ORV: 4D Occupancy-centric Robot Video Generation

Xiuyu Yang[1,2*] Bohan Li[3,4*] Shaocong Xu[1] Nan Wang[1] Chongjie Ye[1,5] Zhaoxi Chen[1,6]
Minghan Qin[7] Yikang Ding[8] Zheng Zhu[9] Xin Jin[4,10] Hang Zhao[2] Hao Zhao[1,11]

[1] Beijing Academy of Artificial Intelligence [2] IIIS, Tsinghua University
[3] Shanghai Jiao Tong University [4] Eastern Institute of Technology, Ningbo
[5] The Chinese University of Hong Kong, Shenzhen [6] S-Lab, Nanyang Technological University
[7] ByteDance [8] Kuaishou [9] GigaAI [10] Zhongguancun Academy [11] AIR, Tsinghua University

Figure 1. We condition robot video generation on 4D semantic occupancy sequences and 7-DoF actions collected from real and simulated environments (through methods π and π'). This occupancy-centric conditioning enables faithful, controllable synthesis of single-view, multi-view, and sim-to-real manipulation videos. We also introduce ORV-Data, a curated 4D occupancy dataset for robot manipulation. Across benchmarks and downstream tasks, ORV improves video quality and control alignment, boosting visual planning and policy learning.

Abstract

Recent embodied intelligence suffers from data scarcity, while conventional simulators lack visual realism. Controllable video generation is emerging as a promising data engine, yet current action-conditioned methods still fall short: generated videos are limited in fidelity and temporal consistency, poorly aligned with controls, and often constrained to singleview settings. We attribute these issues to the representational gap between sparse control inputs and dense pixel outputs. Thus, we introduce ORV, a 4D occupancy-centric framework for robot video generation that couples action priors with occupancy-derived visual priors. Concretely, we align chunked 7-DoF actions with video latents via an Action-Expert AdaLN modulation, and inject 2D renderings of 4D semantic occupancy into the generation process as soft guidance. Meanwhile, a central obstacle is the lack of occupancy data for embodied scenarios; we therefore curate ORV-Data, a large-scale, high-quality 4D semantic occupancy dataset of robot manipulation. Across BridgeV2, DROID, and RT-1, ORV improves video generation quality and controllability, achieving 18.8% lower FVD than state of the art, +3.5% success rate on visual planning, and +6.4% success rate on policy learning. Beyond singleview generation, ORV natively supports multiview consistent synthesis and enables simulation-to-real transfer despite significant domain gaps. Code, models, and data will be released upon acceptance.

1. Introduction

Developing realistic simulators for robot manipulation is crucial for scaling embodied learning [45, 50, 60, 70]. While existing simulators [27, 90] enable safe policy training and

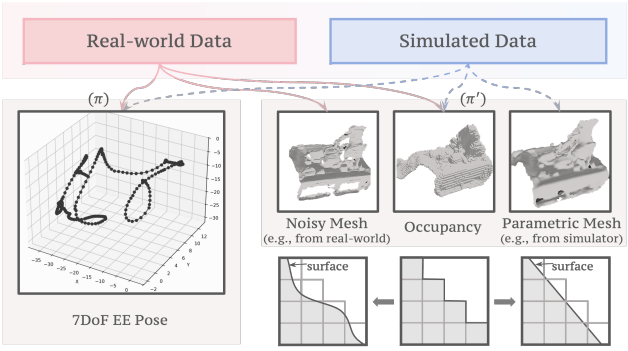

Figure 2. Establishment of non-interactive methods (π, π') both in the real-world environment and physical simulator to collect trajectory priors (7-DoF EE Pose) and visual priors (Occupancy).

efficient data collection, they often struggle to deliver visual realism. Recent progress in generative world models [49, 94, 119], especially action-conditioned video generation, offers a promising alternative by simulating future visual states conditioned on agent actions. These models can render realistic RGB observations responsive to control inputs, yet they still fall short of serving as reliable simulators: generated sequences frequently lack temporal consistency, action alignment, and multiview coherence. Bridging this gap between sparse robot controls and dense visual dynamics remains an open challenge toward building truly *high-fidelity*, *versatile*, and *reliable* generative simulators.

Previous works [2, 80, 99, 102, 136] have advanced action-conditioned video generation using diffusion-based or autoregressive backbones, where robot actions are typically represented as 7-DoF end-effector (EE) poses that guide visual rollout. Other studies [37, 118, 131] instead employ high-level conditioning such as language instructions to drive scene dynamics. Despite these advances, existing approaches remain constrained by three key limitations: (p1) limited visual fidelity and temporal consistency; (p2) drifted or misaligned future predictions that fail to reflect manipulation controls faithfully; and (p3) restriction to singleview observations without enforcing multiview coherence.

We propose ORV, a versatile 4D occupancy-centric framework for robot video generation that produces high-fidelity, action-aligned visual simulations. Our key insight is to incorporate 4D semantic occupancy as visual priors that complement conventional action priors, effectively bridging the representational gap between sparse control trajectories and dense visual dynamics. We think that limitations p2, p3 largely stem from this gap, as also observed in prior works [54, 63, 109, 115] which introduce fine-grained cues such as optical flow, masks, or skeletons to enhance controllability. Furthermore, as illustrated in Fig. 2, occupancy fields demonstrate robustness to geometric noise, providing a natural bridge between simulated and real-world scenarios. Moreover, ORV leverages the generative capabilities of modern video foundation models [49, 94, 119] to boost visual realism and temporal coherence, substantially mitigating issue (p1) while preserving physically consistent dynamics.

The overall framework of ORV is depicted in Fig. 1. Guided by geometric priors from 4D semantic occupancy, ORV enables robot manipulation video generation across diverse object appearances and scenes [3, 63]. Furthermore, view-specific conditioning encourages cross-view coherence, enabling consistent multiview synthesis [1, 4, 26]. Benefiting from the domain-invariant nature of occupancy-derived representations, ORV also facilitates visual transfer from simulation to the real world under varied conditions. To support large-scale training, we curate ORV-Data, a high-quality 4D semantic occupancy dataset for robot manipulation, built through a carefully designed data curation pipeline.

Our contributions can be summarized as follows:

- We propose **ORV**, a *4D occupancy-centric framework*, enabling precise and controllable robot video generation with domain randomization.
- By injecting *occupancy-derived geometric priors* into diffusion noise, ORV achieves temporally consistent and geometrically coherent multiview video generation and simulation-to-real visual transfer.
- We curate **ORV-Data**, a large-scale, high-quality *4D semantic occupancy dataset* of robot manipulation with rich geometric and semantic annotations.
- Experiments across diverse datasets and downstream tasks demonstrate that ORV consistently enhances controllable video generation, visual planning, and data-driven policy learning, achieving state-of-the-art performance.

2. Related Work

Generative Models for World Modeling. Recent advances in video generation [7, 8, 49, 94, 119, 121, 133] have greatly improved the realism of world modeling, benefiting robotics [3, 9, 23, 43, 63, 68, 75, 109, 131, 136], autonomous driving [24, 53, 71, 107], and general scene synthesis [58, 59, 79, 130]. ReCamMaster [8] and SynCamMaster [7] achieve video synthesis of novel trajectories, while IRASim [136] enables action-to-video prediction, and VAP [109] employs visual prompts for precise control in robotics. For autonomous driving, more recent works adopt 3D occupancy as efficient scene representations [15, 39, 40, 51, 52, 98, 103, 111, 112, 114, 132]. For instance, UniScene [53] leverages hierarchical occupancy priors for multimodal scene generation. Beyond explicit video synthesis, implicit generative models have also been adopted for complex interactions and decision making [19, 69, 74].

World Models for Embodied Intelligence. Progress in simulating dynamic environments has fueled the development of world models for robotics [9, 11, 14, 17, 20, 25, 28, 37, 67, 118, 124, 134, 135], where TesserAct [131] performs 4D scene synthesis via appearance–geometry joint modeling and EnerVerse [37] forecasts future environments through a

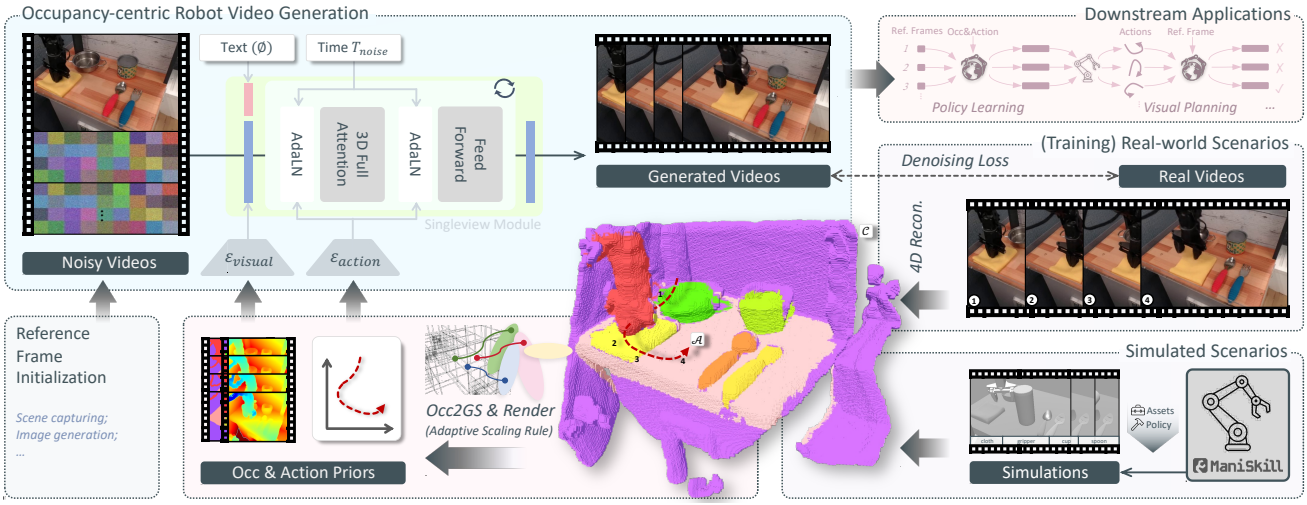

Figure 3. Overview of **ORV framework**. Centered on occupancy representation $\mathcal{C}$, along with actions $\mathcal{A}$, which are extracted from physical simulators (*e.g.*, ManiSkill [27]) or real-world data (*e.g.*, Bridge [93]), we leverage the soft visual priors to enable robot video generation with high visual quality and control alignment. Furthermore, we design a data curation pipeline to construct the robot occupancy data for training purposes. ORV, as a powerful neural simulator, can greatly boost downstream applications (*e.g.*, policy learning, visual planning, etc).

simulation pipeline. iVideoGPT [113] and Vid2World [38] explore action-conditioned visual prediction with autoregressive frameworks. For data augmentation, Cosmos-Transfer [3] and RoboTransfer [63] condition robot video generation on scene maps (*e.g.*, depth and normal), while RoboEngine [123] achieves scene augmentation through the segmentation toolkit. Meanwhile, WorldSimBench [76] establishes unified evaluation benchmarks for world models.

3. ORV: Methodology

We first formulate the robot video generation task (Sec. 3.1). Then we elaborate on the specific architecture of ORV and how these designs can largely improve the robot video generation (Sec. 3.2). Finally, we introduce our robot occupancy dataset curated for the training process (Sec. 3.3) and explain how ORV helps with the robot manipulations.

3.1. Problem Formulation

A generative world model for robot manipulation aims to provide a photorealistic and physically consistent simulation of the environment that mirrors real-world dynamics. Given the context $(\mathcal{S}, \mathcal{O}, \phi, \rho)$, the goal of the model $\mathcal{M}$ is to predict future states $s_{t:t+\Delta T} \in \mathcal{S}$ and corresponding observations $o_{t:t+\Delta t} \in \mathcal{O}$, where $o_t = \phi(s_t)$ denotes the rendered observation from state s_t. Here, ρ defines the underlying rules governing state transitions, leading to the transition probability $p(s_{t:t+\Delta t}, o_{t:t+\Delta t} \mid s_{1:t}, o_{1:t})$.

We formulate $\mathcal{O}$ in RGB space (*e.g.*, images or videos). Conventional text-to-video models [57, 94, 119] condition on $\rho_1 := \texttt{Embed}(\texttt{text})$, yet linguistic abstraction often hinders accurate physical simulation. Recent action-conditioned video generation [80, 99, 109, 113] extends this to $\rho_2 := \texttt{Embed}(a_{t:t+\Delta t} \sim \pi(s_{1:t}))$. Building upon this progression, our model introduces $\rho_3 := \texttt{Embed}(c_{t:t+\Delta t} \sim \pi'(s_{1:t}), a_{t:t+\Delta t} \sim \pi(s_{1:t}))$, where a denotes agent actions and c represents occupancy fields. We denote by π and π' the extraction processes for (a, c) given states s.

As illustrated in Fig. 2, both extraction methods can be established either in the real world (*e.g.*, human teleoperation) or within simulators (*e.g.*, ManiSkill [27], MuJoCo [90]). Notably, we employ π and π' in a *non-interactive* manner—these priors are collected entirely in a single offline pass before being used. Moreover, the motivation for leveraging occupancies lies in their robustness for representing both noisy and parametric scene surfaces (Fig. 2). And the coordinate-based formulation of occupancies enables seamless integration with online occupancy generations [129].

3.2. Occupancy-centric Robot Video Generation

To avoid a costly large-scale pretraining process (as previous works [113, 135]) and reduce the training cost, we build ORV model upon the pretrained open-source models (*e.g.*, we use CogVideoX-2B [119]), which also aligns with our non-interactive purpose (using a bidirectional diffusion model). CogVideoX incorporates the architecture of diffusion transformer (DiT) and achieves incredible performance. Then, we propose a two-stage supervised finetuning (SFT) to inject both action and visual cues into video generations. We aim to address three key aspects: 1) overall quality of generated videos (*e.g.*, consistency of frames and realism), 2) alignment with the instructions ρ_3, and 3) computation efficiency.

Chunk-level Action Conditioning. The 7-DoF action sequences (*e.g.*, $\mathcal{A} \in R^{T \times D_a}$ derived from end-effector pose sequences and $D_a = 7$) serve a high-level control signals in robot video generation. Drawing inspiration from [128, 136],

we inject these 3D action controls through adaptive layer normalization (Action Expert AdaLN) to directly modulate the video latents within each DiT block. More efficiently, as illustrated in Fig. 4, we propose a chunk-level scheme for temporal alignment between high-dimensional actions and videos in modulation.

Specifically, following the temporal compression in 3D VAE [49, 94, 119], we pad zero actions as the placeholders of reference frames. Then an additional shallow MLP (ε_{action} in Fig. 3) is used to map every consecutive r actions into a single token: $\mathcal{A} \in R^{T\times D_a} \rightarrow \texttt{MLP}(\texttt{Pad}(\mathcal{A})) \in R^{(\frac{T}{r}+1)\times D}$, where r denotes the chunk-size and D represents the feature size. Furthermore, we let Action Expert AdaLN reuse the parameters of pretrained Vision Expert AdaLN, eliminating the unnecessary computation cost (as each AdaLN accounts for $\sim 1/3$ of the total parameters).

Occupancy-derived Visual Conditioning. Translating abstract 3D action signals into 2D pixels presents a great challenge; thus, we introduce *soft* and *pixel-level* visual conditionings derived from occupancy fields. However, directly projecting voxels onto 2D planes will cause mutations on pixels between adjacent frames and viewpoints. We further propose to assign each grid with non-learnable Gaussian splatting [46], then render them from certain views (Fig. 3), which greatly improves the conditions quality and saves memory.

Moreover, we propose an *adaptive scaling mechanism* on Gaussians to solve the perspective distortion during rendering (see Sec. 10.1.2 in Suppl. for derivations). Specifically, the scale follows $\sigma = k_2 \cdot \hat{z}^{k_1}$, where $\hat{z} \in [1, 2)$ denotes the *normalized depths in canonical space*, and exponential term k_1, base scale term k_2 control the scaling behavior of Gaussians in the near and far plane, respectively.

To inject such occupancy-derived visual conditionings, we deploy an additional encoder MLP (ε_{visual} in Fig. 3), then augment it with the input images, after which another zero-initialized projector adds the visual conditionings to the input noise: $z_{\text{in}} = \texttt{Zero-MLP}(z_{\text{in}} + \texttt{MLP}(\mathcal{C})) + z_{\text{in}}$. The previous ControlNet-like [126] methods, though demonstrating accurate controls, suffer from a serious computation cost (see Sec. 10.2 in Suppl.). Furthermore, such layer-wise control injection tends to corrupt the video latents when conditions are *soft*—that is, not pixel-level alignment with ground truth.

3.2.1. ORV-MV: Multiview Robot Video Generation

A complete and high-fidelity 4D world, typically formed from multiview observations, greatly benefits robot learnings [65, 75]. Leveraging the 4D occupancy-centric design, ORV(-MV) generates multiview robot manipulation videos well. Some prior works [109, 131, 136], however, capture only a single surface of the scenes, resulting in noticeable artifacts and empty regions in shifted views.

As shown in Fig. 5, ORV-MV introduces an additional view attention (multiview module) prior to the temporal attention (singleview module), inspired by [7, 16]. Both inherit the 3D (2D+1D) attention layers of the pretrained model, with 2D over pixels $H\times W$ and 1D over views V or frames F. The former processes the latents $\mathcal{F}_V \in R^{B_V\times S_V\times D}$, where $S_V = VHW$ denotes patch tokens across all views. While the latter handles $\mathcal{F}_P \in R^{B_P\times S_P\times D}$, where $S_P = THW$ denotes tokens across all times of each view.

We then apply different controls for the two modules. Specifically, singleview modules are conditioned on text, actions, and occmaps. While multiview ones exclude action priors, as they focus on view correspondences. Additionally, details on handling multiview occupancy map data for training purposes are provided in Sec. 8 in Suppl.

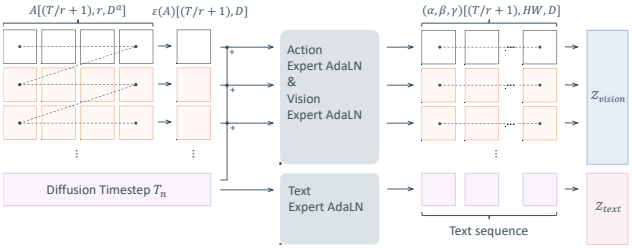

Figure 4. Illustration of three modulations (Expert AdaLN) and injecting actions in our DiT block. And indicates the action paddings serving as the placeholders for reference frames, where ε encodes actions and α, β, γ are modulation vectors. We use $[\cdot]$ to indicate the dimensions for simplicity.

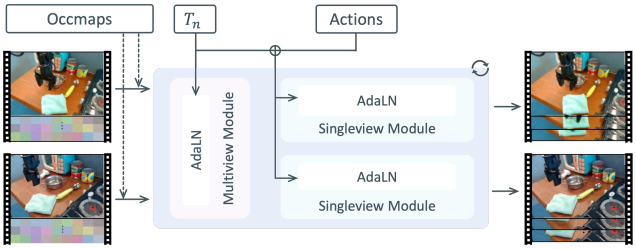

Figure 5. Architecture of ORV-MV, which generates multiview robot manipulation videos with cross-view consistency.

3.2.2. ORV-S2R: Bridge Simulation-to-Real Transfer

The occupancy-derived visual priors (*e.g.*, depth maps) also enable ORV(-S2R) to generate realistic videos from such appearance-agnostic information, which is crucial for alleviating the *visual realism* gap between simulated and real data in robotics. As shown in Fig. 3, physical simulators (*e.g.*, ManiSkill [27], MuJoCo [90]) can readily provide such priors at a low cost.

Previous works, *e.g.*, Cosmos-Transfer [3], RoboTransfer [63], have also demonstrated success in transferring multi-modal data to significantly mitigate the data scarcity problem in robotics. However, as described in Sec. 3.1 and Fig. 2, the occupancy-derived condition maps further exhibit robustness to geometric noise, providing a natural bridge between real-world noisy surfaces and parametric ones in simulation. Experiments in Sec. 4.4 have validated the effectiveness of this design.

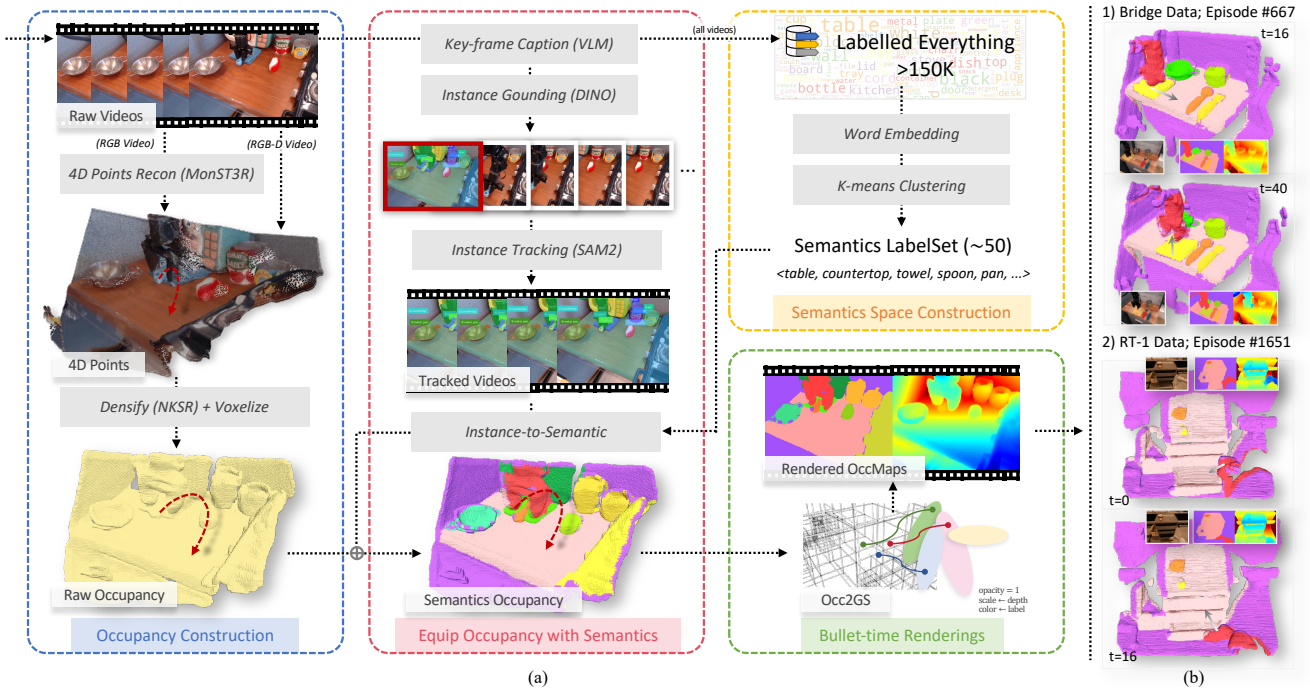

Figure 6. (a) Overview of **Training Dataset Curation Pipeline**, which consists of four steps: 1) semantics space construction, 2) occupancy construction, 3) equip occupancy with semantics, and 4) bullet-time occupancy-to-Gaussian renderings in practical usage. (b) **Occupancy examples** of BridgeData V2 (BridgeV2) [93] and RT-1 [13]. Better to zoom in. Refer to Supplementary Materials for more examples.

3.3. 4D Occupancy Dataset of Robot Manipulation

To train ORV model, we establish a 4D occupancy dataset of robot manipulation through the data curation pipeline shown in Fig. 6(a). The occupancy data are derived from existing popular robot datasets (BridgeData V2 [93], DROID [47], RT-1 [13]). Some examples are shown in Fig. 6(b). More details are provided in Sec. 10.1 and Sec. 11.1 in Suppl.

Semantics Labeling. Complex semantic understanding remains essential in robot manipulation, as predicting next-state dynamics requires recognizing objects—rigid, articulated, or deformable—that exhibit distinct physical behaviors. To this end, we construct the dataset-level semantic space through vision-language-model (VLM) [6] for captioning and K-means [66] clustering over ∼150K labels. For each video, we then extract temporally consistent instances across frames using Grounding DINO [64] and SAM2 [78], after which they are mapped to coherent semantics.

4D Occupancy Generation. This process involves two steps: 1) occupancy construction and 2) semantic enrichment. We first reconstruct sparse 4D points with MonST3R [125] and then densify them by NKSR [36], which greatly fills holes and is robust to noise. Note that for those videos with a depth channel, the reconstruction is not needed. Then, dense points are voxelized to 4D occupancy in canonical space, after which semantics are assigned by majority voting for points with projected semantic labels within each voxel. Finally, we filter the occupancy-rendered data with poor inter-frame consistency (through RAFT [85]).

4. Experiments

In this section, we conduct comprehensive experiments to validate ORV model on multiple tasks, including *controllable video generation*, *visual planning*, and *policy learning*. They are expected to answer these questions: *1) What is the quality of the videos generated by ORV? 2) To what extent is the generative capability of ORV? 3) How can generated videos benefit robot learning tasks?* Additionally, we provide dataset details, experiment details, and more results in Sec. 7, Sec. 10, Sec. 11 in Suppl., respectively.

4.1. Conditional Video Generation

Setup. We evaluate the video generation of ORV on three real-world datasets, their embodiments, views of each episode, and volume are as below:

- BridgeV2 [93]: WidowX, 1∼3 views, ∼60K episodes;
- DROID [47]: Franka Panda, 2 views, ∼76K episodes;
- RT-1 [13]: Google Robot, 1 view, ∼120K episodes;

Please refer to Sec. 7 in Suppl. for more dataset details. For the action-conditioned base model setup, we train ORV for ∼30K steps from the pretrained backbone. For occupancy maps-guided finetuning and multiview video generation, we have additional ∼20K gradient steps of training.

Table 1. Evaluation results of *Conditional Video Generation* on three datasets. Top-1 performance within all variants and each type of model is represented with **bold text** and underlines.

Method	BridgeData V2 [93]				DROID [47]				RT-1 [13]			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
Text-conditioned Generation Models												
CogVideoX [119]	19.432	0.752	7.509	83.561	19.238	0.701	6.341	71.536	20.457	0.816	6.243	42.169
Action-conditioned Generation Models												
AVID [80]	-	-	-	-	-	-	-	-	25.600	0.852	2.965*	24.200
HMA [99]	23.636	0.808	8.849	67.096	21.435	0.821	**3.108**	47.383	25.424	0.840	7.306	84.165
IRASim [136]	25.276	0.833	10.510	20.910	21.632	0.820	5.395	41.031	26.048	0.833	5.600	25.580
ORV (Ours)	25.631	0.873	3.821	17.682	22.034	0.838	4.921	37.094	27.086	0.863	4.210	20.031
Occupancy&Action-conditioned Generation Models												
IRASim† [136]	27.352	0.862	9.413	22.503	22.005	0.827	7.892	44.309	27.213	0.847	5.311	42.130
ORV (Ours)	**28.258**	**0.899**	**3.418**	**16.525**	**22.310**	**0.841**	3.222	**34.603**	**28.214**	**0.878**	**4.013**	**19.931**

* FID Scores of AVID [80] have been computed not in evaluation mode according to the official codes and lead to incorrect results. Thus, we ignore it.
† We incorporate the same occupancy&action conditions to IRASim.

Table 2. Evaluation results of *Visual Planning* on VP^2 [86] Benchmark. Top-1 performance across 8 tasks and the average success rate are highlighted accordingly. We provide the mean and standard deviation of the success rate (in %) on average over 3 runs. The best and second-best performances are represented with **bold text** and underlines, respectively.

Method	Robosuite Push	Flat Block	Open Drawer	Open Slide	Blue Button	Green Button	Red Button	Upright Block	Avg. Success	
Simulator	$93.5^{\pm 2.2}$	$13.3^{\pm 0.1}$	$76.7^{\pm 0.0}$	$71.7^{\pm 1.2}$	$100.0^{\pm 0.0}$	$96.7^{\pm 0.0}$	$90.0^{\pm 0.0}$	$90.0^{\pm 0.0}$	88.4	*
MCVD [92]	$77.3^{\pm 2.6}$	$4.0^{\pm 1.1}$	$11.7^{\pm 1.2}$	$18.3^{\pm 1.0}$	$95.0^{\pm 3.6}$	$83.3^{\pm 0.4}$	$73.3^{\pm 2.6}$	$56.7^{\pm 2.4}$	59.4	67.2
FitVid [5]	$67.7^{\pm 5.3}$	$\mathbf{9.2}^{\pm 4.0}$	$25.3^{\pm 6.9}$	$\mathbf{35.3}^{\pm 4.5}$	$94.0^{\pm 4.6}$	$84.0^{\pm 5.3}$	$58.7^{\pm 5.1}$	$51.3^{\pm 2.7}$	59.5	67.3
MaskViT [29]	$\mathbf{82.6}^{\pm 2.3}$	$4.0^{\pm 3.9}$	$4.0^{\pm 4.5}$	$8.7^{\pm 5.7}$	$94.7^{\pm 2.0}$	$64.0^{\pm 4.3}$	$24.0^{\pm 7.5}$	$\mathbf{62.2}^{\pm 8.6}$	48.6	55.0
iVideoGPT[113]	$78.3^{\pm 0.4}$	$3.3^{\pm 0.7}$	$37.5^{\pm 1.5}$	$16.1^{\pm 2.5}$	$95.6^{\pm 2.9}$	$82.5^{\pm 3.1}$	$92.2^{\pm 1.5}$	$44.7^{\pm 1.7}$	63.9	72.2
ORV (Ours)	$81.4^{\pm 1.7}$	$6.1^{\pm 2.0}$	$\mathbf{40.5}^{\pm 1.1}$	$19.9^{\pm 3.4}$	$\mathbf{96.7}^{\pm 2.5}$	$\mathbf{85.6}^{\pm 3.0}$	$\mathbf{93.2}^{\pm 1.9}$	$44.8^{\pm 1.4}$	**66.0**	**74.7**

* Values in this column are normalized by the simulator's average success rate.

Comparison with Baselines. To comprehensively demonstrate the superiority of ORV model, we compare ORV with original CogVideoX [119] and action-conditioned methods AVID [80], HMA [99], IRASim [136] and more baselines augmented with our occupancy priors (*e.g.*, IRASim). We report the quantitative results of *controllable video generation* in Table 1, where ORV outperforms all baselines across most of the metrics. Moreover, as highlighted (white arrows) in the BridgeV2 example of the singleview generation in Fig. 7, the baseline fails to faithfully infer the dynamics of objects manipulated by the robotic gripper. More details about the baselines and comparison results are in Sec. 11.2 in Suppl.

Multiview Robot Video Generation. We show an example of multiview robot video generation performed by ORV in Fig. 7. The example shows the robot arm performing a cloth-folding task across *three* views, where the outputs maintain exceptional cross-view consistency. This high-fidelity multi-view generation enables efficient downstream applications, including photorealistic scene reconstruction and robotics imitation learning. Note that there exists a lighting discrepancy issue in the original data.

Sim-to-Real Transfer. Fig. 7 illustrates examples of sim-to-real generation through ORV-S2R, as described in Sec. 3.2.2. Details of simulation environment setup and dynamics data generation are provided in Sec. 9 of the Supplementary. Leveraging an additional image generator (ControlNet [126]), we first produce diverse initial frames and then extend them to high-quality, realistic manipulation videos. In this case, using simulator-derived occupancy maps consistent with training preserves the consistent performance. Moreover, thanks to the robustness of occupancy representations, even condition maps with various granularity (e.g., parametric maps from the simulator) yield only minor performance degradation (see results in Sec. 4.4).

4.2. Visual Planning

Setup. We further evaluate the controllability of ORV on VP^2 [87], a visual planning by action controls benchmark. Following [86, 102, 113], we train ORV on 5K trajectories for Robosuite [137] and 35K for RoboDesk [44].

Results. Tab. 2 presents the success rates of ORV compared to the baselines over 9 tasks. ORV outperforms all baselines

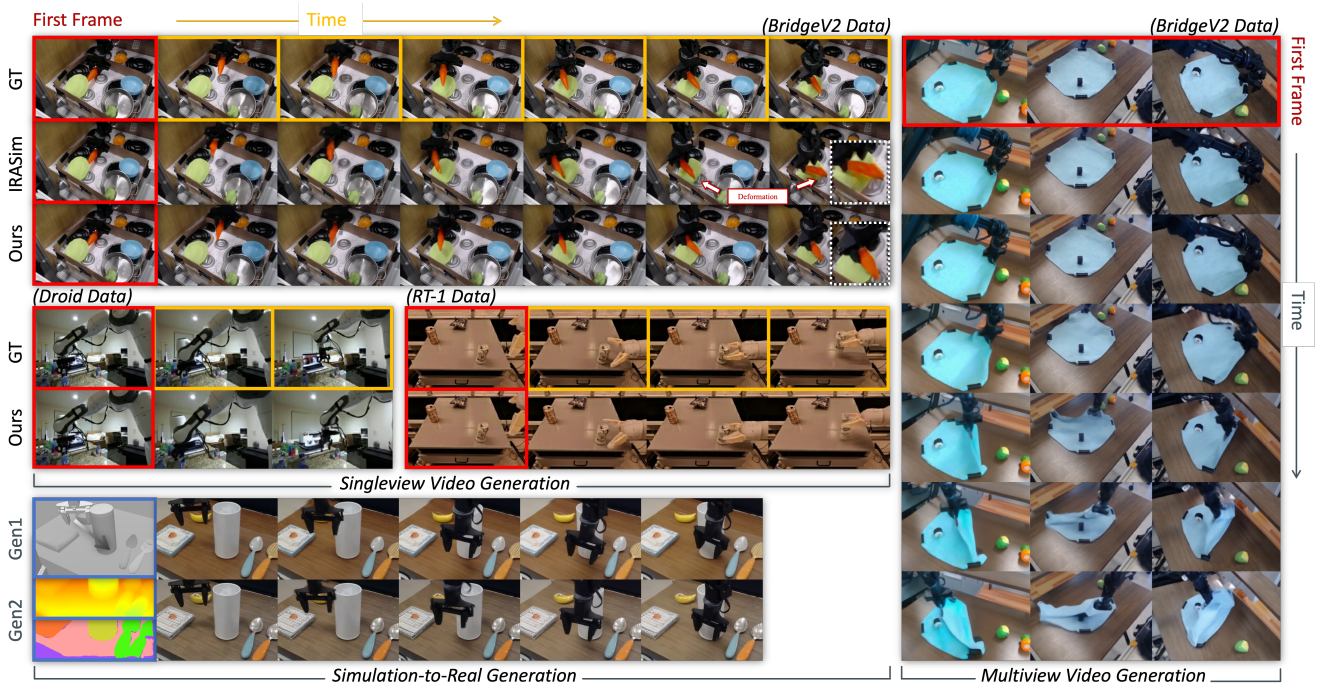

Figure 7. Qualitative results of versatile **Video Generation** with full conditions. Given one-frame observation, ORV predict subsequent 15 frames on validation split of Bridge [93], DROID [47], RT-1 [13] datasets. Red boxes denotes the first frame input of the video generation; Orange boxes denotes the ground-truth of the subsequence frames.

Table 3. Evaluation results on SimplerEnv-WidowX [56] across four manipulation tasks. "+Finetune" indicates the additional finetuning on domain-specific dataset; and "+ORV" indicates that we augment the finetuning dataset with ORV-synthesized data.

Method	Spoon on Towel	Carrot on Plate	Stack Cube	Eggplant in Basket	Avg. Success
RoboVLM*† [62]	18.6%	22.9%	8.1%	0.0%	12.4%
+Finetune*	27.6%	26.7%	12.1%	52.8%	29.8%
+ORV	32.2%	29.6%	15.7%	57.9%	**33.9%**
ΔImprovement	*+4.6%*	*+2.9%*	*+3.6%*	*+5.1%*	***+4.1%***
SpatialVLA*† [77]	12.5%	20.8%	20.8%	58.3%	28.1%
+Finetune*	12.8%	26.1%	26.5%	79.3%	36.2%
+ORV	14.7%	28.4%	27.8%	83.0%	**38.5%**
ΔImprovement	*+1.9%*	*+2.3%*	*+1.3%*	*+3.7%*	***+2.3%***

* The results are reproduced locally for fully fair comparisons.
† Zero-shot performance (Pretraining).

in four RoboDesk tasks and achieves second-best results in the other four tasks, indicating its capability to predict *high-fidelity* future observations, which is fully *controllable*.

4.3. Policy Learning

Setup. To improve policy learning, we employ it as a powerful data engine to augment existing data. Similar to ORV-S2R, we leverage another image generator (ControlNet) to generate diverse initial frames and then extend them to videos, with some examples with appearance randomizations are shown in Fig. 8. For our evaluations, we use post-finetuning after cross-embodiment pre-training as the setup, and leverage ORV to generate additional $\sim 30K$ samples

Table 4. Ablation results of *Video Generation* and *Visual Planning* on approaches of priors injection.

Variants	PSNR↑	SSIM↑	FID↓	FVD↓	Success↑
CogVideoX	19.432	0.752	7.509	83.561	-
Action Conditions					
w/ Text Expert	20.424	0.772	4.104	23.586	52.9
No Chunks	24.813	0.850	3.793	19.944	70.6
Ours (base)	25.631	0.873	3.821	17.682	74.7
Occupancy Map Conditions					
ControlNet	26.974	0.865	3.613	20.069	-
Ours (full)	28.258	0.899	3.418	16.525	-

(refer to Sec. 10.3 in Suppl. for more details). We evaluate the recent open-sourced policy models RoboVLM [62] and SpatialVLA [32] on SimplerEnv-WidowX [56] with BridgeData V2 [93] as the test suite. Each policy model is finetuned both on the original data and the augmented data, following the official instructions. For more discussions about the data augmentation, please refer to Sec. 12.3 in Suppl.

Results. Tab. 3 shows that the augmented data from ORV improves policy learning performance. We keep the original and augmented finetuning data the same in size, with the synthetic data accounting for ∼25% of the augmented data as a practical choice. With the augmentation (the row of "+ORV" in Tab. 3), we achieve gains of ∼13.7% (29.8% → 33.9%) for RoboVLM [62] and ∼6.5% (36.2% → 38.5%) for SpatialVLA [77], which significantly demonstrates the effectiveness of ORV-augmented data for policy learning.

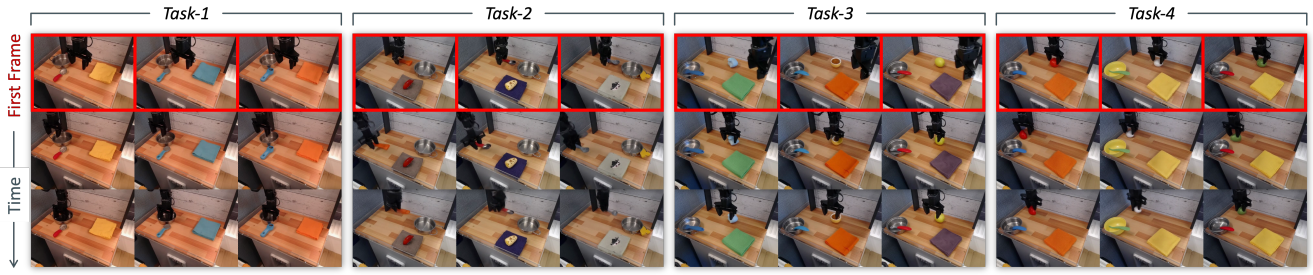

Figure 8. Illustrations of **Appearance Randomization** powered by ORV, generating diverse manipulation videos of four tasks. For each manipulation task in Bridge Data [93], we present three examples with distinct visual appearances, demonstrating that ORV generalizes well to varied context inputs and thereby alleviates the challenge of data collection of robot learning.

Table 5. Ablation results of *Conditional Video Generation* on occupancy conditioning resources and training strategies.

Variants	Source	PSNR↑	SSIM↑	FID↓	FVD↓
Conditioning Resources					
w/o cond. (base)	-	25.631	0.873	3.821	17.682
w/ depth	Fine	30.288	0.919	3.061	14.321
	Coarse	28.031	0.896	4.522	18.548
w/ sem.	Fine	28.896	0.901	3.259	16.171
	Coarse	27.911	0.896	3.467	17.053
Full cond.	Fine	30.431	0.920	2.998	14.301
	Coarse	28.258	0.899	3.418	16.525
Training Strategies (w/o Occupancy Conditionings)					
From scratch	-	23.518	0.811	19.357	84.831
From CogVideoX2B	-	25.631	0.873	3.821	17.682

Table 6. Ablation results of *Multiview Video Generation* on occupancy conditionings on BridgeData V2 [93] with 3 views. Numbers are reported as "with / without" visual priors.

Views	PSNR↑	SSIM↑	FID↓	FVD↓
View0 (anchor)	25.77 / 28.25	0.87 / 0.89	3.20 / 3.11	14.05 / 12.54
View1	23.04 / 25.87	0.79 / 0.85	3.31 / 3.18	16.36 / 13.67
View2	22.90 / 25.79	0.78 / 0.85	3.32 / 3.19	15.97 / 13.62

Table 7. Ablation results of zero-shot *Conditional Video Generation* on different occupancy conditioning resources.

Train	Val	PSNR↑	SSIM↑	FID↓	FVD↓
Coarse	Coarse	28.031	0.896	4.522	18.548
Coarse	Fine	26.608 (-1.423)	0.872 (-0.024)	4.932 (+0.410)	24.134 (+5.586)
Fine	Fine	30.288	0.919	3.061	14.321
Fine	Coarse	19.048 (-11.240)	0.754 (-0.165)	22.893 (+19.832)	132.685 (+109.792)

4.4. Ablation Study and Analysis

Effect of Conditioning Approaches. Tab. 4 ablates our action conditioning (Fig. 4). Altering the Action Expert AdaLN (*e.g.*, combining Vision and Text Experts) significantly degrades performance. Similarly, omitting temporal chunking (directly encoding discrete actions) drops PSNR by 3.2% and success rate by 5.5%. For visual conditioning, injecting occupancy-derived coarse controls into deep layers rather than initial noise reduces PSNR by 4.5

Effect of Control Signals. Tab. 5 evaluates conditioning resources (Coarse: occupancy-rendered; Fine: pixel-level) and types (depth and semantic). The results demonstrate that introducing visual priors leads to significant improvements, with gains of 18.72% (25.621→30.431) and 10.24% (25.621→28.258). Moreover, coarse condition maps achieve performance comparable to their fine counterparts. Additionally, Tab. 6 demonstrates improvements in three-view generation (BridgeData V2 [93]) using view0 as the anchor for multiview priors (details in Suppl. 8).

Effect of Pretraining. We further test the benefits of the pretraining process. As shown in Tab. 5, models trained from the CogVideoX have superior performance compared to those from scratch, particularly on FID and FVD metrics.

Robustness of Occupancy Representations. To validate the robustness of occupancy representations used in ORV model, as described in Sec. 3.1 and Sec. 4.1. We examine the zero-shot performance of models through training and evaluate them under different condition settings, as illustrated in Tab. 7 (refer to Fig. 17 in Suppl. for more qualitative details). The results reveal that models trained on occupancy-derived coarse visual conditions generalize better across conditions of varying granularity. In contrast, ORV models trained on pixel-aligned conditions suffer a dramatic performance drop on coarse inputs. This imposes a major constraint on deploying the model in more diverse scenarios (*e.g.*, from simulation to real-world), necessitating condition maps that accurately align with ground truths. Therefore, previous works [3, 63] are sensitive to inaccuracies in the conditioning, whereas ORV is not.

More Discussions. We have additional broader discussions for a better understanding of our work in Sec. 12 in Suppl.

5. Conclusion

We propose ORV, an occupancy-centric framework for robot video generation that couples action priors with occupancy-derived visual priors. With such an occupancy-centric design, ORV achieves high-quality robot video generation and consistent multiview synthesis. The robustness of occupancy representations further enables ORV to achieve superior visual transfer between simulated and real-world scenarios. Experiments on controllable video generation, visual planning, and policy learning demonstrate the effectiveness and versatility of ORV for advancing robotics research.

References

[1] Cihan Acar, Kuluhan Binici, Alp Tekirdağ, and Yan Wu. Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks. *IEEE Robotics and Automation Letters*, 9(1):691–698, 2023. 2

[2] 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. 2, 9, 11

[3] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025. 2, 3, 4, 8, 11

[4] Ehsan Asali, Prashant Doshi, and Jin Sun. Mvsa-net: Multi-view state-action recognition for robust and deployable trajectory generation. *arXiv preprint arXiv:2311.08393*, 2023. 2

[5] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021. 6

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5, 2, 11, 14

[7] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760*, 2024. 2, 4

[8] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2

[9] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 2

[10] Åke Björck. *Numerical methods for least squares problems*. SIAM, 2024. 1

[11] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 2

[12] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*, 2024. 12

[13] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 5, 6, 7, 1, 3, 14

[14] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 2

[15] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022. 2

[16] Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. *arXiv preprint arXiv:2411.16157*, 2024. 4

[17] 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. 2

[18] Zhuoguang Chen, Kenan Li, Xiuyu Yang, Tao Jiang, Yiming Li, and Hang Zhao. Trackocc: Camera-based 4d panoptic occupancy tracking. *arXiv preprint arXiv:2503.08471*, 2025. 11

[19] Daniil Cherniavskii, Phillip Lippe, Andrii Zadaianchuk, and Efstratios Gavves. Stream: Embodied reasoning through code generation. In *Multi-modal Foundation Model meets Embodied AI Workshop@ ICML2024*, 2024. 2

[20] Xiaowei Chi, Hengyuan Zhang, Chun-Kai Fan, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-min Chan, Wei Xue, Wenhan Luo, Shanghang Zhang, et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461*, 2024. 2

[21] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023. 7

[22] Zhehao Dong, Xiaofeng Wang, Zheng Zhu, Yirui Wang, Yang Wang, Yukun Zhou, Boyuan Wang, Chaojun Ni, Runqi Ouyang, Wenkang Qin, et al. Emma: Generalizing real-world robot manipulation via generative visual transfer. *arXiv preprint arXiv:2509.22407*, 2025. 6, 12

[23] Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control, 2025. 2, 12

[24] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2024. 2

[25] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025. 2

[26] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023. 2

[27] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 1, 3, 4, 14

[28] 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. 2, 11, 12, 13

[29] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022. 6

[30] Songhao Han, Boxiang Qiu, Yue Liao, Siyuan Huang, Chen Gao, Shuicheng Yan, and Si Liu. Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. *arXiv preprint arXiv:2506.06677*, 2025. 13

[31] Shu Han, Xubo Zhu, Ji Wu, Ximeng Cai, Wen Yang, Huai Yu, and Gui-Song Xia. Unicalib: Targetless lidar-camera calibration via probabilistic flow on unified depth representations. *arXiv preprint arXiv:2504.01416*, 2025. 11

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 2015. 7

[33] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 9

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 8

[35] 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. 9

[36] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023. 5, 11, 14

[37] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025. 2

[38] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025. 3

[39] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023. 2, 11

[40] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pages 376–393. Springer, 2024. 2

[41] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 8

[42] 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 neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025. 6

[43] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024. 2

[44] Harini Kannan, Danijar Hafner, Chelsea Finn, and Dumitru Erhan. Robodesk: A multi-task reinforcement learning benchmark. https://github.com/google-research/robodesk, 2021. 6, 14

[45] Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6672–6679. IEEE, 2024. 1

[46] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4

[47] 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. 5, 6, 7, 1, 3, 13, 14

[48] 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. 7

[49] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. 2, 4

[50] Tabitha E Lee, Shivam Vats, Siddharth Girdhar, and Oliver Kroemer. Scale: Causal learning and discovery of robot manipulation skills using simulation. 2023. 1

[51] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 2

[52] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pages 131–148. Springer, 2024. 2

[53] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. *arXiv preprint arXiv:2412.05435*, 2024. 2, 11

[54] Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories, 2025. 2

[55] Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, et al. Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. *arXiv preprint arXiv:2509.22199*, 2025. 6

[56] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 7, 2, 8, 11, 12, 14

[57] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 3

[58] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *arXiv preprint arXiv:2501.18590*, 2025. 2

[59] Chih-Hao Lin, Zian Wang, Ruofan Liang, Yuxuan Zhang, Sanja Fidler, Shenlong Wang, and Zan Gojcic. Controllable weather synthesis and removal with video diffusion models. *arXiv preprint arXiv:2505.00704*, 2025. 2

[60] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024. 1

[61] 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. 7

[62] Huaping Liu, Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, and Hanbo Zhang. Towards generalist robot policies: What matters in building vision-language-action models. 2025. 7, 8, 12

[63] 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 preprint arXiv:2505.23171*, 2025. 2, 3, 4, 8, 11, 12

[64] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 5, 14

[65] Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4d video generation for robot manipulation. *arXiv preprint arXiv:2507.01099*, 2025. 4, 12

[66] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 5

[67] Junfeng Long, Junli Ren, Moji Shi, Zirui Wang, Tao Huang, Ping Luo, and Jiangmiao Pang. Learning humanoid locomotion with perceptive internal model. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9997–10003. IEEE, 2025. 2

[68] Yunhao Luo and Yilun Du. Grounding video models to actions through goal conditioned exploration. *arXiv preprint arXiv:2411.07223*, 2024. 2

[69] Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. *arXiv preprint arXiv:2503.16806*, 2025. 2

[70] Z Mandi, H Bharadhwaj, V Moens, S Song, A Rajeswaran, and V Kumar. Cacti: 256 a framework for scalable multi-task multi-scene visual imitation learning. arxiv preprint 257. *arXiv preprint arXiv:2212.05711*, 258, 2022. 1

[71] Jiageng Mao, Boyi Li, Boris Ivanovic, Yuxiao Chen, Yan Wang, Yurong You, Chaowei Xiao, Danfei Xu, Marco Pavone, and Yue Wang. Dreamdrive: Generative 4d scene modeling from street view images. *arXiv preprint arXiv:2501.00601*, 2024. 2

[72] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3): 7327–7334, 2022. 13

[73] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021. 2

[74] Aviv Netanyahu, Yilun Du, Antonia Bronars, Jyothish Pari, Josh Tenenbaum, Tianmin Shu, and Pulkit Agrawal. Few-shot task learning through inverse generative modeling. *Advances in Neural Information Processing Systems*, 37:98445–98477, 2024. 2

[75] Zezhong Qian, Xiaowei Chi, Yuming Li, Shizun Wang, Zhiyuan Qin, Xiaozhu Ju, Sirui Han, and Shanghang Zhang. Wristworld: Generating wrist-views via 4d world models for robotic manipulation. *arXiv preprint arXiv:2510.07313*, 2025. 2, 4

[76] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024. 3

[77] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 7, 8, 11, 12

[78] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 11, 14

[79] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751*, 2025. 2

[80] Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024. 2, 3, 6

[81] Yinghao Shuai, Ran Yu, Yuantao Chen, Zijian Jiang, Xiaowei Song, Nan Wang, Jv Zheng, Jianzhu Ma, Meng Yang, Zhicheng Wang, et al. Pugs: Zero-shot physical understanding with gaussian splatting. *arXiv preprint arXiv:2502.12231*, 2025. 2

[82] BAAI RoboBrain Team. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 13

[83] GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv e-prints*, pages arXiv–2510, 2025. 6

[84] Qwen Team. Qwen2.5: A party of foundation models, 2024. 6, 7, 14

[85] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5, 14

[86] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. *arXiv preprint arXiv:2304.13723*, 2023. 6, 8

[87] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. In *International Conference on Learning Representations*, 2023. 6

[88] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NeurIPS*, 2024. 11

[89] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[90] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012. 1, 3, 4

[91] 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. 8

[92] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv:2205.09853*, 2022. 6

[93] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. 3, 5, 6, 7, 8, 1, 9, 10, 14

[94] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. 2, 3, 4

[95] Boyuan Wang, Xinpan Meng, Xiaofeng Wang, Zheng Zhu, Angen Ye, Yang Wang, Zhiqin Yang, Chaojun Ni, Guan Huang, and Xingang Wang. Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. *arXiv preprint arXiv:2507.05198*, 2025. 6

[96] Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2404.15014*, 2024. 11

[97] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 1, 2, 3, 14

[98] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024. 2, 11

[99] Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning real-world action-video dynamics with heterogeneous masked autoregression. *arXiv preprint arXiv:2502.04296*, 2025. 2, 3, 6, 9, 11, 12

[100] Nan Wang, Yuantao Chen, Lixing Xiao, Weiqing Xiao, Bohan Li, Zhaoxi Chen, Chongjie Ye, Shaocong Xu, Saining Zhang, Ziyang Yan, et al. Unifying appearance codes and bilateral grids for driving scene gaussian splatting. *arXiv preprint arXiv:2506.05280*, 2025. 11

[101] Nan Wang, Xiaohan Yan, Xiaowei Song, and Zhicheng Wang. Semantic-guided gaussian splatting with deferred

rendering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[102] Sen Wang, Jingyi Tian, Le Wang, Zhimin Liao, Jiayi Li, Huaiyi Dong, Kun Xia, Sanping Zhou, Wei Tang, and Hua Gang. Sampo: Scale-wise autoregression with motion prompt for generative world models. *arXiv preprint arXiv:2509.15536*, 2025. 2, 6, 12

[103] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 2

[104] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*, 2020. 2, 14

[105] Wenbo Wang, Fangyun Wei, Lei Zhou, Xi Chen, Lin Luo, Xiaohan Yi, Yizhong Zhang, Yaobo Liang, Chang Xu, Yan Lu, et al. Unigrasptransformer: Simplified policy distillation for scalable dexterous robotic grasping. *arXiv preprint arXiv:2412.02699*, 2024. 2

[106] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *ICCV*, 2023. 11

[107] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 2

[108] Yunshen Wang, Yicheng Liu, Tianyuan Yuan, Yucheng Mao, Yingshi Liang, Xiuyu Yang, Honggang Zhang, and Hang Zhao. Diffusion-based generative models for 3d occupancy prediction in autonomous driving. *arXiv preprint arXiv:2505.23115*, 2025. 11

[109] Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. *arXiv preprint arXiv:2508.13104*, 2025. 2, 3, 4, 13

[110] 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. 8

[111] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024. 2, 11

[112] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023. 2, 11

[113] 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. 3, 6, 12

[114] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. *arXiv preprint arXiv:2412.04380*, 2024. 2

[115] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024. 2, 11

[116] Yuan Xu, Jiabing Yang, Xiaofeng Wang, Yixiang Chen, Zheng Zhu, Bowen Fang, Guan Huang, Xinze Chen, Yun Ye, Qiang Zhang, et al. Egodemogen: Novel egocentric demonstration generation enables viewpoint-robust manipulation. *arXiv preprint arXiv:2509.22578*, 2025. 6

[117] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 12

[118] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 2

[119] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 4, 6, 14

[120] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3, 2025. 2

[121] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 2

[122] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1

[123] Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025. 3

[124] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: compositional world models for embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*, 2024. 2

[125] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 5, 1, 2, 3, 11, 14

[126] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4, 6, 2

[127] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *ICCV*, 2023. 11

[128] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024. 3, 11

[129] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Hengle Ren, Renjing Xu, and Jian Tang. Roboocc: Enhancing the geometric and semantic scene understanding for robots. *arXiv preprint arXiv:2504.14604*, 2025. 3

[130] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. *arXiv e-prints*, pages arXiv–2503, 2025. 2

[131] 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. 2, 4, 3

[132] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023. 2, 11

[133] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2

[134] 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. 2

[135] 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*, 2025. 2, 3

[136] 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*, 2024. 2, 3, 4, 6, 9, 11, 12

[137] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, Yifeng Zhu, and Kevin Lin. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020. 6, 14