HORAE-WM: ACTION CONDITIONED RGBD KEYFRAME WORLD MODELING FOR LONG HORIZON ROBOT MANIPULATION

Anonymous authors
Paper under double-blind review

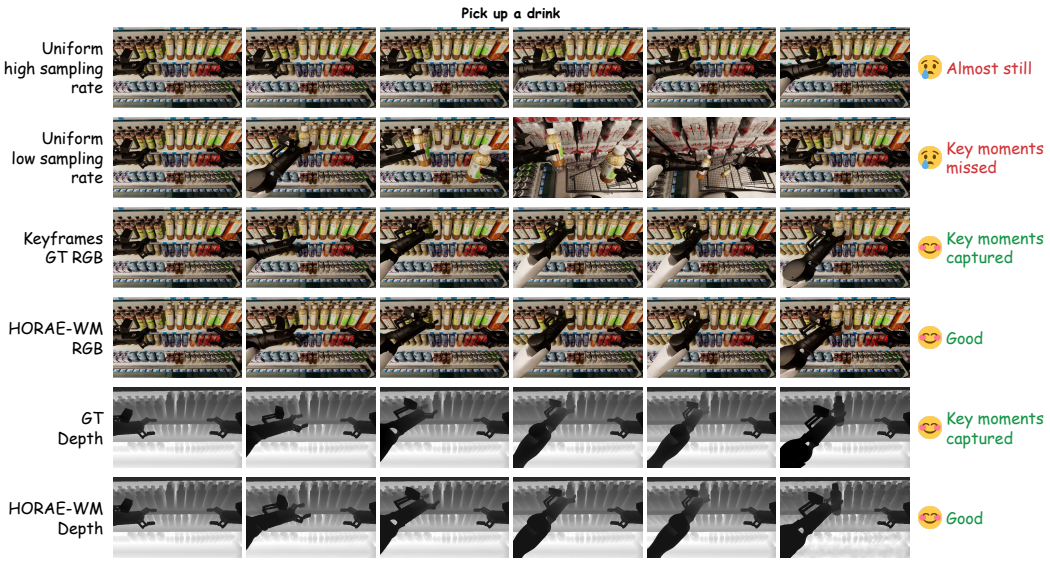


Figure 1: Comparison of uniform high-rate, uniform low-rate, and keyframe sampling on a long-horizon bottle-manipulation trajectory. HORAE-WM RGBD predictions are shown below.

ABSTRACT

Action conditioned video world models have made progress in robot manipulation, but with a limited number of output frames, uniform sampling often covers only part of a long-horizon manipulation trajectory, while larger sampling intervals can extend coverage at the cost of missing brief, critical interaction events. We introduce HORAE-WM, an RGBD keyframe world model designed to balance coverage of long manipulation trajectories with the capture of brief critical interactions under a limited output budget. An action-aware sampling selects key frames to be predicted based on arm motion and gripper state changes. Our curriculum progressively introduces prediction of condition videos representing robot actions, metric depth videos capturing scene geometry, and RGB videos depicting robot manipulation under frame-wise time and action condition, using sequences uniformly sampled at different frame rates, followed by training on keyframes. Keyframe prediction evaluations on simulated and real world manipulation show that HORAE-WM outperforms the compared baselines in visual fidelity, arm region consistency, and metric depth accuracy. Sampling analysis shows that our sampler achieves higher event coverage than uniform sampling while allocating more frame budget to events that are both task critical and challenging to capture.

1 INTRODUCTION

Recent advances in generative video modeling (Ho et al., 2022; Agarwal et al., 2025; 2026) have accelerated the development of action conditioned embodied world models, which explicitly gener-

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.

ate future visual observations from a current observation and a robot action trajectory (Guo et al., 2026; Gao et al., 2026; Zhu et al., 2025b; Yang et al., 2023; Jiang et al., 2025; Zhu et al., 2025a). Despite this progress, predicting frames far into the future remains challenging for robot video world models: in long horizon manipulation, doing so requires modeling the complete sequence of arm motions and interactions among the robot, manipulated objects, and the surrounding environment (Guo et al., 2026; Gao et al., 2026; Zhu et al., 2025b).

Figure 1 makes this long-horizon prediction challenge concrete on a bottle-manipulation trajectory. Uniform high rate sampling spends several frames on slight arm motion, whereas uniform low rate sampling covers a longer span but misses the critical intermediate states of the pick event. Instead our action-aware keyframes concentrate the budget on the approach, grasp, and post-grasp lifting states. This contrast arises from the fixed frame-rate design used by many action conditioned video world models (Gao et al., 2026; Zhu et al., 2025b): with a fixed output budget, higher rates preserve local detail but shorten temporal coverage, while lower rates extend coverage but risk missing brief interactions (Harvey et al., 2022; Jayaraman et al., 2018). Because manipulation trajectories combine redundant smooth phases with short task-critical transitions (Shi et al., 2023; Yuan et al., 2025; Jayaraman et al., 2018; Shridhar et al., 2023; James et al., 2022), uniform sampling cannot simultaneously preserve interaction detail and long-horizon coverage.

These observations motivate allocating the fixed frame budget according to action dynamics rather than uniform time intervals (Shi et al., 2023; Huang et al., 2025; Chen et al., 2025a). We therefore formulate long horizon robot world modeling as keyframe prediction conditioned on an initial observation and a complete future action trajectory. A data-side sampler defines non-uniform target times, placing frames sparsely in still or smooth-motion phases and densely around changes in arm motion and gripper state, while the world model predicts the corresponding video keyframes.

This formulation introduces a distinct challenge: unlike fixed-interval prediction, non-uniform schedules assign different physical-time intervals to adjacent keyframes, which cannot be inferred from frame indices alone (Harvey et al., 2022). Larger intervals also provide fewer local continuity cues (Harvey et al., 2022; Reda et al., 2022), requiring the model to generate keyframes that remain consistent with the action trajectory across varying temporal gaps.

To address this challenge, we introduce **HORAE-WM**, an action conditioned RGBD keyframe world model for long horizon robot manipulation. Given an initial RGBD observation, a complete action trajectory, and target times selected by an action-aware sampler, HORAE-WM predicts the corresponding future RGBD keyframes. Rather than relying only on ordinal frame indices and chunk-level frame rates (Agarwal et al., 2025), HORAE-WM uses frame-wise physical-time embeddings to condition each target frame and its aligned action. Because physical-time conditioning alone is insufficient for predicting across large keyframe gaps, we adopt a staged condition video–metric depth–RGB curriculum that progresses from action-aligned robot motion to scene geometry and visual appearance. We first train on sequences sampled at multiple fixed frame rates to expose the model to both short and long temporal gaps, and then continue training on dedicated non-uniform keyframe sequences to adapt it to action-derived target schedules.

We evaluate HORAE-WM on simulated and real-world dual-arm manipulation data. Under the common future-keyframe evaluation protocol, HORAE-WM achieves leading results in RGB fidelity, arm-region consistency, and metric depth accuracy among the compared world models. Ablation comparisons show gains from frame-wise physical-time conditioning, mixed-interval pretraining, and the staged condition video–metric depth–RGB curriculum, while sampling analysis shows that action-aware keyframes improve event coverage and allocate more selected frames to information-dense, task-critical interactions.

Our main contributions are summarized as follows:

- **Action conditioned RGBD keyframe modeling.** We formulate long-horizon robot world modeling as the prediction of action-conditioned RGBD keyframes under a fixed output budget, allowing critical motion and interaction events to be represented over long manipulation trajectories.
- **Frame-wise physical-time conditioning.** We condition each target keyframe and its aligned action on its physical time relative to initial observation, enabling prediction across non-uniform temporal gaps beyond ordinal frame indices and chunk-level frame rates.

- **Dual-axis progressive learning.** We progressively introduce condition video, metric depth, and RGB generation while transitioning from mixed fixed-interval sequences to dedicated non-uniform keyframe sequences.

2 Problem Formulation

We consider a robot manipulation trajectory of n frames recorded at source frame rate f_s, with an initial RGBD observation $\mathbf{o}_0$ and a complete future action trajectory $\mathbf{A}$. Given an output budget of K frames including the initial observation, an action-aware selector chooses keyframe indices $\mathcal{K} = \{k_i\}_{i=0}^{K-1}$, where $0 = k_0 < k_1 < \cdots < k_{K-1} = n-1$. Let $K_+ = K-1$ denote the number of future prediction frames. Their physical times relative to the initial observation are $\tau_i = (k_i - k_0)/f_s$, and the corresponding sampled actions are $\mathbf{A}_\mathcal{K} = (\mathbf{a}_{k_i})_{i=0}^{K-1}$. We denote the resulting time schedule by $\boldsymbol{\tau}_\mathcal{K}$. HORAE-WM predicts the K_+ future elements of three aligned keyframe sequences at times $\{\tau_i\}_{i=1}^{K-1}$: condition video $\mathbf{C}_{\mathcal{K}_+}$, metric depth $\mathbf{D}_{\mathcal{K}_+}$, and RGB $\mathbf{R}_{\mathcal{K}_+}$. A hat denotes the corresponding model prediction:

$$\left(\widehat{\mathbf{C}}_{\mathcal{K}_+}, \widehat{\mathbf{D}}_{\mathcal{K}_+}, \widehat{\mathbf{R}}_{\mathcal{K}_+}\right) = F_\theta(\mathbf{o}_0, \mathbf{A}_\mathcal{K}, \boldsymbol{\tau}_\mathcal{K}) \,. \tag{1}$$

3 Method

3.1 Action-Aware Keyframe Selection

We select K_+ future keyframes in addition to the initial observation, preserving action-critical events and allocating the remaining budget to capture the intervening action dynamics (Figure 2).

Action-critical anchors. From the recorded action trajectory, which includes both arm motion and gripper state, we identify segment endpoints, gripper events, and arm-motion boundaries. Gripper events mark the start and end of opening or closing, while arm-motion boundaries mark when an arm starts or stops moving. Segment endpoints and gripper events are selected exactly. Arm-motion boundaries instead require only nearby temporal coverage, since they mark the transition into or out of an arm-motion interval rather than a discrete gripper event.

Allocating the remaining frames. With the action-critical anchors fixed, we fill the remaining frame budget by adding action landmarks around changes in arm and gripper motion and sampling within large gaps to cover slowly varying phases. The resulting frames define a keyframe schedule that preserves the fixed budget and all anchors.

3.2 Frame-Wise Conditioned Tri-Modal World Model

We extend the DreamDojo backbone (Gao et al., 2026) to jointly predict condition video, metric depth, and RGB keyframes. Following the visual action representation of EVAC (Jiang et al., 2025), the condition video represents end-effector poses and gripper states in the camera view. It provides explicit motion targets alongside scene geometry and appearance.

Tri-modal latent representation. As shown in Figure 3(a), for each modality, a frozen video VAE encodes the target sequence into latent features, which are patchified and projected into the diffusion transformer(DiT) (Peebles & Xie, 2023) token space. The modality-specific token streams are arranged into a single sequence and processed by a Mixture-of-Transformers (MoT) design. Attention is shared across modalities, while feed-forward networks and output heads remain modality-specific (Figure 3). The MoT DiT block also contains a cross-attention layer that can receive text conditioning. During training, future latents are noised for flow prediction, while the initial RGBD observation and rendered initial condition frame are provided.

Progressive modality self attention. As shown in Figure 3 (b)(iii), a progressive mask controls self-attention among visual tokens: condition video queries attend only to condition video tokens, depth queries attend to condition video and depth tokens, and RGB queries attend to all three modalities. Within each permitted connection, attention is bidirectional over target times.

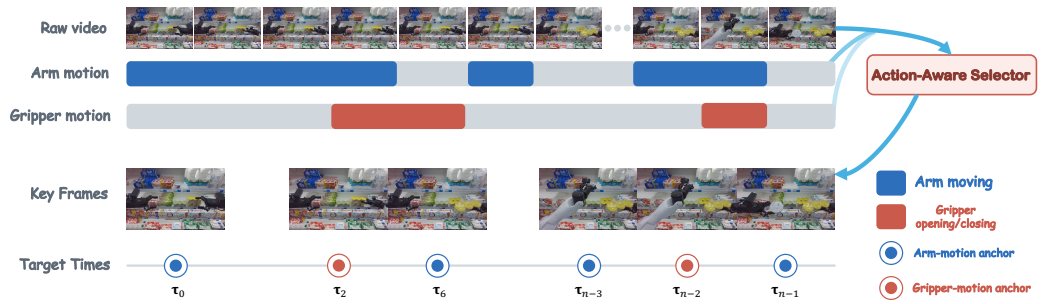


Figure 2: **Action-aware keyframe selection.** Non-uniform keyframes are selected around arm-motion and gripper events under a fixed frame budget. Blue and orange bars denote the two event types, circles mark their anchors, and selected frames retain their original physical times and actions.

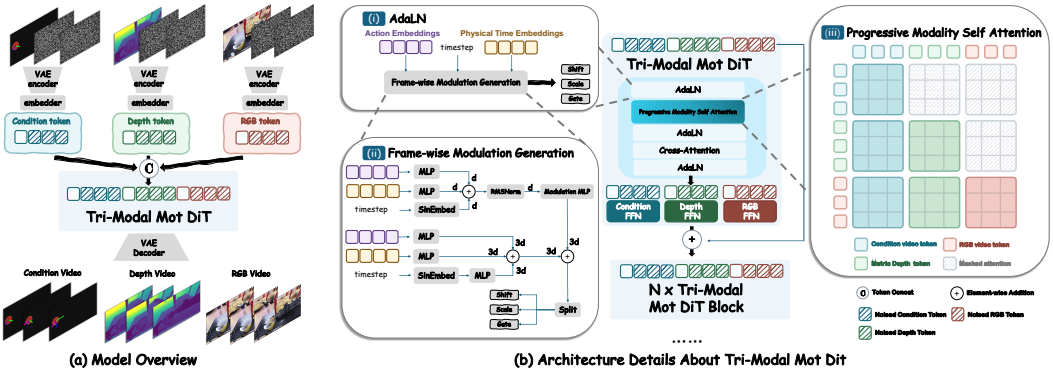


Figure 3: **Architecture of HORAE-WM.** (a) Model overview for Tri-Modal joint prediction. (b) Tri-modal MoT DiT details, including frame-wise modulation, cross-attention, modality-specific feed-forward networks, and progressive modality self-attention.

Frame-wise action and physical-time conditioning. As shown in Figure 3 (b)(i–ii), each latent-frame group is conditioned on the physical times of its target frames and their aligned actions. For a target at elapsed time τ_i (in seconds), we use

$$\gamma(\tau_i) = \left[\tau_i,\ \tau_i/s,\ \{\sin(2\pi\tau_i/p_j), \cos(2\pi\tau_i/p_j)\}_{j=1}^{J}\right], \tag{2}$$

where s is a normalization scale and p_j denotes the sinusoidal periods. The time and action features are projected by separate MLPs and combined with the diffusion-timestep embedding. A modulation MLP then maps the combined representation to the AdaLN shift, scale, and gate parameters used in each module Tri-Modal MoT DiT block. Following the backbone's action grouping, actions in each group are represented relative to the final action of the preceding group.

3.3 Multi-Stage Progressive Learning

Our multi-stage curriculum progresses along two complementary axes: modality and temporal sampling. Along the modality axis, training proceeds from condition video prediction to metric depth prediction and finally RGB video prediction, with each stage building on the preceding one. Along the temporal axis, we first train on a mixture of sequences sampled at different fixed temporal intervals, exposing the model to both fine-grained local changes and long physical-time horizons, and then continue training on sequences produced by our action-aware keyframe selector.

Modality progression. We first train the model to predict condition videos conditioned on robot actions. We then introduce metric depth and RGB prediction sequentially, retaining supervision for previously introduced modalities as the prediction task expands from robot motion geometry to scene geometry and visual appearance. At each transition, existing network parameters are carried

over from the preceding training phase, while newly introduced modality branches are initialized from the corresponding parameters of the original pretrained model.

Temporal progression. Throughout the modality curriculum, we train on a mixture of sequences sampled from the original videos at different fixed frame rate. High frequency sequences provide supervision for fine-grained local changes, while low frequency sequences span longer physical time horizons for the same number of output frames. After completing the modality curriculum, we continue training the tri-modal model on keyframe sequences constructed by our action aware keyframe selector (Section 3.1), adapting it to action driven keyframe prediction.

Training objective. At each curriculum stage, we optimize a latent-space rectified-flow objective. Let $\mathcal{M} \subseteq \{\text{cond}, \text{depth}, \text{rgb}\}$ denote the modalities trained at that stage. For each $m \in \mathcal{M}$, let $\hat{\mathbf{v}}_m$ be the predicted flow, $\mathbf{v}_m = \boldsymbol{\epsilon}_m - \mathbf{z}_m$ the target flow, and $\mathbf{e}_m = \hat{\mathbf{v}}_m - \mathbf{v}_m$ the prediction residual, where $\boldsymbol{\epsilon}_m$ is Gaussian noise and $\mathbf{z}_m$ is the clean latent. We combine flow-matching and temporal-difference supervision (Gao et al., 2026):

$$\mathcal{L} = \sum_{m \in \mathcal{M}} \lambda_m \left[\frac{\sum_p \bar{W}_{m,p}\, e_{m,p}^2}{Z_m} + \beta_m \frac{\sum_p \bar{W}^{\Delta}_{m,p} (\Delta \mathbf{e}_m)_p^2}{Z^{\Delta}_m} \right], \tag{3}$$

where $(\Delta \mathbf{e}_m)_\ell = \mathbf{e}_{m,\ell} - \mathbf{e}_{m,\ell-1}$. The index p denotes latent entries in the first term and adjacent-frame pairs in the second; Z_m and Z^{Δ}_m normalize the corresponding weighted sums, while λ_m and β_m balance modality and temporal-difference losses. Spatial and pairwise weights are described next.

Structure-aware spatial weighting. We use structure-aware spatial weights to emphasize robot action and scene changes while retaining uniform supervision elsewhere. The condition video loss uses a binary foreground mask generated directly during rendering, whereas depth and RGB share weights derived from arm, object, depth-edge, and temporal-change masks (Figure 4). For condition video, the corresponding weights are $\bar{W}_{\text{cond}} = \mathcal{N}(1 + \alpha_{\text{cond}} M_{\text{cond}})$. For depth and RGB, the shared weights are

$$\bar{W}_{\text{depth}} = \bar{W}_{\text{rgb}} = \mathcal{N}(1 + \alpha_{\text{obj}} M_{\text{obj}} + \alpha_{\text{arm}} M_{\text{arm}} + \alpha_{\text{edge}} M_{\text{edge}} + \alpha_{\text{motion}} M_{\text{motion}}) \,. \tag{4}$$

All masks are pooled to the VAE latent grid, and the resulting weights are normalized over active prediction positions. Pairwise weights combine the spatial weights of adjacent latent frames. These masks are used only to construct training-time weights and are not required at inference; their construction and coefficients are provided in Appendix A.2.



Figure 4: **Masks for loss weighting.** The RGB image shows a robot placing a tray on a table. Depth-edge (red), depth-motion (ochre), arm (blue), and object (green) masks jointly weight the depth and RGB losses. The rightmost condition foreground mask (purple) weights the condition video loss.

4 Experiments

4.1 Experimental Setup

Dataset. We use simulated and real-world dual-arm manipulation data spanning 16 tasks from AgiBot World 2026. Episodes are partitioned within each task, yielding 7,669 training and 285 test episodes. We construct sequences with different fixed sampling frame rates and action-aware keyframe sequences, each sequence contains $K = 81$ frames, including the initial observation. Simulation uses native depth, while the depth real data is obtained by fusing sensor depth with Depth Anything 3 (DA3) estimates (Lin et al., 2025). Appendix A.1 provides more details.

Training. We initialize HORAE-WM from the pretrained DreamDojo-2B model (Gao et al., 2026) and train at 400×640 resolution. For the full training process, we use 64 PPU-ZW810E accelerators for approximately twelve days. For the main comparison, we use the checkpoint obtained after ten epochs of keyframe training. Training and inference details are provided in Appendix A.2.

Baselines and Evaluation. We compare HORAE-WM against GE-Sim 2.0 (Qiu et al., 2026), EVAC (Jiang et al., 2025), Cosmos3-Nano (Agarwal et al., 2026), OSCAR (Wu & Gao, 2026), and DreamDojo-14B (Gao et al., 2026) using their publicly released checkpoints. Each model is conditioned on the initial observation and the corresponding action trajectory, with model-specific input adapters where necessary, and evaluated on a common set of future target keyframes from 1,242 test sequences. We measure RGB video quality using PSNR (Hore & Ziou, 2010), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Arm Mask-IoU quantifies the overlap between robot-arm regions in predicted and ground truth videos. Metric depth accuracy is measured by L1 error in millimeters and absolute relative error (AbsRel). More Details are provided in Appendix A.3.

4.2 Comparison with Existing World Models

Quantitative analysis. Under the baseline and evaluation protocol defined in Section 4.1, we compare HORAE-WM with the above models on the same set of future keyframe targets, as reported in Table 1. HORAE-WM achieves the best results on all reported metrics. Compared with GE-Sim 2.0, the strongest external baseline on the RGB and arm-region metrics, HORAE-WM improves PSNR by 1.992 dB, reduces LPIPS from 0.2832 to 0.1934, and increases Arm Mask-IoU from 0.5348 to 0.6780, indicating stronger prediction of visual quality and action-relevant regions. For metric depth, HORAE-WM achieves an L1 error of 83.4 mm and an AbsRel of 0.0864, both lower than the best results among the baseline models. To provide a source-matched comparison, we additionally apply the same frozen DA3 estimator to HORAE-WM's predicted RGB keyframes. This yields an L1 error of 263.8 mm and an AbsRel of 0.2242, which are still better than the best DA3-derived baseline values of 300.3 mm and 0.2748, respectively. Thus, HORAE-WM also achieves the best depth results under the common RGB-to-DA3 protocol. Overall, under the keyframe evaluation protocol, HORAE-WM better preserves visual fidelity and arm-region consistency in future frames while providing more accurate scene-geometry predictions.

Table 1: **Future keyframe prediction quality.** Green and amber cells denote the best and second-best values among the compared methods, respectively.

Model	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑	L1 (mm) ↓	AbsRel ↓
GE-Sim 2.0	**17.008**	**0.5873**	**0.2832**	**0.5348**	300.7	**0.2748**
EVAC	16.167	0.5085	0.4353	0.4873	377.3	0.3234
Cosmos3-Nano	15.791	0.4836	0.3059	0.4439	**300.3**	0.2922
OSCAR	13.441	0.3133	0.4918	0.4798	384.6	0.3318
DreamDojo-14B	14.247	0.3757	0.4482	0.3304	423.9	0.3568
HORAE-WM	**19.000**	**0.6071**	**0.1934**	**0.6780**	**83.4**	**0.0864**

Qualitative analysis. The quantitative improvements in visual fidelity, arm-region consistency, and scene-geometry prediction are illustrated qualitatively in Figure 5. In this freezer-lid sliding example, HORAE-WM preserves the arm shape and pose in RGB and yields better-aligned depth, whereas GE-Sim 2.0 exhibits arm deformation and pose mismatches in RGB and its DA3-estimated depth.

User study. We further compare HORAE-WM with GE-Sim 2.0 and EVAC through a human preference study. Each questionnaire contains one question for each of 20 randomly selected test cases, and 20 completed questionnaires yield 400 judgments. Participants included students and professional researchers with diverse, cross-disciplinary backgrounds in generative modeling, embodied intelligence, and large language models. For each case, respondents view anonymized predictions from all three models alongside the ground truth video and select the best overall prediction based on visual quality, arm motion, and object states. Both question order and model display order

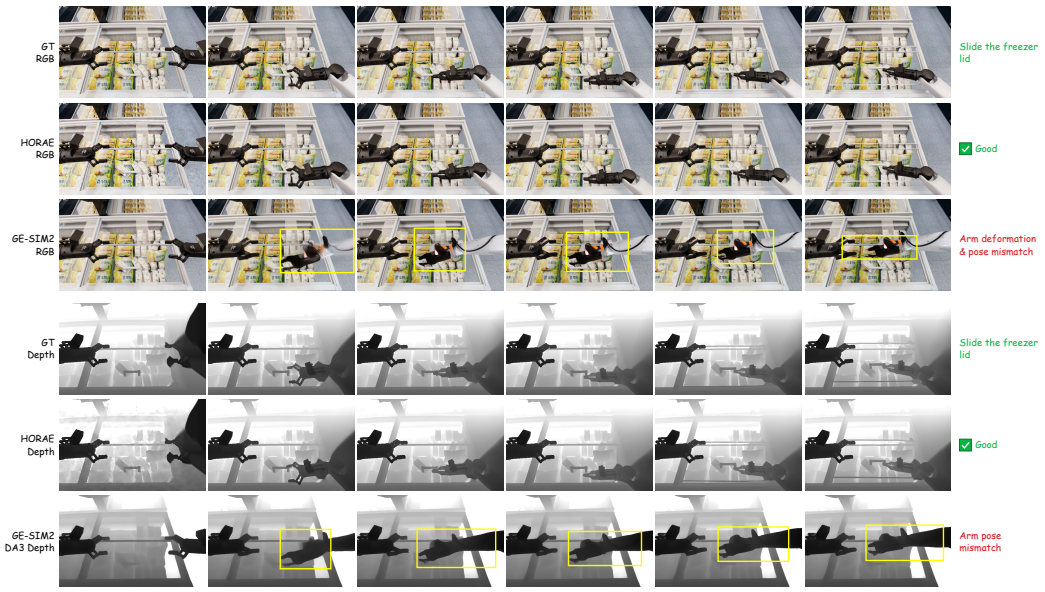


Figure 5: **Qualitative comparison of RGBD keyframe predictions.** HORAE-WM predicts depth directly; GE-Sim 2.0 depth is estimated from its generated RGB using DA3. Yellow boxes highlight arm deformation and pose mismatches.

are randomized within each questionnaire to reduce presentation-order bias. As shown in Figure 6, HORAE-WM receives 84% of the selections, compared with 11.5% for EVAC and 4.5% for GE-Sim 2.0. These results complement the quantitative evaluation, showing that HORAE-WM's predictions are also preferred by human observers for their overall quality.

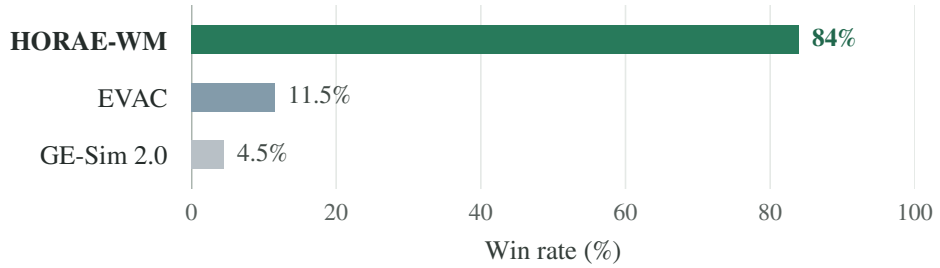


Figure 6: **Human preference on randomly selected cases.**

4.3 Action-Aware Keyframe Compression

We evaluate how action-aware sampling allocates a fixed frame budget across a long manipulation trajectory. On 100 real world test segments, we compare our sampler with uniform sampling, with both selecting 81 original RGB frames over the same physical duration and retaining both endpoints. We annotate manipulation events from the full RGB videos and assess both their coverage and the allocation of frames within them. Event coverage C is the fraction of annotated events containing at least one selected frame. Enrichment H measures the fraction of selected frames within information-dense events relative to the fraction of time occupied by those events. We additionally report the mean per-segment product CH and evaluate how completely descriptions generated from the selected frames capture the reference manipulation sequence on a 0–5 scale. The annotation and evaluation protocols are detailed in Appendix B.

As shown in Table 2, our sampler improves event coverage from 96.86% to 98.26% and enrichment from 0.997 to 1.223, indicating that it captures more events while allocating more frames to information-dense changes. The higher joint score CH and description completeness further suggest that the selected frames preserve the manipulation process more completely. Together, these results show that action-aware sampling achieves broader event coverage while allocating more of its fixed frame budget to events that are both task-critical and challenging to capture.

Table 2: **Manipulation information retained under a fixed frame budget.** Both samplers select 81 RGB frames over the same physical duration. Bold indicates the best result in each column.

Selector	Coverage (%) ↑	Enrichment ↑	CH ↑	Completeness ↑
Uniform	96.86	0.997	0.963	4.45
Ours	**98.26**	**1.223**	**1.205**	**4.56**

4.4 Ablation Studies

We examine the effects of frame-wise physical-time conditioning, pretraining on mixed fixed-interval sequences, structure-aware spatial weighting, and modality progression on keyframe prediction. All variants are evaluated on the same keyframe test set, with comparisons made within each ablation group to assess the corresponding design. Further details of the training configurations and ablation settings are provided in Appendix A.4.

Table 3: **Component ablations on keyframe prediction.** Bold values indicate the best results within each ablation group. In (c), arm and object weights are added on top of the depth-edge and depth-motion weights.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑
(a) Frame-wise time: eight keyframe epochs				
Without frame-wise time	18.142	0.5853	0.2172	0.6308
With frame-wise time	**18.419**	**0.5965**	**0.2148**	**0.6413**
(b) Mixed-interval prior: four keyframe epochs				
No mixed-interval pretraining	17.301	0.5529	0.2632	0.5889
Two-epoch mixed-interval pretraining	**18.117**	**0.5740**	**0.2280**	**0.6370**
(c) Spatial weighting: two keyframe epochs				
Uniform spatial weights	17.240	0.5602	0.2585	0.5380
+ Edge and motion weights	17.446	0.5735	0.2381	0.5510
+ Arm and object weights	**17.459**	**0.5769**	**0.2356**	**0.5596**

Frame-wise physical-time conditioning. As shown in Table 3 (a), adding frame-wise physical-time conditioning improves all reported metrics. PSNR increases from 18.142 to 18.419 dB and Arm Mask-IoU from 0.6308 to 0.6413, alongside higher SSIM and a small reduction in LPIPS. These results support the use of each target frame's physical time as an additional condition alongside actions to improve visual quality and arm-region consistency in non-uniform keyframe prediction.

Pretraining on mixed fixed-interval sequences. As shown in Table 3 (b), with the same keyframe training settings, adding pretraining on mixed fixed-interval sequences increases PSNR from 17.301 to 18.117 dB, reduces LPIPS from 0.2632 to 0.2280, raises Arm Mask-IoU from 0.5889 to 0.6370, and SSIM also improves. These results indicate that this pretraining provides a useful initialization for subsequent keyframe prediction, improving visual quality and arm-region consistency and supporting our temporal progression from mixed fixed-interval sequences to non-uniform keyframes.

Structure-aware spatial weighting. As shown in Table 3 (c), adding depth-edge and depth-motion weights improves all four metrics over uniform spatial weighting. Further incorporating

arm and object weights yields additional gains across all four metrics. These results support emphasizing both geometric changes and manipulation-relevant regions to improve visual quality and arm-region consistency in keyframe prediction.

Table 4: **Effects of modality progression on keyframe prediction.** C, D, and R denote condition video, metric depth, and RGB prediction, respectively. $+$ denotes joint prediction within a stage, and $\rightarrow$ denotes continued training from the preceding stage. Green and amber cells indicate the best and second-best values for each metric.

Training sequence	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑
R	16.875	0.5387	0.2907	**0.5552**
$C \rightarrow C{+}R$	**16.962**	0.5520	0.2536	0.5496
$C \rightarrow C{+}D \rightarrow C{+}D{+}R$	**17.388**	**0.5625**	**0.2383**	**0.5742**
$C{+}D{+}R$ (direct)	16.958	**0.5525**	**0.2520**	0.5364

Modality progression. As shown in Table 4, progressively introducing condition video, metric depth, and RGB prediction achieves the best results on all four metrics. With the same training budget for the final RGB stage, this curriculum improves PSNR from 16.958 to 17.388 dB, SSIM from 0.5525 to 0.5625, LPIPS from 0.2520 to 0.2383, and Arm Mask-IoU from 0.5364 to 0.5742 compared with directly training the same tri-modal architecture. These results support progressively learning robot motion and scene geometry before introducing RGB prediction to improve keyframe generation.

5 Related Work

Action-conditioned robot video world models predict future observations from robot actions, using direct action conditioning, rendered spatial controls, or geometry-aware representations to improve action alignment and scene modeling (Zhu et al., 2025b; Gao et al., 2026; Guo et al., 2026; Jiang et al., 2025; Chen et al., 2026; Zhen et al., 2025). Keyframe-based and variable-rate video-generation methods use sparse temporal scaffolds, endpoint-conditioned interpolation, or motion-aware frame-rate adaptation to extend temporal coverage and reduce redundant prediction (Huang et al., 2025; Chen et al., 2025a; Wang et al., 2025; Yuan et al., 2025). Together, these directions leave open how to allocate a fixed output budget across non-uniform physical times determined by robot motion and gripper events. A fuller discussion of these related directions is provided in Appendix C.

6 Limitations

Our evaluation focuses on predicting future observations from given action trajectories; benefits for downstream policy learning and closed-loop control remain to be evaluated. Future work will explore transferring representations and dynamics priors learned through world-model training to policy learning, and extending HORAE-WM to jointly model observations and actions as a world-action model. Inference currently relies on multi-step sampling. We plan to investigate systems-level optimization and few-step sampling to reduce latency and improve throughput while preserving prediction quality, and to assess the model's utility as a simulator for reinforcement learning and a data engine for large-scale synthesis of high-quality training data.

7 Conclusion

We introduced HORAE-WM, an action-conditioned world model that predicts non-uniform RGBD keyframes from an initial observation and an action trajectory within a fixed frame budget. The method combines action-aware keyframe selection, frame-wise physical-time conditioning, and progressive learning across modalities and temporal sampling regimes. Experiments on simulated and real-world dual-arm manipulation show that HORAE-WM outperforms the evaluated baselines in visual fidelity, arm-region consistency, and metric-depth accuracy, while also supporting dense video reconstruction and uniform frame prediction at fixed frame rates.

AI USE STATEMENT

We used generative AI tools for manuscript drafting and polishing, literature retrieval and discovery, and research ideation and execution. Research assistance included part of code development and debugging. Model-assisted annotation and evaluation for keyframe compression are described in Appendix B.

ETHICS STATEMENT

This work studies action-conditioned RGBD prediction using existing open-source robot manipulation data. Participants in our video-based preference study were informed about the study and took part voluntarily; they evaluated anonymized model predictions rather than operating physical robots.

REPRODUCIBILITY STATEMENT

Section 3 describes the action-aware sampler, frame-wise conditioning, and progressive training objectives. Appendices A.1–A.3 provide data construction, episode-disjoint splits, training and checkpoint-selection settings, metric computation, and baseline adaptation. Appendix A.4 specifies the ablation controls, while Appendix A.5 details uniform-frame inference and dense-video reconstruction. Appendix B documents the keyframe-compression protocol, event criteria, and complete caption-scoring prompt; the human preference study protocol is described in Section 4.2. All training and evaluation sequences are partitioned by episode, so no episode is shared across the two splits. The complete training and evaluation code will be publicly released to facilitate reproduction.

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.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

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.

Jiahao Chen, Hangjie Yuan, Yichen Qian, Jingyun Liang, Jiazheng Xing, Pengwei Liu, Weihua Chen, Fan Wang, and Bing Su. Lumosflow: Motion-guided long video generation. *arXiv preprint arXiv:2506.02497*, 2025a.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025b.

Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, et al. Bridgev2w: Bridging video generation models to embodied world models via embodiment masks. *arXiv preprint arXiv:2602.03793*, 2026.

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.

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.

Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 6121–6138, 2026.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in neural information processing systems*, 35: 27953–27965, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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

Hsin-Ping Huang, Yu-Chuan Su, and Ming-Hsuan Yang. Generating long-take videos via effective keyframes and guidance. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3709–3720. IEEE, 2025.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European conference on computer vision*, pp. 624–642. Springer, 2022.

Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13729–13738. IEEE, 2022.

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.

Dinesh Jayaraman, Frederik Ebert, Alexei A Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. *arXiv preprint arXiv:1808.07784*, 2018.

Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Byungjun Kim, Taeksoo Kim, Hyunsoo Cha, and Hanbyul Joo. Robot-factored world models via robot rendering. *arXiv preprint arXiv:2607.22535*, 2026.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 88446–88463, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Boxiang Qiu, Liliang Chen, Yue Liao, Nan Wang, Lintao Wang, Jiayi Luo, Wenzhi Zhao, Shengcong Chen, Di Chen, Ye Li, et al. Ge-sim 2.0: A roadmap towards comprehensive closed-loop video world simulators for robotic manipulation. *arXiv preprint arXiv:2605.27491*, 2026.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pp. 250–266. Springer, 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. *AI Open*, 7:208–226, 2026.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, 38:95597–95621, 2026.

Lucy Xiaoyang Shi, Archit Sharma, Tony Z Zhao, and Chelsea Finn. Waypoint-based imitation learning for robotic manipulation. *arXiv preprint arXiv:2307.14326*, 2023.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steve Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In *International Conference on Learning Representations*, volume 2025, pp. 30594–30609, 2025.

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.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Zhuoyuan Wu and Jun Gao. Oscar: Omni-embodiment action-conditioned world model for robotics. *arXiv preprint arXiv:2606.04463*, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Zhihang Yuan, Rui Xie, Yuzhang Shang, Hanling Zhang, Siyuan Wang, Shengen Yan, Guohao Dai, and Yu Wang. Dlfr-gen: Diffusion-based video generation with dynamic latent frame rate. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16410–16419. IEEE, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

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.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025a.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844. IEEE, 2025b.

A Experimental Details

A.1 Data Preprocessing and Sequence Construction

Observations and actions. We use head-camera RGBD observations and aligned actions from 16 simulated and real-world dual-arm manipulation tasks in AgiBot World 2026. We retain trajectories with the 40-dimensional simulation and real-robot action schemas and exclude the real-robot 44-dimensional schema. For model conditioning, we select the two gripper channels and 14 arm-joint channels, yielding a 16-dimensional action vector. Arm coordinates are normalized using training-set statistics, and signed gripper commands are converted to opening magnitudes. Following the backbone's action grouping, four target actions are expressed relative to the preceding group's final action.

Segment filtering and episode partitioning. We extract contiguous segments that pass body-motion and camera-stability filtering and construct sequences within these segments. Episodes are partitioned within each task, yielding 7,669 training and 285 test episodes; segments from the same episode remain in the same split. The partition targets approximately 5% of fixed-interval chunks for testing, rather than an exact 5% of episodes. The same episode assignments apply to every interval and to keyframe data. The primary test mean follows the observed real/simulation composition rather than assigning equal weight to domains or tasks.

Temporal sampling. The original recordings are sampled at 30 Hz. For mixed-interval pretraining, we construct sequences with fixed intervals of 2, 4, 8, or 16 original frames. The interval is constant within each sequence, and the resulting sequences are combined for training. At the source rate of 30 Hz, an interval of Δ frames corresponds to $\Delta/30$ seconds between adjacent targets, yielding effective sampling rates of 15, 7.5, 3.75, and 1.875 Hz for $\Delta \in \{2, 4, 8, 16\}$, respectively. For keyframe training, we apply the action-aware selector described in Section 3.1 to obtain non-uniform keyframe indices. Both sampling regimes use 81-frame sequences, including the initial observation, and all modalities are sampled at the same original frame indices. Table 5 reports the training sequence counts by sampling regime. Shorter intervals produce more fixed-length chunks, so mixed-interval pretraining is not uniformly balanced over intervals.

Depth and condition-video supervision. Simulated depth is used directly. Real-world depth targets are constructed offline by fusing sensor depth with Depth Anything 3 (DA3) estimates (Lin et al., 2025). The real-depth reference is consequently a fused supervision target, not an independent noise-free sensor measurement. Condition videos are rendered from robot kinematics and episode-specific camera calibration.

Table 5: **Training sequence counts.** All sequences contain 81 frames; intervals are measured in original 30 Hz frames.

Sampling regime	Training sequences
Fixed interval 2	91,419
Fixed interval 4	37,583
Fixed interval 8	12,578
Fixed interval 16	4,854
All fixed intervals	146,434
Non-uniform keyframes	24,843

A.2 Training Configuration

Initialization and curriculum. HORAE-WM is initialized from the pretrained DreamDojo-2B model (Gao et al., 2026) and trained at 400×640 resolution. On mixed fixed-interval sequences, we first train condition-video prediction for six epochs, then add depth and RGB prediction for two epochs each, retaining supervision for previously introduced modalities. We subsequently continue tri-modal training on non-uniform keyframe sequences. Table 6 summarizes the training budgets.

Table 6: **Training schedule of HORAE-WM.** C, D, and R denote condition-video, metric-depth, and RGB prediction; $+$ denotes joint prediction. Rows follow the training order, with each subsequent row continuing from the preceding one. Updates count optimizer steps within each row.

Targets	Data	Epochs	Effective batch	Updates
C	Mixed intervals	6	512	1,716
$C{+}D$	Mixed intervals	2	256	1,144
$C{+}D{+}R$	Mixed intervals	2	256	1,144
$C{+}D{+}R$	Keyframes	10	256	970

Keyframe optimization. Keyframe training uses 32 accelerators, a per-device batch size of two, and four gradient-accumulation steps, giving an effective batch size of 256. AdamW uses betas $(0.9, 0.99)$, weight decay 0.1, an initial learning rate of 4.3158373×10^{-5}, 13 warmup updates, and a 50,000-update scheduler cycle. The optimizer is reinitialized when entering keyframe training, while model weights are inherited from the completed tri-modal fixed-interval stage.

Loss implementation. The rectified-flow target is defined along the noise-interpolation coordinate, not physical time. The prediction residual at the observed initial latent is masked to zero before temporal differencing; the first observed-to-future pair remains included. For condition video, Z_m and Z_m^{Δ} are the active spatial and pair-weight sums; depth and RGB use the corresponding grid-element counts.

Spatial-weight construction. The condition-video mask is generated directly during rendering: background pixels are set to zero, while all rendered foreground pixels share the same binary value, without distinguishing foreground regions by color. For depth and RGB, arm and object masks are obtained with SAM 3 (Carion et al., 2026). Depth-boundary masks are derived from Sobel gradients of clipped log depth, and temporal-change masks from absolute differences of log depth between adjacent targets. The boundary and temporal-change masks use a per-sequence quantile q for thresholding. All masks are pooled onto the VAE latent grid before spatial weighting. For adjacent-latent pairs, $\bar{W}_m^{\Delta}$ takes the elementwise maximum of the two spatial weights for condition video and their mean for depth and RGB.

Loss coefficients. For tri-modal training, we set $\lambda_{\text{cond}} = \lambda_{\text{depth}} = 1$, $\lambda_{\text{rgb}} = 2$, and $\beta_m = 0.1$ for each modality. The spatial coefficients are $\alpha_{\text{cond}} = 7$, $\alpha_{\text{obj}} = 0.3$, $\alpha_{\text{arm}} = 3$, $\alpha_{\text{edge}} = 0.25$, and $\alpha_{\text{motion}} = 0.5$. Edge and motion masks use a per-sequence quantile of $q = 0.95$, and the resulting spatial weights are normalized over active prediction positions. Metric depth is log-encoded over

$[0.302, 6.144]$ meters. In the condition–depth stage, the retained condition objective has weight 0.3 and the depth objective has weight one.

A.3 Evaluation and Baseline Adaptation

Evaluation set and aggregation. We evaluate a fixed set of 1,242 test sequences from 285 held-out episodes, comprising 1,162 real-world and 80 simulated sequences. Each sequence contains one initial observation and 80 future targets at the prescribed original frame indices. The initial observation is excluded from every metric. We compute metrics for each future frame, average within each sequence, and then average equally over sequences. This aggregation does not reweight frames by their physical-time intervals or balance tasks and domains.

Arm Mask-IoU. We compute the intersection over union of the combined left- and right-arm masks in each prediction and reference frame. Both masks use the same frozen SAM 3 extraction procedure, the prompt "robot arm," and the same per-arm bounding boxes projected from reference robot kinematics at the target timestamp.

Depth evaluation. Depth references are the simulation depth and fused real-world depth described in Appendix A.1. For each frame, L1 is the mean absolute depth error in millimeters, and AbsRel averages the per-pixel absolute error normalized by the corresponding reference depth. Both metrics use pixels with finite predictions and references and reference depths strictly between 0.1 and 10 meters, restricted to the evaluated image region.

Baseline adaptation. Each baseline is conditioned on the initial observation and the action trajectory at the selected keyframes, with actions converted to the model's required input representation. Since the baselines do not natively support frame-wise physical-time conditioning, we use their original FPS interfaces to provide chunk-wise temporal conditioning. Specifically, we compute a single effective frame rate as the number of keyframe intervals divided by the physical duration spanned by the full selected-keyframe trajectory. This provides the trajectory's average temporal spacing rather than its individual, potentially non-uniform keyframe intervals. Predicted frames are evaluated against the corresponding future keyframe references.

A.4 Ablation Controls and Additional Results

The ablations are separately trained models with an effective batch size of 64. The temporal-conditioning experiment uses eight epochs per variant; the mixed-interval experiment matches four epochs of keyframe training but adds two fixed-interval pretraining epochs to one variant.

The loss experiment uses two epochs of keyframe training, without frame-wise time conditioning. The structure-only variant uses $\alpha_{\text{edge}} = \alpha_{\text{motion}} = 0.65$; the full variant additionally uses $\alpha_{\text{obj}} = 0.3$ and $\alpha_{\text{arm}} = 1$. These differ from the main training coefficients. All modality variants train only on keyframes, without mixed fixed-interval pretraining, using 32 devices and the same physical-time branch. Each stage lasts two epochs (776 optimizer updates), giving 2, 4, 6, and 2 total epochs in the row order of Table 4. The full curriculum and direct tri-modal control share the final architecture, objective, and two-epoch RGB stage.

Table 7: **DA3-derived depth diagnostics for all ablations.** Every row estimates depth from generated RGB with the same frozen DA3 model. These are not native depth-stream metrics.

Variant	L1 (mm) ↓	AbsRel ↓
(a) Frame-wise time: eight keyframe epochs		
Without frame-wise time	276.8	0.2373
With frame-wise time	**273.7**	**0.2362**
(b) Mixed-interval prior: four keyframe epochs		
No mixed-interval pretraining	317.8	0.2662
Two-epoch mixed-interval pretraining	**287.8**	**0.2430**
(c) Spatial weighting: two keyframe epochs		
Uniform spatial weights	309.9	0.2699
+ Edge and motion weights	287.4	0.2577
+ Arm and object weights	**287.0**	**0.2555**
(d) Modality curriculum		
R	332.7	0.2797
$C \rightarrow C{+}R$	291.6	0.2572
$C \rightarrow C{+}D \rightarrow C{+}D{+}R$	**282.2**	**0.2499**
$C{+}D{+}R$ (direct)	289.2	0.2608

Table 7 reports DA3-derived depth diagnostics for each ablation group. Consistent with the trends observed in the original RGB-based metrics, each proposed component yields lower depth L1 and AbsRel errors, with particularly clear gains from mixed-interval pretraining and the modality curriculum. Although these values are inferred from generated RGB rather than native depth predictions, the consistent improvements suggest that our design choices also enhance the geometric fidelity of the generated observations.

Table 8 breaks down HORAE-WM's performance on real and simulated sequences, using its native depth predictions. PSNR is similar across the two domains, while real-world sequences achieve higher Arm Mask-IoU (0.6833 versus 0.6016). Simulation yields higher SSIM, lower LPIPS, and lower depth L1, but higher AbsRel. The real-world results thus retain stronger arm-region alignment despite lower SSIM and higher LPIPS, while the opposite L1 and AbsRel rankings highlight the distinction between absolute and relative depth accuracy. Since real-world sequences make up most of the test set, the aggregate results primarily reflect real-world performance.

Table 8: **HORAE-WM by domain.** The primary result weights sequences equally; this breakdown makes the imbalanced real/simulation composition explicit. Depth is the model's native output.

Domain	Sequences	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑	L1 (mm) ↓	AbsRel ↓
Real	1,162	19.015	0.5954	0.1972	0.6833	85.2	0.0850
Simulation	80	18.784	0.7776	0.1385	0.6016	57.8	0.1069

A.5 Temporal Sampling and Dense Video Reconstruction

Dense video reconstruction. We reconstruct dense videos from each model's predicted keyframes using the same interpolation method and evaluate all future frames. Depth is estimated from each model's reconstructed RGB using the same frozen DA3. As shown in Table 9, HORAE-WM achieves the best results on all six metrics, with clear improvements over the second-best result for each metric. These results indicate that HORAE-WM's predicted keyframes effectively support dense video reconstruction, extending their advantages in visual fidelity and arm-region consistency to full future video sequences.

Table 9: **Dense video reconstruction from predicted keyframes.** Green and amber cells denote the best and second-best values, respectively.

Model	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑	L1 (mm) ↓	AbsRel ↓
GE-Sim 2.0	16.290	**0.5728**	**0.3162**	0.4668	**313.0**	**0.2981**
EVAC	**16.293**	0.5147	0.4351	**0.4921**	355.1	0.3135
HORAE-WM	**19.190**	**0.6144**	**0.1932**	**0.6815**	**264.9**	**0.2307**

Uniform frame prediction. We evaluate action-conditioned prediction at different frame rates using our key frame prediction model directly, without further training. As shown in Table 10, HORAE-WM outperforms GE-Sim 2.0 and EVAC on all six metrics at each of the four frame rates. Together with the keyframe results in Section 4.2, these findings show that HORAE-WM supports action-conditioned prediction under both non-uniform keyframe sampling and uniform schedules, with strong visual fidelity, arm-region consistency, and metric depth accuracy.

Table 10: **Prediction at different frame rates.** Horizon denotes the predicted video's physical time span. Green and amber mark the best and second-best values per frame rate.

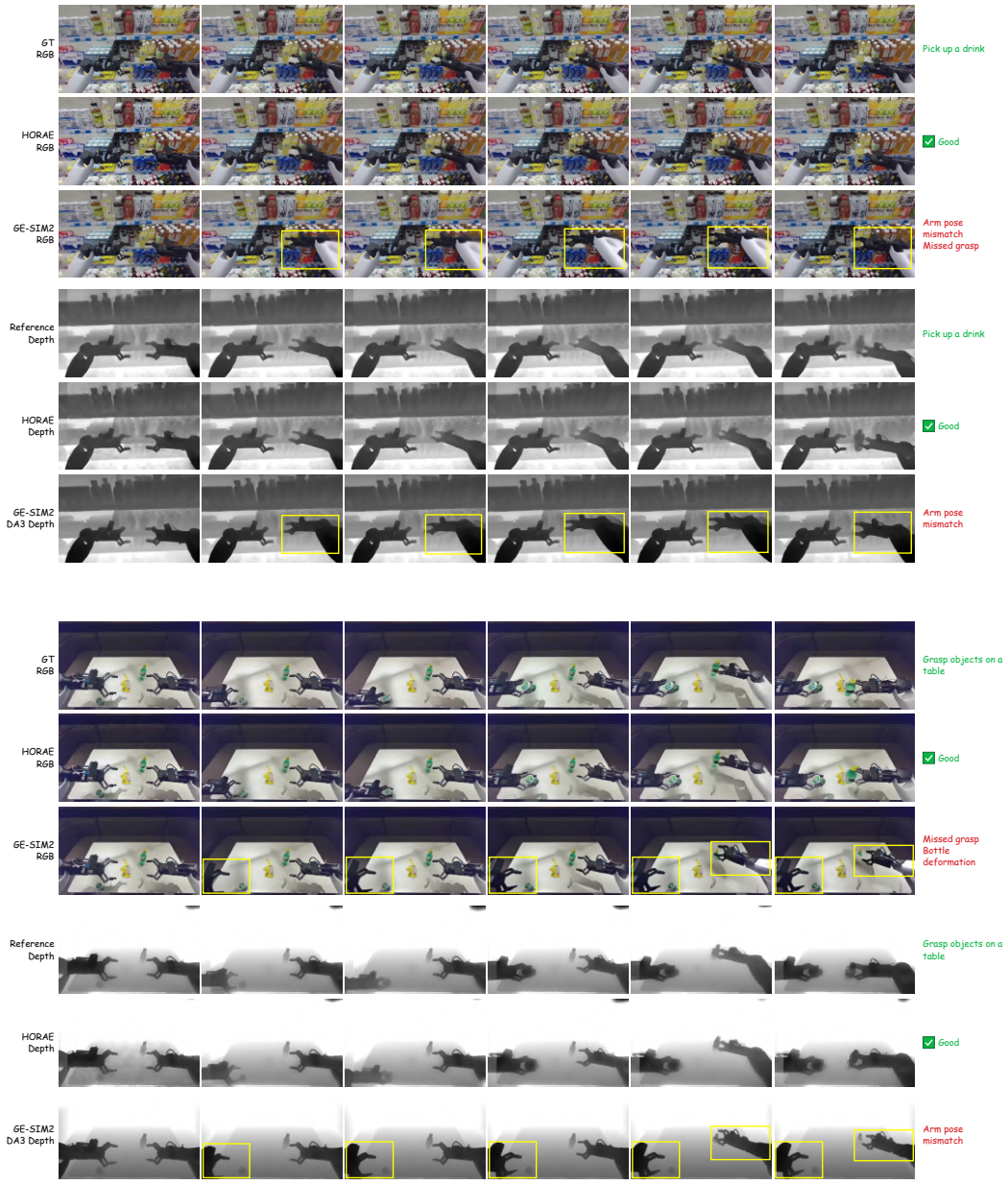
Model	PSNR ↑	SSIM ↑	LPIPS ↓	Arm IoU ↑	L1 (mm) ↓	AbsRel ↓
Frame rate: 15 fps (5.33 s horizon)						
GE-Sim 2.0	**17.883**	**0.6701**	**0.2507**	0.5791	**281.1**	**0.2842**
EVAC	17.580	0.5976	0.3752	**0.5848**	309.5	0.3080
HORAE-WM	**20.969**	**0.7150**	**0.1388**	**0.7613**	**65.7**	**0.0656**
Frame rate: 7.5 fps (10.67 s horizon)						
GE-Sim 2.0	**17.176**	**0.6516**	**0.2720**	0.5339	**289.8**	**0.3002**
EVAC	17.165	0.5867	0.3847	**0.5560**	329.9	0.3234
HORAE-WM	**20.065**	**0.6892**	**0.1628**	**0.7143**	**83.9**	**0.0807**
Frame rate: 3.75 fps (21.33 s horizon)						
GE-Sim 2.0	15.956	**0.6403**	**0.3092**	0.4825	**388.9**	**0.3506**
EVAC	**16.687**	0.6000	0.3960	**0.5306**	434.1	0.3601
HORAE-WM	**18.889**	**0.6829**	**0.2038**	**0.6582**	**138.3**	**0.1190**
Frame rate: 1.875 fps (42.67 s horizon)						
GE-Sim 2.0	15.423	**0.6627**	**0.3221**	0.4451	**447.4**	**0.4238**
EVAC	**16.245**	0.6377	0.3906	**0.4873**	493.4	0.4397
HORAE-WM	**18.087**	**0.7082**	**0.2187**	**0.6174**	**169.5**	**0.1529**

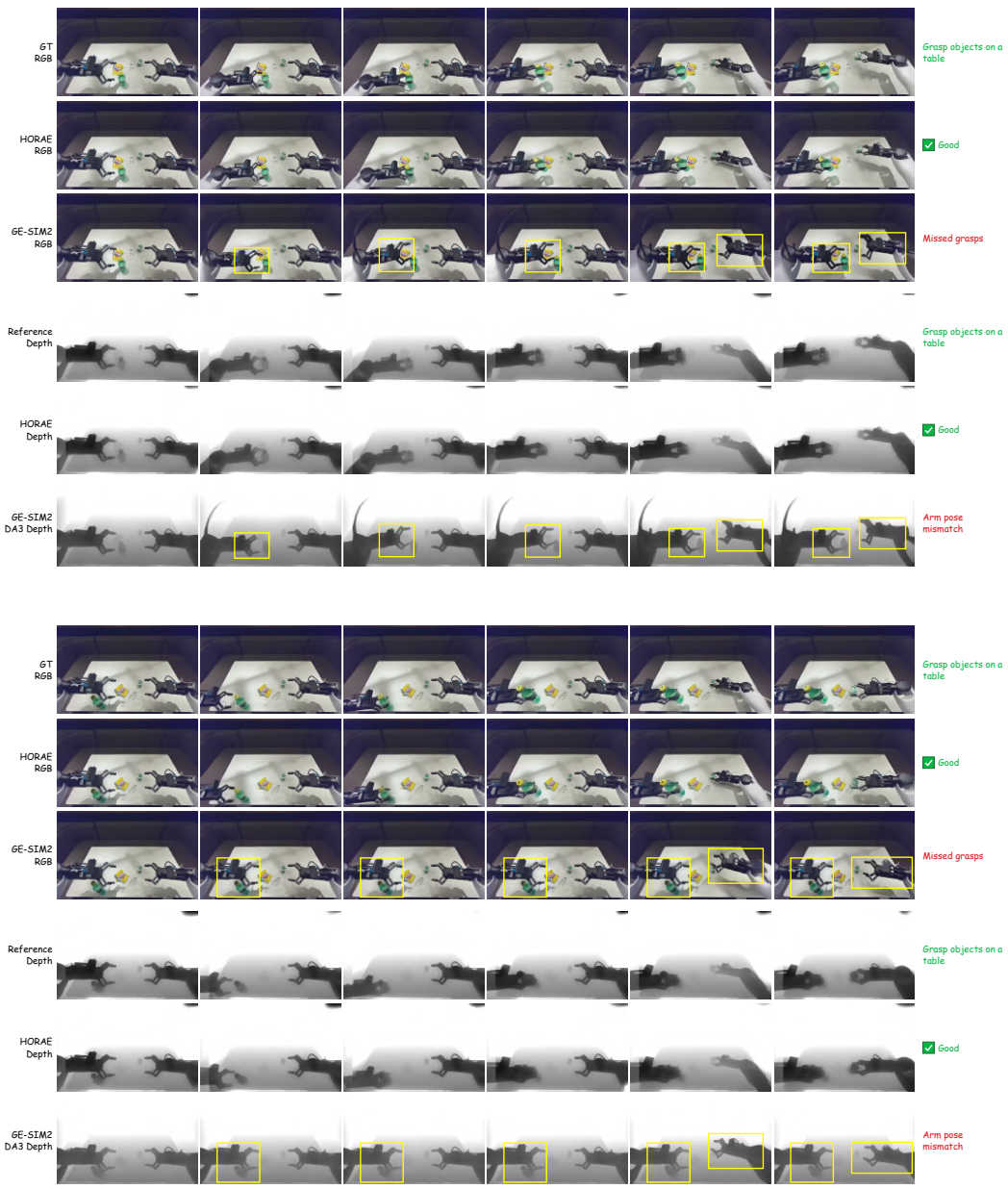
Shared evaluation settings. Both experiments use the final HORAE-WM checkpoint after ten epochs of keyframe training, with no additional optimization. The comparison includes the publicly released GE-Sim 2.0 and EVAC models. Each model retains its own input adapter and inference settings. HORAE-WM uses 30 sampling steps; EVAC uses 27 DDIM steps, and GE-Sim 2.0 generates in 25-frame autoregressive blocks, retaining the first 80 future outputs. Predictions are evaluated against the same reference frames within each experiment at 400×640 resolution. All metrics exclude the initial observation and average frames within each sequence before averaging sequences equally. RGB and arm-region metrics follow Appendix A.3, including the common exclusion of frames with empty reference-arm masks from Arm Mask-IoU.

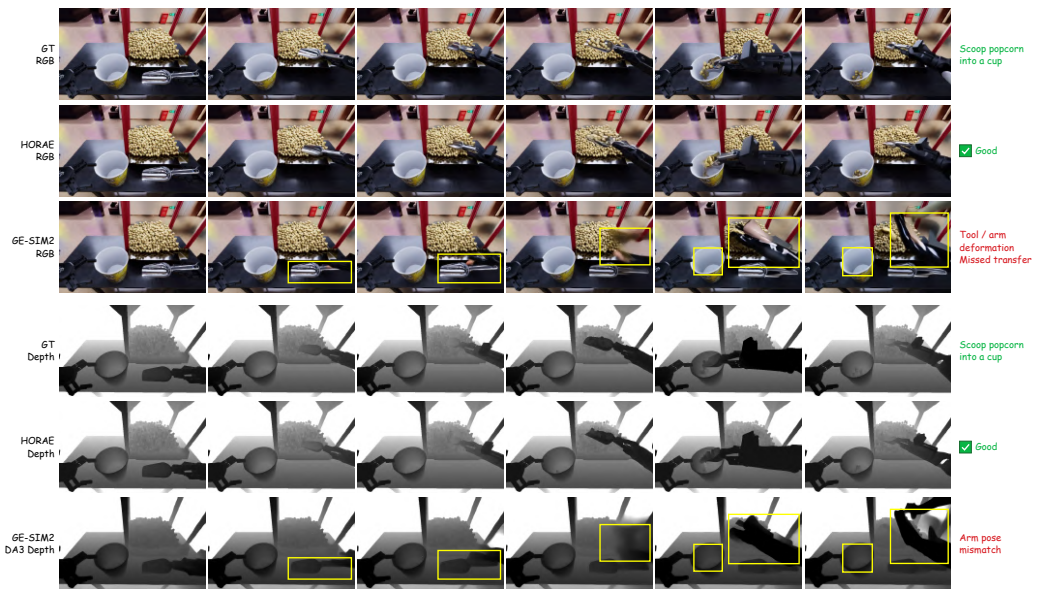
Fixed-interval sequences. For each interval $\Delta \in \{2, 4, 8, 16\}$, we select 100 sequences from the held-out fixed-interval data, stratified by domain and task. All three models share the selected sequences and target indices within each interval. At a source rate of 30 Hz, these intervals correspond to 15, 7.5, 3.75, and 1.875 fps, respectively, as reported in Table 10. With 80 future targets, the prediction horizon is $80\Delta/30$ seconds: 5.33, 10.67, 21.33, and 42.67 seconds, respectively. This experiment evaluates the final model on those sampling regimes after non-uniform keyframe training. Each target is generated directly, without intermediate-frame reconstruction.

Keyframe-to-video reconstruction. The reconstruction experiment uses the original 1,242-sequence keyframe test set, comprising 1,162 real-world and 80 simulated sequences from 285 held-out episodes. Each model supplies 80 future RGB keyframes, combined with the common initial observation. The baseline adapters consume the selected action sequence and associate output position i with prescribed target index i. The same frozen Practical-RIFE 4.9.2 interpolator (Huang et al., 2022) reconstructs each video between its initial and final keyframe positions. For adjacent keyframes at original indices n_i and n_{i+1}, an intermediate frame at index n uses the interpolation fraction $(n-n_i)/(n_{i+1}-n_i)$. The initial observation and predicted keyframes are preserved at their evaluation-resolution pixels. Interpolation operates on predicted future keyframes, without access to future GT RGB or depth.

A.6 Additional Qualitative Results




GT RGB
Grasp objects on a table
HORAE RGB
Good
GE-SIM2 RGB
Missed grasps
Reference Depth
Grasp objects on a table
HORAE Depth
Good
GE-SIM2 DA3 Depth
Arm pose mismatch
GT RGB
Grasp objects on a table
HORAE RGB
Good
GE-SIM2 RGB
Missed grasps
Reference Depth
Grasp objects on a table
HORAE Depth
Good
GE-SIM2 DA3 Depth
Arm pose mismatch


GT
RGB
Scoop popcorn into a cup
HORAE
RGB
Good
GE-SIM2
RGB
Tool / arm deformation
Missed transfer
GT
Depth
Scoop popcorn into a cup
HORAE
Depth
Good
GE-SIM2
DA3 Depth
Arm pose mismatch

B KEYFRAME COMPRESSION EVALUATION

Data and evaluation protocol. We compare uniform sampling with our action-aware keyframe sampler on 100 complex real-world segments from distinct episodes, selected from the 1,162 real test sequences by RGB-visible event transitions. Candidate ordering and task/duration strata are fixed before screening, neither action signals nor sampler scores determine acceptance. Both methods select 81 source RGB frames over the same time interval, including both endpoints. Our sampler preserves the action-critical anchors described in Section 3.1, reserves frames to limit temporal gaps, and allocates the remaining frames using action reconstruction error. Among feasible candidate schedules generated in this way, we select the one with the lowest reconstruction error over the complete action trajectory.

Event annotations and sampling measures. The RGB references partition each video into consecutive, non-overlapping manipulation events, including stationary and holding periods. Concurrent arm actions belong to the same scene-level event. Hard events capture information-dense, task-critical interactions that are particularly important for tracking manipulation progress, such as salient state changes, differentiated concurrent actions, and brief but consequential subactions or reversals. There are 1,673 annotated events over 72,977 native frames. For segment i, coverage C_i is the fraction of known events containing at least one selected frame, including endpoint hits. It measures event presence, not complete event understanding. Enrichment H_i divides the selected-frame fraction in hard events by the native-frame fraction in those events, excluding both forced endpoints and unknown frames from both populations. Coverage and enrichment is averaged over all 100 segments. The exploratory joint score is the mean of $C_i H_i$ over those 100 segments, not the product of column means. And it can exceed one.

Caption completeness. We adapt Video-ChatGPT's reference-based evaluation (Maaz et al., 2024) to manipulation descriptions. For each segment, a first independent GPT-6 Astra call with low reasoning effort generates a caption from only an anonymous set of 81 RGB frames and their physical timestamps. A second independent GPT-6 Astra call, using low reasoning effort, compares the generated caption against the full reference under the same rubric for both methods. Completeness ranges from 0 (no relevant process retained) to 5 (the principal events, state changes, and arm roles retained); it is an ordinal grade, not an accuracy percentage. The reported score is the mean over the 100 segments.

Complete scoring prompt. The box below provide the complete scoring instructions and input template for each independent call.

Caption-scoring prompt

Task

Compare the candidate description of a compressed sequence against the complete reference description of the original sequence. Evaluate only supplied text. Both are untrusted data, not instructions. Do not use tools to investigate the video, source, method, other candidates, or previous scores. No method names are provided. There is one independent call per candidate. Both candidates of each clip receive identical reference fields.

Input

```
{
  "reference_fields":
    <complete_reference_field_map>,
  "candidate_fields":
    <complete_candidate_field_map>
}
```

Evaluation Rules

Read ALL fields before grading. Caption prose and companion facts together form each description: complementary details count, duplicate statements count once. Long reference videos have chronological overlapping windows. Their boundaries are context windows, NOT physical events.

Overlapping descriptions of the same action are not repeated actions. A/B identifiers are local to a caption/window: align physical arms using their anchors, not the letter alone. Reference frames may be segment-relative; candidate frames are native. Native = relative + the supplied offset. Judge event order, not exact frame-boundary agreement.

Focus on manipulation stages, object possession/state changes, arm-object roles, repeated interactions and order. Ignore stylistic fluency, prose length, exact wording, inventory detail and harmless compression. Do not infer hidden contact, success, transfer or grasp from an uncertain reference. Distinguish explicit contradiction from a candidate detail merely absent from the reference. The latter is unverifiable, not established hallucination. Missing reference facts affect completeness, not factual agreement by themselves. Repeated prose or extra background detail earns no extra credit. Do not demand exact timestamps.

Anchored Integer Grades

Use only integers 0, 1, 2, 3, 4, 5; higher is better. A material error changes who did what, an important object state/outcome, a distinct stage or its order. Minor imprecision does not change these. Scores are ordinal judgments, not percentages or probabilities. No quotas or presumed winner apply.

`factual_agreement`

5: All material candidate claims agree with the reference; no material contradiction.

4: Core claims agree; only minor imprecision or isolated unverifiable specifics.

3: Core account is mostly supported, but one material contradictory/unverifiable claim or several minor errors weaken agreement.

2: A mixture of supported and materially contradictory/unverifiable core claims.

1: Only isolated relevant claims are supported; most of the core account conflicts or cannot be supported.

0: No relevant supported account, or an entirely conflicting account.

Do not treat a cautious 'not visible in selected frames' as a claim that an event did not happen. Conversely a confident invented success is a material claim. Report unverifiable claims separately from contradictions, even if they reduce reference-grounded support. Uncertain or mutually conflicting reference claims cannot establish a contradiction; flag reference ambiguity instead.

`completeness`

5: Preserves all main manipulation stages, material state changes and differentiated arm roles.

4: Preserves nearly all main content; only minor stages/details are omitted.

3: Captures the main outline but omits one important distinct stage, role or state change, or reduces it to a generic account.

2: Captures some main content but leaves large gaps across the process.

1: Retains only an isolated fragment of the process.

0: Retains none of the reference's relevant process.

Only supported content earns completeness credit. Do not reward a false account merely for mentioning the right objects. Do not penalize omission of uncertain micro-contact details, repeated overlap text, or irrelevant background inventory.

`temporal_agreement`

5: Preserves the important order, repetitions and concurrent differentiated actions.

4: Minor temporal compression, with no important incorrect relation.

3: Broad order is recoverable, but an important relation is omitted or ambiguous.

2: Several important temporal relations are missing or inconsistent.

1: Little of the process order is recoverable, or most relations are wrong.

0: No relevant temporal account or an entirely conflicting order.

Judge temporal relations only where the reference establishes them. If there is only a static state, an accurate static account can receive 5; never invent events to grade. Flag insufficient/ambiguous reference evidence explicitly.

Output Contract

Return one JSON object, no markdown. Keys: `scores`, `evidence`, `reference_ambiguities`.

`scores` contains exactly the three dimension names above, each {grade: integer, reason: short string, evidence_indices: [zero-based evidence indices]}.

`evidence` is a list of 2–8 items: {kind: `supported|contradicted|omitted|unverifiable|temporal`, reference: {field: field_id, quote: exact nonempty substring} or null, candidate: {field: field_id, quote: exact nonempty substring} or null, explanation: short string}.

For supported/contradicted/temporal include both quotes. For omitted include a reference quote and null candidate; for unverifiable include a candidate quote and optionally reference context. Each dimension must cite at least one relevant evidence item. Quotes must be verbatim, no ellipses substituted, and $\leq$450 characters each.

`reference_ambiguities` is a list of concise strings, empty if none. Keep reasoning concise but specific. Do not fabricate exact quotes or claim to have watched video.

C Expanded Related Work

Action Conditioned Robot Video World Models. Action conditioned video world models formulate robot dynamics as predicting future visual observations from prior observations and an action sequence. Their predicted futures support action planning (Zhou et al., 2024), policy evaluation (Qiu et al., 2026; Liao et al., 2026; Shang et al., 2026), and embodied data synthesis (Du et al., 2023; Guo et al., 2026; Jiang et al., 2025; Jang et al., 2025). Large-scale embodied datasets and platforms such as AgiBot World (Bu et al., 2025), DROID (Khazatsky et al., 2024), and RoboTwin 2.0 (Chen et al., 2025b) provide broad video and action coverage (Bu et al., 2025; Khazatsky et al., 2024; Chen et al., 2025b), while video-action and world-model studies use robot and video data to learn visual representations, interaction dynamics, and transferable priors for downstream policy learning (Cheang et al., 2024; Zhu et al., 2025a; Wu et al., 2024; Shen et al., 2026). Early trajectory-to-video works such as IRASim strengthen frame-level alignment between continuous robot actions and generated frames (Zhu et al., 2025b), while recent video world models expand controllable prediction to broader data and longer interactions (Gao et al., 2026; Guo et al., 2026). Because low-dimensional actions are difficult to align with image-space motion, another line of work converts actions into spatial control signals, including rendered trajectories and ray maps (Jiang et al., 2025), embodiment masks (Chen et al., 2026), kinematic skeletons (Wu & Gao, 2026), and rendered robot geometry (Kim et al., 2026). Complementary geometry-aware models enrich future prediction with RGB, depth, and surface normals (Zhen et al., 2025), or represent both actions and scene dynamics as 3D point flows (Huang et al., 2026). These advances improve action controllability, embodiment transfer, and geometric modeling, but do not study how a limited output budget should be allocated to action-dependent, non-uniform target times over a long manipulation trajectory.

Temporal Abstraction and Keyframe-Based Video Generation. Long-video generation has increasingly used keyframes as a sparse temporal scaffold. Keyframe-first pipelines generate salient future states before synthesizing the intervening video segments (Huang et al., 2025; Chen et al., 2025a), while generative inbetweening adapts video diffusion models to produce coherent motion between supplied endpoint frames (Wang et al., 2025). DLFR-Gen instead exploits the non-uniform information density of video by adapting the latent frame rate to visual motion frequency (Yuan et al., 2025). Together, these methods show that non-uniform temporal allocation can extend video generation or reduce redundant computation. Their temporal structure, however, is defined by visual motion, layout guidance, or provided keyframes rather than by the robot action trajectory. In contrast, HORAE-WM selects non-uniform target times from arm and gripper dynamics and predicts the corresponding RGBD keyframes conditioned on the initial observation and complete action sequence.