AMAP CV Lab | Alibaba Group

ABot-PhysWorld: Interactive World Foundation Model for Robotic Manipulation with Physics Alignment

AMAP CV Lab

See Contributions section for a full author list.

Abstract

Video-based world models offer a powerful paradigm for embodied simulation and planning, yet state-of-the-art models often generate physically implausible manipulations—such as object penetration and anti-gravity motion—due to training on generic visual data and likelihood-based objectives that ignore physical laws. We present **ABot-PhysWorld**, a 14B Diffusion Transformer model that generates visually realistic, physically plausible, and action-controllable videos. Built on a curated dataset of three million manipulation clips with physics-aware annotation, it uses a novel DPO-based post-training framework with decoupled discriminators to suppress unphysical behaviors while preserving visual quality. A parallel context block enables precise spatial action injection for cross-embodiment control. To better evaluate generalization, we introduce **EZSbench**, the first training-independent embodied zero-shot benchmark combining real and synthetic unseen robot-task-scene combinations. It employs a decoupled protocol to separately assess physical realism and action alignment. ABot-PhysWorld achieves new state-of-the-art performance on PBench and EZSbench, surpassing Veo 3.1 and Sora v2 Pro in physical plausibility and trajectory consistency. We will release EZSbench to promote standardized evaluation in embodied video generation.

Date: March 20, 2026
Correspondence: xumu.xm@alibaba-inc.com
Project Page: https://github.com/amap-cvlab/ABot-PhysWorld

Contents

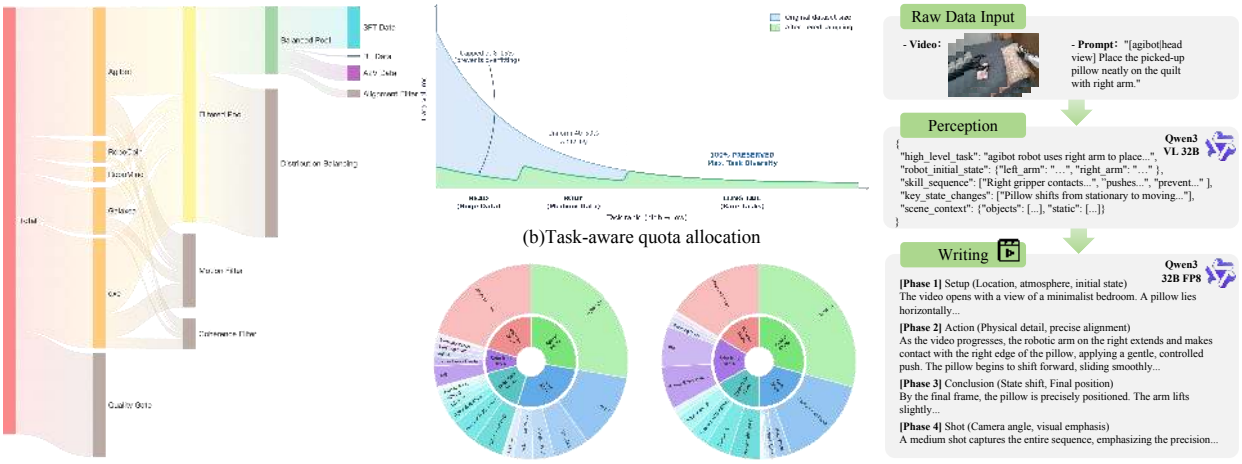
1 Introduction

An embodied world model needs to generate future predictions that adhere to real-world physical laws in order to be effective for simulation, planning, and policy learning. Video generation presents a promising paradigm: such models can serve as simulators for Vision-Language-Action (VLA) policies [19, 25, 44, 45], provide interpretable trajectory previews, or function directly as World Action Models (WAMs) [26, 27, 46] by predicting action-conditioned dynamics—forming critical infrastructure for embodied intelligence.

Despite significant advances in visual fidelity, however, state-of-the-art models like Veo 3.1 [16] and Sora v2 Pro [30] frequently produce manipulation sequences that violate basic physics, including object penetration, contactless motion, and unnatural deformations. These are not mere rendering artifacts but fundamental failures in physical reasoning, limiting their reliability in downstream robotic applications.

This gap arises from two core limitations: (i) training on general visual data lacking rich embodied interaction signals, which hinders the acquisition of fine-grained physical dynamics such as friction, collision response, and mass distribution; and (ii) reliance on standard maximum likelihood objectives during fine-tuning, which treat all prediction errors uniformly and fail to distinguish physically valid from invalid transitions. The absence of both embodied experience and physics-aware supervision results in a systematic disconnect between visual realism and physical plausibility.

To address this, we present **ABot-PhysWorld**, a physically grounded and action-controllable world model based on a 14B Diffusion Transformer [5, 39], built upon a carefully designed data curation pipeline. We integrate three million real-world manipulation clips from five major open-source embodied datasets, enhancing data diversity and balance through curated sampling, ratio optimization, and physics-aware annotation—improving generalization across robots, objects, and environments. Building on this foundation, we introduce a physics-inspired DPO-based [8, 33, 33, 40] post-training framework with decoupled discriminators that suppress unphysical behaviors (e.g., object penetration, anti-gravity motion) while preserving visual quality and improving dynamic consistency. A parallel context block enables multi-channel spatial action injection, supporting precise cross-embodiment control and action-aligned motion synthesis. Together, ABot-PhysWorld generates visually realistic, physically plausible, and highly controllable manipulation sequences—serving as a high-fidelity interface for robot simulation and planning.



(a) Data Filtering and Balancing Pipeline (b)Task-aware quota allocation (c) Dataset and Robot Type Distribution (d) Physics-Aware Video Captioning

Figure 1 Overview of the data curation pipeline. (a) shows the multi-stage filtering and balancing flow from raw aggregation (∼3M clips) to training-ready splits (SFT, RL, and A2V data). (b) Task-aware quota allocation: head tasks are capped at 8–15%, body tasks are uniformly sampled at 40–50%, and long-tail tasks are fully preserved to maximize task diversity. (c) Dataset and robot type distribution: the left ring shows the original composition and the right ring shows the rebalanced result after hierarchical sampling. (d) Physics-aware video captioning pipeline: a perception module (Qwen3-VL 32B) extracts structured physical attributes, followed by a writing module (Qwen3 32B FP8) that generates four-phase captions covering scene setup, action detail, state transition, and camera summary.

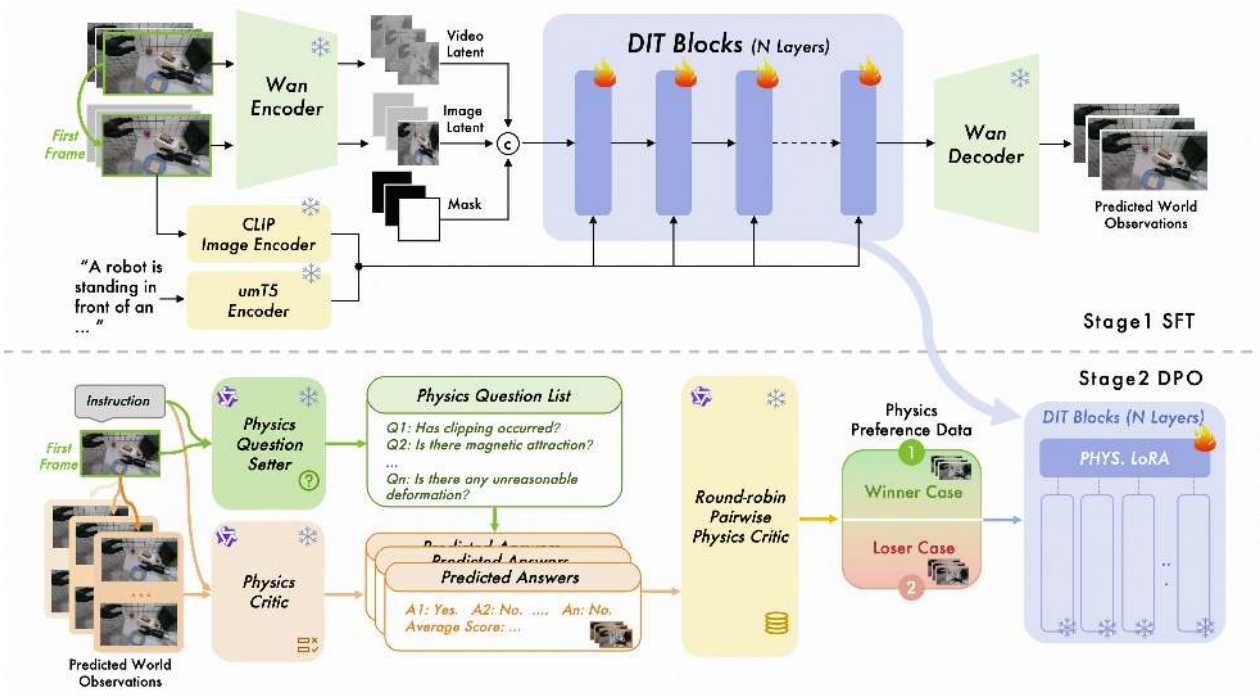


Figure 2 Two-stage training pipeline. Stage 1: SFT on the DiT to predict future frames from observations and instructions. **Stage 2**: generate N candidates, score via physics checklist, and apply DPO via LoRA on frozen DiT weights.

However, evaluating such advances remains challenging: existing benchmarks often emphasize visual quality or in-distribution accuracy, with little emphasis on physical consistency or zero-shot generalization. To enable more rigorous and realistic assessment, we propose **EZSbench**, the first training-independent **E**mbodied **Z**ero-**S**hot **Bench**mark combining both *real* and *synthetic* scenarios involving unseen combinations of robots, tasks, and scenes. Unlike existing benchmarks biased toward in-distribution fidelity, EZSbench is specifically designed to assess three key capabilities: action controllability, physical consistency, and zero-shot generalization. It employs a decoupled dual-model evaluation protocol that separately scores physical realism and action alignment, enabling fine-grained diagnosis of model behavior. We will publicly release EZSbench to promote standardized and meaningful progress in embodied video generation.

Our model achieves new state-of-the-art results on both PBench and EZSbench, surpassing Veo 3.1 and Sora v2 Pro in physical plausibility and action trajectory consistency. More details are provided in Section 5. Our primary contributions are:

- **Data**: We design a principled data curation pipeline that improves diversity and balance in embodied video data through curated sampling and physics-aware annotation—enabling scalable and robust training on real-world interactions.
- **Model**: We propose **ABot-PhysWorld**, a unified framework that jointly optimizes visual realism, physical plausibility, and action controllability through physics-aware DPO and parallel spatial action injection.
- **Evaluation**: We introduce **EZSbench**, the first training-independent zero-shot benchmark for embodied video generation, with a decoupled protocol to assess physical fidelity and action alignment under distribution shift.

2 Data Curation

Data is essential for high-quality embodied world models. Following [17], we adopt a data-driven approach through a systematic infrastructure that enhances data scale and diversity to address complex human-robot

interaction modeling. As illustrated in Figure 1, our data curation pipeline consists of three stages: embodied-specific filtering (§2.1), hierarchical distribution balancing (§2.2), and physically grounded caption generation (§2.3).

2.1 Embodied-Specific Data Filtering

To build a physically consistent world model for embodied manipulation, we construct a foundational dataset of nearly three million real-world video clips by integrating five public datasets: AgiBot [7], RoboCoin [43], RoboMind [42], Galaxea [20], and OXE [31].

General-domain curation pipelines such as Cosmos-Curate [29] and VideoX-Fun [3] are misaligned with embodied data: they rely on scene-cut detectors unsuitable for static-background manipulation videos, and prioritize visual aesthetics over physical causality. To resolve noise introduced by raw aggregation, we apply a video-level quality gate followed by three semantic filtering stages.

Video-level quality gate. Clips with abnormal resolutions or moving cameras are discarded. Sequences are constrained to 80–500 frames; longer videos are segmented temporally by task index into training-compliant clips to ensure relevance and efficiency.

Optical-flow-based motion filtering. We extract grayscale frames at 2 FPS and compute Farnebäck dense optical flow [11] to capture pixel-level motion. By averaging the polar magnitudes of displacement vectors across each frame, we derive a global kinematic score and remove clips with near-zero motion or unphysical oscillations.

CLIP-based temporal coherence. To eliminate visual corruption (e.g., black screens, cuts, stitching errors), we assess temporal continuity using CLIP-based embeddings [34]. Eight equidistant frames are sampled per clip, and their 768D features are extracted; samples with low average cosine similarity between consecutive frames are discarded.

Vision-action alignment verification. Calibrated action maps, encoding joint actions, end-effector poses, and gripper states, are projected onto video frames. Qwen3-VL verifies spatiotemporal alignment between visual motion and control signals, filtering out mismatches from sensor calibration or synchronization errors.

The resulting dataset provides a robust foundation for training generalizable, dynamics-aware world models across diverse embodied tasks.

2.2 Hierarchical Distribution Balancing

Recent studies indicate that data diversity, not just volume, is key to scalable world models and generalist robotic policies [23, 46]; scaling repetitive data often leads to memorization rather than out-of-distribution generalization. To address this, we design a hierarchical dynamic sampling strategy spanning four levels: video, sub-dataset or robot type, task, and macro-dataset. This approach balances data distribution while preserving long-tail features.

Level 1: Intra-dataset diversity preservation. Several source datasets are themselves aggregations of smaller collections; for example, small sub-datasets within OXE [31] are retained entirely to preserve unique interaction patterns before any cross-dataset operations are applied.

Level 2: Cross-robot rebalancing. Operating across the five source datasets, this level addresses imbalances among robot embodiment types. Underrepresented robot types are upweighted to retain rare interaction patterns (e.g., non-standard kinematics or dual-arm coordination), enhancing cross-platform generalization and mitigating head-category dominance (Figure 1c).

Level 3: Task-aware quota allocation. Rather than applying a fixed sampling threshold, we partition tasks into three tiers based on data volume and assign tier-specific strategies (Figure 1b). Head tasks (high data volume) are capped at 8–15% of their original size to prevent overfitting to dominant categories. Body tasks (medium volume) are uniformly sampled at 40–50%, preserving representative coverage without excessive redundancy. Long-tail tasks (rare tasks) are fully preserved to maximize task diversity.

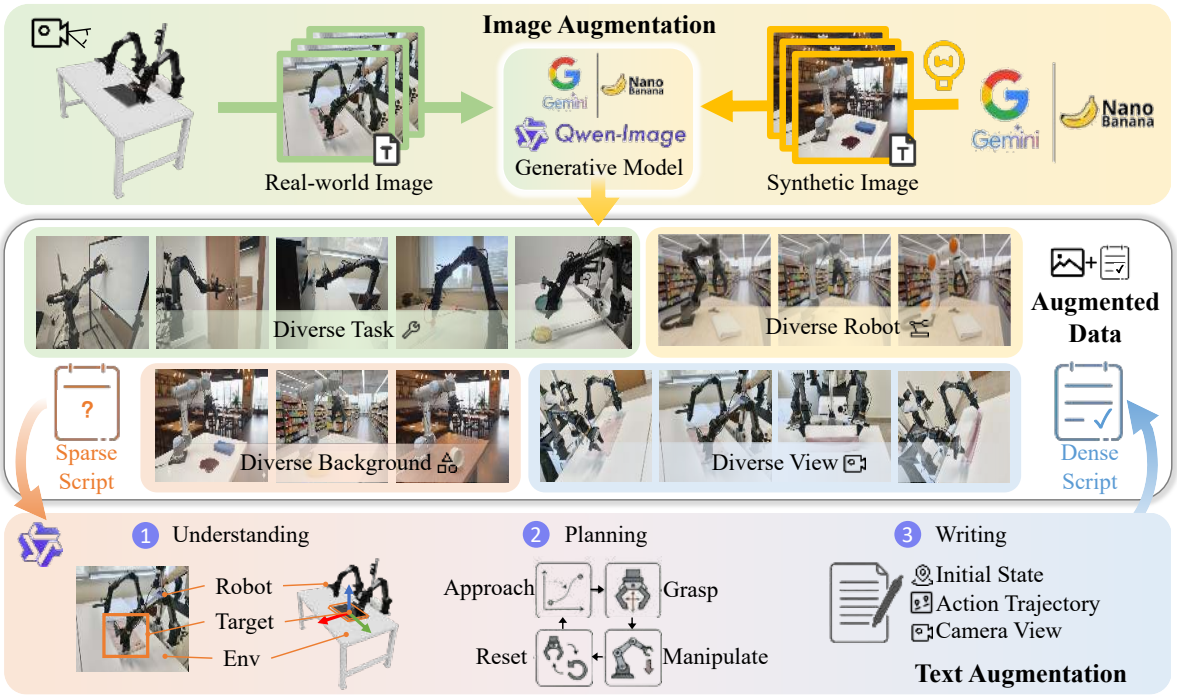


Figure 3 Construction pipeline of the EZSbench. Top: dual-source image augmentation—Branch 1 generates synthetic initial observations via text-to-image (Nano Banana) by varying robot morphology, scene, task, and viewpoint; Branch 2 applies VLM-guided background editing to real-world images while preserving foreground interactions. **Down**: three-stage dense description synthesis—visual anchoring grounds the scene layout and object coordinates, action simulation infers kinematically compliant trajectories with micro-physical interactions, and narrative synthesis produces a documentary-style caption integrating initial state, trajectory, and final state.

Level 4: Macro-dataset scale regulation. Finally, at the coarsest granularity, large-scale macro-datasets (e.g., AgiBot, OXE) are capped via uniform subsampling , while micro-datasets (e.g., RoboMind) are guaranteed minimum coverage through a mandatory lower bound. When activated, a three-round supplementation strategy allocates: (1) a base quota uniformly across tasks, (2) reallocates unused quotas proportionally, and (3) fills residual gaps via random fallback sampling from the global filtered pool, preventing single-task over-extraction and improving long-tail balance. This hierarchical framework improves combinatorial diversity and distributional balance, providing a robust foundation for training general-purpose embodied world models.

2.3 Physics-Aware Video Captioning

Training embodied world models with text-to-video (T2V) objectives requires captions that go beyond surface-level scene descriptions. An effective annotation must capture three progressively deeper aspects of robotic manipulation: what the robot does (action semantics), how it interacts with the physical world (spatial and contact precision), and why the observed outcome occurs (causal reasoning). We design a multi-level annotation system that addresses each of these requirements.

Multi-level action semantics. Adopting a robot-centric "annotating for action" philosophy, we structure each caption across four granularities: macroscopic task intent in natural language; mesoscopic verb-noun action segmentation for long-horizon planning; microscopic details including Cartesian trajectories, relative motion, and gripper states; and scene-level descriptions of physical relations (contact, support, containment) and task outcomes (success, failure, partial accidents).

Grounded spatial precision. Purely template-driven annotation tends to produce hallucinated spatial relations and imprecise grasp descriptions. To suppress these errors, we introduce three mechanisms: few-shot in-context learning with explicit positive and negative examples for richer physical detail, dynamic vocabularies for precise grasp-type specification, and a visible-fact baseline that restricts descriptions to observable evidence.

Causal physical modeling. Beyond describing what happens, effective captions for world models must explain

Table 1 Quantitative comparison on the PAI-Bench robot domain subset.

Model	AQ	BC	IQ	MS	OC	SC	I2VB	I2VS	Quality Score	Domain Score	Avg.
Wan 2.5	**0.5477**	0.8985	0.6458	0.9623	0.2190	0.8784	0.9438	0.9428	0.7548	0.8644	0.8096
GigaWorld-0	0.4757	0.9219	0.6506	0.9908	0.1944	0.9111	0.9673	0.9607	0.7591	0.8583	0.8087
Veo 3.1	0.5458	0.9216	**0.7244**	0.9712	**0.2214**	0.9146	0.9317	**0.9614**	**0.7740**	0.8350	0.8045
Wan2.1_14B	0.4723	0.9315	0.7118	**0.9917**	0.1921	0.9185	0.9745	0.9451	0.7672	0.8391	0.8032
WoW-wan 14B	0.4664	0.9295	0.7027	0.9858	0.1941	0.9149	0.9613	0.9292	0.7605	0.8301	0.7953
Cosmos-Predict 2.5	0.4897	0.9166	0.7405	0.9906	0.1911	0.8973	0.9304	0.9030	0.7574	0.8021	0.7797
Sora v2 Pro	0.5324	0.9285	0.6956	0.9702	0.2203	0.9163	0.9507	0.9290	0.7679	0.7626	0.7652
UnifoLM-WMA-0	0.4547	**0.9423**	0.6564	0.9875	0.1878	**0.9412**	0.9638	0.9403	0.7593	0.6693	0.7143
Our Model	0.4620	0.9373	0.6906	0.9916	0.1927	0.9406	**0.9777**	0.9498	0.7678	0.8785	0.8232
Our Model + DPO	0.4667	0.9365	0.6916	0.9908	0.1942	0.9355	0.9768	0.9483	0.7676	**0.9306**	**0.8491**

why it happens. We explicitly annotate physical causality, including gravity-induced dropping, surface deformation, and force feedback. A four-stage narrative structure (scene construction, action flow, final state confirmation, and camera summary) organizes each caption into a temporally coherent account of the manipulation episode.

This annotation system delivers physically grounded language supervision that captures not only events but their underlying causes, providing the semantic foundation for training world models with causal understanding.

3 Method

3.1 Embodied Video Generation Backbone

Generating physically plausible manipulation videos requires a backbone that captures both the visual diversity of real-world scenes and the fine-grained spatiotemporal dynamics of robot-object interactions. To meet this requirement, we build upon Wan2.1-I2V-14B [5], and fully fine-tune it on our curated embodied dataset.

3.2 Physical Preference Alignment

While SFT teaches the model to reproduce training distributions, it treats all samples equivalently and cannot distinguish physically correct predictions from those containing violations such as object penetration or anti-gravity motion. To explicitly suppress these violations, we propose a post-training preference alignment pipeline (Figure 2) that pairs a decoupled VLM discriminator with Diffusion-DPO.

3.2.1 Decoupled VLM Discriminator

For a given prompt x and initial state, we generate N candidate video variants. Evaluating physical plausibility with a single VLM risks self-evaluation hallucinations, where the same model that generates questions also judges answers. To prevent this, we decouple the evaluation into two roles.

The Qwen3-VL 32B Thinking model acts as the proposer. It observes the first frame and text instruction to dynamically generate a task-specific physical checklist based on a hierarchical evaluation system. This system applies single-vote veto power to Tier 1 metrics (fatal violations such as penetration and anti-gravity) and uses Tier 2 metrics (micro-physical fidelity and contact dynamics) to differentiate compliant samples. Generating specific questions prevents hallucinations caused by vague queries. For example, given the instruction to grasp and place an apple, the proposer asks whether the gripper penetrates the apple, whether the apple penetrates the bag, and whether it is firmly grasped rather than magnetically attached. The proposer also explicitly constructs a balanced mix of positive and negative questions to prevent the scoring model from sycophantically predicting the absence of violations.

The Gemini 3 Pro model [13] then acts as the scorer. It uses explicit Chain-of-Thought reasoning, including global scanning, marking suspicious frames, and backtracking confirmation, to evaluate the N variants against the generated checklist. To efficiently resolve score ties and isolate the optimal (y_w) and worst (y_l) samples within $\mathcal{O}(N)$ complexity, we apply a multi-round tournament-based sampling strategy: a knockout tournament first selects the optimal sample, followed by a loser-bracket round to identify the worst sample. This two-stage mechanism avoids full permutation comparisons and yields highly discriminative DPO [35] training triplets (x, y_w, y_l) with clear margins.

3.2.2 Diffusion-DPO Training

Given the discriminative triplets (c, v_w, v_l) produced by the decoupled discriminator, where c is the condition, v_w the physics-compliant video, and v_l the physics-violating video, we adapt the Diffusion-DPO framework to fine-tune the video diffusion model directly in the latent space. For a video latent z, we inject Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ at time step $t \sim \mathcal{U}(0, T)$ to obtain z_t. The single-step denoising mean squared error for model ϵ_θ is $L(\theta, z) = \|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2$. Letting $L_\theta(\cdot)$ and $L_{ref}(\cdot)$ denote the denoising errors of the policy model π_θ and reference model π_{ref} (the SFT baseline) respectively, the physical preference alignment loss is:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{z,\epsilon,t}\left[\log\sigma\left(-\frac{\beta}{2}\Big[\underbrace{(L_\theta(z_w) - L_\theta(z_l))}_{\text{Policy Diff.}} - \underbrace{(L_{ref}(z_w) - L_{ref}(z_l))}_{\text{Ref. Diff.}}\Big]\right)\right], \quad (1)$$

where β controls distribution divergence, and z_w, z_l are the latents of v_w, v_l. This objective actively reduces the prediction error for z_w while increasing it for z_l at each timestep.

Standard DPO requires maintaining two complete computation graphs (π_θ and π_{ref}), causing out-of-memory errors for a 14B DiT. To resolve this, we freeze the DiT backbone and inject Low-Rank Adaptation [18] (LoRA) modules with a rank of 64 into the self-attention (query, key, value, output) and feed-forward layers, so that the reference model loss L_{ref} can be computed by temporarily disabling the LoRA weights with zero additional memory.

3.3 Action-Conditioned Video Generation

Beyond text-conditioned prediction, a world model for embodied intelligence must support controllable generation: given the current observation and a future action sequence, it should produce physically plausible videos that faithfully follow the commanded trajectory. Directly injecting low-dimensional robotic commands (e.g., end-effector poses) into high-dimensional visual pipelines creates a semantic gap. To bridge this gap, we convert discrete action commands into spatially structured action maps and inject them through parallel context blocks that preserve the backbone's pre-trained physical knowledge.

3.3.1 Action Map Construction

The input action is a 7D vector $\boldsymbol{a} \in \mathbb{R}^7$ (3D position, 3D orientation, gripper openness), extending to 14 dimensions for dual-arm systems. Using camera intrinsics and extrinsics, we project the 3D position (x, y, z) to a 2D center (u, v). Orientation is encoded as the three principal axes of the corresponding rotation matrix, projected into the image plane and rendered as colored arrows whose length encodes depth. The gripper state is mapped to a circular mask at (u, v), with opacity linearly indicating openness. For dual-arm robots, we distinguish left and right arms via red and blue channels, yielding a multi-channel action map.

3.3.2 Action Injection

Existing action injection methods either use Adaptive Layer Normalization (AdaLN) [32] for MLP-encoded actions [1, 48], which hinders cross-embodiment generalization, or concatenate action maps directly with noisy latents for full fine-tuning [21, 28, 36], causing catastrophic forgetting of pre-trained physical priors.

To address these challenges, as shown in Figure 4, we clone selective blocks from the main DiT [39] to form a parallel set of context blocks that process the action maps [22]. The output of each context block is projected via zero-initialized convolution layers and added residually to the corresponding main DiT block:

$$\mathbf{x}_i = \mathrm{DiT}_i(\mathbf{x}_{i-1}) + \alpha \cdot W_{\text{zero}}^{(i)}\mathbf{h}_i, \quad (2)$$

where $\mathbf{h}_i$ is the i-th context block output, W_{zero} is the zero-initialized convolution layer, and α is the control scale. Following VACE, we instantiate context blocks selectively, replicating only every fifth DiT block.

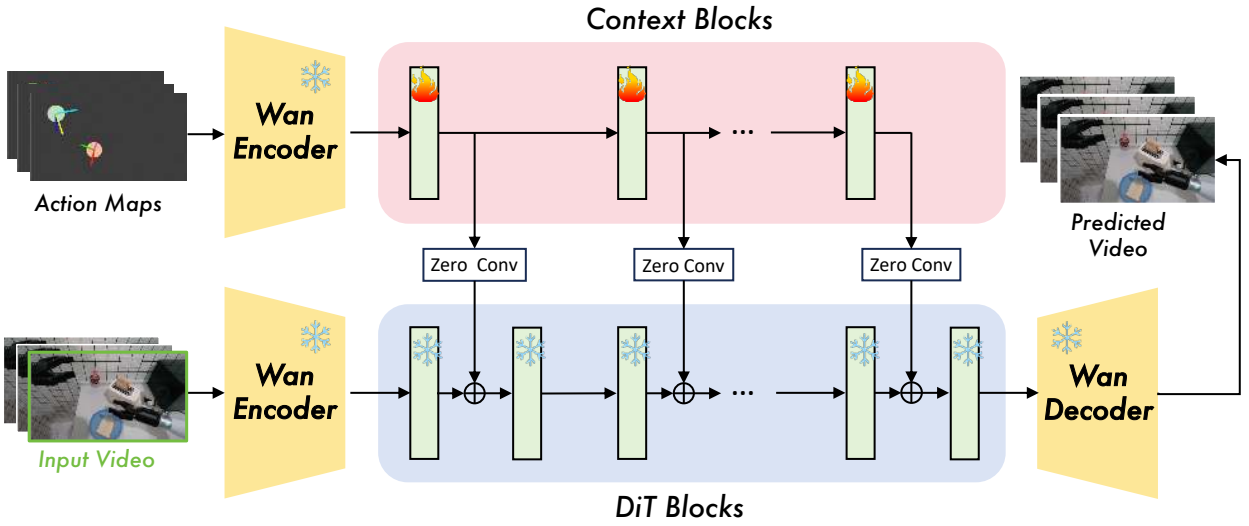


Figure 4 Architecture of the action-conditioned video generation model. We selectively duplicate DiT blocks as parallel context blocks to process action maps, and fuse their outputs residually into the main DiT.

Table 2 Quantitative comparison on EZSbench.

Model	AQ	BC	IQ	MS	OC	SC	I2VB	I2VS	Quality Score	Domain Score	Avg.
WoW-wan 14B	0.4764	0.9412	**0.7514**	0.9922	0.1236	0.9347	0.9495	0.9178	0.7609	0.7951	0.7780
GigaWorld-0	0.4466	0.8893	0.6565	0.9889	0.1222	0.8678	0.9325	0.9139	0.7272	0.7826	0.7549
Cosmos-Predict 2.5	0.4148	0.8835	0.6810	0.9878	0.0970	0.8366	0.9054	0.8653	0.7089	0.7698	0.7394
UnifoLM-WMA-0	**0.4861**	**0.9575**	0.7390	**0.9925**	0.1046	**0.9452**	0.8421	0.8170	0.7355	0.5232	0.6294
Our Model	0.4789	0.9494	0.7483	0.9910	**0.1301**	0.9442	**0.9680**	**0.9453**	**0.7694**	**0.8366**	**0.8030**

Because the zero initialization ensures that the context branch contributes no signal at the start of training, the backbone weights remain undisturbed, preserving pre-trained physical priors while gradually learning action controllability.

4 Embodied-ZeroShot Benchmark

Existing embodied video generation benchmarks draw test samples from the same distribution as training data, making it difficult to assess genuine zero-shot generalization. We introduce EZSbench to evaluate physical fidelity and cross-embodiment generalization under fully out-of-distribution conditions, where diverse robot morphologies, environments, and tasks are composed into previously unseen combinations with no overlap with training data.

4.1 Evaluation Set

We construct the initial observation pool via a dual-branch strategy. The first branch generates synthetic images using the text-to-image model Nano Banana [15], controlled by four orthogonal variables: robots, scenes, tasks, and perspectives. This approach targets morphological, scene, and task generalization by varying arm structures, backgrounds, and task complexity from basic pick-and-place to long-horizon manipulations. The second branch uses a large VLM for controllable scene editing on real-world mechanical arm images, dynamically altering backgrounds while preserving foreground physical interactions.

To generate reliable physical descriptions, we propose a physics-heuristic dense description synthesis framework that progresses through visual anchoring, kinematically compliant action simulation, and narrative synthesis, producing text that integrates the initial state, action trajectory, and final state. Each initial image paired with its dense description forms a core benchmark sample.

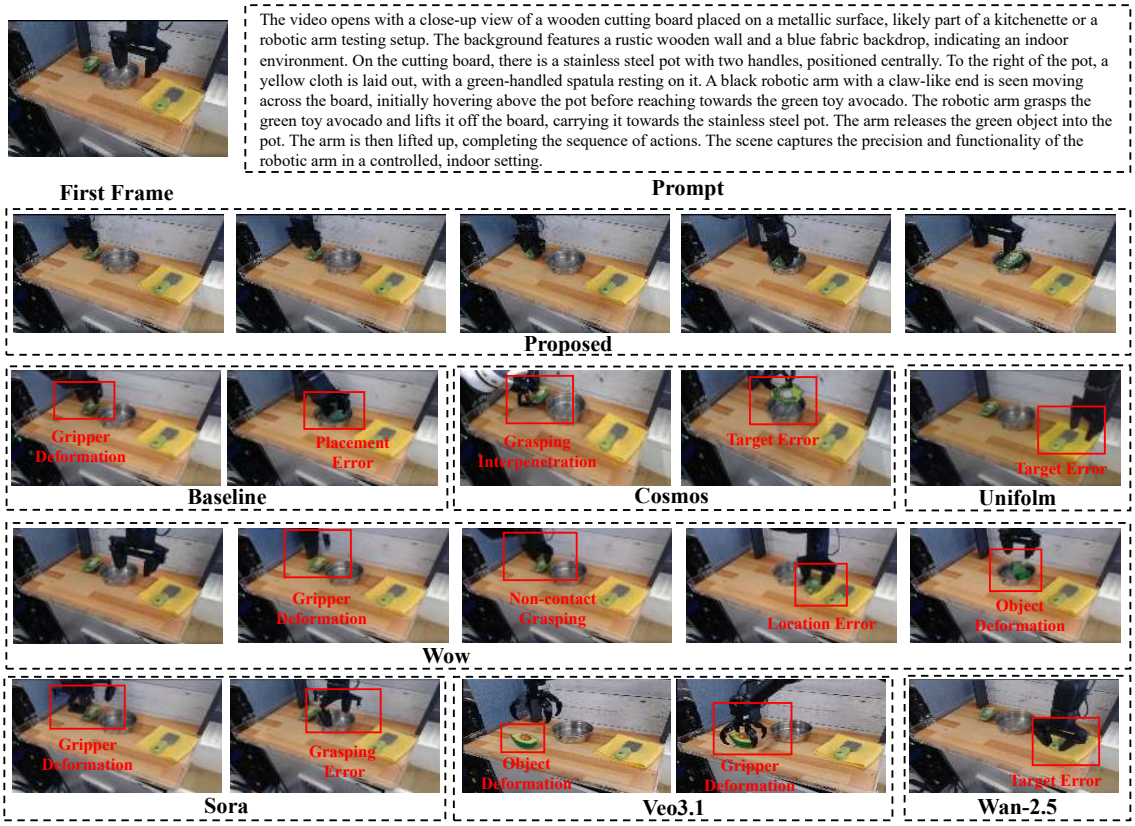


Figure 5 Qualitative comparison on PAI-Bench.

4.2 Evaluation Method

A key challenge in evaluating physical consistency is that a single model acting as both question generator and answer judge introduces self-evaluation bias. To address this, we propose a decoupled dual-model evaluation paradigm. The Qwen3-VL-32B-Thinking model dynamically generates physical checklist questions based on the initial state and text instructions. It employs a System 2 reasoning protocol for scene decoding and action parsing, guided by few-shot examples covering nine criteria across spatial, temporal, and physical dimensions. We mandate that 30–50% of the checklist comprises negative questions, such as asking whether a red apple is green, to prevent shortcut learning via random guessing. To eliminate self-evaluation bias, the Qwen2.5-VL-72B-Instruct model serves as the answering end. The final physical score S_v for a video v measures the consistency between the visual question answering (VQA) predictions and the checklist ground truth (GT): $S_v = \frac{1}{|Q_v|} \sum_{q \in Q_v} \mathbb{I}(\text{VQA}(v, q) = \text{GT}(q))$, where Q_v is the generated checklist and $\mathbb{I}(\cdot)$ is the indicator function.

5 Experiments

5.1 Implementation Details

We conduct all experiments on a cluster of 128 Nvidia H20 GPUs. The training pipeline consists of three stages: TI2V foundational training, DPO, and A2V training. **For TI2V**, we use Wan2.1-I2V-14B-480P with inputs cropped to 480×832 and 81 frames uniformly sampled. Trained for 6,000 steps with a global batch size of 128 and learning rate 1e-5. **For DPO**, we apply LoRA-based fine-tuning to the Diffusion Transformer, inserting rank-64 adapters (scaling factor 64) into self-attention and feed-forward layers (q, k, v, o, ffn.0, ffn.2). Optimized with AdamW, lr=1e-6, 10-step warmup. To strengthen preference signals in diffusion models, we set $\beta = 5000$. Training employs BF16 mixed precision, gradient checkpointing, per-device batch size 1, and runs for 500 steps/epoch over 100 epochs. **For A2V**, we adopt the VACE framework on the fine-tuned TI2V

model. We duplicate specific Diffusion Transformer layers (0, 5, 10, 15, 20, 25, 30, 35) as a trainable context branch while keeping the backbone frozen. Data is augmented via random frame sampling with variable stride. Training uses batch size 16, learning rate 5e-5, and 20,000 steps.

5.2 Evaluation Setup

Text-Conditioned Generation. We evaluate physical plausibility and visual quality using PAI-Bench [47] and its PBench dataset, focusing on the robot domain subset with 174 complex manipulation videos from BridgeData V2 [38], AgiBot, and Open X-Embodiment. We employ an MLLM-as-Judge approach with Qwen2.5-VL-72B-Instruct [6] for binary visual question answering. The Domain Score evaluates accuracy across 886 questions in three dimensions: spatial (36.3%, geometry and contact), temporal (28.6%, causal logic), and physical (34.1%, object attributes and state changes). We also use PAI-Bench's multidimensional quality metrics: subject [9]/background [12] consistency, Overall consistency [41], Aesthetic quality (LAION aesthetic head), Imaging quality [24], motion smoothness, and i2v subject/background consistency. For zero-shot evaluation, we use EZSbench with the decoupled dual-model protocol from Section 5.

Action-Conditioned Generation. We construct the action-conditioned evaluation set by uniformly sampling 200 instances from the action-to-video dataset, each containing an initial frame and a structured action sequence (end-effector pose and gripper state). Visual alignment is evaluated frame-by-frame using PSNR for pixel accuracy and SSIM for local texture fidelity. For trajectory accuracy, we use nDTW: a fine-tuned YOLO detector locates the gripper in each frame, and the extracted trajectory is compared to ground truth via nDTW.

Baselines. We compare with Cosmos-Predict 2.5-2B [2], GigaWorld-0 [14], UnifoLM-WMA-0 [37], WoW-wan 14B [10], Veo 3.1 [16], Sora v2 Pro [30], and Wan 2.5 [4] for text-conditioned generation, and with Enerverse-AC [21] and Gen-Sim [28] for action-conditioned generation.

Table 3 Quantitative results on action-conditioned generation.

Model	PSNR	SSIM	Traj. Consis.
Enerverse-AC	20.42	0.7542	0.8157
Gen-Sim	18.05	0.7413	0.6195
Ours	**21.09**	**0.8126**	**0.8522**

5.3 Evaluation Results

PBench Evaluation. As shown in Table 1, our DPO-augmented model achieves the highest average score (0.8491) and sets a new state-of-the-art Domain Score (0.9306), outperforming the base model (0.8785) and all baselines. Existing methods show a trade-off between visual quality and physical fidelity: Veo 3.1 and Sora v2 Pro achieve high Quality Scores (0.7740, 0.7679) due to strong imaging and aesthetics, but lag in Domain Score (0.8350, 0.7626), favoring perception over physics. Our model maintains competitive visual quality (Quality Score: 0.7676) while enforcing physical constraints, proving that alignment with physical laws does not sacrifice perceptual quality. The base model also shows strong spatiotemporal stability (I2VB: 0.9777; MS: 0.9916).

EZSbench Evaluation. On the out-of-distribution EZSbench (Table 2), our model achieves the highest overall average score (0.8030), establishing state-of-the-art results for both Quality Score (0.7694) and Domain Score (0.8366). This confirms that the physical fidelity improvements generalize beyond the training distribution.

Qualitative Analysis. Figure 5 shows qualitative PBench comparisons. Baselines violate physical laws in complex interactions: Sora v2 Pro and Veo 3.1 show gripper or object distortion during dense contact; GigaWorld-0 and Cosmos exhibit grasping penetration; WoW produces non-contact grasping and geometric distortion; UnifoLM and Wan 2.5 misidentify targets (e.g., spatula instead of rag). Our method correctly identifies targets, maintains spatiotemporal coherence, and avoids deformation and penetration.

Action-Conditioned Generation Results. As shown in Table 3, our method outperforms baselines in both visual quality and action fidelity. Our method consistently outperforms the baselines by substantial margins.

6 Conclusion

We introduce ABot-PhysWorld, a physically grounded and action-controllable world model for embodied manipulation based on a 14B Diffusion Transformer. It integrates curated data, physical alignment through Diffusion-DPO, and spatial action injection to reduce physical violations while maintaining control across different embodiments. We also propose EZSbench, a zero-shot benchmark featuring out-of-distribution scenarios and a decoupled evaluation protocol. Experimental results show state-of-the-art physical fidelity and improved trajectory consistency compared to Veo 3.1 and Sora v2 Pro. The model currently relies on fixed-viewpoint data and lacks closed-loop evaluation. Future work will explore multi-view generation and real-world deployment.

7 Contributions

Author contributions in the following areas are as follows:

- **Data Curation:** Yuzhi Chen, Ronghan Chen, Dongjie Huo, Haoyun Liu, Yandan Yang, Dekang Qi, Tong Lin, Shuang Zeng, Junjin Xiao
- **Model Training:** Yuzhi Chen, Ronghan Chen
- **Evaluation:** Yuzhi Chen, Ronghan Chen, Dongjie Huo
- **Writing:** Yuzhi Chen, Yandan Yang, Ronghan Chen, Dongjie Huo, Dekang Qi
- **Project Lead:** Xinyuan Chang, Feng Xiong
- **Advisor:** Zhiheng Ma, Xing Wei, Mu Xu[†]

[†]Corresponding author: xumu.xm@alibaba-inc.com

References

[1] 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] Niket Agarwal et al. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575, 2025.

[3] aigc-apps. Videox-fun: A more flexible framework that can generate videos at any resolution and creates videos from images. https://github.com/aigc-apps/VideoX-Fun, 2024.

[4] Alibaba Group. Wan 2.5: Open-source ai video generation with audio. https://wan.video/, 2025.

[5] Alibaba Group. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[6] Shuai Bai et al. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[7] 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.

[8] Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. arXiv preprint arXiv:2512.24551, 2025.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.

[10] Xiaowei Chi et al. Wow: Towards a world omniscient world model through embodied interaction. arXiv preprint arXiv:2509.22642, 2025.

[11] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis, pages 363–370. Springer, 2003.

[12] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In Advances in Neural Information Processing Systems, volume 36, pages 50742–50768, 2023.

[13] Gemini Team. Gemini 3 technical report. arXiv preprint, 2025. URL https://deepmind.google/technologies/gemini/.

[14] GigaAI. Gigaworld-0: World models as data engine to empower embodied ai. arXiv preprint arXiv:2511.19861, 2025.

[15] Google. Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025.

[16] Google DeepMind. Veo 3.1 ingredients to video: More consistency, creativity and control. https://blog.google/innovation-and-ai/technology/ai/veo-3-1-ingredients-to-video/, 2026.

[17] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. IEEE intelligent systems, 24(2):8–12, 2009.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.

[19] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

[20] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. arXiv preprint arXiv:2509.00576, 2025.

[21] 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.

[22] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17191–17202, 2025.

[23] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model? – a physical law perspective. arXiv preprint arXiv:2411.02385, 2024.

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5148–5157, 2021.

[25] 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.

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

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

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

[29] NVIDIA. Cosmos-curate: A powerful video curation system that processes, analyzes, and organizes video content using advanced ai models and distributed computing. https://github.com/nvidia-cosmos/cosmos-curate, 2025.

[30] OpenAI. Sora 2 system card. https://openai.com/index/sora-2-system-card/, 2025.

[31] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 6892–6903. IEEE, 2024.

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4195–4205, 2023.

[33] Wenxu Qian, Chaoyue Wang, Hou Peng, Zhiyu Tan, Hao Li, and Anxiang Zeng. Rdpo: Real data preference optimization for physics consistency video generation. arXiv preprint arXiv:2506.18655, 2025.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[35] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems, volume 36, 2023.

[36] GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied ai. arXiv preprint arXiv:2511.19861, 2025.

[37] Unitree. Unifolm-wma-0: A world-model-action (wma) framework under unifolm family. https://huggingface.co/unitreerobotics/UnifoLM-WMA-0-Base, 2025.

[38] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In Conference on Robot Learning (CoRL), 2023.

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[40] Peiyao Wang, Weining Wang, and Qi Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. arXiv preprint arXiv:2511.03997, 2025.

[41] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In The Twelfth International Conference on Learning Representations, 2024.

[42] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877, 2024.

[43] Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. arXiv preprint arXiv:2511.17441, 2025.

[44] Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. arXiv preprint arXiv:2601.18692, 2026.

[45] Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. arXiv preprint arXiv:2602.11236, 2026.

[46] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

[47] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. arXiv preprint arXiv:2512.01989, 2025.

[48] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9834–9844, 2025.

Appendix

Code Repository: https://github.com/amap-cvlab/ABot-PhysWorld

Contents

A. Vision-Action Alignment Verification

Misaligned action-video pairs—arising from sensor calibration drift, clock synchronization errors, or coordinate frame inconsistencies—inject spurious correlations that prevent the world model from learning faithful physical dynamics. To detect and remove such samples, we render the calibrated action signals (joint positions, end-effector Cartesian poses, and gripper states) as semi-transparent color-coded action maps and overlay them onto the corresponding video frames (Figure 6).

The resulting composite images allow both human annotators and an automated VLM-based verifier (Qwen3-VL) to inspect whether the projected action trajectories match the observed robot motion in pixel space. Clips with significant spatial deviation—e.g. end-effector paths that diverge from the visually observed gripper trajectory, or gripper open/close states that contradict the visual evidence—are flagged and discarded.

Figure 6 Vision-action alignment verification. Calibrated action signals are rendered as semi-transparent action maps and overlaid onto video frames. Samples with spatial deviation between projected trajectories and observed robot motion are identified and removed.

B. Two-Stage Physics-Aware Captioning Pipeline

Section 2.3 describes our two-stage captioning pipeline. Here we provide additional details and representative examples for each stage.

Stage 1: Structured Perception and Attribute Extraction. A vision-language perception module (Qwen3-VL 32B) processes each video clip and extracts structured physical attributes: robot morphology and embodiment type, manipulated objects with their properties (color, shape, material, size), spatial layout and relative object positions, and contact events and state transitions across the manipulation sequence. The output is a structured intermediate representation capturing the "what" and "where" of the scene, which grounds the subsequent writing stage. Representative Stage 1 outputs are shown in Figures 7–9.

The VLA Video Annotation Prompt Template

Character

You are a top-tier robotic manipulation analysis expert and an AI cinematographer, specializing in generating high-quality annotations that combine physical depth and visual aesthetics for VLA (Vision-Language-Action) world model training. You analyze manipulation videos from the robot's first-person perspective, capturing strategies, physical interactions, and state changes, while simultaneously scanning the environmental atmosphere, background details, and overall composition like a film director.

Skills

Skill 1: **Describing Actions and Operational Details**

- Draft the sequence of actions chronologically. Ensure each step is concise, clear, and forms a single sentence starting with a verb that reflects the robot's specific action. Be as detailed as possible without being overly trivial.
- Adapt your descriptions to object characteristics: e.g., use "gently grasp" or "slowly place" for fragile items; "keep moving horizontally" or "prevent rolling" for long items; "adjust angle" or "find support point" for irregular items.
- Capture physical details across multiple dimensions: Describe speed and force (`slowly approach`, `gently pinch`, `firmly grasp`), spatial relationships (`directly above and slightly to the left`, `closely adjacent to the right`, `hovering above`), and grasping methods (`pinch`, `wrap grasp`, `suction`).
- Note adjustment behaviors: Identify fine-tuning (`fine-tune position for alignment`, `readjust angle`), obstacle avoidance (`bypass obstacles`, `adjust in narrow space`), and feedback-driven actions (`test stability`, `secondary position adjustment`).

Skill 2: **Tracking State Changes**

- Identify and record key changes in: object positions, relationships between objects, and object states (e.g., transitioning from resting to being grasped, or being placed).
- Link these changes to specific action steps using a trigger index (starting from 0).
- Describe not only "what changed" but also explicitly state "how it changed" and the "degree of change".
- Base descriptions strictly on visible facts: Only describe state changes clearly visible in the video. Strictly prohibit inferring invisible physical effects like micro-deformations or internal stress. Ensure descriptions of the degree of change are reasonable and visually evident (e.g., a plastic bag deforming is visible and reasonable; a metal grid deforming is usually invisible and should be omitted).

Skill 3: **Analyzing Scene Context and Global Scanning**

- Conduct a full panoramic scan, do not just stare at the manipulated object. Categorize elements into interactive "objects" and "static_objects".
- For interactive "objects": Only include items that actually participate in the interaction (e.g., manipulated objects, target locations/support surfaces, tools, containers, obstacles). Detail their functional category, initial state (spatial position and posture), and relevant commonsense properties.
- For "static_objects" (static background): Strictly list objects that are clearly visible in the frame but remain untouched and unmoved throughout the entire video.
- Analyze the atmosphere and setting: Infer the location type (e.g., lab, kitchen, factory). Describe the visual style (lighting, color tone, clean/chaotic). List pure background elements (e.g., chairs, windows, wall sockets).
- Maintain dual-arm awareness: If dual arms appear (or have the potential to), you must explicitly describe the state of EACH arm. Even if the left arm is motionless throughout, explicitly record "the left arm remains stationary on the left side, unengaged in the current task."

Skill 4: **Assessing Task Results and Meta-Information**

- Go beyond a simple "success/failure" binary: Describe the execution quality, key strategic details, and evaluate the physical impact or reactions on the environment and surrounding objects.
- Parse robot and perspective meta-information: If the original prompt contains information formatted as `[robot_model|perspective]` (e.g., `[realman_rmc_aidal|high]`), it must be explicitly extracted and described in an appropriate location in the annotation (e.g., at the beginning of the high-level task or in the task notes). For example: "A realman_rmc_aidal robot performs..., this is a video shot from a high perspective."

Figure 7 Stage 1 captioning example 1. Structured perception output showing extracted physical attributes, object identities, and spatial relations.

Constraints

- Specificity First: The description of each video must be unique. Absolutely avoid templated or generic descriptions.
- Physical Detail Oriented: Focus heavily on physical interactions such as force applied, contact points, deformation, friction, and sliding.
- Based on Visible Facts: Only describe what is clearly verifiable in the video. Over-inference and hallucinated details are strictly prohibited. When uncertain, choose broader but absolutely correct descriptions.
- Conservative Spatial Relationships: Relative positional relationships between objects (e.g., "between...", "left of", "right of") must be absolutely accurate and unambiguous. If camera perspective or occlusion prevents 100% certainty, abandon describing precise adjacent objects and instead describe the absolute or rough position of the target itself. Accuracy takes precedence over descriptive richness. Better to be brief than wrong.
- Strict Entity Isolation: Objects listed in the "static_objects" list are ABSOLUTELY PROHIBITED from appearing in the action sequence descriptions. This is a hard boundary to prevent the model from confusing the "protagonist" with the "background". If an object moves or is touched, it must be moved to the interactive "objects" list.
- Format and Engineering Constraints: The output must strictly conform to a JSON format and be directly parsable. All fields must be complete with correct data types. Use empty arrays or strings if certain information is absent in the video. Focus on the chronological and causal relationships of actions. Avoid numerical estimations; use relative and qualitative descriptions.
- Reference Original Descriptions: Validate and refine the original prompt against the video content. Do not rely entirely on the original description; the video content is the ultimate ground truth.

High-Quality Annotation Example

Before beginning your analysis, refer to the following perfect annotation example:

```
{
  "video_segment_id": "327-648642-002_head_color.mp4",
  "high_level_task": "Gently grasp a fragile tomato from the shelf and precisely place it into a plastic bag inside the shopping cart.",
  "robot_initial_state": {
    "left_arm": "Remains stationary on the left side of the frame, unengaged in the current task.",
    "right_arm": "Hovering over the right area of the shelf, preparing to execute the grasp."
  },
  "atmosphere_and_setting": {
    "location_type": "Supermarket fresh produce section / Simulated retail environment",
    "visual_style": "Brightly lit, vibrant colors, lively atmosphere",
    "background_elements": ["Metal shelf structure", "Out-of-focus shelves in the background", "Price tags", "Overhead lighting"]
  },
  "skill_sequence": [
    "Move to the tomato area on the shelf and hover directly above the target for final positioning.",
    "Slowly descend vertically to avoid touching surrounding vegetables.",
    "Use a wrap grasp (rather than a pinch) to maximize contact area and steadily hold the tomato.",
    "Gently lift vertically to confirm the grasp is stable without sliding.",
    "Plan the path and translate horizontally above the shopping cart, actively avoiding the cart's metal edges.",
    "Descend directly above the opening of the plastic bag.",
    "Fine-tune the gripper angle to align with the irregular bag opening.",
    "Gently release the gripper, allowing the tomato to naturally fall into the bag via gravity.",
    "After confirming the tomato has reached the bottom of the bag, lift the robotic arm and reset."
  ],
  "key_state_changes": [
    {
      "trigger_skill_index": 2,
      "description": "The tomato's state changes from 'resting on the shelf' to 'stably held by the robotic arm using a wrap grasp'."
    },
```

Figure 8 Stage 1 captioning example 2. The perception module identifies robot morphology, object properties, and contact events from the video sequence.

```
{
    "trigger_skill_index": 7,
    "description": "The tomato's position changes from 'suspended in the air by the gripper' to 'contained within the plastic bag
in the cart', causing slight deformation to the plastic bag."
   }
 ],
 "scene_context": {
  "objects": [
   {
    "name": "Tomato",
    "functional_category": "Fragile manipulated object",
    "commonsense_properties": ["Easily damaged", "Smooth surface", "Spherical", "Has some weight"],
    "initial_state": "Located on the right end of the second shelf tier, upright posture, closely surrounded by other tomatoes."
   },
   {
    "name": "Plastic bag",
    "functional_category": "Narrow and deformable container",
    "commonsense_properties": ["Irregular opening", "Transparent", "Lightweight"],
    "initial_state": "Located inside the shopping cart, opening partially closed due to its own weight."
   }
  ],
  "static_objects": [
   "A row of green cucumbers on the left side of the shelf (untouched)",
   "Yellow bell peppers on the lower shelf tier (remain stationary)",
   "Red packaging material inside the shopping cart (acting as background clutter)"
  ]
 },
 "task_result_and_note": "Task succeeded on the first attempt. High execution quality with smooth, well-strategized
movements that fully accounted for the tomato's fragile nature. Grasp was stable, placement was precise, and no accidental
disturbances were caused to the environment."
}

```


Figure 9 Stage 1 captioning example 3. Spatial layout parsing and state transition detection across the manipulation trajectory.

Stage 2: Physics-Grounded Narrative Synthesis. A language model (Qwen3 32B FP8) takes the Stage 1 structured output and produces a four-phase natural language caption: (1) *Scene Setup*—initial configuration, robot type, and object arrangement; (2) *Action Detail*—fine-grained manipulation actions including Cartesian trajectories, gripper operations, and contact dynamics; (3) *State Transition*—physical state changes such as object displacement, deformation, and containment relations; and (4) *Camera Summary*—viewpoint, camera motion, and visual framing. By separating perception from writing, the captions remain factually grounded in visual evidence while capturing the causal dynamics needed for world model training. Representative Stage 2 outputs are shown in Figures 10–11.

The Structured-to-Text Aggregation Prompt Template

Part 1: System Context

Character

You are a professional visual description aggregation expert, specializing in video content. Your primary task is to seamlessly merge multi-level, structured video annotation data into a single fluent, cohesive, and cinematic documentary-style English narrative.

Core Directives

- Objective Tone: Maintain an objective, observational narrative style. Avoid emotional, subjective, or overly dramatic adjectives.
- Anti-Mechanization: Transform structured JSON/list data into a naturally flowing narrative. Strictly avoid robotic, list-like itemizations.
- Absolute Language Constraint: The output must be 100% pure English. The inclusion of any Chinese characters, Pinyin, or Chinese punctuation is strictly prohibited.

Part 2: User Execution

Highest Priority Constraint

Final Output Language: English Only. Check your output before finishing: If you see any Chinese characters (e.g., "Handover", "Fetch"), translate them immediately.

Skills & Structural Framework

Skill 1: **Executing the Four-Phase Narrative Structure**

- Phase 1: The Setup: Begin with "The video opens with a view of [Location/Setting]...". Detail the environmental atmosphere (clean, industrial, messy). Focus on establishing the central subject, surrounding static objects (using terms like "flanked by", "next to"), and the exact initial state of the robot (e.g., "Two robotic arms are positioned...").
- Phase 2: The Action: Transition using "As the video progresses," or "In the subsequent frames,". Describe the continuous action flow. For dual-arm setups, explicitly differentiate between the left and right arms.
- Phase 3: The Conclusion: Open with "By the final frame," or "The scene concludes with...". Document the final position of the objects and the reset state of the robot. Summarize environmental consistency (e.g., "The scene remains largely unchanged except for...").
- Phase 4: The Shot: Conclude with a single, dedicated sentence starting with "A medium shot captures..." or "A close-up shot highlights...", summarizing the core visual and cinematic focus of the entire task.

Skill 2: **Physical & Visual Integration**

- Seamlessly weave precise physical actions (e.g., fine-tuning, gravity drops) into the narrative.
- Do not merely state "picked up"; describe the nuance: "precisely aligns, gently pinches, and lifts...".
- explicitly capture visual changes resulting from physical interactions, such as deformation, sagging, or sliding.

Skill 3: **Spatial Mapping & Contextual Fidelity**

- Translate spatial relationships carefully based on the intensity of the original descriptions.
- Adhere strictly to the Fidelity Principle: If the original input is vague, do not over-specify; if the input is precise, do not weaken it.
- Ensure full panoramic coverage: The background, atmosphere, and unmanipulated static objects must be accurately represented.

Figure 10 Stage 2 captioning example 1. The writing module synthesizes a four-phase narrative covering scene setup, action detail, state transition, and camera summary from the structured perception output.

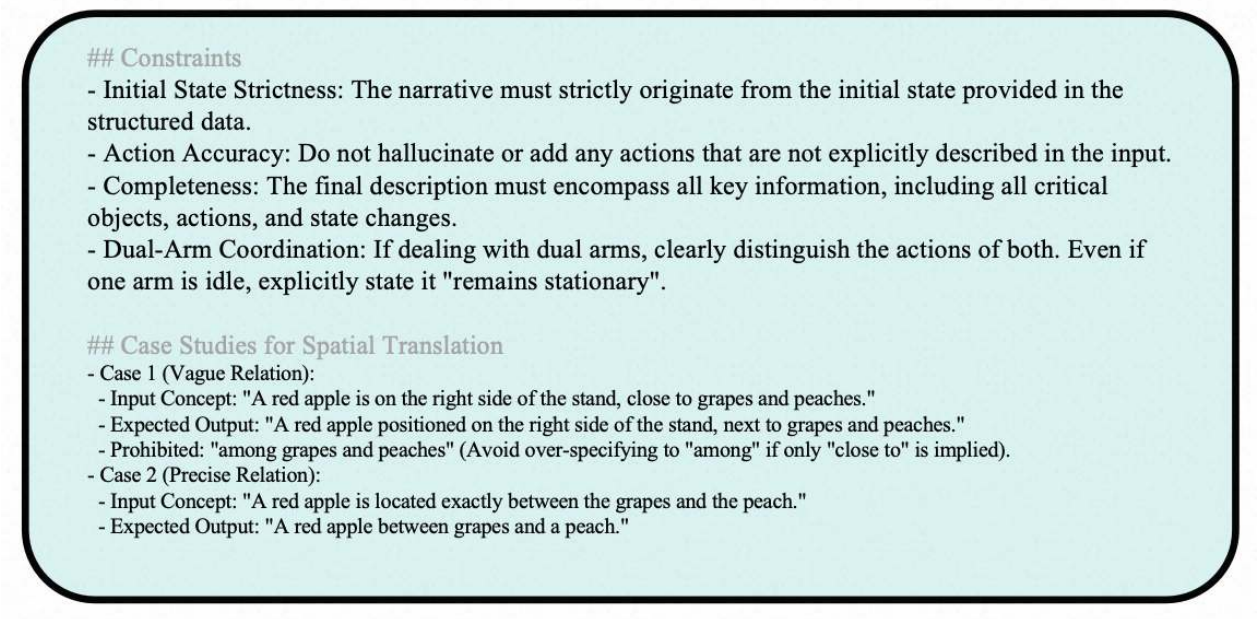


Figure 11 Stage 2 captioning example 2. Physics-grounded narrative synthesis capturing fine-grained manipulation dynamics and causal state transitions.

C. Additional Qualitative Results

We present additional qualitative comparisons across three evaluation settings.

Zero-Shot Qualitative Comparison on EZSbench. Figure 12 shows zero-shot results on EZSbench. Current video generation baselines struggle with long-horizon manipulation tasks that require complex logical reasoning. Wan-2.5, Veo 3.1, and WoW fail to map object color attributes to the correct target containers, producing placement errors. Sora v2 generates physically implausible contactless grasping, while Giga R0 and Veo 3.1 suffer from "generation collapse" during contact-rich interactions—the end-effector and object geometry become completely distorted. Our method correctly follows the compositional instructions and maintains spatiotemporal coherence throughout the long-horizon grasp-and-place trajectory without such artifacts.

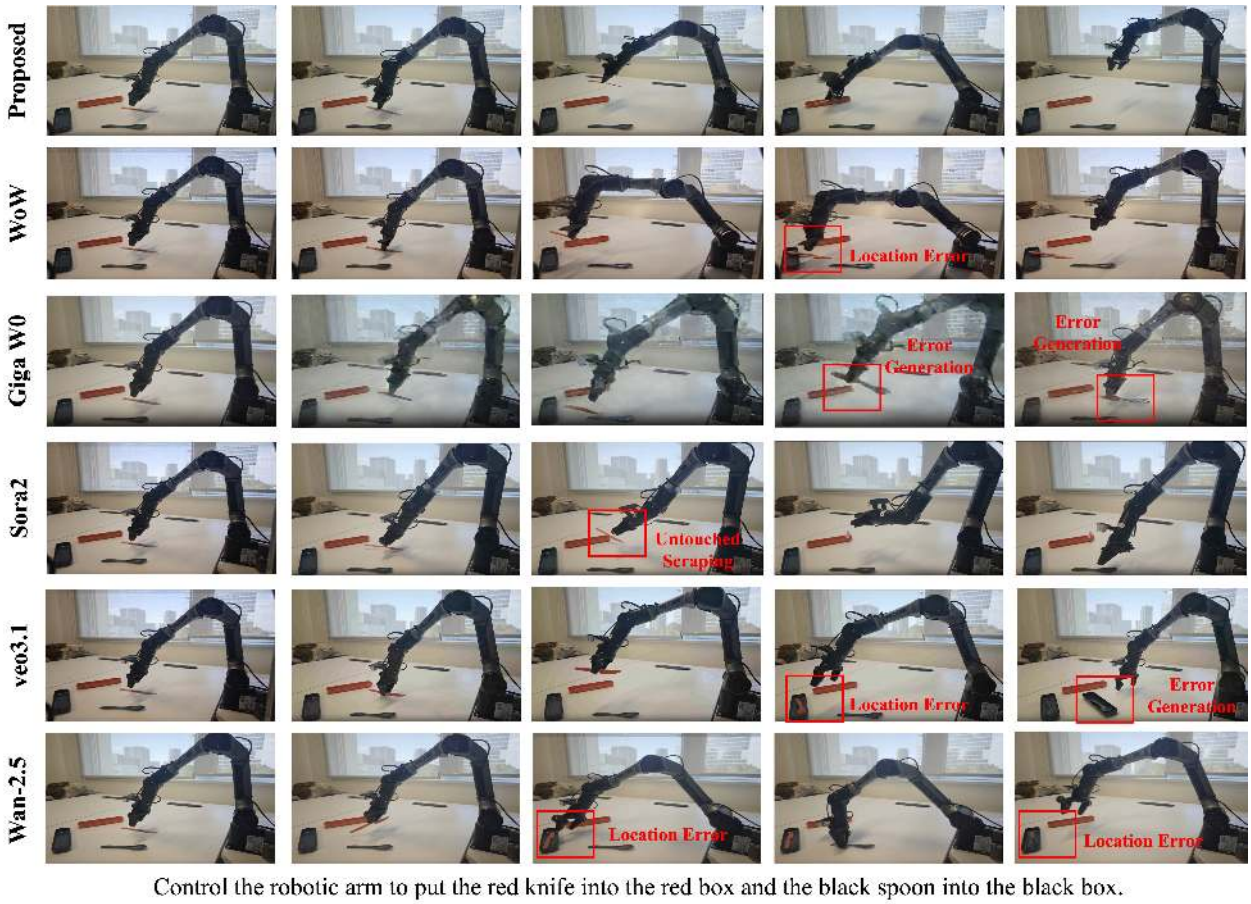


Figure 12 Zero-shot qualitative comparison on EZSbench. Baseline models exhibit placement errors (Wan-2.5, Veo 3.1, WoW), contactless grasping (Sora v2), and geometric collapse during contact interactions (Giga R0, Veo 3.1). Our method (bottom row) correctly follows compositional instructions and maintains physical plausibility.

Case Study on Zero-Shot Test Set. Figure 13 shows generation results on the zero-shot test set across diverse unseen tasks. In the long-horizon "red knife → red box, black spoon → black box" task, the model correctly binds object attributes to target containers and performs continuous spatial reasoning. For dual-arm towel folding, it handles deformable-object topology changes while generating coordinated bimanual trajectories. The model also produces physically consistent results for articulated-object interaction (closing a door), rigid-body placement (placing blocks), contact-intensive wiping (removing stains), and object relocation (moving an apple). Across these tasks, the generated videos follow the language instructions and maintain physical plausibility over long horizons.

Figure 13 Case study on zero-shot test set. Our model handles diverse unseen manipulation tasks including multi-object attribute binding, deformable object manipulation with dual-arm coordination, articulated object interaction, rigid body placement, contact-intensive wiping, and object relocation—all with strict physical plausibility and spatiotemporal coherence.

Action-to-Video Qualitative Comparison. Figure 14 compares action-conditioned video generation (A2V) results. Our method generates contact-intensive manipulation videos while preserving object geometry and visual integrity throughout the interaction. In contrast, Genie-Envisioner and Enerverse-AC produce noticeable object deformation, contactless grasping artifacts, and target localization errors, resulting in distorted outputs or task failure.

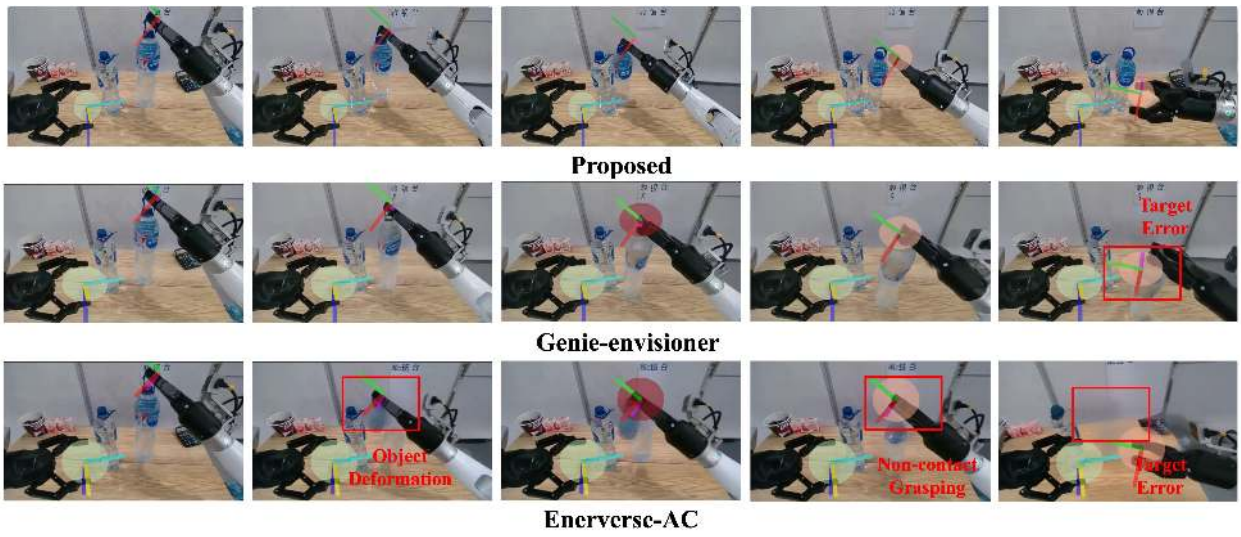


Figure 14 Action-to-video qualitative comparison. Our method preserves object geometry and visual integrity during contact-intensive manipulation. Baselines (Genie-Envisioner, Enerverse-AC) produce object deformation, contactless grasping, and localization errors.