VIEWWEAVER: A CONDITION-ENHANCED MULTI-VIEW EMBODIED WORLD MODEL

Anonymous authors
Paper under double-blind review

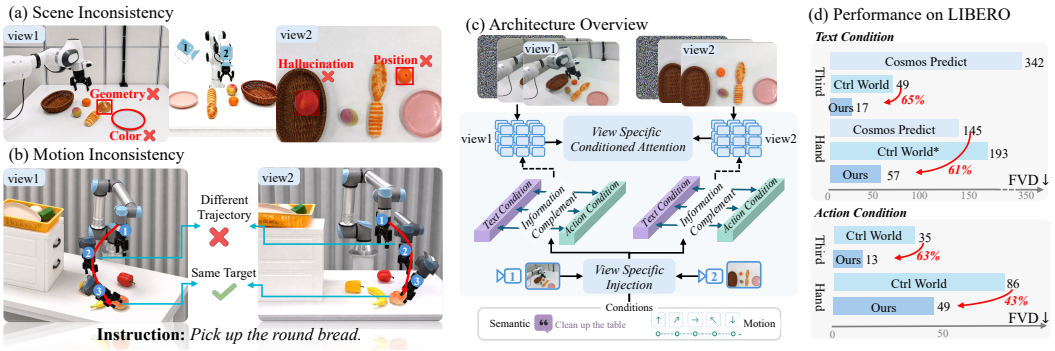

Figure 1: (a) Views may suffer from scene inconsistency, disagreeing in color, geometry, position and presence. (b) Under the same instruction, views may depict different trajectories, leading to motion inconsistency. (c) ViewWeaver supports an arbitrary number of views and processes all views in parallel, incorporates view-specific information into text or action conditions, aligns complementary semantic and motion cues, and applys view-specific conditioned attention to multi-view features. (d) ViewWeaver outperforms representative baselines under both text and action conditions.

ABSTRACT

Video world models have been increasingly adopted for future prediction and world simulation in embodied intelligence. As embodied policies increasingly rely on multi-view observations, generating coherent futures across views becomes essential. However, methods that simply concatenate multi-view inputs or rely on attention to implicitly infer cross-view relationships often produce inconsistencies in scene appearance and robot motion. We propose **ViewWeaver**, a condition-enhanced multi-view world model comprising two key components. First, View-aware Semantic-Motion Conditioning (VSMC) incorporates each view's initial observation and camera information into conditions. It further aligns text and action conditions in a shared latent space to encourage complementarity between task semantics and fine-grained motion cues. Second, Condition-Modulated Cross-View Attention (CM-CVA) uses the view-specific conditions to modulate each view's features before selectively aggregating complementary information from the other views. We evaluate ViewWeaver on the LIBERO-90 dataset and the AgiBot-World dataset under both text- and action-conditioned settings. ViewWeaver achieves the strongest overall performance across views. In particular, it reduces third-view FVD from 48.96 to 16.60 on text-conditioned LIBERO-90, a relative reduction of 66.1%, and reduces left- and right-view FVD by 19.2% and 20.6% on action-conditioned AgiBot-World. Extensive ablation studies and qualitative comparisons further demonstrate that ViewWeaver improves cross-view consistency in both scene appearance and robot motion. Code is provided in the supplementary material.

1 INTRODUCTION

Video world models Mei et al. (2026) can generate high-fidelity, physically plausible visual predictions. This capability has enabled their widespread application in embodied intelligence, including text-conditioned future prediction for data augmentation Xie et al. (2026); Liu et al. (2025), visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

planning Du et al. (2023); Hu et al. (2024), and action-conditioned world simulation for policy evaluation Tseng et al. (2025) and reinforcement learning Zhu et al. (2025); Liu et al. (2026). Modern embodied policies Kim et al. (2024); Intelligence et al. (2025) commonly rely on multi-view observations for comprehensive environmental perception. This highlights the need for multi-view embodied world models that can generate synchronized predictions across views.

Existing multi-view embodied world models Guo et al. (2025); Zhou et al. (2026) typically concatenate observations from multiple views and rely on attention mechanisms to implicitly capture cross-view relationships. However, without explicit correspondences, they must infer both which image regions correspond to the same 3D content and how a shared physical motion manifests across different views. As a result, they often struggle to produce future predictions that are consistent across views in terms of both scene appearance and robot motion.

As illustrated in Figure 1(a), generated observations may exhibit scene inconsistencies in object geometry, color, location, and presence. Without explicit cross-view constraints, the model may resolve generative ambiguities independently for each view, leading to divergent predictions of the same scene. Figure 1(b) further demonstrates motion inconsistency: although both generated views depict the robot successfully grasping the target object, they exhibit different motion trajectories. This arises because text provides only high-level semantic guidance about *what* the robot should accomplish, while lacking the fine-grained motion constraints needed to specify *how* the task should be executed.

To address these challenges, we propose ViewWeaver, a condition-enhanced multi-view world model for embodied intelligence that flexibly accommodates an arbitrary number of views. As shown in Figure 1 (c), ViewWeaver unifies text-conditioned future prediction and action-conditioned world simulation within a single framework. It augments both text and action conditions with view-specific information, promotes semantic-motion complementarity between the two conditioning modalities, and performs condition-modulated cross-view attention over multi-view features.

Specifically, we design a View-aware Semantic-Motion Conditioning (VSMC) module to construct view-aware, modality-consistent conditions. First, VSMC incorporates each view's initial observation and camera information into the corresponding text or action conditions, providing view-specific guidance. Second, it aligns text and action conditions in a shared representation space to encourage the resulting representations to encode complementary semantic and motion information. As a result, action conditions are enriched with high-level task semantics, while text conditions are grounded in fine-grained motion cues.

ViewWeaver introduces a Condition-Modulated Cross-View Attention (CM-CVA) module as a side branch of a diffusion-based video generation model. CM-CVA first uses the view-specific conditions produced by VSMC to modulate the corresponding view features. The modulated features then attend to features from the other views, enabling condition-guided information exchange across views. Finally, the resulting features are fused with the outputs of the original model blocks and progressively denoised to generate coherent future multi-view observations.

We evaluate ViewWeaver on the simulated LIBERO-90 Liu et al. (2023) dataset and the real-world AgiBot-World Beta Shi et al. (2025) dataset under both text- and action-conditioned settings. As shown in Figure 1 (d), ViewWeaver achieves the best overall performance across all views under both conditioning settings. Specifically, under text guidance on LIBERO-90, ViewWeaver achieves an FVD_{third} of 16.60, outperforming Ctrl-World (48.96) by 66.1%. Under action conditioning on AgiBot-World Beta, it achieves FVD_{left} and FVD_{right} of 76.91 and 96.95, outperforming Ctrl-World by 19.2% and 20.6%, respectively. It also achieves the lowest MEt3R of 0.545 (vs. 0.648 for Ctrl-World) on action-conditioned LIBERO-90 and the highest Trajectory Accuracy of 80.01 (vs. 77.67 for Ctrl-World) on action-conditioned AgiBot-World Beta, respectively, demonstrating better cross-view consistency and adherence to the prescribed action trajectory. Ablation studies validate the effectiveness of both VSMC and CM-CVA. Qualitative results demonstrate substantial improvements in cross-view consistency in both scene appearance and robot motion.

2 Related Works

2.1 Video World Models for Embodied Intelligence

Recent advances in video generation models Ali et al. (2025); Team et al. (2025b); Xiang et al. (2024); Chi et al. (2025) have enabled the synthesis of physical realistic content, facilitating their

adoption as high-fidelity world models for embodied intelligence Mei et al. (2026). Existing approaches can be broadly categorized into two groups: instruction-conditioned world models Du et al. (2024); Black et al. (2024); Ko et al. (2024); Zhou et al. (2024); Cheang et al. (2024); Yang et al. (2025); Jang et al. (2025) and action-conditioned world models Finn & Levine (2017); Finn et al. (2016); Rigter et al. (2024); Zhu et al. (2024); Guo et al. (2026); Gao et al. (2026).

Instruction-conditioned models use task labels or language to guide future observation generation. These models are commonly used as policy backbones, producing actions through inverse dynamics models Bu et al. (2024); Feng et al. (2025); Zhang et al. (2026), dedicated action prediction modules Jia et al. (2025); Pai et al. (2025), or shared backbones for action and video modeling Kim et al. (2026). Action-conditioned world models instead use robot control signals as explicit conditions for visual dynamics prediction. They are primarily used for policy evaluation Li et al. (2025); Agarwal et al. (2025); Quevedo et al. (2025); Team et al. (2025a) and for dynamics or reward modeling in model-based reinforcement learning Wu et al. (2023); Xiao et al. (2025); Hafner et al. (2025).

2.2 Multi-View Video Generation

Multi-view video generation focuses on cross-view consistency, including novel-view synthesis from images or monocular videos Yu et al. (2024); You et al. (2025); Xie et al. (2025b); Wu et al. (2025); Zhang et al. (2024); Van Hoorick et al. (2026), joint completion under predefined camera layouts Kuang et al. (2024); Xu et al. (2025); Li et al. (2024), object-centric generation Li et al. (2024); Wu et al. (2025); Zhang et al. (2024), and panoramic expansion Xie et al. (2025a). Unlike these methods, multi-view embodied world models predict full-scene dynamics from current multi-view observations, capturing embodied agents, robot-object interactions from a mixture of agent-mounted moving and static cameras.

2.3 Multi-View Embodied World Models

Multi-view embodied world modeling remains underexplored. For data generation, some methods perform view synthesis using 4D Gaussian Splatting or semantic occupancy representations Huang et al. (2026); Yang et al. (2026); Wang et al. (2026). For future prediction, some methods concatenate all views into a single input Guo et al. (2025); Li et al. (2026); Ye et al. (2026). MTV-World Su et al. (2025) independently generates each view, conditioned on view-specific actions. Genie Envisioner Liao et al. (2025) introduces cross-view attention to facilitate information exchange across views. In contrast, ViewWeaver enables both text- and action-conditioned multi-view future prediction through view-aware conditioning and semantic-motion alignment, together with condition-guided cross-view attention for selective information exchange across views.

3 Method

3.1 Problem Formulation

A multi-view world model predicts future observations from current and historical observations under either text or action conditioning. At timestep t, the observation sequence is denoted by $o_{1:t} \in \mathbb{R}^{V\times t\times h\times w\times 3}$, where V is the number of views. The future observations are predicted as

$$\hat{o}_{t+1:T} = \mathcal{W}(o_{1:t}; \mathcal{C}), \tag{1}$$

where the condition $\mathcal{C}$ is defined as

$$\mathcal{C} = \begin{cases} (\mathcal{T}, c_{1:t}), & \text{text-conditioned,} \\ (a_{1:T}, c_{1:T}), & \text{action-conditioned.} \end{cases} \tag{2}$$

Here, $\mathcal{T}$ denotes the text instruction, $a_{1:T} \in \mathbb{R}^{T\times d_a}$ denotes the action sequence. $c_{1:t} \in \mathbb{R}^{V\times t\times 25}$ is the camera parameters, including both intrinsic and extrinsic parameters.

3.2 Architecture

ViewWeaver is built upon a text-conditioned video diffusion model, which comprises a VAE and a diffusion backbone. Following Ctrl-World Guo et al. (2025), we extend the conditioning mechanism

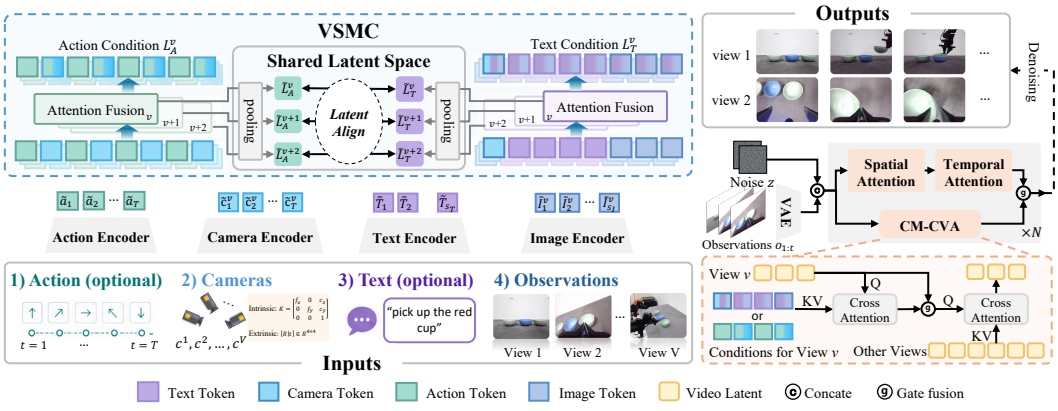

Figure 2: Overview of ViewWeaver, a multi-view video world model that unifies text- and action-conditioned generation from multi-view observations, camera parameters, text instructions, and robot actions. VSMC augments action conditions with camera tokens and text conditions with image and camera tokens, transforming them into view-aware condition latents. A latent alignment loss maps these latents into a shared representation space, encouraging the condition representations to encode complementary semantic and motion. CM-CVA further uses the resulting view-aware latents to modulate view-specific queries and condition cross-view attention, selectively aggregating complementary information from other views.

to additionally support action-conditioned generation. The resulting shared backbone supports both conditioning settings, with only one type of condition used at inference time.

The observation sequence $o_{1:t}$ is first encoded by the VAE into latent $z_{1:t} \in \mathbb{R}^{V\times t\times N_h\times N_w\times d_{\text{vae}}}$. At each denoising step, let $z \in \mathbb{R}^{V\times(T-t)\times N_h\times N_w\times d_{\text{vae}}}$ denote the noisy latent of the future observations. We concatenate $z_{1:t}$ and z along the temporal dimension. Then we repeat the current latent z_t along the temporal dimension into z'_t and concatenate it with the latent sequence along the channel:

$$\bar{z} = [[z_{1:t}; z]_{\text{time}}; z'_t]_{\text{feat}}\,. \tag{3}$$

For clarity, we omit the channel projection that maps $\bar{z}$ to the input dimension of the diffusion backbone. The backbone processes $\bar{z}$ at each step to predict the noise for iterative denoising Rombach et al. (2022), and the denoised future latents are decoded by the VAE to produce synchronized multi-view observations $\hat{o}_{t+1:T}$.

To construct view-specific conditions that capture both high-level task semantics and fine-grained motion cues, we introduce the View-aware Semantic-Motion Conditioning module (VSMC). It first encodes the text instruction, action sequence, camera parameters, and current multi-view observation into token sequences $\tilde{T} \in \mathbb{R}^{s_T\times d}$, $\tilde{a} \in \mathbb{R}^{T\times d}$, $\tilde{c} \in \mathbb{R}^{V\times T\times d}$, and $\tilde{I} \in \mathbb{R}^{V\times s_I\times d}$, respectively. VSMC fuses $\tilde{T}$, $\tilde{I}$, and the current camera token $\tilde{c}_t$ to obtain the view-aware text condition L_T. It also fuses $\tilde{a}$ with $\tilde{c}$ to construct the view-aware action condition L_A. A cosine-similarity loss is further applied between L_T and L_A to encourage semantic-motion complementarity.

ViewWeaver processes the latent stream of each view independently while introducing a parallel Condition-Modulated Cross-View Attention branch (CM-CVA) for condition-relevant information exchange. This view-number-agnostic design allows ViewWeaver to naturally support an arbitrary number of input views. CM-CVA modulates the query of each view with its corresponding condition latent and selectively aggregates complementary features from the other views through cross-attention. The resulting cross-view features are then integrated with the outputs of the original block of the backbone through a gated fusion mechanism.

3.3 View-aware Semantic-Motion Conditioning

To provide view-specific conditional guidance and leverage the complementary semantic and motion information from text and actions, we introduce VSMC.

In the action branch, the action tokens $\tilde{a}$ specify fine-grained motion but are shared across views. Since the same action induces different image-space motion under different cameras, we fuse $\tilde{a}$ with

the temporally aligned camera tokens $\tilde{c}$ through camera-aware attention:

$$\hat{c}_A = \mathcal{A}_c^A(\tilde{c}, \tilde{a}) \in \mathbb{R}^{V\times T\times d}. \tag{4}$$

The resulting camera-aware features are concatenated with the action tokens along the sequence dimension to form the view-specific action condition $L_A \in \mathbb{R}^{V\times T\times 2\times d}$.

In the text branch, text instructions convey high-level task semantics but lack concrete motion details. Through subsequent alignment, we use fine-grained action cues to make the text condition more concrete. However, a text instruction cannot be directly associated with a concrete action without being grounded in the current scene. We therefore first fuse the text tokens $\tilde{T}$ with the current-image tokens $\tilde{I}$:

$$F_T = \mathcal{A}_I(\tilde{T}, \tilde{I}) \in \mathbb{R}^{V\times(s_T+s_I)\times d}. \tag{5}$$

As in the action branch, we fuse F_T with current camera token $\tilde{c}_t$ through camera-aware attention:

$$\hat{c}_T = \mathcal{A}_c^T(\tilde{c}_t, F_T) \in \mathbb{R}^{V\times d}. \tag{6}$$

The resulting camera-aware feature is concatenated with F_T along the sequence dimension to form the view-specific text condition $L_T \in \mathbb{R}^{V\times(s_T+s_I+1)\times d}$. Details of the camera-aware attention modules $\mathcal{A}_c^A$, $\mathcal{A}_c^T$ and the image-text attention module $\mathcal{A}_I$ are provided in the appendix.

Finally, we align the paired text and action conditions corresponding to the same execution. We pool them into view-level representations $\tilde{L}_T, \tilde{L}_A \in \mathbb{R}^{V\times d}$ and apply the cosine-similarity loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{V}\sum_{v=1}^{V}\frac{\left\langle \tilde{L}_T^v, \tilde{L}_A^v \right\rangle}{\|\tilde{L}_T^v\|_2 \, \|\tilde{L}_A^v\|_2}. \tag{7}$$

This alignment allows action cues to make the text condition more concrete and text description to enable the action condition to obtain more task-level guidance, enabling the two conditions to complement each other.

3.4 Condition-Modulated Cross-View Attention

Existing multi-view generation methods typically use standard attention mechanisms to exchange information across views. In embodied scenes, however, large portions of the observations, such as backgrounds and task-irrelevant objects, remain static during task execution. In contrast, task-relevant regions, such as manipulated objects and regions undergoing motion, provide critical cues for predicting future dynamics and should receive greater attention during cross-view interaction. Without explicit guidance, standard attention may propagate redundant features from task-irrelevant regions and introduce additional noise.

To address this, we augment each attention block of the diffusion backbone with a parallel Condition-Modulated Cross-View Attention (CM-CVA) for information exchange across views.

As the latent input $\bar{z}$ from Equation (3) passes through the attention blocks of the diffusion backbone, it is progressively transformed into a sequence of intermediate hidden states $h_1, h_2, \cdots$. For an arbitrary attention block, we denote its hidden state by $h \in \mathbb{R}^{V\times f_t\times f_h\times f_w\times d}$, where f_t, f_h, and f_w denote the numbers of tokens along the temporal, height, and width dimensions, respectively. CM-CVA performs attention among spatial tokens across different views. We flatten the spatial grid to obtain $h' \in \mathbb{R}^{V\times f_t\times f_h f_w\times d}$ and project the flattened features into queries, keys, and values:

$$q = \mathrm{LN}(W_q h'), \quad k = \mathrm{LN}(W_k h'), \quad v = W_v h', \tag{8}$$

where $q, k, v \in \mathbb{R}^{V\times f_t\times f_h f_w\times d}$; W_q, W_k, and W_v are learnable projections; and LN denotes layer normalization.

Before exchanging information across views, CM-CVA incorporates the corresponding view-specific VSMC condition into the query:

$$\tilde{q} = q + \gamma_c \odot \text{Attn}\left(q, L_*, L_*\right), \tag{9}$$

where L_* denotes either the text condition L_T or the action condition L_A, and $\gamma_c \in \mathbb{R}^d$ is a learnable channel-wise gate. The cross-view features are then computed as

$$h_{\text{cva}} = \text{CrossAttn}\left(\tilde{q}, k, v; \mathbf{I}_V\right), \tag{10}$$

where $\mathbf{I}_V$ is an inverted identity matrix operating as an attention mask to prevent self-attention within the same view.

Finally, we fuse h_{cva} with the output of the corresponding backbone attention block through a learnable channel-wise gate. The gated fusion retains the original backbone output as the base representation while adaptively controlling the contribution of the cross-view features.

3.5 Training Paradigm

We progressively train ViewWeaver in three stages using both text- and action-conditioned samples.

Given a clean latent x_0, its noisy version x_τ at diffusion timestep τ, noise ϵ, and either the text or action condition $\mathcal{C}$ defined in Equation (2), the diffusion objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\mathcal{C},\epsilon,\tau}\left[\|\epsilon - \epsilon_\theta(x_\tau, \tau, \mathcal{C})\|_2^2\right]. \tag{11}$$

In Stage 1, we jointly train the video diffusion backbone and VSMC without camera-aware fusion. At this stage, the camera information in $\mathcal{C}$ is omitted. The primary goal is to align the text and action condition representations associated with the same execution, thereby encouraging each representation to capture both semantic content and motion dynamics. The training objective is $\mathcal{L}_{\text{diff}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$.

In Stage 2, we enable camera-aware fusion and continue optimizing the backbone and VSMC. The camera information in $\mathcal{C}$ is incorporated into the condition representations through VSMC. The training objective is $\mathcal{L}_{\text{diff}}$.

In Stage 3, we freeze the backbone and VSMC and optimize only CM-CVA. This stage learns condition-modulated information exchange across views without disrupting the condition alignment and per-view embodied generation capabilities acquired in the previous stages. The training objective is $\mathcal{L}_{\text{diff}}$.

4 Experiments

Implementation Details. We initialize ViewWeaver from Stable Video Diffusion Blattmann et al. (2023) and optimize it using AdamW. The learning rate is linearly warmed up to 1×10^{-5} over the first 500 iterations and remains constant thereafter. We condition the model on 6 historical frames, including 4 frames sampled backward with a temporal interval and 2 frames at $t = 0$. The weight of the alignment loss is set to $\lambda_{\text{align}} = 0.1$. Stage 1 is trained for $150{,}000$ iterations with a batch size of 4, while Stages 2 and 3 are each trained for $100{,}000$ iterations with a batch size of 16. All experiments are conducted on 4 NVIDIA A800 GPUs.

Datasets. We conduct experiments on LIBERO and AgiBot-World Beta to evaluate the world-modeling capabilities of our model in both simulated and real-world scenarios. Additional details are provided in the supplementary material.

Metrics. We use four common metrics: PSNR Hore & Ziou (2010), LPIPS Zhang et al. (2018), SSIM Wang et al. (2004), and FVD Unterthiner et al. (2018), together with MEt3R Asim et al. (2025) to measure multi-view 3D consistency. For action-conditioned generation, we additionally use Trajectory Accuracy (TA) from WorldArena Shang et al. (2026) to measure adherence to the prescribed action.

Baselines. For action-conditioned generation, we select iVideoGPT Wu et al. (2024), IRASim Zhu et al. (2024), and Ctrl-World Guo et al. (2025) as baselines. For text-conditioned generation, we compare against iVideoGPT Wu et al. (2024), Genie Envisioner Liao et al. (2025), Cosmos Predict2 Agarwal et al. (2025), and Ctrl-World* (a modified version for text conditioning).

4.1 Main Results

4.1.1 Quantitative Results on LIBERO.

Under text conditioning, ViewWeaver reduces FVD_{third} from 48.96 to 16.60 (66.1%) and FVD_{hand} from 144.92 to 57.18 (60.5%). It also achieves the lowest MEt3R score of 0.542, compared with

Table 1: Quantitative comparison on LIBERO under text- and action-conditioned settings.

	Third				Hand				MEt3R↓	TA↑
	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓		
Text Conditioned										
iVideoGPT	195.44	16.99	0.714	0.217	251.34	15.10	0.561	0.432	0.665	-
Genie Envisioner	267.93	20.52	0.829	0.143	172.06	18.29	0.688	0.348	0.599	-
Cosmos Predict	341.55	20.55	0.802	0.134	144.92	18.13	0.629	0.328	0.601	-
Ctrl-World*	48.96	21.48	0.845	0.106	192.71	18.71	0.698	0.317	0.592	-
ViewWeaver	**16.60**	**22.09**	**0.856**	**0.091**	**57.18**	**19.00**	**0.704**	**0.279**	**0.542**	-
Action Conditioned										
iVideoGPT	229.27	20.39	0.790	0.141	270.12	17.41	0.631	0.337	0.673	68.08
IRASim	886.88	22.52	0.798	0.149	412.27	19.58	0.685	0.297	0.629	64.49
Ctrl-World	34.57	23.49	0.858	0.123	86.11	20.40	0.742	0.265	0.648	73.65
ViewWeaver	**12.90**	**25.93**	**0.906**	**0.056**	**48.94**	**21.38**	**0.761**	**0.203**	**0.545**	**81.05**

Table 2: Quantitative comparison on AgiBot-World Beta under text-conditioned and action-conditioned settings.

	Head				Left Hand				Right Hand				MEt3R↓	TA↑
	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓		
Text Conditioned														
iVideoGPT	575.57	18.95	0.660	0.283	520.86	17.13	0.579	0.356	567.13	18.22	0.623	0.368	0.669	-
Genie Envisioner	77.47	23.38	0.855	0.159	111.34	22.53	0.801	0.215	128.92	24.04	0.801	0.241	0.623	-
Cosmos Predict	167.94	22.12	0.817	0.166	142.14	21.07	0.736	0.230	129.57	22.30	0.733	0.261	0.615	-
Ctrl-World*	75.63	23.94	0.858	0.127	103.81	22.71	0.774	0.196	133.64	24.15	0.780	0.227	0.614	-
ViewWeaver	**58.32**	**24.85**	**0.870**	**0.109**	**75.76**	**24.12**	**0.825**	**0.164**	**98.56**	**25.61**	**0.818**	**0.194**	**0.606**	-
Action Conditioned														
iVideoGPT	588.74	19.03	0.660	0.282	514.07	17.24	0.588	0.354	572.47	18.61	0.632	0.362	0.668	26.02
IRASim	1140.21	19.70	0.700	0.259	909.42	18.77	0.647	0.319	854.81	20.15	0.665	0.339	0.609	51.72
Ctrl-World	67.05	24.37	0.858	0.113	95.14	23.76	0.817	0.165	122.09	25.43	**0.817**	0.200	0.642	77.67
ViewWeaver	**57.18**	**24.94**	**0.871**	**0.108**	**76.91**	**24.20**	**0.827**	**0.162**	**96.95**	**25.60**	**0.817**	**0.194**	**0.607**	**80.01**

0.592 for the best baseline. Under action conditioning, ViewWeaver reduces FVD_{third} from 34.57 to 12.90 (62.7%) and FVD_{hand} from 86.11 to 48.94 (43.2%), respectively. It also improves PSNR_{third} from 23.49 to 25.93 and LPIPS_{third} from 0.123 to 0.056. Its MEt3R score is the lowest at 0.545, compared with 0.629 for the best baseline, demonstrating better cross-view consistency. It also achieves the highest TA of 81.05 (vs. 73.65), indicating that it more faithfully follows the prescribed action trajectory.

It is worth noting that existing baselines exhibit clear performance imbalances across views. Under text conditioning, Ctrl-World* obtains FVD_{third} and FVD_{hand} scores of 48.96 and 192.71, respectively, resulting in a gap of 143.75. Cosmos Predict exhibits the opposite trend, scoring 341.55 and 144.92, with an even larger gap of 196.63. In contrast, ViewWeaver achieves the best results on both views (16.60 and 57.18), with a much smaller gap of 40.58, demonstrating more stable cross-view dynamics modeling.

4.1.2 Quantitative Results on AgiBot-World.

On AgiBot-World, ViewWeaver generalizes well to real-world multi-view robotic data. Under text conditioning, it reduces FVD_{head} from 75.63 to 58.32, FVD_{left} from 103.81 to 75.76, and FVD_{right} from 128.92 to 98.56. It also reduces LPIPS_{head} from 0.127 to 0.109 and achieves the lowest available MEt3R score of 0.606. Under action conditioning, ViewWeaver reduces FVD_{head} from 67.05 to 57.18, FVD_{left} from 95.14 to 76.91, and FVD_{right} from 122.09 to 96.95. It also achieves the lowest MEt3R score of 0.607, further demonstrating better cross-view consistency. ViewWeaver also attains the highest TA of 80.01 (vs. 77.67), indicating more faithful adherence to the prescribed action trajectory in real-world settings. We exclude GE-Sim Liao et al. (2025) from this comparison because its training code is not publicly available, preventing evaluation under our setting.

ViewWeaver's improvements on AgiBot-World are particularly pronounced in FVD. Because static backgrounds dominate much of each frame, frame-level metrics are strongly influenced by static regions, while the main prediction challenges occur in localized dynamic regions, such as the robot arm, manipulated objects, contact events, and occlusion boundaries. Errors in these regions may

have little impact on individual-frame scores but can accumulate into trajectory drift and motion discontinuities; FVD is more sensitive to such errors because it evaluates sequences using spatio-temporal video features. By constructing view-aware semantic-motion conditions with VSMC and effective feature exchange with CM-CVA, ViewWeaver better models these critical dynamics, improving trajectory continuity and the spatio-temporal consistency of multi-view predictions.

4.1.3 Bridging the Gap Between Text and Action Conditions.

Text instructions convey high-level task semantics but lack the fine-grained trajectory constraints of action sequences, making text-conditioned prediction more prone to motion inconsistencies and generally weaker than action-conditioned prediction. On LIBERO, Ctrl-World exhibits text–action gaps of 14.39 in FVD_{third} (48.96 vs. 34.57) and 106.60 in FVD_{hand} (192.71 vs. 86.11). ViewWeaver narrows these gaps to 3.70 (16.60 vs. 12.90) and 8.24 (57.18 vs. 48.94), corresponding to relative reductions of 74.3% and 92.3%, respectively. These results show that VSMC's semantic-motion alignment grounds high-level text semantics in scene-dependent motion cues, improving text-conditioned prediction and substantially narrowing the gap between text- and action-conditioned generation.

4.2 Qualitative Results

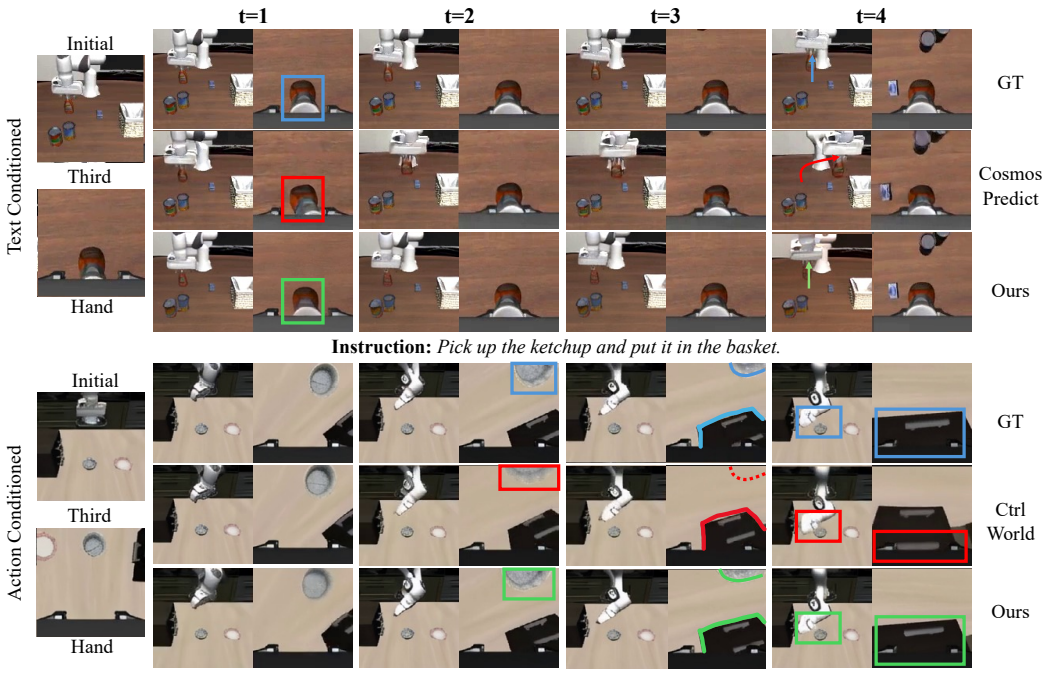

Figure 3: Qualitative comparisons on LIBERO under text- and action-conditioned settings. ViewWeaver better preserves scene appearance and cross-view robot motion than baselines.

Figure 3 presents two examples from LIBERO under the text- and action-conditioned settings, comparing ViewWeaver with the corresponding baselines. In the text-conditioned example, at $t = 1$, Cosmos Predict generates a deformed and tilted can in the Hand view that is inconsistent with the ground truth (GT), whereas ViewWeaver preserves its correct appearance. Over the subsequent three time steps, Cosmos Predict follows the GT motion in the Third view but generates a rightward trajectory in the Hand view, rather than the upward motion observed in GT. Although each predicted trajectory may independently satisfy the text instruction, the two views depict incompatible executions and therefore exhibit clear cross-view motion inconsistency. In contrast, ViewWeaver produces consistent results across both views, benefiting from the complementary semantic and motion representations learned by VSMC.

In the action-conditioned example, Ctrl-World exhibits several scene inconsistencies, including an incorrect bowl shape in the Hand view at $t = 2$, as well as inaccurate cabinet geometry and bowl presence at $t = 3$. At $t = 4$, its predicted end-effector is lower than its GT position in the Third view

Table 3: Ablation study of ViewWeaver on LIBERO under text- and action-conditioned settings.

	Text Conditioned				Action Conditioned			
	$\text{FVD}_{third}\downarrow$	$\text{PSNR}_{third}\uparrow$	$\text{FVD}_{hand}\downarrow$	$\text{PSNR}_{hand}\uparrow$	$\text{FVD}_{third}\downarrow$	$\text{PSNR}_{third}\uparrow$	$\text{FVD}_{hand}\downarrow$	$\text{PSNR}_{hand}\uparrow$
Stable Video Diffusion	80.20	21.42	230.33	18.77	34.78	22.95	156.48	19.68
+Alignment	43.68	21.27	165.15	18.42	25.36	23.48	116.65	19.69
+Camera (VSMC)	36.86	21.55	145.20	18.78	21.68	24.19	99.24	20.40
+CM-CVA (ViewWeaver)	**30.33**	**21.88**	**116.74**	**18.93**	**18.49**	**25.19**	**85.61**	**21.04**
All-in-All	36.69	21.15	121.71	18.36	20.13	24.21	90.31	20.39

but higher than its GT position in the Hand view, revealing a clear cross-view motion inconsistency. By contrast, ViewWeaver remains consistent with GT in both scene appearance and end-effector motion across the two views.

More visualizations are provided in the supplementary material.

4.3 Ablation Study

To improve training efficiency, we use a batch size of 4 in all ablation experiments.

Effectiveness of VSMC. We evaluate the two key components of VSMC: text–action alignment and camera-aware fusion. Adding text–action alignment (+Alignment) substantially reduces FVD_{third} and FVD_{hand} from 80.20 and 230.33 to 43.68 and 165.15 under text conditioning, and from 34.78 and 156.48 to 25.36 and 116.65 under action conditioning. In contrast, the corresponding PSNR changes are modest: from 21.42 and 18.77 to 21.27 and 18.42 under text conditioning, and from 22.95 and 19.68 to 23.48 and 19.69 under action conditioning. This suggests that alignment improves distributional and temporal quality by learning complementary semantic and motion representations. Completing VSMC with camera-aware fusion (+Camera) further reduces FVD_{third} and FVD_{hand} to 36.86 and 145.20 under text conditioning, and to 21.68 and 99.24 under action conditioning, while increasing the corresponding PSNR scores to 21.55 and 18.78, and to 24.19 and 20.40, respectively. These gains show that and camera parameters provide view-specific guidance for modeling appearance and dynamics, particularly for the more challenging Hand view.

Effectiveness of CM-CVA Adding CM-CVA (+CM-CVA) yields the full ViewWeaver model and the best overall results. Compared with +Camera, it reduces FVD_{third} and FVD_{hand} from 36.86 and 145.20 to 30.33 and 116.74 under text conditioning, and from 21.68 and 99.24 to 18.49 and 85.61 under action conditioning; the corresponding PSNR_{third} and PSNR_{hand} scores also increase from 21.55 and 18.78 to 21.88 and 18.93, and from 24.19 and 20.40 to 25.19 and 21.04, respectively. These results show that condition-modulated cross-view attention selectively aggregates complementary information across views, while gated fusion preserves the backbone representations to improve prediction quality and cross-view consistency.

Effectiveness of the Training Strategy. To evaluate the progressive training strategy, we compare the full ViewWeaver model with the All-in-All configuration, which jointly trains the backbone, VSMC, and CM-CVA without stage-wise separation. Compared with All-in-All, progressive training reduces the FVD_{third} and FVD_{hand} from 36.69 and 121.71 to 30.33 and 116.74 under text conditioning, and from 20.13 and 90.31 to 18.49 and 85.61 under action conditioning, while consistently improving PSNR. These results demonstrate that stage-wise optimization improves both distributional quality and frame-level fidelity.

5 Conclusion

We presented ViewWeaver, a condition-enhanced multi-view video world model that unifies text-conditioned future prediction and action-conditioned world simulation for embodied intelligence. VSMC integrates camera cues and aligns text and action representations to construct view-specific semantic-motion conditions, while CM-CVA uses these conditions to guide gated information exchange across views. On LIBERO-90 and an AgiBot-World Beta, ViewWeaver achieves the strongest overall performance among the evaluated baselines under both conditioning modes. Ablations support the effectiveness of both modules, while qualitative comparisons indicate improved cross-view consistency in scene appearance and robot motion.

AI USE STATEMENT

In this work, we have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning or reformatting research datasets, supporting qualitative or thematic data analysis, or interpreting results, and formulating mathematical claims and proof ingredients, writing proofs, proposing and refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work. All AI-generated wording was reviewed and confirmed by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 33431–33452, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. *Advances in Neural Information Processing Systems*, 37:139002–139029, 2024.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Kaelbling, et al. Video language planning. In *International Conference on Learning Representations*, volume 2024, pp. 31138–31155, 2024.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Jun Guo, Xiaojian Ma, Yikai Wang, Min Yang, Huaping Liu, and Qing Li. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *IEEE Robotics and Automation Letters*, 11(3):2466–2473, 2026.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Yue Liao, Zhengkai Jiang, Yue Hu, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *Advances in Neural Information Processing Systems*, 38:37693–37720, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Yueru Jia, Jiaming Liu, Shengbang Liu, Rui Zhou, Wanhe Yu, Yuyang Yan, Xiaowei Chi, Yandong Guo, Boxin Shi, and Shanghang Zhang. Video2act: A dual-system video diffusion policy with robotic spatio-motional modeling. *arXiv preprint arXiv:2512.03044*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations*, volume 2024, pp. 40938–40958, 2024.

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *Advances in Neural Information Processing Systems*, 2024.

Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. In *Advances in Neural Information Processing Systems*, 2024.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiaxiong Qiu, Zheng Zhu, Guan Huang, and Zhizhong Su. Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. *arXiv e-prints*, pp. arXiv–2505, 2025.

Xiaokang Liu, Zechen Bai, Hai Ci, Kevin Yuchen Ma, and Mike Zheng Shou. World-vla-loop: Closed-loop learning of video world model and vla policy. *arXiv preprint arXiv:2602.06508*, 2026.

Zhiting Mei, Tenny Yin, Ola Shorinwa, Apurva Badithela, Zhonghe Zheng, Joseph Bruno, Madison Bland, Lihan Zha, Asher Hancock, Jaime Fernández Fisac, et al. Video generation models in robotics-applications, research challenges, future directions. *arXiv preprint arXiv:2601.07823*, 2026.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

Modi Shi, Yuxiang Lu, Huijie Wang, Chengen Xie, and Qingwen Bu. Introducing agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. `https://opendrivelab.com/AgiBot-World/`, March 2025. Blog post.

Taiyi Su, Jian Zhu, Yaxuan Li, Chong Ma, Jianjun Zhang, Zitai Huang, Hanli Wang, and Yi Xu. Towards high-consistency embodied world model with multi-view trajectory videos. *arXiv preprint arXiv:2511.12882*, 2025.

Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025a.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025b.

Wei-Cheng Tseng, Jinwei Gu, Qinsheng Zhang, Hanzi Mao, Ming-Yu Liu, Florian Shkurti, and Lin Yen-Chen. Scalable policy evaluation with video world models. *arXiv preprint arXiv:2511.11520*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Basile Van Hoorick, Dian Chen, Shun Iwase, Pavel Tokmakov, Muhammad Zubair Irshad, Igor Vasiljevic, Swati Gupta, Fangzhou Cheng, Sergey Zakharov, and Vitor Campagnolo Guizilini. AnyView: Synthesizing any novel view in dynamic scenes. *arXiv preprint arXiv:2601.16982*, 2026.

Guanzhi Wang, Yecheng Jason Wu, Haochen Wu, Chengyang He, Zixuan Wang, and Shuran Song. RoboVIP: Towards generalist robot policies through multi-view video prediction. *arXiv preprint arXiv:2602.09817*, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-env: Leveraging world model as a virtual environment for vla post-training. *arXiv preprint arXiv:2509.24948*, 2025.

Kevin Xie, Amirmojtaba Sabour, Jiahui Huang, Despoina Paschalidou, Greg Klar, Umar Iqbal, Sanja Fidler, and Xiaohui Zeng. VideoPanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*, 2025a.

Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11078–11086, 2026.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D: Dynamic 3d content generation with multi-frame and multi-view consistency. In *International Conference on Learning Representations*, 2025b.

Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. In *International Conference on Machine Learning*, 2025.

Liudi Yang, Yang Bai, George Eskandar, Fengyi Shen, Mohammad Altillawi, Dong Chen, Ziyuan Liu, and Abhinav Valada. Covar: Co-generation of video and action for robotic manipulation via multi-modal diffusion. *arXiv preprint arXiv:2512.16023*, 2025.

Zhen Yang, Zhaoxi Chen, Minghan Qin, Yikang Ding, Xin Jin, Hang Zhao, and Hao Zhao. ORV: 4d occupancy-centric robot video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. NVS-Solver: Video diffusion model as zero-shot novel view synthesizer. In *International Conference on Learning Representations*, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. In *Advances in Neural Information Processing Systems*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

Wenyao Zhang, Bozhou Zhang, Zekun Qi, Wenjun Zeng, Xin Jin, and Li Zhang. Disentangled robot learning via separate forward and inverse dynamics pretraining. *arXiv preprint arXiv:2604.16391*, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Xinyu Zhou, Ruihai Qin, Yanjun Ji, Joy Hsu, Xiaodong Yang, Hao Luo, and Xiaolong Wang. τ_0-WM: A unified video-action world model for robotic manipulation. *arXiv preprint arXiv:2606.01027*, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 1(2):3, 2024.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. *arXiv preprint arXiv:2511.09515*, 2025.

Appendix

A Details for Method

A.0.1 Historical Observation Sampling

We retain the notation in the main paper, where $o_{1:t}$ denotes all observations available at the current time step and $c_{1:t}$ denotes their camera parameters. The implementation does not feed the entire history to the model. Instead, it applies a sparse six-frame selector $\mathcal{S}_6(\cdot)$ to $o_{1:t}$. Using the one-based notation of the main paper, the ordered selector is $\boldsymbol{\iota}_t = (1, 1, t-12, t-9, t-6, t-3)$, with indices clipped at the beginning of a trajectory when necessary. Thus, the initial observation is included twice and four recent historical observations are sampled at a three-frame interval (the first two entries are absolute initial-frame indices, while the remaining entries are offsets from t). We use the same indices to select the corresponding historical camera parameters and actions. Thus, $o_{1:t}$ and $c_{1:t}$ continue to denote the available sequences, whereas $\mathcal{S}_6(o_{1:t})$ and $\mathcal{S}_6(c_{1:t})$ denote the actual sparse inputs. The denoising objective supervises only the future observations $o_{t+1:T}$.

A.0.2 Camera Information

For view v at time t, the raw camera vector concatenates the flattened extrinsic $E_t^v \in \mathbb{R}^{4\times4}$ and intrinsic matrices $K_t^v \in \mathbb{R}^{3\times3}$:

$$c_t^v = [\mathrm{vec}(E_t^v); \mathrm{vec}(K_t^v)] \in \mathbb{R}^{25}. \tag{12}$$

The two conditioning modes use the camera information available at inference time. In the text-conditioned setting, the current observation provides the current camera parameters c_t, so VSMC uses only the encoded token $\tilde{c}_t$ rather than the full historical sequence $c_{1:t}$. In the action-conditioned setting, the prescribed action sequence $a_{1:T}$ is accompanied by its camera trajectory $c_{1:T}$; VSMC therefore uses the temporally aligned encoded sequence $\tilde{c}_{1:T}$. When historical camera context is needed, only $\mathcal{S}_6(c_{1:t})$ is materialized, following the sparse observation selector described in the implementation details. This historical selection is distinct from the text branch, which uses the current token c_t.

A.0.3 Image-Text Fusion

Let $\tilde{T} \in \mathbb{R}^{s_T\times d}$ and $\tilde{I} \in \mathbb{R}^{V\times s_I\times d}$ denote the encoded tokens. The text sequence is repeated for every view and concatenated with the image sequence:

$$\begin{aligned} \tilde{T}_V &= \mathrm{Repeat}_V(\tilde{T}), & X_I &= [\tilde{I}; \tilde{T}_V]_{\mathrm{seq}}, \\ F_T &= \mathcal{A}_I(\tilde{T}, \tilde{I}) = \mathrm{MHA}(X_I, X_I, X_I). \end{aligned} \tag{13}$$

Here MHA is multi-head attention. The attention is self-attention within each view.

A.0.4 Camera Fusion

For the action branch, the action token at each time step is repeated over views. The camera token is the query and the aligned action token is the key/value:

$$\hat{c}_{A,v}^\tau = \left[\mathcal{A}_c^A(\tilde{c}, \tilde{a}_V)\right]_{v,\tau} = \tilde{c}_v^\tau + \boldsymbol{g}_A \odot \mathrm{MHA}(\tilde{c}_v^\tau, \tilde{a}_\tau, \tilde{a}_\tau). \tag{14}$$

For the text branch, only the current camera token is available. It queries the view-specific text–image sequence:

$$\hat{c}_T^v = \left[\mathcal{A}_c^T(\tilde{c}_t, F_T)\right]_v = \tilde{c}_t^v + \boldsymbol{g}_T \odot \mathrm{MHA}(\tilde{c}_t^v, F_T^v, F_T^v). \tag{15}$$

Here $\boldsymbol{g}_A, \boldsymbol{g}_T \in \mathbb{R}^d$ are learnable channel-wise gates, initialized to zero, and the multiplication is elementwise.

A.0.5 CONDITION TENSOR SHAPES

For readability, the batch dimension is omitted below and the condition width is denoted by $d = 1024$. The action sequence starts as $a_{1:T} \in \mathbb{R}^{T\times d_a}$ and is encoded as $\tilde{a} \in \mathbb{R}^{T\times d}$. After repeating it over the V views and inserting a singleton sequence dimension, it becomes $\mathbb{R}^{V\times T\times 1\times d}$. Equivalently,

$$\tilde{a}_V = \mathrm{Unsqueeze}_S(\mathrm{Repeat}_V(\tilde{a})), \qquad \tilde{a}_V \in \mathbb{R}^{V\times T\times 1\times d}. \tag{16}$$

The camera sequence starts as $c_{1:T} \in \mathbb{R}^{V\times T\times 25}$ in the action branch (the text branch uses its current token c_t), and is encoded as $\tilde{c} \in \mathbb{R}^{V\times T\times d}$. Thus,

$$\tilde{c}_S = \mathrm{Unsqueeze}_S(\tilde{c}), \qquad \tilde{c}_S \in \mathbb{R}^{V\times T\times 1\times d}. \tag{17}$$

After camera fusion, concatenating the two along the sequence axis produces the action condition

$$L_A \in \mathbb{R}^{V\times T\times 2\times d}, \tag{18}$$

where the two tokens at time τ are the action token and its camera-aware token.

The encoded text sequence has shape $\tilde{T} \in \mathbb{R}^{s_T\times d}$. We repeat it over views, while the encoded current multi-view observation has shape $\tilde{I} \in \mathbb{R}^{V\times s_I\times d}$. Their image–text fusion is performed independently for each view and yields $F_T \in \mathbb{R}^{V\times(s_T+s_I)\times d}$. The current camera token is inserted as one additional sequence token and camera–text fusion yields

$$L_T \in \mathbb{R}^{V\times(s_T+s_I+1)\times d}. \tag{19}$$

For use by the diffusion backbone,

$$\begin{aligned} L_T^{\mathrm{time}} &= \mathrm{Broadcast}_T(\mathrm{Unsqueeze}_{\mathrm{time}}(L_T)), \\ L_T^{\mathrm{time}} &\in \mathbb{R}^{V\times T\times(s_T+s_I+1)\times d}. \end{aligned} \tag{20}$$

Thus, both conditioning paths have the common layout $V \times T \times S \times d$, but their sequence lengths differ: the action path uses $S = 2$, whereas the text path uses $S = s_T + s_I + 1$. Consequently, each action time step receives its own action token, while all time steps share the same text condition.

A.0.6 SIDE-BRANCH CM-CVA PLACEMENT

Let h be the hidden state immediately before the spatial transformer block at one resolution. With the CM-CVA side branch enabled, the model saves this tensor before executing the original spatial block. The original spatial block and the temporal block then run unchanged; CM-CVA is computed from the saved tensor and fused after the temporal/spatial AlphaBlender output. Thus, CM-CVA is parallel to the original block rather than a replacement for it.

A.0.7 CROSS-VIEW ATTENTION SCOPE

For target view i, the keys and values are taken only from the other views:

$$\begin{aligned} K_{\neg i} &= [k_j]_{j\neq i}, \qquad V_{\neg i} = [v_j]_{j\neq i}, \\ h_{\mathrm{cva},i} &= \mathrm{MHA}(\tilde{q}_i, K_{\neg i}, V_{\neg i}). \end{aligned} \tag{21}$$

The spatial-only setting reshapes the sequence by time and performs this operation independently at each temporal index. A target view cannot attend to its own spatial tokens, and no additional temporal mixing is introduced by CM-CVA. The operation is defined by the view axis and therefore does not depend on the number of cameras.

A.0.8 GATED FUSION

At the ℓ-th block, after the cross-view h_{cva} output is projected back to the local channel width, it is reshaped to the original spatio-temporal layout and added to the original spatio-temporal output h_{st} with a learned gate vector g_ℓ:

$$h_{\mathrm{out}} = h_{\mathrm{st}} + g_\ell \odot h_{\mathrm{cva}}, \qquad g_\ell \in \mathbb{R}^d. \tag{22}$$

All gate vectors are initialized to zero.

Task	Domain	Environment	Interaction
Brew coffee with a capsule machine	Catering	Kitchen	Beverage preparation
Open the fridge to get food	Household	Kitchen	Object retrieval
Open the wardrobe and hang the clothes	Household	Bedroom	Clothing organization
Pick up items in the supermarket	Commercial	Supermarket	Item collection
Place items from the meeting-room table into the storage box	Office	Meeting room	Object rearrangement
Place the pen into the pen holder	Office	Desk	Object placement
Pour water	Catering	Dining area	Liquid pouring
Sort items in the warehouse	Industrial	Warehouse	Item sorting

Table 4: Tasks selected in Agibot-World Beta.

A.0.9 EDM DENOISING OBJECTIVE

The diffusion loss $\mathcal{L}_{\text{diff}}$ in the main paper is implemented with EDM-style supervised fine-tuning. For clean future latent x_0 corresponding to $o_{t+1:T}$, sample $\log\sigma \sim \mathcal{N}(0.7, 1.6^2)$ and set $x_\sigma = x_0 + \sigma\epsilon$. The network input and the reconstructed clean latent are

$$\begin{aligned} c_{\text{in}} &= (1+\sigma^2)^{-1/2}, \quad c_{\text{skip}} = (1+\sigma^2)^{-1}, \\ c_{\text{out}} &= -\sigma(1+\sigma^2)^{-1/2}, \quad c_{\text{noise}} = \log(\sigma)/4, \\ \hat{x}_0 &= c_{\text{skip}} x_\sigma + c_{\text{out}}\, D_\theta(c_{\text{in}} x_\sigma, c_{\text{noise}}, \mathcal{C}). \end{aligned} \tag{23}$$

The concrete objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\frac{1+\sigma^2}{\sigma^2}\,\|\hat{x}_0 - x_0\|_2^2\right]. \tag{24}$$

It is evaluated only on the future latents; the six sparsely selected historical observations provide context and are not prediction targets. Each batch chooses text or action conditioning with probability $1/2$, so both modes optimize the same denoiser.

B DETAILS FOR EXPERIMENT

B.1 DATASETS

B.1.1 LIBERO-90

LIBERO-90 Liu et al. (2023) is a simulated robotic manipulation benchmark comprising 90 language-conditioned tasks. It provides 100 trajectories per task. We use a 4:1 trajectory-level training–test split. Training clips use the six-frame history described above and 81 model frames at 128×128; this corresponds to the 80-frame future horizon after the current frame reported in the main paper. The two standard synchronized views are evaluated separately as Third and Hand.

B.1.2 AGIBOT-WORLD BETA

AgiBot-World Beta Shi et al. (2025) is a large-scale real-world robotic manipulation dataset containing synchronized multi-view observations and action trajectories collected with dual-arm robots. We construct a subset covering five application domains—household, catering, industrial, commercial, and office—with diverse scene contexts and interaction types, as summarized in Table 4. We sample 90 trajectories per task for training and 5 for evaluation. Clips are resized to 192×256 and contain six history frames and 57 model frames, corresponding to the 56-frame future horizon used for this benchmark. The synchronized views are Head, Left Hand, and Right Hand.

B.2 EVALUATION

B.2.1 EVALUATION PROTOCOL

At test time, videos are sliced every 40 frames, and each clip retains the history length and model-frame length of its training configuration. All diffusion-based methods use ten denoising steps at inference.

Hyperparameter	Stage 1	Stage 2	Stage 3
Global batch size	4	16	16
Training steps	150,000	100,000	100,000
Learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.01	0.01	0.01
Adam betas	$(0.9, 0.999)$	$(0.9, 0.999)$	$(0.9, 0.999)$
Adam ϵ	10^{-8}	10^{-8}	10^{-8}
Learning-rate schedule	Constant	Constant	Constant
Warm-up steps	500	500	500
Model dtype	FP32	FP32	FP32
Mixed precision	FP16	FP16	FP16
Gradient accumulation	1	1	1
Gradient clipping norm	1.0	1.0	1.0
Gradient clipping start step	5,000	5,000	5,000
Gradient checkpointing	False	False	False
Random seed	42	42	42
Data-loader workers	8	8	8
Prefetch factor	2	2	2

Table 5: Training Configuration. Stage-wise training hyperparameters.

B.2.2 METRICS

We compute PSNR Hore & Ziou (2010), SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), and FVD Unterthiner et al. (2018) separately for each view: PSNR measures pixel-level fidelity, SSIM measures local structural agreement, LPIPS measures perceptual distance in a pretrained deep-feature space, and FVD compares the distributions of spatiotemporal video features from generated and reference clips. For multi-view consistency, MEt3R Asim et al. (2025) is a reference-free metric that estimates dense cross-view correspondences for each synchronized generated view pair and measures the agreement of geometry-aware features under those correspondences. We average MEt3R over all view pairs, so it evaluates whether the generated views support a coherent shared 3D scene rather than only per-view quality. Concretely, we compute one score for each synchronized view pair and report their arithmetic mean. For action-conditioned generation, we additionally report Trajectory Accuracy (TA) Shang et al. (2026), which measures adherence to the prescribed action trajectory. Higher PSNR, SSIM, and TA are better, whereas lower LPIPS, FVD, and MEt3R are better.

B.3 BASELINES

iVideoGPT. iVideoGPT Wu et al. (2024) is an autoregressive interactive video world model that supports both text-conditioned future prediction and action-conditioned rollout generation.

Genie Envisioner. Genie Envisioner Liao et al. (2025) is a unified world foundation platform for robotic manipulation that uses cross attention to exchange information across synchronized views.

Cosmos Predict2. Cosmos Predict2 Agarwal et al. (2025) is a general world foundation model for physical AI; we use its text-conditioned video-prediction variant.

IRASim. IRASim Zhu et al. (2024) is an action-conditioned simulator that predicts visual robot-interaction rollouts from control sequences.

Ctrl-World. Ctrl-World Guo et al. (2025) is a controllable generative world model for robot manipulation that generates future observations from prescribed actions.

Ctrl-World*. Ctrl-World* denotes our text-conditioned adaptation of Ctrl-World Guo et al. (2025). A CLIP encoder maps the instruction to a text condition, which replaces the original action condition.

C Implementation Details

C.1 Model Architecture

C.1.1 Text and Image Encoders

The instruction is tokenized by CLIP ViT-B/32, and the current image from each view is encoded by SigLIP-B/16 at 224×224. CLIP has hidden width 512 and a maximum sequence length of 77, whereas SigLIP produces 196 patch tokens of width 768. The text projector is a two-layer MLP $512 \to 512 \to 1024$, and the image projector is $768 \to 512 \to 1024$. Each projector uses GELU with the tanh approximation between its two linear layers.

C.1.2 Action Encoder

Each action $a_i \in \mathbb{R}^{d_a}$ is encoded independently by a three-linear-layer MLP. For LIBERO configuration, $d_a = 7$, for AgiBot-World configuration, $d_a = 16$:

$$d_a \to 1024 \to 1024 \to 1024. \tag{25}$$

A SiLU activation follows each of the first two linear layers, and no normalization layer is used. The same architecture accepts the dataset- specific action width used by the real-world experiments. Applying the encoder to $a_{1:T}$ gives $\tilde{a} \in \mathbb{R}^{T\times 1024}$, which is repeated across views before camera–action fusion.

C.1.3 Camera Encoder

Each 25-dimensional camera vector is encoded by the same MLP architecture:

$$25 \to 1024 \to 1024 \to 1024. \tag{26}$$

It likewise uses SiLU after the first two layers and no LayerNorm. The encoder is applied independently to every view and time step, producing $\tilde{c} \in \mathbb{R}^{V\times T\times 1024}$. The text and action branches share this camera encoder but use separate subsequent camera-fusion modules.

C.1.4 Image-Text Fusion block

The fusion block consists of two eight-head self-attention blocks. Each block has width 1024 and a feed-forward network $1024 \to 2048 \to 1024$ with GELU. The residual branches use pre-normalization with $\mathrm{LayerNorm}(1024, \epsilon = 10^{-6})$; the same pair of LayerNorm modules is reused by both blocks. The attention layer uses linear query, key, value, and output projections, RMSNorm with $\epsilon = 10^{-6}$ on its queries and keys, and one-dimensional rotary embeddings.

C.1.5 Camera Fusion Block

Both camera-fusion operators use an independent one-block attention module with hidden width $d = 1024$ and eight heads. In the action branch, the aligned action token and camera token are concatenated along the sequence dimension at each time step; the camera token queries the action token, as in Eq. 14. In the text branch, the current camera token $\tilde{c}_t$ is appended to the view-specific text–image sequence F_T and queries that sequence, as in Eq. 15. Only the camera token is updated, while the action or text–image tokens are passed through unchanged. The block uses pre-attention LayerNorm with $\epsilon = 10^{-6}$, RMSNorm on the query and key projections, linear value and output projections, and a zero-initialized channel-wise residual gate. It has no additional feed-forward network or rotary position embedding. The text and action branches use separate copies of this block.

C.2 Training Configuration

Table 5 summarizes the stage-wise training hyperparameters.

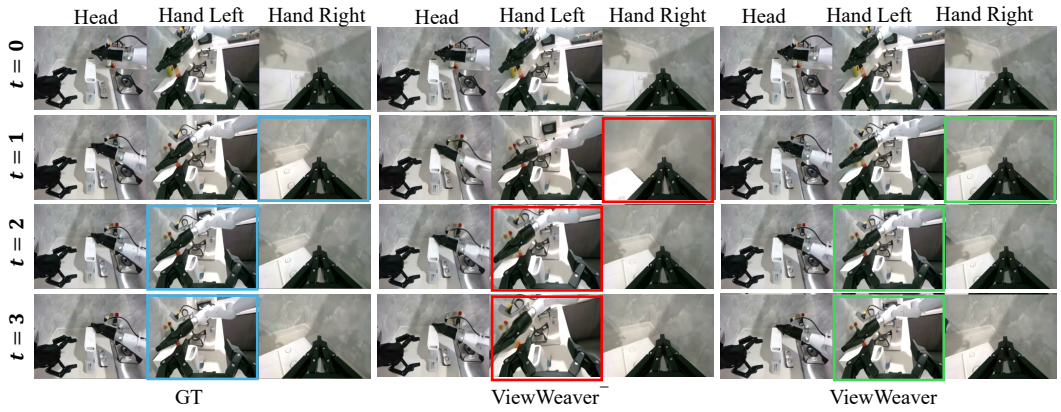

Figure 4: (Action-conditioned) Push the button of the capsule machine. ViewWeaver$^-$ misplaces the coffee machine at $t = 1$ and produces inconsistent right-arm motion across views at $t = 2$–3.

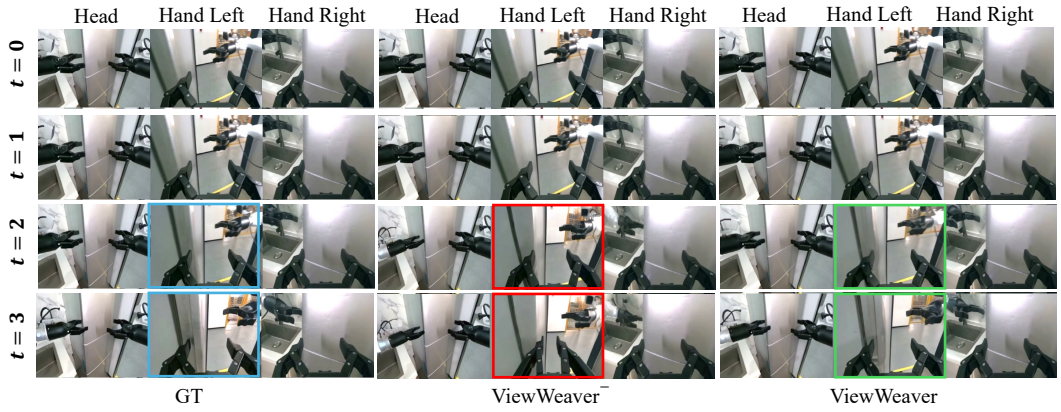

Figure 5: (Action-conditioned) Open the fridge. ViewWeaver$^-$ generates a Hand Left gripper closure that is inconsistent with the other views.

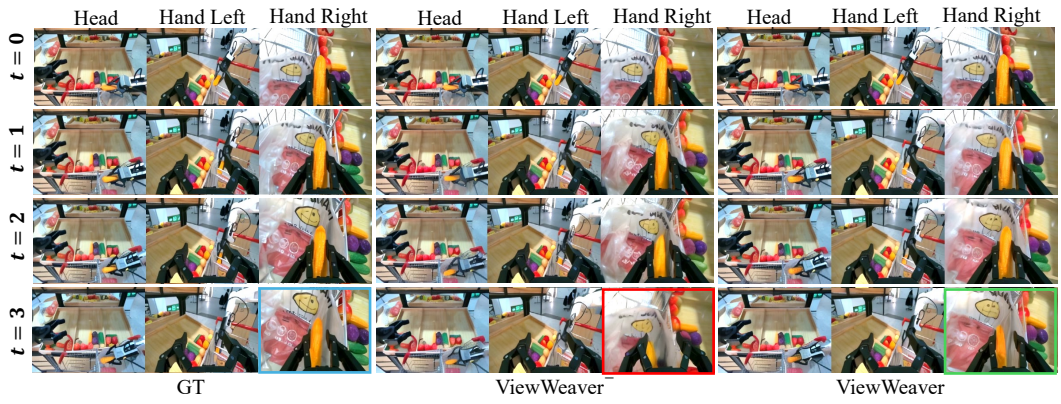

Figure 6: (Action-conditioned) Put the items in the shopping cart. ViewWeaver$^-$ predicts the item falling into the cart earlier in the Hand Left view than in the other views.

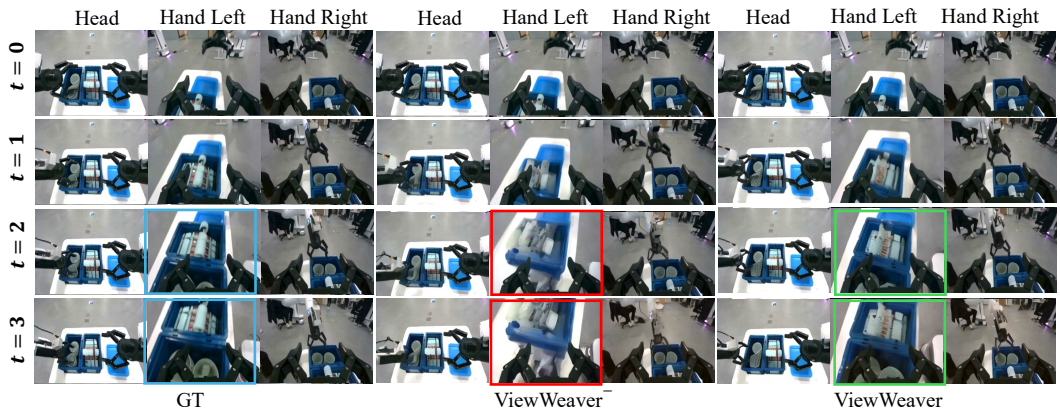

Figure 7: (Text-conditioned) Pick up the cup in the blue plastic box. ViewWeaver$^-$ produces blurred and hallucinated details for newly revealed objects as the Hand Left camera moves downward.

D More Visualizations

We provide additional qualitative comparisons on AgiBot-World Beta in Figures 4 to 7. All examples use the same observations and conditions, and compare the full ViewWeaver model with an ablated variant, denoted as ViewWeaver$^-$, that removes both VSMC and CM-CVA. The comparisons highlight the role of the two modules in preserving cross-view consistency in scene appearance and robot motion.

Figure 4 shows an action-conditioned example of pushing the button of a capsule coffee machine. At $t = 1$, ViewWeaver$^-$ misplaces the coffee machine in the Hand Right view relative to the ground truth (GT). Moreover, between $t = 2$ and $t = 3$, the right arm moves in the Hand Left view, while remaining nearly static in the other views. In contrast, ViewWeaver preserves the object location and the arm motion consistently across views. In Figure 5, the gripper in the Hand Left view of ViewWeaver$^-$ exhibits a closing behavior that is inconsistent with the corresponding gripper states in the other views, whereas ViewWeaver maintains synchronized gripper motion. Figure 6 presents another action-conditioned example in which an item falls into the shopping cart earlier in the Hand Left view of ViewWeaver$^-$ than in the other views. ViewWeaver synchronizes this event across all views. Finally, in the text-conditioned example in Figure 7, as the Hand Left camera moves downward, newly revealed objects become noticeably blurred and contain hallucinated details in ViewWeaver$^-$. ViewWeaver instead generates more plausible appearances.

E Codes

We provide the code for training and inference.

F Demo Videos

We provide eight demonstration videos on AgiBot-World and LIBERO-90, covering both action-conditioned and text-conditioned generation:

- (AgiBot-World, action-conditioned) The robot pours water into a container.
- (AgiBot-World, action-conditioned) The robot opens a refrigerator door.
- (AgiBot-World, text-conditioned) The robot opens a wardrobe in response to the instruction.
- (AgiBot-World, text-conditioned) The robot follows the instruction to push the button on a capsule coffee machine.
- (LIBERO-90, action-conditioned) The robot picks up the book in the middle and places it on a cabinet shelf.

- (LIBERO-90, action-conditioned) The robot places a white mug on a plate.
- (LIBERO-90, text-conditioned) The robot places a white mug on a plate following the instruction.
- (LIBERO-90, text-conditioned) The robot picks up the alphabet soup and places it in a basket following the instruction.

G FUTURE WORK

ViewWeaver's unified support for text and action conditioning makes it applicable to a broad range of downstream embodied tasks. Future work could explore two complementary directions. First, its text-conditioned generation capability could be extended toward a robot policy: combined with an inverse dynamics model (IDM), or augmented with an action-output module, the model could generate executable robot control signals. Second, its action-conditioned generation capability could be used as a world simulator for model-based reinforcement learning of VLA policies, providing simulated rollouts to support policy training. The same simulator could also serve as a tool for policy evaluation, enabling systematic assessment of candidate policies under diverse task conditions.