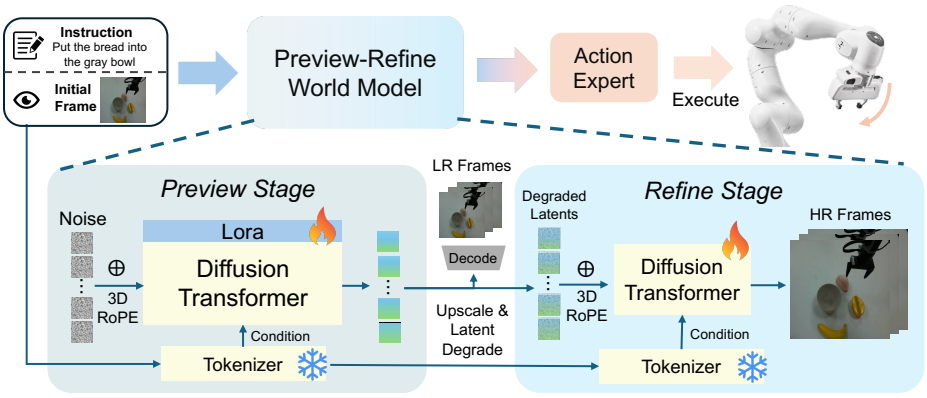

Fig. 2: Pipeline of DVG-WM. The video world model consists of two main stages: 1) a preview stage that learns dynamics and generates low-resolution trajectories conditioned on language instructions and initial frames; 2) a refinement stage that upsamples the latent dynamics to produce high-fidelity video predictions with minimal additional computation. Finally, the inverse dynamics model extracts the action sequence from the predicted video for manipulation tasks.

3 Methodology

3.1 Problem Formulation

Embodied world models aim to predict the dynamics of the environment by forecasting the future observations. Mathematically, given an initial observation image $x_0 \in \mathcal{X}$ and the condition c, the embodied world model learns a probabilistic generative model $p(x_{1:N}|x_0, c)$ to predict the future observations $x_{1:N}$ with a sequence length of N. Depending on the applications, the conditioning variable c may take the form of actions to simulate the action-conditioned outcomes [12,53], or goals (*e.g.*, language [24], goal images [51]) to plan the transition towards the expected future states. In this work, we focus on the latter setting conditioned on language instructions, enabling the world model to function as a language-guided planner for high-level task reasoning and interactive manipulation.

Existing video-based world models typically generate future frames $x_{1:N}$ at the original resolution (*e.g.*, 720p, 1080p), which is computationally intensive due to the entanglement of dynamics modeling and high-fidelity visual synthesis. To enable efficient world modeling, we leverage 3D causal VAE [43] to compress video pixels $x_T \in \mathbb{R}^{H\times W\times T}$ into latent features $z \in \mathbb{Q}^{h\times w\times t}$, where $h = H/8, w = W/8$ and $t = (T-1)/4 + 1$. Our model is designed to predict videos with 49 frames at 720p resolution. As shown in Fig. 2, we then propose a disentangled two-stage pipeline to learn the low-resolution (LR) dynamics and high-resolution (HR) visual synthesis separately, where each stage is optimized with tailored network architectures and training strategies for computational

efficiency. The preview and refinement stages $g_{\mathrm{p}}, g_{\mathrm{r}}$ can be formulated as:

$$g_{\mathrm{p}}(z_{\mathrm{lr}}|x_0^{\mathrm{d}}, c); \quad g_{\mathrm{r}}(z_{\mathrm{hr}}|z_{\mathrm{lr}}, x_0, c) \tag{1}$$

where x_0^{d} is the downsampled initial frame for the preview stage. The low-resolution dynamics can be optionally previewed by the 3D VAE decoder $\mathcal{E}$ as $x_{\mathrm{lr}} = \mathcal{E}(z_{\mathrm{lr}})$, while the high-resolution videos can be similarly obtained. The horizon N of z and x is omitted for brevity. The following sections provide detailed descriptions of each stage.

3.2 Low-Resolution Preview Stage

In the preview stage, the goal is to generate video latents z_{lr} that align with the language instruction and captures the essential dynamics of the interaction. To achieve this, we adopt CogVideoX-5B [43] as the video generation backbone. For computation efficiency, we discard the learnable positional embedding and employ flexible 3D rotary positional embedding [34] to allow for variable output resolutions. All other configurations such as language and image embedding, and denoising scheduler are the same as [43]. Then, we apply LoRA [16] with rank 128 on attention layers, FFN, and layer normalization to adapt the CogVideoX-5B to a lower resolution. Compared to full-parameter tuning, the LoRA-based adaptation significantly reduces the training cost while capturing the essential dynamic patterns, as demonstrated by the ablation studies in Sec. 4.3. In this manner, the preview stage provides a physically meaningful initialization for the subsequent refinement stage, as detailed in the following section.

3.3 High-Resolution Refinement Stage

For the fine-grained visual synthesis, we employ a smaller model CogVideoX-2B due to the observation that embodied manipulation scenes are typically more structured, and their state transitions tend to be slower and more locally constrained compared to in-the-wild videos. As shown in Fig. 2, the refinement stage takes the low-resolution dynamics z_{lr} as input and generates high-resolution video latents z_{hr} with minimal NFEs.

Flow Matching for Cascaded Generation. Given the low-resolution dynamics z_{lr} as the input, the intuitive solution to generate high-resolution videos is to apply the diffusion process on another sequence of random noise conditioned on the dynamics. However, this approach consumes

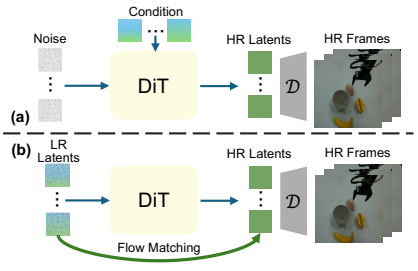

Fig. 3: Comparison of refinement approaches. (a) Conventional DDPM treating low-resolution latents as conditioning input. (b) DVG-WM uses flow matching to directly map the low-resolution dynamics to high-resolution video latents.

significant computation loads since the denoising process needs to learn to generate the entire video from scratch, which is redundant given that the preview stage already provides a reasonable initialization.

To solve this issue, we leverage flow matching [25, 27] to directly establish the mapping from the low-resolution dynamics to the high-resolution video in the latent space. In our scenario, the flow matching learns the vector field $v_\theta(z_\tau, \tau \mid z_{\mathrm{lr}}, x_0, c)$ that specifies a step-by-step update direction in latent space, where z_τ denotes an intermediate latent state at interpolation step $\tau \in [0, 1]$, formulated as:

$$z_\tau = (1-\tau)\,\tilde{z}_{\mathrm{lr}} + \tau\, z_{\mathrm{hr}}, \qquad \tau \sim \mathcal{U}(0,1). \tag{2}$$

where $\tilde{z}_{\mathrm{lr}}$ is the linearly upsampled version of z_{lr} to ensure identical sequence length. In this manner, a straight-line probability path (*i.e.*, an ODE trajectory for diffusion) is established, starting from the upscaled initialized latent $\tilde{z}_{\mathrm{lr}}$ and progressively following these directions to refine it into the final high-resolution latent z_{hr}.

The path in Eq. (2) linearly interpolates between the coarse initialization $\tilde{z}_{\mathrm{lr}}$ (at $\tau = 0$) and the ground-truth latent z_{hr} (at $\tau = 1$). As τ increases, the latent should move from $\tilde{z}_{\mathrm{lr}}$ toward z_{hr} along this straight line. Therefore, the desired update direction $u^\star(z_\tau, \tau)$ (*i.e.*, velocity) at any intermediate point z_τ is simply the displacement from the start to the target:

$$u^\star(z_\tau, \tau) \triangleq \frac{dz_\tau}{d\tau} = z_{\mathrm{hr}} - \tilde{z}_{\mathrm{lr}} \tag{3}$$

We train a neural vector field v_θ to predict this velocity, and the flow-matching objective can be formulated as:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\big\| v_\theta(z_\tau, \tau \mid z_{\mathrm{lr}}, x_0, c) - u^\star(z_\tau, \tau)\big\|_2^2\Big], \tag{4}$$

where the expectation $\mathbb{E}$ is computed over the training samples $(\tilde{z}_{\mathrm{lr}}, z_{\mathrm{hr}}, z_\tau, c)$, and $\tau \sim \mathcal{U}(0,1)$.

At inference, refinement starts from the initialized latent $z^0 = \tilde{z}_{\mathrm{lr}}$ and applies limited NFEs through a deterministic flow process:

$$z^{k+1} = z^k + \Delta\tau\; v_\theta\big(z^k, \tau_k \mid z_{\mathrm{lr}}, x_0, c\big), \qquad k = 0, \ldots, K-1, \tag{5}$$

where τ_k denotes the discretized interpolation step and K is typically 4 in our experiments, and the final output is $z_{\mathrm{hr}} = z^K$ that can be decoded into high-resolution videos by the 3D VAE decoder $\mathcal{D}$.

Latent Degradation for Contact-Rich Manipulation. To train the refinement stage, the straightforward approach is to degrade the high-resolution video within the pixel level and encode it to form z_{lr} so that the two stages can be trained independently, which is commonly used in video super-resolution [41,52]. However, this pixel-level degradation retains tight correlation between low- and high-resolution videos, constraining the world model from regenerating detailed structures for the interactions between the end-effector and objects. To address

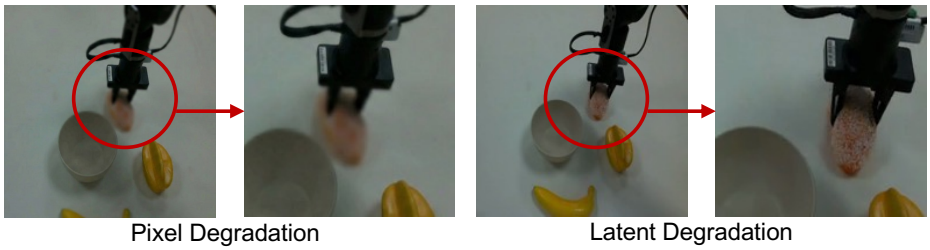

Fig. 4: Comparison of pixel (left) and latent degradation (right) for training the refinement stage. Latent degradation generates contact-rich details in zoomed-in view.

this issue, we use the latent degradation to perturb the upscaled output of the preview stage $\tilde{z}_{\text{lr}}$, allowing for **reconstructing fine-grained contact details** rather than upscaling pixels. More details can be found in the supplementary materials. As shown in Fig. 4 of video prediction for a real-world task, the latent degradation successfully generates contact-rich details in the zoomed-in view, preserving the structures of the gripper and the target object, while the pixel degradation presents blur around the interaction region.

3.4 Action Generation for Imitation Learning

Given the high-resolution video prediction x_{hr}, we equip DVG-WM with an action expert to convert imagined futures into executable actions. We instantiate the action expert as a vision-only Diffusion Policy (DP) [10], since robot states (*e.g.*, joint angle) are unavailable for video prediction. The predicted frames of different timestamps are separately encoded as the condition h_t of DP, which denoises a noise sequence into actions through the standard diffusion objective:

$$\mathcal{L}_{\text{DP}} = \mathbb{E}\Big[\big\|\epsilon - \epsilon_\phi(a_\sigma, \sigma \mid h_t)\big\|_2^2\Big], \tag{6}$$

where a_σ is the noised action sequence at diffusion level σ and $\epsilon \sim \mathcal{N}(0, I)$. Note that the action expert can also use diffusion-based models [24, 51] and vision-language-action (VLA) models like NORA [19] that do not rely on robot state, which is unavailable for predicted future frames.

Training Strategy. We train the full system in two stages for stability. We first train the world model alone using the objectives in Eq. (4), obtaining a reliable video predictor. Afterwards, we attach the action expert and optimize it on demonstrations while jointly fine-tuning the world model using only $\mathcal{L}_{\text{DP}}$. The gradients from the imitation learning loss propagate through both the action generation expert and the world model, aligning the visual prediction with downstream action planning.

Table 1: Quantitative comparison on video quality and object-level accuracy.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	Object ↑
CogVideoX-5B	19.286	0.761	0.138	171.24	76%
Wan2.1-14B	18.964	0.732	0.162	198.54	68%
LongScape	19.977	**0.788**	0.123	153.72	-
LVP-14B	19.582	0.765	0.134	187.69	80%
DVG-WM (ours)	**20.019**	0.783	**0.120**	**152.36**	**89%**

4 Experiments

4.1 Experimental Settings

Dataset Preparations. The proposed DVG-WM is trained and evaluated on the simulation LIBERO dataset [26] and a self-collected real-world dataset containing 7K trajectories, aligning with the real-world platform setup in Sec. 4.4. We evaluate the DVG-WM for video quality on LIBERO, while the real-world dataset is used to evaluate the performance of DVG-WM as an action planner. For comprehensive evaluation with varied instructions, we split the official demonstrations according to the tasks with a ratio of 8:2 for training and testing. Additionally, we preprocess the dataset by replaying trajectories, removing dummy episodes and resizing, following [6, 23], obtaining around 5K trajectories in total. As for the real-world dataset, the trajectories with less than 49 frames are discarded. More details about the dataset are in supplementary materials.

Implementation Details. All videos are resized to 256×384 for the preview stage and 480×720 for refinement. For the preview stage, the model is trained using LoRA [16] for 10,000 iterations with a batch size of 4 on a single A100-80G GPU. AdamW optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.95$, a weight decay of $1e-4$, gradient clipping setting to 0.1. For the refinement stage, the model is trained for 10 epochs with a batch size of 6 on LIBERO, taking around 24 hours on 8 A100-80G GPUs. The training hyper-parameters are kept the same as [43]. More training details are in Sec. 4.4 and supplementary materials.

Baselines. We compare our approach with advanced video generation world models, including CogVideoX-5B [43], Wan2.1-14B I2V [36], LongScape [32], and large video planner (LVP) [8]. These models are fine-tuned on the same training set. Note that Wan2.1 and LVP are based on different resolutions, so an additional training set with the same trajectories is prepared.

Evaluation Metrics. For video quality evaluation, we use the standard metrics including PSNR, SSIM [38], LPIPS [48] and FVD. We also use a metric of object-level accuracy following [23] where 10% of trajectories are manually checked to report the ratio of the robot operating on the proper object. For the real-world tasks in Sec. 4.4, we additionally evaluate two metrics to test the world model's understanding of the physical interaction process: 1) Mask-IoU between embodiment regions [9] in generated and ground-truth frames, obtained via Grounded

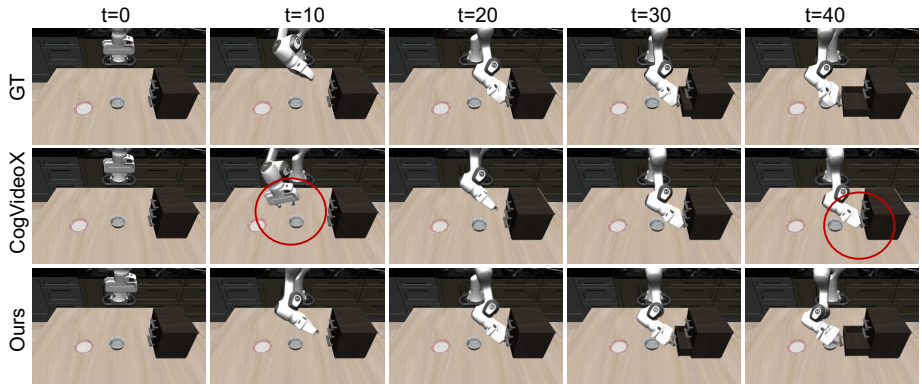

Fig. 5: Qualitative comparison of video predictions from DVG-WM and CogVideoX on LIBERO tasks. Red circles highlight wrong critical interaction details of CogVideoX.

SAM [30] with "robotic arm" as the textual prompt. 2) Success rates of the executed actions planned by DVG-WM equipped with an action expert on nine real-world tasks across two platforms.

4.2 Main Results

Quantitative Comparison. In Tab. 1, our DVG-WM delivers stronger video prediction on LIBERO. Our method achieves superior visual quality metrics with an average improvement of 3.7%, while showcasing impressive efficiency. Although SSIM is slightly lower than LongScape, it remains competitive, which reveals that the refinement stage improves high-frequency details without introducing severe structural distortion. Note that the original metrics of LongScape are reported and object-level accuracy is inaccessible. Beyond low-level similarity, DVG-WM substantially improves object-level accuracy to 89%, showing that the model more reliably grounds the interaction on the correct target objects, which is critical for manipulation-centric tasks. Achieving both high fidelity and reliable instruction following underscores the benefit of our two-stage design, where the preview supplies contact-rich cues that steer refinement towards physically meaningful, high-quality generation.

Qualitative Comparison. As shown in Fig. 5, CogVideoX is compared with our method since they share the same resolution. Obviously, our DVG-WM generates high-quality videos, and forecasts the complete trajectory for the open-drawer task, compared to CogVideoX. Despite the satisfactory visual quality, CogVideoX fails to accurately follow the instruction, where the robot arm shows the intention to interact with the wrong object (the bowl) instead of the drawer at $t = 10$. Additionally, the final interaction stage of pulling the drawer is missing at $t = 40$ in CogVideoX, which is critical for the task completion. In contrast, DVG-WM successfully captures the contact-rich details of the interaction, such as the gripper's contact with the drawer and the subsequent movement of the

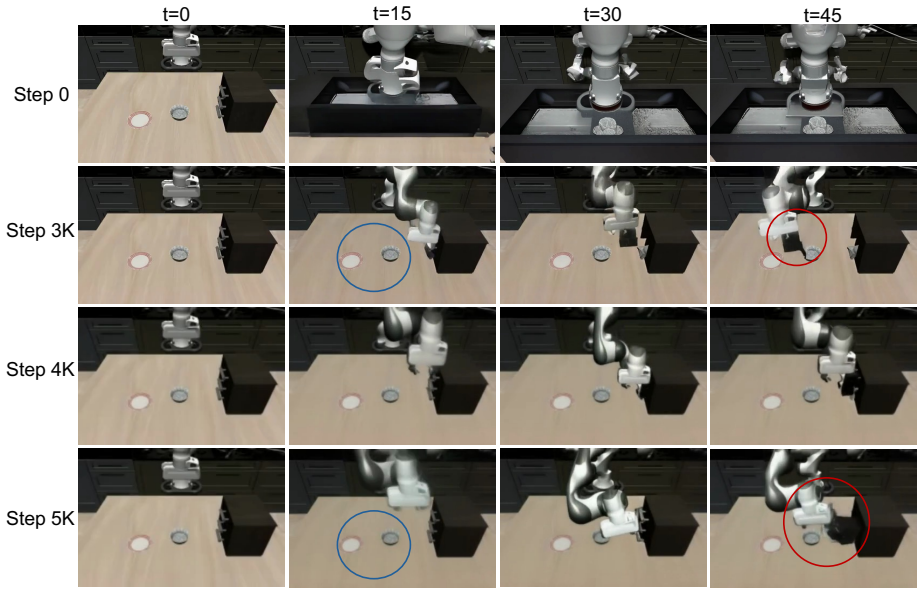

Fig. 6: Comparison with generated videos of preview stage with different LoRA fine-tuning steps. Blue circles highlight the texture of the bowls, while the red circles highlight the contact details between the gripper and the drawer.

drawer, demonstrating its superior capability to model the physical dynamics and follow instructions accurately.

Efficiency Comparison. Our DVG-WM achieves a significant speedup of up to 3.97× compared to LVP, since only 4 denoising steps is needed for refinement and the 50 denoising steps are conducted at the much lower resolution, which is crucial for real-world iterative planning. The efficiency can be attributed to the disentangled design that learns dynamics and visual synthesis separately in their optimal resolution settings.

Table 2: Comparison of inference time for the video world models.

Method	Inference time
CogVideoX-5B	236.8s
Wan2.1-14B	312.0s
LVP-14B	354.2s
DVG-WM (ours)	**88.7s**

4.3 Ablation Studies

Diving into the Disentangled Design. To further validate the effectiveness of the disentangled model design, we conduct a qualitative comparison to analyze the effect of the LoRA fine-tuning for the preview stage in Fig. 6. For step 0, as well as the zero-shot inference for CogVideoX-5B, the layouts of the scene are totally changed and the robot arm is missing, even though the texture of the scene is fine-grained. This is expected since the original model is pretrained on in-the-wild videos, not including embodied data.

Table 3: Component analysis of our DVG-WM on LIBERO.

#Stages	Method	PSNR ↑	SSIM ↑	LPIPS↓	FVD ↓
Two	DVG-WM (Ours)	**20.019**	**0.783**	**0.120**	**152.36**
	Pixel Degradation	19.421	0.768	0.128	168.54
	Naive Cascading	19.580	0.765	0.134	187.69
	360p Preview + Refine	18.812	0.756	0.146	188.47
One	Only Preview	14.714	0.721	0.341	238.72
	Only Refine	19.277	0.760	0.135	169.29

As the LoRA iteration step increases, the preview stage fails to restore the original textures of the scene, including the objects and robot arms, as highlighted in blue circles. This blur is even be aggravated as the iteration step increases. This loss of texture synthesis is expected since the preview stage is optimized for dynamics modeling rather than visual synthesis. However, the preview stage successfully captures the contact-rich details of the interaction, such as the gripper's contact with the drawer and the subsequent movement of the drawer, as highlighted in red circles. For step 3K, the drawer is unreasonably pulled out, while the model at step 5K successfully models this interaction, revealing the understanding of the instruction and interaction process. This observation validates that the preview stage learns to capture the essential dynamic patterns for physical interactions, which provides a meaningful initialization for the refinement stage.

Meanwhile, the CogVideoX that is used for the refinement stage successfully restores the texture details of the scene, as in the second row of Fig. 5, but fails to plan the correct trajectory. Combined with the dynamics preview, the full model successfully generates high-fidelity videos with accurate interaction dynamics, demonstrating the effectiveness of the disentangled design, as in the last row of Fig. 5.

Component Analysis. We conduct component ablations on LIBERO to validate the design choices of DVG-WM in Tab. 3. We have multiple observations: 1) Replacing latent degradation with pixel-level degradation consistently degrades quality (SSIM 0.783→0.768), supporting our intuition that pixel-space pairing causes tight correlation between the source and target, leading to shortcut upscaling. Conversely, latent degradation reconstructs contact-relevant structures with much higher vi-

Fig. 7: Real-world setup for Galaxea A1 Arm Platform.

sual and structural quality. 2) Naive cascading that treats the low-resolution dynamics as conditioning input for the refinement stage (in Fig. 3 (a)) leads to a performance drop, validateing the effectiveness of flow matching in directly mapping the dynamics to high-resolution video latents, rather than generating from random noise. 3) The preview and refinement stages are both necessary. Removing the preview stage reduces all metrics, confirming that the low-resolution latents provide informative cues that guide refinement. Similarly, removing the refinement stage leads to a severe collapse in perceptual quality, showing that dynamics-only low-resolution predictions are insufficient to preserve fine-grained details. 4) Lowering the preview resolution to 360×240 further worsens the performance, which may be attributed to the failure to leverage the pretraining knowledge of CogVideoX that is large-scale pretrained on 256×384 videos.

Number of Refinement Steps. We further study the effect of the number of refinement steps K in the refinement stage on LIBERO, as reported in Tab. 4. Increasing K from 1 to 4 consistently improves all metrics, since additional denoising steps progressively recover fine-grained textures from the low-resolution dynamics. However, doubling K from 4 to 8 yields only marginal gains (*e.g.*, FVD 152.36→150.42) while nearly doubling the refinement computation. We therefore adopt K=4 by default, striking a favorable balance between visual fidelity and inference efficiency.

Table 4: Ablation over the number of refinement steps K on LIBERO.

K	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
1	16.514	0.681	0.218	245.83
2	18.702	0.745	0.165	198.41
4	**20.019**	**0.783**	**0.120**	**152.36**
8	20.124	0.785	0.118	150.42

4.4 Real-World Experiments

Platform Setup. We conduct real-world experiments on two robot platforms, a Galaxea A1 robotic arm and a UR7e robotic arm, each equipped with a parallel gripper. The observation is provided by a RealSense L515 RGBD camera, as shown in Fig. 7. The layouts are aligned with the real-world dataset used for training the world model.

Real-World Tasks and Results. We evaluate DVG-WM on nine real-world manipulation tasks spanning two robot platforms (Galaxea A1 and UR7e): move banana, pick-and-place bread, sweep rubbish, stack cube, adapter insertion, push cube, open drawer, assemble gears, and make a toast. For each task, around 50 demonstrations are prepared for imitation learning in Eq. (6). To evaluate the models, we compare the video quality sharing the protocol in Tab. 1, the mask-IoU for the embodiment regions between the generated and ground-truth videos, and the actual success rates for the tasks. The dataset is resized to 480×720 to train the CogVideoX and our DVG-WM, along with the action expert of DP, whose input horizon is set to 1 to adapt to the iterative planning. We roll out the learned policy 20 times for each task under four distinct settings with various tabletop layouts, instance locations, and task horizons, yielding $20 \times 4 = 80$ rollouts per task, and report the average success rate with standard deviation.

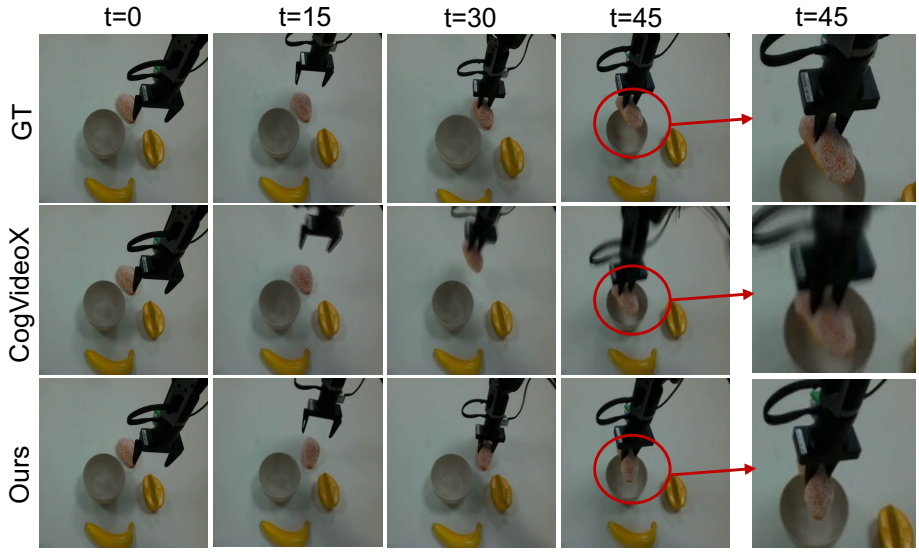

Fig. 8: Qualitative results on the real-world manipulation task of placing bread into bowl. DVG-WM generates high-fidelity video predictions that accurately capture the interaction details and follow language instructions for pick-and-place.

Table 5: Quantitative comparison on video quality for real-world scenarios.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	Mask-IoU ↑
CogVideoX-5B	18.642	0.745	0.151	189.32	0.54
DVG-WM (ours)	**19.287**	**0.771**	**0.128**	**164.78**	**0.57**

As shown in Tab. 5, DVG-WM outperforms CogVideoX-5B across all video quality metrics, demonstrating its superior capability to generate high-fidelity videos that accurately capture the dynamics of real-world interactions. The improvement in Mask-IoU indicates that DVG-WM better perceives the embodiment regions, which is crucial for understanding the physical interactions between the robot and objects.

For visualization, DVG-WM successfully predicts realistic future interactions on all two tasks in Fig. 8, demonstrating its capability to generalize from simulation to real-world scenarios. The predicted videos of CogVideoX present slight ghosting artifacts and motion blur, as revealed in [32]. Conversely, our DVG-WM exhibits fine-grained contact details between the gripper and objects and proper task completion sequences, even the contact area and texture of the bread preserved, demonstrating its capability to be utilized as a real-world video planner.

To verify if the DVG-WM can be used for practical manipulation tasks, we equipped it with a vision-only DP as the action expert and jointly train

Table 6: Success rates (%) on nine real-world tasks across two platforms (Galaxea A1 & UR7e), each evaluated with 20×4=80 rollouts. **Avg.** is the mean over all nine tasks.

Method	move banana	pick&place bread	sweep rubbish	stack cube	adapter insertion
CogVideoX	48.4±11.6	43.2±6.8	38.6±11.4	26.4±8.6	18.2±6.8
DVG-WM (ours)	**73.2±12.8**	**67.8±7.2**	**72.6±7.6**	**57.8±12.2**	**52.4±7.6**

Method	push cube	open drawer	assemble gears	make a toast	**Avg.**
CogVideoX	52.6±7.4	41.4±8.6	22.6±7.4	14.4±5.6	34.0
DVG-WM (ours)	**78.4±6.6**	**68.2±11.8**	**51.4±8.6**	**38.2±11.8**	**62.2**

them through imitation learning based on the provided demonstrations of the dataset. The same protocol is applied to every task, and the average success rates with standard deviation are reported in Tab. 6. We observe that DVG-WM achieves an average success rate of 62.2% across the nine tasks, outperforming the CogVideoX world model (34.0%) by 28.2% on average. The margin is most pronounced on contact-rich and long-horizon tasks such as adapter insertion, assemble gears, and make a toast, where accurate dynamics prediction is critical for completion. These consistent gains across diverse tasks and two platforms validate its effectiveness as an embodied world model for real-world robotic manipulation. The demos and more details of the real-world experiments are in supplementary materials.

5 Conclusion

In this paper, we identify the bottleneck of inefficient video-based embodied world models: modeling dynamics and synthesizing high-resolution observations are tightly entangled. We then present DVG-WM, a disentangled video generation world model that addresses this limitation by decomposing world modeling into a preview stage that captures dynamics and a high-resolution refinement stage restoring fine-grained details with minimal computation. To efficiently connect the two stages, we use flow matching to directly map the dynamics to video latents, together with a latent degradation mechanism that encourages reconstructing contact-rich structures rather than performing shortcut upscaling. Experiments on LIBERO and real-world platforms demonstrate that DVG-WM improves video prediction quality and instruction-following accuracy while reducing inference time with up to 3.97× acceleration. DVG-WM further achieves higher real-world success rates with 28.2% improvement. We believe disentangled video generation provides a promising direction for embodied world models.

Acknowledgments

This work was supported by MoE AcRF Tier 2 under Grant MOE-T2EP50125-0004, and Singapore National Robotics Programme Research Project DS-RFM M25N4N2009, DEM M25N4N2150, MoE Tier 1 Seed Project RS17/24.

References

1. 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 preprint arXiv:2501.03575 (2025) 4
2. Bi, H., Tan, H., Xie, S., Wang, Z., Huang, S., Liu, H., Zhao, R., Feng, Y., Xiang, C., Rong, Y., et al.: Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030 (2025) 4
3. 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 preprint arXiv:2311.15127 (2023) 4
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024) 4
5. Cen, J., Huang, S., Yuan, Y., Li, K., Yuan, H., Yu, C., Jiang, Y., Guo, J., Li, X., Luo, H., et al.: Rynnvla-002: A unified vision-language-action and world model. arXiv preprint arXiv:2511.17502 (2025) 4
6. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025) 4, 9
7. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024) 4
8. Chen, B., Zhang, T., Geng, H., Song, K., Zhang, C., Li, P., Freeman, W.T., Malik, J., Abbeel, P., Tedrake, R., et al.: Large video planner enables generalizable robot control. arXiv preprint arXiv:2512.15840 (2025) 2, 4, 9
9. Chen, Y., Li, P., Yang, J., He, K., Wu, X., Xu, Y., Wang, K., Liu, J., Liu, N., Huang, Y., et al.: Bridgev2w: Bridging video generation models to embodied world models via embodiment masks. arXiv preprint arXiv:2602.03793 (2026) 2, 4, 9
10. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025) 3, 4, 8
11. Chi, X., Jia, P., Fan, C.K., Ju, X., Mi, W., Zhang, K., Qin, Z., Tian, W., Ge, K., Li, H., et al.: Wow: Towards a world omniscient world model through embodied interaction. arXiv preprint arXiv:2509.22642 (2025) 4
12. Guo, Y., Shi, L.X., Chen, J., Finn, C.: Ctrl-world: A controllable generative world model for robot manipulation. arXiv preprint arXiv:2510.10125 (2025) 1, 4, 5
13. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: Venhancer: Generative space-time enhancement for video generation. arXiv preprint arXiv:2407.07667 (2024) 4
14. 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 preprint arXiv:2210.02303 (2022) 4
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3, 4
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 6, 9

17. Huang, S., Chen, L., Zhou, P., Chen, S., Jiang, Z., Hu, Y., Liao, Y., Gao, P., Li, H., Yao, M., et al.: Enerverse: Envisioning embodied future space for robotics manipulation. arXiv preprint arXiv:2501.01895 (2025) 4
18. Huang, W., Chao, Y.W., Mousavian, A., Liu, M.Y., Fox, D., Mo, K., Fei-Fei, L.: Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. arXiv preprint arXiv:2601.03782 (2026) 4
19. Hung, C.Y., Sun, Q., Hong, P., Zadeh, A., Li, C., Tan, U., Majumder, N., Poria, S., et al.: Nora: A small open-sourced generalist vision language action model for embodied tasks. arXiv preprint arXiv:2504.19854 (2025) 8
20. Jiang, Z., Liu, K., Qin, Y., Tian, S., Zheng, Y., Zhou, M., Yu, C., Li, H., Zhao, D.: World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. arXiv preprint arXiv:2509.19080 (2025) 4
21. Kim, M.J., Gao, Y., Lin, T.Y., Lin, Y.C., Ge, Y., Lam, G., Liang, P., Song, S., Liu, M.Y., Finn, C., et al.: Cosmos policy: Fine-tuning video models for visuomotor control and planning. arXiv preprint arXiv:2601.16163 (2026) 4
22. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., et al.: Causal world modeling for robot control. arXiv preprint arXiv:2601.21998 (2026) 4
23. Li, S., Hao, Q., Shang, Y., Li, Y.: Keyworld: Key frame reasoning enables effective and efficient world models. arXiv preprint arXiv:2509.21027 (2025) 2, 4, 9
24. Liao, Y., Zhou, P., Huang, S., Yang, D., Chen, S., Jiang, Y., Hu, Y., Cai, J., Liu, S., Luo, J., et al.: Genie envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint arXiv:2508.05635 (2025) 2, 4, 5, 8
25. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 4, 7
26. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023) 3, 9
27. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) 7
28. Lu, G., Jia, B., Li, P., Chen, Y., Wang, Z., Tang, Y., Huang, S.: Gwm: Towards scalable gaussian world models for robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9263–9274 (2025) 2, 4
29. Qian, Z., Chi, X., Li, Y., Wang, S., Qin, Z., Ju, X., Han, S., Zhang, S.: Wristworld: Generating wrist-views via 4d world models for robotic manipulation. arXiv preprint arXiv:2510.07313 (2025) 4
30. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) 10
31. Ruhe, D., Heek, J., Salimans, T., Hoogeboom, E.: Rolling diffusion models. arXiv preprint arXiv:2402.09470 (2024) 4
32. Shang, Y., Jin, L., Ma, Y., Zhang, X., Gao, C., Wu, W., Li, Y.: Longscape: Advancing long-horizon embodied world models with context-aware moe. arXiv preprint arXiv:2509.21790 (2025) 2, 4, 9, 14
33. Shen, Y., Wei, F., Du, Z., Liang, Y., Lu, Y., Yang, J., Zheng, N., Guo, B.: Videovla: Video generators can be generalizable robot manipulators. arXiv preprint arXiv:2512.06963 (2025) 4
34. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 6

35. Team, G.R., Choromanski, K., Devin, C., Du, Y., Dwibedi, D., Gao, R., Jindal, A., Kipf, T., Kirmani, S., Leal, I., et al.: Evaluating gemini robotics policies in a veo world simulator. arXiv preprint arXiv:2512.10675 (2025) 4
36. 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 preprint arXiv:2503.20314 (2025) 2, 9
37. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025) 4
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9
39. Wu, J., Yin, S., Feng, N., He, X., Li, D., Hao, J., Long, M.: ivideogpt: Interactive videogpts are scalable world models. Advances in Neural Information Processing Systems **37**, 68082–68119 (2024) 4
40. Xiao, J., Yang, Y., Chang, X., Chen, R., Xiong, F., Xu, M., Zheng, W.S., Zhang, Q.: World-env: Leveraging world model as a virtual environment for vla post-training. arXiv preprint arXiv:2509.24948 (2025) 2, 4
41. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17108–17118 (2025) 2, 4, 7
42. Yang, J., Lin, K., Li, J., Zhang, W., Lin, T., Wu, L., Su, Z., Zhao, H., Zhang, Y.Q., Chen, L., et al.: Rise: Self-improving robot policy with compositional world model. arXiv preprint arXiv:2602.11075 (2026) 4
43. 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. arXiv preprint arXiv:2408.06072 (2024) 2, 5, 6, 9
44. Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y.L., Zhu, C., Xiang, J., et al.: World action models are zero-shot policies. arXiv preprint arXiv:2602.15922 (2026) 4
45. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. arXiv preprint arXiv:2403.03954 (2024) 3, 4
46. Zhang, C., Wu, Z., Lu, G., Tang, Y., Wang, Z.: imowm: Taming interactive multimodal world model for robotic manipulation. arXiv preprint arXiv:2510.09036 (2025) 4
47. Zhang, P.F., Cheng, Y., Sun, X., Wang, S., Li, F., Zhu, L., Shen, H.T.: A step toward world models: A survey on robotic manipulation. arXiv preprint arXiv:2511.02097 (2025) 1
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 9
49. Zhang, S., Li, W., Chen, S., Ge, C., Sun, P., Zhang, Y., Jiang, Y., Yuan, Z., Peng, B., Luo, P.: Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. arXiv preprint arXiv:2502.05179 (2025) 2, 4
50. Zhang, W., Hu, T., Zhang, H., Qiao, Y., Qin, Y., Li, Y., Liu, J., Kong, T., Liu, L., Ma, X.: Chain-of-action: Trajectory autoregressive modeling for robotic manipulation. arXiv preprint arXiv:2506.09990 (2025) 4

51. Zhou, P., Chen, L., Chen, S., Chen, D., Zhao, W., Jin, R., Ren, G., Luo, J.: Act2goal: From world model to general goal-conditioned policy. arXiv preprint arXiv:2512.23541 (2025) 2, 4, 5, 8
52. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2535–2545 (2024) 4, 7
53. Zhu, F., Wu, H., Guo, S., Liu, Y., Cheang, C., Kong, T.: Irasim: A fine-grained world model for robot manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9834–9844 (2025) 2, 4, 5
54. Zhu, F., Yan, Z., Hong, Z., Shou, Q., Ma, X., Guo, S.: Wmpo: World model-based policy optimization for vision-language-action models. arXiv preprint arXiv:2511.09515 (2025) 4