

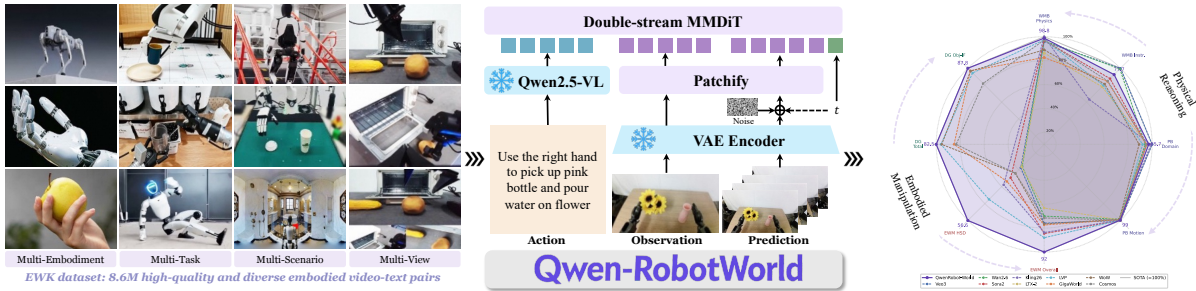
Qwen-RobotWorld Technical Report: Unifying Embodied World Modeling through Language-Conditioned Video Generation

Qwen Team

<https://qwen.ai/blog?id=qwen-robotworld>

Abstract

We introduce **QWEN-ROBOTWORLD**, a language-conditioned video world model for embodied intelligence. With natural language as a unified action interface, it predicts physically grounded future visual trajectories from current observations across robotic manipulation, autonomous driving, indoor navigation, and human-to-robot transfer. This unified formulation provides three promising application directions: synthetic data generation for policy training augmentation, scalable virtual environments for policy evaluation, and language-guided planning signals for downstream robot control. This is achieved through a three-part design: a) **Double-Stream MMDiT with MLLM Action Encoding**, where a 60-layer double-stream diffusion transformer couples frozen Qwen2.5-VL semantics with video-VAE latents through layer-wise joint attention; b) **Embodied World Knowledge (EWK)**, an 8.6M video-text corpus (200M+ frames) with action-language mapping over 20+ embodiments and 500+ action categories; and c) **General+Expert Progressive Curriculum**, a two-stage training strategy that first learns general visual priors and then injects embodied specialization under a shared language interface. Extensive results show strong competitiveness: ranks 1st overall on EWM-Bench and DreamGen Bench, outperforms all open-source models on WorldModelBench and PBench. Additional zero-shot analyses on RoboTwin-IF benchmark further support robust generalization and multi-view consistency.



1 Introduction

Embodied intelligence requires agents to perceive, reason, and act within physical environments—spanning robotic manipulation at tabletop scale, autonomous navigation through urban traffic, and wayfinding across indoor spaces. Training such systems directly in the real world is costly, inefficient, and fraught with safety risks. World models offer a scalable alternative: by learning environment dynamics from observational data, they serve as interactive training platforms that allow embodied agents to acquire and refine behaviors without physical deployment.

A world model can be formalized as a *state transition function*: given a current state s_t and an action a_t , it predicts the resulting state $s_{t+1} = f(s_t, a_t)$ Ye et al. (2026). In video-based world models, states are visual observations (video frames or their latent representations), and the model generates future visual trajectories conditioned on the current observation and an action signal. The action a_t can take various forms—low-level motor commands, high-level waypoint trajectories, or natural language instructions. Among these, **natural language is the most general and accessible action representation** Ye et al. (2026): a single instruction such as “pick up the red cup and place it on the shelf” implicitly encodes the complete action sequence, goal state, and physical constraints, without requiring robot-specific control interfaces. Language actions can furthermore be utilized in two complementary directions: as an explicit *input* fused into the model’s condition signal to govern state transitions, or as an *output* inferred post-hoc from generated video to serve as an action label. This flexibility positions language-conditioned world models as universal simulation backbones that generalize across embodied platforms without interface redesign.

However, a fundamental tension currently limits world model effectiveness. *General video generation models* OpenAI (2024); Google DeepMind (2025) learn rich visual priors from internet-scale data but

fail to accurately model embodied physics—contact dynamics, rigid-body structural constraints, and action-consequence relationships that are critical for physically plausible state transitions. *Domain-specific embodied models* Agarwal et al. (2025); Chen et al. (2025a); Team et al. (2025), conversely, are tailored to individual scenarios (e.g., tabletop manipulation or driving); they rely on structured, robot-specific action representations such as joint angles or waypoints, which cannot generalize across embodiment types or task categories, fundamentally limiting their utility as cross-platform simulation environments.

Bridging this gap requires grounding diverse embodied experiences in general visual priors, with natural language as the unified action interface that enables cross-scenario and cross-task integration. Different embodied domains provide complementary physical knowledge that collectively enriches the world model’s state transition function: *manipulation* teaches fine-grained contact physics and object-state transformations within confined workspaces; *autonomous driving* teaches large-scale multi-agent dynamics and 3D scene geometry through ego-motion parallax and scene-scale transitions; *indoor navigation* teaches room-scale spatial reasoning, where language instructions must be grounded into spatially coherent visual trajectories over extended horizons. Because these domains share a common language interface, they can be trained jointly—with each domain’s physical knowledge reinforcing the others rather than conflicting. Furthermore, translating human demonstrations into robot executions through video editing opens a practical pathway to scale embodied training data beyond the limits of physical robot collection.

We present **QWEN-ROBOTWORLD**, a language-conditioned video world model in the Qwen series that realizes this vision through tightly coupled innovations in architecture, data, and training. Beyond high-fidelity action-conditioned prediction, the model serves as a unified backbone that, with task-specific adaptation, can support three representative embodied world model applications: a *synthetic data engine*, a *policy evaluation environment*, and an *action planner*.

Architecture: Double-Stream MMDiT with MLLM Action Encoding (§3). To implement language-conditioned state transitions, we adopt a double-stream Multimodal Diffusion Transformer (MMDiT) backbone. An *understanding stream* processes rich semantic features extracted by a frozen Qwen2.5-VL encoder, representing the action a_t ; a *generation stream* processes visual latents from a video-compatible VAE, representing the visual state s_t . The two streams interact via joint attention at every layer, enabling bidirectional cross-modal fusion throughout the denoising process. Using an MLLM as the action encoder—rather than lightweight encoders such as T5 Raffel et al. (2020) or CLIP Radford et al. (2021)—yields two key advantages: (1) its deep language understanding accurately parses complex, compositional instructions into precise condition signals that govern fine-grained state transitions; (2) its internalized world knowledge (e.g., that robot arms are rigid bodies with fixed link lengths and joint constraints) implicitly constrains the space of physically plausible transitions, and—combined with T2I co-training—prevents object deformation across video frames without requiring explicit geometric prompts, a common failure mode in models lacking such semantic grounding.

Data: Embodied World Knowledge Dataset (§2). To train a state transition function that generalizes across embodied domains, we construct the Embodied World Knowledge (EWK) dataset—approximately 8.6M video-text pairs comprising over 200M observation frames. The corpus spans four embodied domains alongside general video data (30% of the total): **manipulation** (~5.9M samples, 20+ robot morphologies, 1300+ skills) provides the core embodied foundation; **autonomous driving** (~200K samples from Waymo, NVIDIA PhysicalAI-AD, Bench2Drive, and Sekai) contributes large-scale ego-motion and multi-agent dynamics; **indoor navigation** (6K+ language-guided episodes from VLNVerse) provides room-scale spatial reasoning grounded in continuous trajectories; and **human-to-robot transfer data**—generated via an automated MANO Romero et al. (2017)-to-robot pipeline across 14 robot morphologies—enables cross-embodiment video editing. A central methodological contribution is our *action-language mapping framework*, which standardizes actions across 20+ robot embodiment types and 500+ action categories into a unified natural language interface, yielding approximately 8.6M high-quality cross-scenario, cross-task embodied video-text pairs. This is complemented by task-aware temporal segmentation (ensuring each sample captures a complete, well-defined state transition) and a hierarchical five-layer viewpoint-aware annotation pipeline that substantially improves caption specificity and downstream instruction-following.

Training: From General Priors to Embodied Specialization (§4). We adopt a two-stage progressive training curriculum. In pretraining, joint training across T2I, T2V, and T2V tasks over general-domain data builds foundational visual priors, with T2I specifically anchoring geometrically correct object morphology that transfers to video generation. In the SFT stage, embodied data is introduced progressively (70% embodied, 30% general) through a four-phase mixing schedule: single-view manipulation → multi-view expansion → multi-view concatenated generation → complex tasks and cross-domain data. Within the embodied portion, manipulation dominates at ~90% sampling weight to ensure depth of physical grounding, while multi-view concatenation and navigation/driving data each receive ~5% to provide breadth. This *general + expert* joint training paradigm—unified under the natural language action

interface—enables stable co-training across diverse scenarios and tasks, with each domain’s physical knowledge mutually reinforcing the others. Asymmetric 3D RoPE positional encoding and multi-view concatenation training enable geometrically consistent synthesis across synchronized camera views without architectural modification.

Evaluated on four established benchmarks, QWEN-ROBOTWORLD achieves competitive performance across cross-scenario and cross-task settings. It outperforms all open-source models on WorldModelBench (8.99, 3rd overall), attaining perfect physics adherence scores across Newton’s laws, mass conservation, fluid dynamics, and gravity—on par with leading closed-source models—while achieving strong instruction following (2.33/3.0). It ranks **1st overall** on EWMBench (4.60), with substantially leading motion fidelity in HSD (0.566, +33% over the runner-up) and top scene consistency (0.914). On DreamGen Bench, the model ranks **1st overall** (4.952) across three robotic embodiment subsets, excelling in object-level compositional generalization. On PBench, it outperforms all open-source models (0.804), with domain understanding placing 3rd overall (0.857) and motion smoothness ranking 2nd among open-source models (0.990). Qualitative results further showcase generalization across cross-task video editing—including human-to-robot transfer, where the model synthesizes realistic robot execution from a human demonstration video without robot-specific prompting—as well as autonomous driving scene synthesis and room-scale indoor navigation generation; additional zero-shot performance on RoboTwin-IF benchmark further support robust transfer under complex instructions.

Our contributions are summarized as follows:

- **Framework.** We propose QWEN-ROBOTWORLD, a language-conditioned video world model that treats natural language as a universal action interface to unify cross-scenario and cross-task embodied capabilities. By jointly training manipulation, driving, navigation, and human-to-robot transfer under a shared language interface, the model achieves complementary physical generalization that no single-domain model can match.
- **Data.** We propose an action-language mapping framework that standardizes 20+ robot embodiment types and 500+ action categories into a unified natural language interface, and construct approximately 8.6M high-quality, cross-scenario, cross-task embodied video-text pairs constituting the EWK dataset.
- **Training.** We propose a general + expert joint training paradigm that, under the unified natural language interface, equips the model with both broad world modeling capability and deep embodied domain expertise, enabling stable and scalable co-training across diverse scenarios and tasks.
- **Performance.** QWEN-ROBOTWORLD achieves comprehensive improvements on cross-scenario and cross-task embodied evaluation metrics, ranking 1st overall on EWMBench and DreamGen Bench and outperforming all open-source models on WorldModelBench and PBench.

2 Data

The central challenge in training a universal embodied world model is not data scale alone, but representational heterogeneity: robotic manipulation actions are expressed as joint angles or end-effector waypoints, driving as steering commands and velocity profiles, and navigation as heading vectors—each requiring a separate model or interface per domain. We resolve this through an **action-language mapping** framework that converts heterogeneous actions from 20+ robot embodiment types and 500+ action categories into a unified natural language interface. Under this unified interface, videos from a Franka gripper, an autonomous vehicle, and an indoor navigation agent all become instances of the same language-conditioned video generation task, enabling cross-scenario and cross-task joint training under a single model without any domain-specific control interface. As shown in Figure 1, this framework produces approximately 6M high-quality, cross-scenario, cross-task embodied video-text pairs, which we further augment with general video data (30% of the total) to construct the Embodied World Knowledge (EWK) dataset: a corpus of 8.6M video-text pairs comprising over 200M observation frames.

2.1 Action-Language Mapping

The action-language mapping framework addresses a fundamental asymmetry in embodied data: the *visual* states (video frames) are already in a common pixel space, but the *action* representations are fragmented across incompatible modalities. Our framework resolves this by projecting all action signals onto a shared natural language space, so that the same diffusion transformer can learn $s_{t+1} = f(s_t, a_t)$ regardless of the underlying physical domain.

Why Language as the Unified Action Interface. Unlike low-level action representations—joint angles, end-effector waypoints, force-torque commands—which are hardware-specific and require a separate

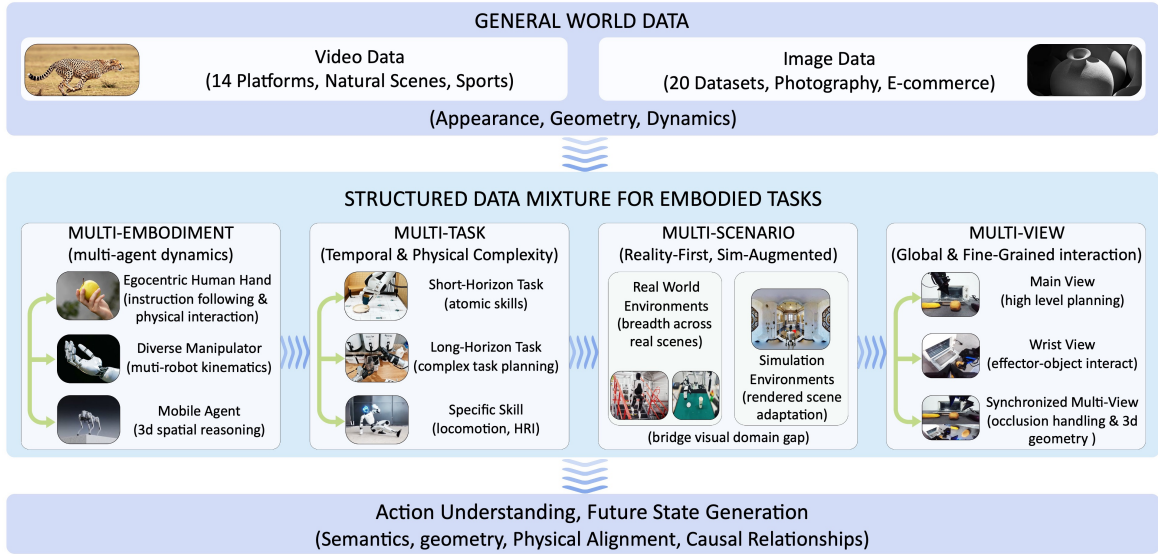


Figure 1: Overview of the Embodied World Knowledge (EWK) training corpus. **General world data** (top) supplies foundational priors on appearance, geometry, and dynamics from internet-scale video and image collections. **Structured embodied data** (middle) is organized along four complementary axes, each targeting a distinct source of physical variation: *Multi-Embodiment* (human hands, diverse robot manipulators, mobile agents); *Multi-Task* (short-horizon atomic skills, long-horizon compositional planning, specific skills such as locomotion and HRI); *Multi-Scenario*, a *reality-first, sim-augmented* design that bridges real captures and the simulators where downstream VLA policies are trained and evaluated; and *Multi-View* (main, wrist, and synchronized multi-view streams covering both global planning and fine-grained effector–object interaction). Jointly, these signals supply the semantics, geometry, physical alignment, and causal relationships (bottom) required for language-conditioned action understanding and future-state generation.

control interface per embodiment, natural language offers a universal, embodiment-agnostic action interface. A single instruction such as “grasp the red cup and lift it vertically” implicitly encodes the full action sequence, goal state, and physical constraints, without any knowledge of the underlying kinematic chain. By training the model to predict the next visual state s_{t+1} from a language action a_t alone, we obtain a simulation backbone that generalizes across embodiments—whether a Franka gripper, an Aloha dual-arm system, or a humanoid—without retraining or re-engineering robot-specific interfaces. This generality, however, places demanding requirements on annotation quality: each caption must function as a *complete, self-contained action specification*, precise enough that the model can predict s_{t+1} from a_t and s_t alone, without access to any robot metadata or proprioceptive signals.

Hierarchical Five-Layer Annotation. To consistently produce such action-rich captions across 20+ robot embodiment types and 500+ action categories, we design a hierarchical annotation framework with five progressive layers. The first three form a structured chain-of-thought that decomposes each visual state transition into interpretable components:

1. **Task Goal Layer**—infer the high-level intent of the transition (what should change between s_t and s_{t+1}), integrating external instructions with observed video content;
2. **Action Detail Layer**—decompose the action a_t into spatio-temporal trajectories, micro-actions, speed, and force, with **mandatory explicit declaration of viewpoint information** (egocentric main view, wrist view, external view, or concatenated multi-view combinations);
3. **Physical Feedback Layer**—describe the observable consequences of the action on the environment (object displacement, deformation, contact state changes), grounding each transition in verifiable physical outcomes.

Based on this analysis, two granularities of action descriptions are generated:

4. **Comprehensive Description** (50–100 words)—fully specifies the viewpoint–agent–action–feedback quadruple, providing a rich action signal for precise state transition prediction;
5. **Concise Description** (15–30 words)—retains only the essential viewpoint–agent–key action elements, enabling the model to handle brief, high-level commands at inference time.

Table 1: Detailed inventory of the Embodied World Knowledge (EWK) training data mixture, organized by domain.

Dataset	Embodiment	Views	Tasks	Contribution
A. Manipulation				
EgoHOD Pei et al. (2025), EPIC-Kitchens Damen et al. (2018), Egocentric-10k Build AI (2025)	Human hands	Egocentric	Daily grasping & kitchen	Dexterity & coordination prior
Bridge V2 Walke et al. (2023), RH20T Fang et al. (2024), Droid Khazatsky et al. (2024)	Single-arm grippers	external + wrist	Tabletop pick-and-place	Interaction primitives
Robomind Wu et al. (2025a), RoboCoin Wu et al. (2025b)	Single/dual-arm, humanoids	Ego + external	Rigid & deformable objects	Cross-embodiment generalization
Agibot-World AgiBot-World-Contributors (2025), Galaxea Galaxea AI (2025)	Single-arm (gripper + dexterous hand)	Synced ego + wrist + external	Long-horizon sequential	Temporal & multi-view consistency
Qwen-Aloha (internal)	Dual-arm grippers	Head + dual wrist	Diverse grasping	Multi-view grasping prior
ActionNet Fourier ActionNet Team & Mu (2025), OpenLoong OpenLoong Baihu Team (2025)	Dexterous hands	Wrist + external	Tool use & in-hand	Fine-grained dexterity
InternData-A1 Tian et al. (2025), Robotwin Chen et al. (2025b), Groot-XE Bjorck et al. (2025), RT1 Brohan et al. (2023)	Mixed arms (simulated)	Variable	Fluids & deformables	Sim-to-real alignment
B. Autonomous Driving				
Waymo E2E Waymo Team (2024), NVIDIA PhysicalAI-AD NVIDIA (2025b)	Ego vehicle	5–8 surround-view	Urban driving & traffic	Large-scale motion & 3D geometry
Bench2Drive Jia et al. (2024)	Ego vehicle (sim)	6 surround-view	9.8K traffic scenarios	Sim diversity & GT annotations
Sekai Sekai Team (2025)	Pedestrian / drone	Egocentric	Urban walking	Pedestrian-scale locomotion
C. Indoor Navigation				
VLNVerse Lin et al. (2025)	Mobile agent	Egocentric	134 indoor scenes, lang-guided	3D reasoning & lang-trajectory align
D. Human-to-Robot Transfer				
Paired H2R dataset	Human → 14 robot arms	Egocentric bimanual	Cross-embodiment manipulation	Video editing supervision

We enforce four quality control principles: *operation focus* (only agent actions and object interactions), *viewpoint definition* (explicit viewpoint type and semantic role), *objectivity* (only visible dynamics), and *physical verifiability* (only visually verifiable outcomes). In training, we sample from comprehensive and concise descriptions with equal probability (50% each), so the model learns to execute both detailed trajectory specifications and brief task-level commands.

Coverage: 20+ Robot Embodiments, 500+ Action Categories. The framework is applied across all data domains. On the embodiment axis, it covers human hands, seven robot arm configurations (single-arm gripper, dual-arm gripper, single-arm dexterous hand, dual-arm dexterous hand, mobile dual-arm, half-humanoid, and full humanoid), ego vehicle (surround-view cameras), pedestrian/drone, and mobile navigation agent—representing 20+ distinct robot embodiments in total, as sourced from RoboCoin (15 robot models across three structural categories), Robomind (4 morphologies), InternData-A1 (4 robot models), Groot-XE, and various other datasets. On the action axis, it spans 500+ action categories derived from the explicit motion primitive vocabularies across our training datasets—Agibot-World alone defines 84 distinct manipulation primitives (grasp, push, pour, fold, wipe, cut, etc.)—supplemented by unique primitives from other manipulation datasets and locomotion/navigation actions (turning, lane-changing, waypoint following, obstacle avoidance, etc.), organized into four tiers: (1) manipulation primitives, (2) long-horizon compositions, (3) locomotion and navigation, and (4) dynamic and deformable interactions. This systematic coverage ensures that the resulting embodied video-text pairs span a semantically rich and physically diverse action space that no single domain could provide.

2.2 Data Collection

2.2.1 General Data

General world data lays the foundation for the model to grasp basic physical laws and form accurate visual representations. This category encompasses diverse videos and still images from the internet. Video data are standardized to 24 FPS and support multiple resolutions and aspect ratios (1:1, 2:3, 3:2, 3:4, 4:3, 9:16, 16:9, etc.). Image data integrates high-quality static photographs, serving as visual quality anchors that establish precise representations of object appearance, material texture, and spatial composition. All general data is annotated with natural language descriptions generated by Qwen2.5-VL Bai et al. (2025); annotations omit viewpoint-specific information to maintain flexibility and generality. Notably, we adopt a conservative stance on AI-generated content (AIGC): general data excludes AI-produced images and videos, as these often introduce visual artifacts, physical inconsistencies, and implicit biases that could undermine the model’s generalization capabilities.

2.2.2 Embodied Manipulation Data

To enable the world model to acquire grounded physical understanding across scenarios and tasks, we build a structured data mixture spanning manipulation, driving, navigation, and cross-embodiment transfer domains, as summarized in Table 1. For the core manipulation domain, we organize the data around four dimensions: Multi-Embodiment, Multi-Task, Multi-Scenario, and Multi-View.

Multi-Embodiment. Manipulation data spans a spectrum of embodiments—human hands, single-arm grippers, dual-arm dexterous systems, mobile manipulators, and full-body humanoids—so the model learns to separate task-level intent from embodiment-specific kinematics. Human manipulation data (EgoHOD Pei et al. (2025), EPIC-Kitchens Damen et al. (2018)) provides a dexterity ceiling: the model observes what physically capable interaction looks like, acquiring priors for fluid hand–eye coordination and tool use. Robot data then teaches the model how those same intents map onto diverse mechanical morphologies. By exposing the model to both human demonstrations and robot executions of overlapping tasks (e.g., Robomind Wu et al. (2025a), RoboCoin Wu et al. (2025b)), it learns embodiment-invariant action semantics—the ability to predict “what should happen next” regardless of whether the actor is a two-finger gripper or a seven-finger dexterous hand.

Multi-Task. The manipulation corpus covers a skill hierarchy from atomic contact-level actions to extended multi-step procedures, teaching the model to operate at multiple temporal granularities. Short-horizon datasets (Bridge V2 Walke et al. (2023), RH20T Fang et al. (2024)) provide dense coverage of fundamental interaction primitives—grasping, pushing, inserting—that ground the model’s understanding of contact physics and object affordances. Long-horizon datasets (Agibot-World AgiBot-World-Contributors (2025), Galaxea Galaxea AI (2025)) chain these primitives into coherent sequences, forcing the model to maintain state tracking and causal reasoning across dozens of steps. Additionally, dynamic-interaction datasets (Humanoid Everyday Zhao et al. (2025)) introduce high-velocity, whole-body motions that test the model’s ability to predict outcomes under significant momentum and balance constraints. Together, this range ensures the model can reason about both “what happens when you press here” and “what happens after ten sequential decisions.”

Multi-Scenario. Multi-scenario coverage advances along two complementary axes: *breadth across real environments*, and *extension to simulator-rendered scenarios*. Along the first axis, physical interaction manifests differently depending on context—a kitchen counter presents different lighting, clutter density, and surface properties than a factory floor or an outdoor worksite. Our manipulation data is therefore predominantly real-world, spanning domestic kitchens, workshops, laboratories, and unstructured outdoor settings, exposing the model to genuine variation in illumination, occlusion, material appearance, and background complexity—so it does not brittly overfit to any single environment. Along the second axis, we incorporate photorealistic simulation data (InternData-A1 Tian et al. (2025)) as a first-class complement. This is motivated by the VLA landscape: a substantial portion of policy models are trained in simulators, and virtually all are evaluated there using standardized benchmarks such as LIBERO Liu et al. (2023), SimplerEnv Li et al. (2024), and RL Bench James et al. (2020). A world model intended as a general simulation backbone must therefore generate faithfully under simulator-style appearances and physics, bridging the visual domain gap between real and synthetic observations so it can serve both sim-to-real transfer and closed-loop evaluation pipelines. The simulation portion additionally supplies precisely controlled variations in lighting, object pose, and camera placement that further strengthen visual robustness.

Multi-View. Single-view data teaches the model to predict plausible futures from a fixed perspective, but many physically critical events are partially or fully occluded from any single camera. Synchronized multi-view recordings (Agibot-World [AgiBot-World-Contributors \(2025\)](#), Robomind [Wu et al. \(2025a\)](#)) expose the model to the same event from head-mounted, wrist-mounted, and external viewpoints simultaneously. This serves two purposes: during training, cross-view correspondence acts as a geometric regularizer, implicitly teaching the model about object shape, depth, and spatial relationships; at inference, the model can generate from any individual viewpoint or compose multi-view outputs that remain mutually consistent. Approximately 1.6M of our 6M embodied samples include synchronized 2–4 view concatenations, providing substantial multi-view supervision without dominating the corpus.

2.2.3 Autonomous Driving Data

While manipulation data captures fine-grained object interactions within a confined workspace, autonomous driving data exposes the model to a substantially larger motion space with diverse maneuvers (turning, lane changing, acceleration) spanning a much wider range of velocities and trajectories. Driving scenes also contain rich multi-agent dynamics—surrounding vehicles, pedestrians, and cyclists interacting under traffic rules—requiring the world model to learn how multiple objects move, occlude, and influence each other over time. Furthermore, the large camera displacement provides dense supervisory signal for 3D scene geometry through parallax and perspective changes, strengthening the model’s capacity for view synthesis and spatial reasoning.

We curate multi-view driving videos from four large-scale datasets: [Waymo E2E Waymo Team \(2024\)](#) (real-world driving, 8 surround-view cameras, 7,044 clips / 11.3h), [NVIDIA PhysicalAI-AD NVIDIA \(2025b\)](#) (real-world driving, 5 cameras with 30°–120° FoV, 1,342,418 clips / 1,715.9h), [Bench2Drive Jia et al. \(2024\)](#) (CARLA-simulated driving under 9,881 diverse traffic scenarios, 6 cameras, 384,948 clips / 511.2h), and [Sekai Sekai Team \(2025\)](#) (egocentric pedestrian walking and drone videos, 9,995 clips / 166.6h with scene and weather annotations). In total, the driving data comprises 1,744,405 clips spanning 2,405 hours. We apply a unified three-stage processing pipeline: (1) frame extraction with trajectory unification into a common waypoint format, (2) action-based clip segmentation (2–8 s) according to ego maneuver transitions, and (3) caption generation combining structured trajectory descriptions with optional VLM augmentation.

2.2.4 Egocentric Indoor Navigation Data

Egocentric indoor navigation data provides a complementary perspective to both manipulation and driving data. Unlike manipulation which focuses on fine-grained object interactions within a confined workspace, and driving which operates in large-scale outdoor environments, indoor navigation requires the model to understand room-scale spatial layouts, obstacle-aware path planning, and the mapping from textual navigation commands to spatially coherent visual trajectories.

Following VLNVerse [Lin et al. \(2025\)](#), we collect physically grounded egocentric navigation data using NVIDIA Isaac Sim [NVIDIA \(2022\)](#) with photorealistic rendering and continuous control. We gather 6,064 successful navigation episodes across 134 indoor scenes, each consisting of an egocentric RGB video (256 × 256 resolution at 10 FPS) paired with natural language navigation instructions. The trajectories average approximately 8.2 m in length (ranging from 4 to 17.5 m), accumulating a total traversal distance of roughly 49.8 km and approximately 5.8 hours of continuous first-person navigation video. The instructions are provided in two formats: single-string step-by-step directives (3,031 episodes, averaging 67.2 words) and multi-granularity descriptions at formal, natural, and casual registers (3,033 episodes).

Video generation models trained on such traversal data can acquire emergent 3D consistency and spatial coherence across frames [Gao et al. \(2026\)](#); [Bar et al. \(2025\)](#), while the physically grounded, action-conditioned nature of each sequence encourages the model to internalize depth reasoning, geometric consistency, and obstacle-aware planning [Shang et al. \(2025\)](#); [Han et al. \(2025\)](#); [Zhen et al. \(2025\)](#). By grounding language instructions in continuous egocentric traversals, this data enables the world model to jointly learn language understanding, 3D spatial reasoning, and embodied action prediction within indoor environments.

2.2.5 Human-to-Robot Transfer Data

To train the model on cross-embodiment visual correspondence without physical robot collection, we curate two complementary sources of human-to-robot transfer data. The first is a large-scale human-robot paired dataset constructed from egocentric bimanual manipulation recordings via an automated pipeline: 3D hand keypoints are extracted through MANO [Romero et al. \(2017\)](#) reconstruction and retargeted to robot end-effector trajectories, human hands are removed via video inpainting, and 14 robot arm models

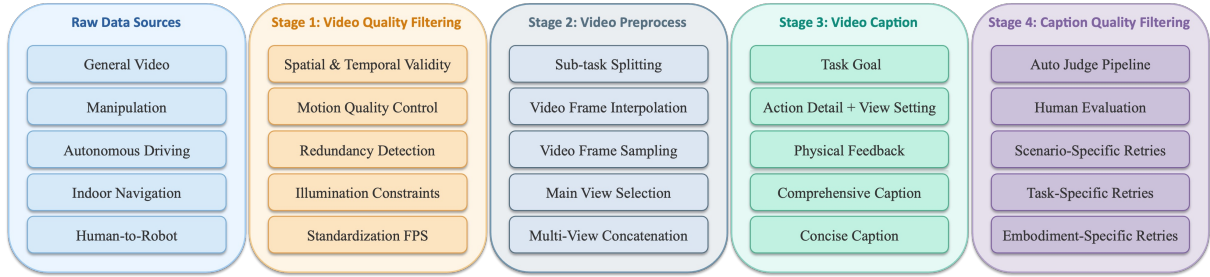


Figure 2: Overview of the unified data processing pipeline. Stage 1 (Raw Data Collection) collects heterogeneous data from five source categories spanning general and embodied domains. Stage 2 (Video Preprocessing) applies domain-adaptive operations—frame extraction, frame interpolation, sub-task splitting, main-view selection, and multi-view concatenation—to produce uniformly structured clips. Stage 3 (Hierarchical Annotation) generates viewpoint-aware captions through a five-layer framework: task goal, action detail, physical feedback, comprehensive caption, and concise caption. Stage 4 (Caption Quality Filtering) combines an automated LLM-based judge with human evaluation; underperforming captions are routed back for scenario-, task-, or embodiment-specific iterative prompt refinement.

are rendered into the inpainted scene using MuJoCo [Todorov et al. \(2012\)](#) inverse kinematics, yielding four aligned video streams per episode (original human video, hand-removed scene, pure simulation, and robot-overlaid scene). The diversity of 14 embodiments within shared scenes ensures the editing capability generalizes across robot morphologies.

The second source addresses a fundamental limitation of direct rendering: simplified renderers ignore scene illumination, cast shadows, and material-dependent specular reflections, creating a photometric gap between rendered and real observations. To bridge this, we build upon the open-sourced InternA1 dataset [Tian et al. \(2025\)](#), which uses NVIDIA Isaac Sim [NVIDIA \(2022\)](#) to provide photorealistic RGB observations with environment lighting and accurate shadows. Using the same dynamics parameters and robot URDFs, we render matched egocentric views in MuJoCo [Todorov et al. \(2012\)](#)—without lighting or shadow effects—producing paired samples that share identical geometry and viewpoint while differing in photometric realism. This paired data enables the model to learn the visual mapping between simplified rendering and photorealistic observations, covering Franka Emika Panda, AgileX Split Aloha, ARX Lift2, and AgiBot Genie1 across single-arm, dual-arm, mobile dual-arm, and humanoid configurations, with approximately 80K episodes spanning pick-and-place, articulated object manipulation, and multi-object rearrangement tasks.

2.3 Data Processing

We design a unified data processing pipeline that transforms heterogeneous raw data from diverse embodied and general video sources into high-quality, consistently formatted training samples. As illustrated in Figure 2, the pipeline consists of four stages: **(1)** Raw Data Collection, **(2)** Video Preprocessing, **(3)** Hierarchical Annotation, and **(4)** Caption Quality Filtering with Iterative Prompt Refinement. Stages 2 and 3 apply domain-adaptive operations depending on source data characteristics, while Stage 4 forms a closed feedback loop that routes underperforming captions back for targeted re-annotation.

2.3.1 Stage 1: Raw Data Collection

The pipeline begins by ingesting raw video data from five source categories spanning both general and embodied domains. **General Video** provides internet-scale visual diversity from documentaries, professional stock libraries, and curated web clips. **Manipulation** data covers a broad spectrum of robot embodiments—single-arm grippers, dual-arm systems, dexterous hands, mobile platforms, and humanoids—from datasets including EgoHOD, Bridge V2, DROID, RoboMind, Agibot-World, and others. **Autonomous Driving** contributes large-scale ego-motion and multi-agent dynamics from Waymo, Bench2Drive, NVIDIA PhysicalAI-AD, and Sekai. **Indoor Navigation** supplies language-guided spatial reasoning episodes from VLNVerse across 134 indoor scenes. **Human-to-Robot Transfer** provides paired human demonstration and robot execution data constructed via our automated MANO-to-robot pipeline across 14 robot types.

2.3.2 Stage 2: Video Preprocessing

Raw videos undergo domain-adaptive preprocessing to produce uniformly structured clips suitable for training. We apply five complementary operations depending on the source data characteristics:

Frame Extraction. For short-horizon task videos (typically single-step manipulations lasting 2–8 s), we extract frames at a target rate that captures the essential phases of the interaction—approach, contact, manipulation, and result—ensuring each sample contains the complete causal chain of the atomic action.

Frame Interpolation. When source videos have insufficient frame rates for smooth motion learning, we apply temporal interpolation to increase frame density, preserving continuous motion trajectories critical for modeling fine-grained contact dynamics and object-state transitions.

Sub-task Splitting. For long-horizon episodes involving multi-step procedures (e.g., sequential pick-and-place, complex assembly), we decompose the video into semantically coherent sub-task segments. Each segment captures a complete atomic action with clear start and end states, preventing the partial-execution artifacts that arise from naive uniform truncation.

Main-View Selection. For multi-camera recordings where only a primary viewpoint is needed (e.g., single-view manipulation training), we select the most informative camera stream—typically the egocentric or external view that best captures the interaction region—discarding redundant angles.

Multi-View Concatenation. Conversely, for multi-view co-training, we concatenate synchronized clips from 2–4 camera viewpoints into a single horizontal layout, preserving temporal alignment across views. This enables the model to learn cross-view geometric consistency and synchronized state transitions without architectural modifications.

2.3.3 Stage 3: Hierarchical Annotation

Hierarchical Annotation Prompt Template

You are an expert embodied-AI annotator. Given a video clip of a {{embodiment_type}} performing a manipulation task captured from a {{viewpoint_type}} viewpoint, produce the following five-layer annotation.

— Analysis Phase —

Layer 1 – Task Goal: Identify the high-level intent of this interaction. What is the agent trying to achieve? Describe the desired state transition from the current observation to the goal state in one sentence.

Layer 2 – Action Detail: Decompose the agent’s actions into a step-by-step sequence. For each step, specify: (a) the motion trajectory and direction, (b) micro-actions (approach, grasp, lift, rotate, release, etc.), (c) estimated speed and force level. **You must explicitly state the viewpoint** (egocentric / wrist / external / multi-view concatenation).

Layer 3 – Physical Feedback: Describe the observable physical consequences of each action on the environment: object displacement, deformation, contact state changes, and any secondary effects (e.g., liquid sloshing, cloth folding). Only include visually verifiable outcomes.

— Generation Phase —

Layer 4 – Comprehensive Caption (50–100 words): Synthesize Layers 1–3 into a cohesive paragraph that fully specifies the *viewpoint–agent–action–feedback* quadruple. Include the camera perspective, embodiment identity, complete action sequence, and physical outcomes.

Layer 5 – Concise Caption (15–30 words): Condense to an instruction-style summary retaining only the essential *viewpoint–agent–key action* elements, suitable as a direct language command for the world model.

Quality Constraints:

- *Operation focus:* describe only agent actions and object interactions; omit background narration.
- *Viewpoint definition:* explicitly name the viewpoint type and its semantic role.
- *Objectivity:* report only visible dynamics; do not infer hidden states.
- *Physical verifiability:* every claimed outcome must be visually confirmable from the video.

Preprocessed videos pass through our five-layer hierarchical annotation framework (Section 2.1), which generates viewpoint-aware captions at two granularities—comprehensive (50–100 words) and concise (15–30 words)—sampled with equal probability during training. The prompt template used by the annotation model is shown above.

2.3.4 Stage 4: Caption Quality Filtering

To ensure annotation quality across the diverse range of scenarios, tasks, and embodiments in our corpus, we implement a closed-loop quality filtering system combining automated assessment with human oversight. Captions that fail quality checks are routed back to Stage 3 for targeted re-annotation, forming an iterative refinement loop.

Judge Pipeline. An automated LLM-based judge assesses each caption along several dimensions, including factual accuracy, specificity, instruction clarity, and viewpoint consistency. Specifically, it evaluates whether the caption correctly describes the video content, provides sufficient detail beyond

generic descriptions, can function as an actionable command, and maintains spatial references consistent with the camera perspective. Captions that do not satisfy any of these criteria are flagged for further review.

Human Evaluation. A subset of captions—particularly those near judgment thresholds or from underrepresented domains—undergoes manual review by human annotators who validate correctness, identify systematic failure patterns, and provide ground-truth corrections that inform subsequent prompt refinements.

Iterative Prompt Refinement. When the judge pipeline identifies consistent underperformance in specific categories, we trigger targeted prompt redesign along three axes: **scenario-specific** retries (e.g., outdoor lighting conditions, kitchen environments), **task-specific** retries (e.g., articulated object manipulation, fluid pouring), and **embodiment-specific** retries (e.g., humanoid bimanual coordination, dexterous hand manipulation). Each retry employs a specialized prompt template tailored to the failure mode, and the refined captions are re-evaluated through the judge pipeline until they meet quality standards. This iterative loop ensures that no scenario, task, or embodiment category suffers from systematically poor annotations due to one-size-fits-all prompting.

Final Corpus Statistics. After the complete four-stage pipeline, the final training corpus comprises approximately 8.6M video-text pairs (over 200M observation frames), with embodied data accounting for 70% and general data for 30%. Within the embodied portion, single-view manipulation data constitutes the majority at ~ 4.3 M samples, followed by ~ 1.6 M multi-view concatenated samples with synchronized 2–4 camera views, and ~ 200 K navigation and driving samples.

3 Model

3.1 Model Architecture

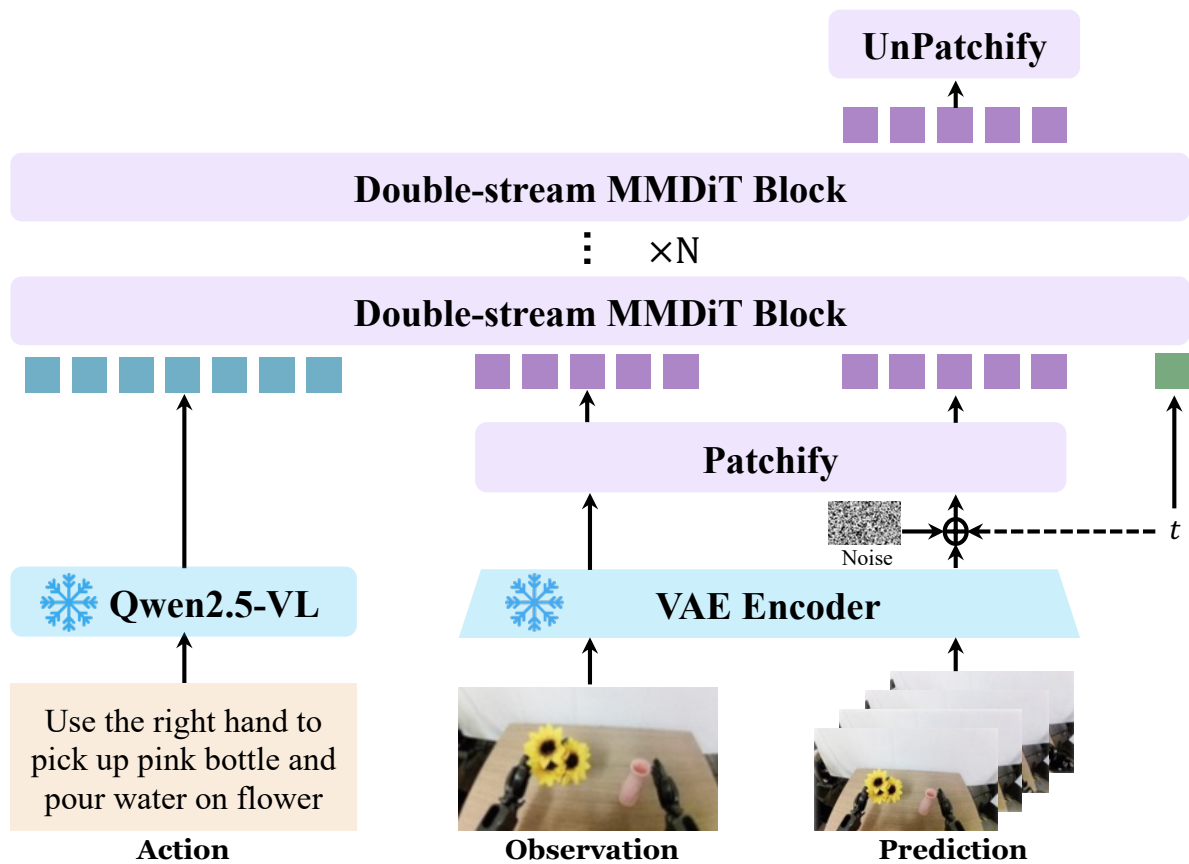


Figure 3: Overview of our video generation architecture with 60-layer double-stream MMDiT backbone.

As shown in Figure 3, the model consists of three components: an MLLM as the action encoder, a VAE as the state encoder/decoder, and an MMDiT Esser et al. (2024) as the transition function, organized in a

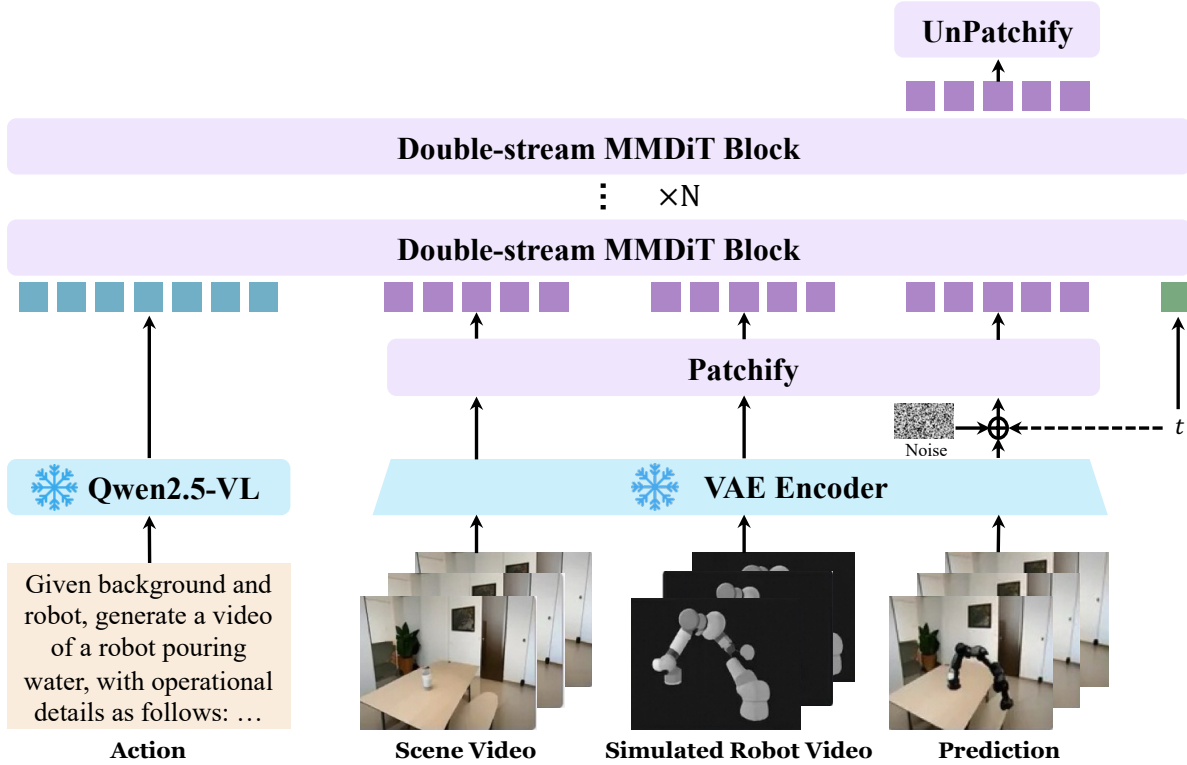


Figure 4: **Scene2Robot: multi-segment conditioning for cross-embodiment video synthesis.** The input sequence is organized as three contiguous segments — scene condition (F frames), robot reference (F frames), and generation (F frames). An index-based mechanism assigns condition tokens to timestep $t = 0$ and excludes them from loss computation, so only the generation segment is trainable. Joint attention at every MMDiT block enables the generation segment to simultaneously attend to scene appearance and robot motion trajectory, producing semantically coherent cross-embodiment synthesis.

double-stream design.

MLLM — Action Encoder. We employ a frozen Qwen2.5-VL Bai et al. (2025) to encode user inputs into condition signals. For a given input text S , it extracts last-layer hidden states $\mathbf{h} = \phi(S)$, serving as the action condition.

VAE — State Encoder/Decoder. The VAE encodes video frames into latent representations $\mathbf{z} = \mathcal{E}(\mathbf{x})$ and decodes predicted latents back into visual observations. We adopt the Wan-VAE Wan et al. (2025) architecture, which handles both image and video modalities.

MMDiT — Transition Function. The MMDiT adopts a double-stream architecture: the understanding stream receives the MLLM encoding \mathbf{h} (projected via a trainable connector), and the generation stream receives noisy state latents from the VAE. At each block, the two streams interact via joint attention. The backbone comprises 60 double-stream blocks with 24 attention heads (head dimension 128), hidden size 3,072, and patch size 2×2 . Total parameters: MLLM 7B, VAE 127M (encoder 54M + decoder 73M), MMDiT 20B. The context length supports up to 48,360 video tokens.

3.2 3D Rotary Position Encoding

We employ 3D RoPE Su et al. (2024); Heo et al. (2024) to independently encode the temporal, spatial height, and spatial width dimensions. Rather than allocating dimensions uniformly, we use an asymmetric split: 16 dimensions for the temporal axis and 56 dimensions each for height and width, totaling 128 dimensions ($pe_axes_dim = [16, 56, 56]$). The temporal axis receives fewer dimensions as adjacent frames are strongly correlated; the spatial axes receive more to capture the greater diversity of object positions and scene layouts. We also apply Scalable RoPE Wan et al. (2025) to support generalization to varying resolutions and durations at inference.

3.3 Scene2Robot

Building upon the double-stream MMDiT architecture (§3.1) and the asymmetric 3D RoPE encoding (§3.2), we design SCENE2ROBOT, a multi-segment conditioning mechanism that repurposes the same backbone for cross-embodiment video synthesis, as illustrated in Figure 4.

First-Frame Conditioning (TI2V Baseline). For standard text-image-to-video tasks, the first frame serves as a fixed visual condition: its VAE latents are assigned timestep $t=0$ in the generation stream and excluded from the denoising loss, while the frozen Qwen2.5-VL encodes the text instruction into the understanding stream. Because the double-stream joint attention (§3.1) fuses both signals at every layer, the generation tokens can simultaneously attend to the visual anchor and the semantic action specification, producing temporally coherent continuations grounded in the language command.

Multi-Segment Extension for Human-to-Robot Transfer. Human-to-robot transfer poses a video editing problem: the model must reference *both* the scene context (background, object layout, lighting) *and* the target robot’s motion trajectory from a simulated demonstration. We address this by extending first-frame conditioning to a three-segment input sequence, all processed within the same VAE-MMDiT pipeline without any architectural modification:

1. **Scene condition** (F frames): the original human demonstration video, with human hands masked out, encoded by the VAE to provide appearance, spatial layout, and object state information.
2. **Robot reference** (F frames): a simulated robot execution rendered via MuJoCo, encoded by the VAE, supplying the target embodiment’s kinematic trajectory and morphology.
3. **Generation** (F frames): noisy latents to be denoised into the final photorealistic robot execution video.

Segments (1) and (2) share the same $t=0$ assignment as first-frame conditioning and are excluded from loss computation; only segment (3) receives gradient updates during training. The 3D RoPE encoding (§3.2) assigns each segment its own temporal index range, allowing the model to distinguish temporal positions across segments. Joint attention in every MMDiT block then enables the generation tokens to simultaneously attend to scene appearance from segment (1), robot motion from segment (2), and the MLLM action semantics from the understanding stream. This tripartite conditioning enables the model to synthesize photorealistic robot executions that faithfully preserve both the scene context and the instructed manipulation behavior.

4 Training

4.1 Training Strategy

We propose a joint training paradigm in which general scene generation and robot manipulation prediction are unified under a single natural language interface as the same conditional video generation task, with the model continuously receiving gradient updates from both data regimes throughout training. This shared formulation allows general world priors and embodied action priors to reinforce each other through a common backbone, enabling stable cross-scenario and cross-task co-training. The curriculum proceeds in two progressive stages: pretraining establishes broad world foundations, and SFT deepens embodied specialization while preserving the general-expert balance.

4.1.1 Pretraining Stage: Establishing General World Foundation

General World Priors. We curate over 200M real-world observation samples from 14 high-quality video platforms, covering natural scenes, daily life, and sports. This breadth allows the model to internalize domain-agnostic world priors—object motion, lighting variation, collision dynamics—that form the general backbone for later embodied generalization. We further incorporate multi-camera synchronized observations with 3D RoPE spatial encoding, establishing preliminary cross-view geometric consistency as a spatial foundation for multi-view embodied generation.

Human Interaction Priors. We introduce large-scale first-person hand manipulation data (Ego4D [Grauman et al. \(2022\)](#), EPIC-Kitchen [Damen et al. \(2018\)](#), etc.). Human demonstration serves as a natural bridge between general and embodied: by learning grasping, tool use, and object manipulation from everyday human behavior, the model builds action priors and affordance understanding that transfer directly to robot operation in later stages.

Multi-Task Joint Training. T2I, T2V, and TI2V tasks are trained jointly on a shared backbone, serving as the core mechanism through which general and embodied capabilities coexist in one model. The T2I task learns sharp visual representations from general image data, acting as a visual quality anchor whose

object morphology knowledge automatically transfers to video generation tasks through the shared backbone, preventing deformation and identity inconsistency. Task ratios gradually shift from pure T2I toward full three-task joint training, so the model operates stably across multiple generation modes by the end of pretraining.

4.1.2 SFT Stage: Embodied Specialization

The SFT stage progressively deepens embodied expertise while keeping general world data in every training batch, ensuring that embodied specialization and general world modeling capability advance together rather than trade off.

Progressive Embodied Knowledge Injection. We adopt a four-phase data mixing schedule. In early training, multi-embodiment robot data and human hand manipulation data co-dominate: human action priors guide the learning of cross-embodiment operation commonalities, while robot data strengthens concrete execution representations. We then gradually increase wrist-view and third-person view data to broaden viewpoint coverage. Building on this, we introduce multi-view concatenated training: synchronized first frames from multiple cameras are spatially concatenated as a single input, requiring the model to jointly generate subsequent frames for all views simultaneously, forcing the attention layers to establish cross-view spatial correspondences and achieve geometrically consistent multi-view generation. In the final phase, scarce high-complexity tasks (pouring, folding, bimanual coordination, multi-material interaction) and long-horizon reasoning data are targeted for supplementation to push the frontier of embodied capability. Throughout this process, general world data continuously participates in every training batch, jointly acting on the same backbone alongside embodied data to ensure that embodied specialization and general world modeling capability advance together.

4.2 Training Objective and Infrastructure

We adopt the flow matching objective Lipman et al. (2023); Liu et al. (2022), where input videos are encoded into latent space via the VAE encoder and noise is sampled from a standard normal distribution. Qwen2.5-VL encodes text inputs as guidance signal. Timesteps are sampled from a log-normal distribution with adaptive shifting based on video sequence length Esser et al. (2024). For T2V tasks, the first-frame timestep is fixed at 0 to ensure that the generation process is conditioned on the given observation frame. Training is conducted with Megatron-LM Shoeybi et al. (2019) using a hybrid parallelism strategy, with selective activation recomputation Korthikanti et al. (2023) applied to a subset of dual-stream blocks to balance memory usage and training throughput.

5 Experiments

We conduct comprehensive evaluations on four benchmarks spanning embodied manipulation, physical reasoning, and general video quality. Across these benchmarks, our model delivers consistently strong results, achieving state-of-the-art performance on EWMBench for embodied world modeling (Overall **4.60**, +0.55 over LVP), ranking **1st overall** on DreamGen Bench (Total **4.952**), and **1st among open-source models** on WorldModelBench (Total **8.99**).

Quantitative Evaluation (§5.1). We evaluate against two categories of baselines: (1) *general video generation models*—Sora2 OpenAI (2024), Veo3 Google DeepMind (2025), Wan2.6 Wan Team (2025), Kling Kuaishou Technology (2024), and LTX-2 Lightricks (2025); and (2) *embodied world models*—Cosmos Agarwal et al. (2025), WoW Chi et al. (2025), LVP Chen et al. (2025a), Vidar Feng et al. (2025), and GigaWorld Team et al. (2025).

Qualitative Analysis (§5.2). We evaluate manipulation capabilities along three progressive dimensions: fine-grained language grounding, generalization across embodiments, tasks, and viewpoints, and zero-shot robustness against strong baselines.

Cross-Domain Generalization (§5.3) further covers human-to-robot transfer, autonomous driving, and indoor navigation as supplementary tasks.

5.1 Quantitative Evaluation

Unless noted otherwise, quantitative tables use **boldface** for the best value in each column and underline for the second best.

5.1.1 EWMBench: Embodied Motion Fidelity

Benchmark. EWMBench Yue et al. (2025) evaluates embodied world models on three dimensions: *scene consistency* (SceneC), *motion correctness* (HSD, Dyn, nDTW), and *semantic alignment* (Diversity, BLEU, CLIP, Logics). The benchmark contains 21 samples across 7 tasks with clear action-ordering constraints.

Table 2: Performance comparison on EWMBench.

Type	Model	Scene	Motion			Semantics				Overall
		SceneC	HSD	Dyn	nDTW	Diversity	BLEU	CLIP	Logics	
General	Veo3	0.8415	0.2130	0.1932	0.1613	0.0221	0.2139	0.8965	0.9474	3.49
	Wan2.6	0.6712	0.2034	0.0900	0.1715	0.0502	0.1616	0.8743	1.0000	3.22
	Kling26	0.8211	0.3272	0.1822	0.3423	0.0173	0.2591	0.9014	1.0000	3.85
	LTX-2	0.7850	0.2076	0.1283	0.2443	0.0120	0.1425	0.8869	0.5000	3.01
	Sora2	0.8526	0.2807	0.3494	0.2754	0.0314	<u>0.2466</u>	0.9100	0.9474	3.89
Embodied	Cosmos	0.7963	0.2500	0.2052	0.2533	0.0803	0.1230	0.8458	0.7333	3.29
	GigaWorld	0.8707	0.3050	0.0849	0.2783	0.0278	0.2048	0.8873	0.9000	3.56
	LVP	0.8795	<u>0.4248</u>	0.0433	<u>0.6226</u>	0.0093	0.2179	0.8995	<u>0.9524</u>	<u>4.05</u>
	Vidar	0.7341	<u>0.1877</u>	0.1520	<u>0.1769</u>	<u>0.0653</u>	0.1607	0.8821	<u>0.9411</u>	3.30
	Wow	<u>0.8866</u>	0.2494	0.0529	0.2566	0.0266	0.1932	0.9001	<u>0.9524</u>	3.52
Ours	0.9142	0.5660	<u>0.3429</u>	0.6708	0.0114	0.2079	0.8834	1.0000	4.60	

Results. Table 2 shows our model ranks **1st** overall with a score of **4.60**, outperforming the runner-up LVP (4.05) by +0.55. We lead in motion fidelity—HSD (**0.566**) surpasses LVP (0.425) by 33%—and achieve top performance in scene consistency (SceneC: **0.914**) and logic constraint satisfaction (Logics: **1.00**).

5.1.2 DreamGen Bench

Benchmark. DreamGen Bench Zhou et al. (2025) evaluates the quality of robot videos generated by video world models, measuring instruction following (IF) and physics alignment (PA) across three subsets of the GR1 robot embodiment: environment generalization (GR1-Env), object generalization (GR1-Object), and behavior generalization (GR1-Behavior). IF is assessed using Qwen2.5-VL Bai et al. (2025) as the evaluator.

Table 3: Performance comparison on DreamGen Bench.

Model	GR1-Env		GR1-Object		GR1-Behavior		Total
	PA	IF	PA	IF	PA	IF	
Cosmos-sft	0.709	0.655	<u>0.775</u>	0.720	0.649	0.621	4.129
LVP	<u>0.810</u>	0.772	0.745	0.829	0.713	0.889	<u>4.758</u>
Vidar	<u>0.445</u>	0.647	0.478	0.726	0.394	0.651	<u>3.341</u>
GigaWorld	0.621	0.933	0.500	<u>0.852</u>	0.426	<u>0.884</u>	4.216
Wow	0.793	<u>0.826</u>	0.755	0.849	0.809	0.696	4.728
Ours	0.828	0.793	0.840	0.878	<u>0.781</u>	0.832	4.952

Results. Table 3 shows our model achieves the highest total score of **4.952**, ranking **1st** overall. We lead in GR1-Object IF (**0.878**, 1st), demonstrating strong object-level compositional generalization, and physics alignment is consistent across all subsets (PA: 0.828/0.840/0.781). GR1-Behavior IF (0.832) slightly trails LVP (**0.889**) and GigaWorld (0.884), indicating long-horizon behavior generalization as a direction for further improvement.

5.1.3 PBench: Physical Behavior Evaluation

Benchmark. PBench NVIDIA (2025a) evaluates models on two complementary aspects: (1) *Domain Score*, which measures physical behavior understanding via QA pairs assessed by Qwen2.5-VL across six domains (AV, Robot, Industry, Physics, Human, Common Sense); and (2) *Quality Score*, which measures visual quality via eight VBench Huang et al. (2024) metrics including image-to-video consistency, aesthetic quality, motion smoothness, and subject consistency. The *Overall Score* is the average of the two.

Results. As shown in Table 4, our model outperforms all among open-source models with an overall score of **0.804**. Domain understanding is our strongest dimension (0.857, **3rd overall**), surpassing most closed-source models. Motion smoothness also stands out (0.990, **2nd among open-source models**), reflecting consistent temporal coherence in generation. Aesthetic quality (0.455) and imaging quality (0.649) are relatively lower, primarily because our model is purpose-built for embodied tasks and operates

Table 4: Performance comparison on PBench.

Type	Model	Quality Metrics (VBench)								Qual.	Domain	Overall
		I2V-Bg	I2V-S	Aes	Img	Bg-Con	Mot	Sub-Con	O-Con			
General	Veo3	0.975	0.980	0.526	0.698	0.938	0.994	0.927	0.128	0.771	0.882	0.827
	Wan2.6	0.856	0.843	0.514	0.719	0.906	0.978	0.843	0.136	0.724	0.832	0.778
	Sora2	<u>0.981</u>	0.973	0.487	0.672	0.961	0.994	<u>0.954</u>	0.129	0.769	0.841	0.805
	Kling26	0.982	0.979	<u>0.521</u>	<u>0.699</u>	0.920	0.990	0.927	0.124	0.768	<u>0.874</u>	<u>0.821</u>
	LTX-2	0.948	0.955	0.506	0.622	0.932	0.986	0.904	0.118	0.746	0.845	0.796
Embodied	Cosmos	0.974	0.973	0.470	0.663	0.940	0.989	0.931	0.160	0.763	0.840	0.802
	LVP	0.979	0.981	0.515	0.679	0.954	<u>0.991</u>	0.962	0.116	0.772	0.812	0.792
	GigaWorld	0.957	0.944	0.495	0.641	0.925	0.984	0.892	0.128	0.746	0.841	0.794
	Vidar	0.935	0.922	0.501	0.573	0.912	0.982	0.863	0.120	0.726	0.810	0.768
	Wow	0.967	0.957	0.517	0.689	0.941	0.980	0.929	0.111	0.761	0.786	0.774
Ours	0.956	0.943	0.455	0.649	<u>0.956</u>	0.990	0.933	0.124	0.751	0.857	0.804	

at a lower output resolution than general-purpose video generators, which reduces VBench’s pixel-level quality scores; nonetheless, this resolution is fully sufficient for downstream robot control tasks.

5.1.4 WorldModelBench: Physical Reasoning and Instruction Following

Benchmark. WorldModelBench Li et al. (2025) evaluates models on three dimensions: *instruction following* (0–3 scale), *common sense* (frame and temporal quality), and *physics adherence* (5 violation types: Newton’s laws, mass conservation, fluid dynamics, penetration, gravity). The benchmark contains 350 instances across 7 domains with 56 subdomains.

Table 5: Performance comparison on WorldModelBench.

Type	Model	Instr.	Common Sense			Physics Adherence					Phys.	Total
		(0-3)	Frame	Temp	Overall	Newton	Mass	Fluid	Penetr.	Grav.	Overall	
General	Veo3	2.52	<u>0.98</u>	<u>0.95</u>	1.93	1.00	0.89	0.99	0.91	1.00	4.80	<u>9.25</u>
	Wan2.6	<u>2.50</u>	0.99	<u>0.95</u>	1.94	1.00	0.89	0.99	<u>0.94</u>	1.00	<u>4.83</u>	9.27
	Sora2	2.21	0.96	0.93	1.88	1.00	0.91	0.99	<u>0.95</u>	1.00	4.84	8.93
	Kling26	1.59	0.97	1.00	1.97	1.00	1.00	1.00	1.00	1.00	5.00	8.55
	LTX-2	1.97	0.69	0.62	1.32	0.99	0.60	1.00	0.73	1.00	4.32	7.61
Embodied	Cosmos	2.14	1.00	0.94	<u>1.94</u>	1.00	0.92	1.00	0.94	1.00	4.86	8.94
	LVP	2.01	0.89	0.91	1.80	1.00	<u>0.93</u>	0.99	0.95	1.00	4.87	8.67
	GigaWorld	2.13	0.59	0.46	1.05	1.00	0.48	0.99	0.69	0.98	4.13	7.31
	Vidar	1.62	0.54	0.45	0.99	1.00	0.56	1.00	0.85	1.00	4.40	7.01
	Wow	2.05	0.76	0.65	1.41	1.00	0.65	0.99	0.81	1.00	4.45	7.91
Ours	2.33	0.87	0.85	1.72	1.00	1.00	1.00	0.94	1.00	<u>4.94</u>	8.99	

Results. Table 5 shows our model outperforms all open-source models (**8.99**, 3rd overall), trailing only closed-source Wan2.6 and Veo3. We achieve perfect physics adherence (1.00) across all four categories and strong instruction following (2.33/3.0), with the common-sense gap attributable to our lower output resolution.

5.2 Qualitative Analysis

5.2.1 Fine-Grained Language Grounding

Precise grounding of language in visual actions is foundational to QWEN-ROBOTWORLD’s design as a language-conditioned world model. Figure 5 evaluates this capability at two levels. (a) *Contrastive pairs*: given identical initial frames, the model produces qualitatively distinct videos when a single keyword differs—target object identity, action type, or spatial placement—demonstrating fine-grained semantic discrimination beyond generic manipulation priors. (b) *Complex instructions*: the model handles long-horizon sequential tasks with multi-step dependencies and abstract goal instructions that require inferring the manipulation sequence from context, decomposing each into a temporally coherent execution without explicit sub-task prompts.

5.2.2 Generalization across Embodiments, Tasks, and Viewpoints

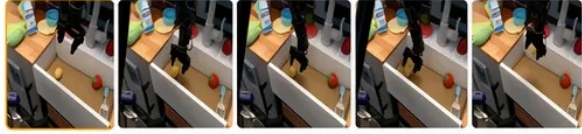
Figure 6 demonstrates three complementary dimensions of QWEN-ROBOTWORLD’s manipulation capability. (A) **Cross-embodiment**: a single instruction drives four distinct robot morphologies—single-arm gripper, dual-arm system, humanoid, and dexterous hand—without embodiment-specific adaptation, validating natural language as a universal action interface; each cell shows three key frames (initial, mid,

(a) Contrastive Instruction Following

Pick up the **red strawberry** from the table.



Pick up the **yellow potato** from the table.



Target object

Pick up the pen from the table and place it on the **wooden tray**.



Pick up the pen from the table and place it on the **white paper**.



Destination

Take the glue stick and **hand it to the person**.



Take the glue stick and **put it in the penholder**.



Action type

(b) Complex Instruction Following

Grasp each Rubik's cube in turn and stack them into a stable two-layer structure on the tray.



① Reach cube 1

② Place on tray

③ Reach cube 2

④ Stack on top

Pick up the red and yellow bell peppers in sequence and place them on the table from left to right from the camera's perspective.



① Pick red pepper

② Place left slot

③ Pick yellow pepper

④ Place right slot

Figure 5: **Fine-grained language grounding.** (a) **Contrastive:** each pair of columns shares an identical initial frame (colored border); only the highlighted keyword differs between the two instructions. *Pair 1:* target object identity. *Pair 2:* destination. *Pair 3:* action type. In every case the generated motion is precisely grounded to the discriminating keyword. (b) **Complex:** two examples requiring multi-step execution or abstract goal inference. Colored labels mark key action milestones within each generated sequence.

final). (B) **Cross-task \times cross-environment:** generations across fruit pick-and-place, bowl retrieval, cloth folding, and human-robot handover each exhibit task-appropriate contact dynamics, reflecting grounded physical knowledge across diverse real-world environments; each row shows an initial frame followed by four evenly-spaced generated frames. (C) **Multi-view consistency:** three synchronized camera streams (main, wrist-left, wrist-right) are jointly generated from the same supermarket pick-and-place episode as (B, row 1), with object identity and motion trajectory remaining geometrically consistent across all viewpoints.

5.2.3 Zero-Shot Robustness on RoboTwin-IF

Building on the single-model capabilities demonstrated above, we next examine whether these gains persist under controlled model-to-model comparisons. Aggregate embodied-world-model scores can entangle three different failure sources: instruction mismatch, cross-view inconsistency, and generic visual degradation. To isolate these factors, we perform a **zero-shot** side-by-side comparison on four Unitree G1 tasks against two strong embodied baselines, LVP and **Cosmos2.5-14B**. Figure 7 shows that

(A) Cross-Embodiment

same instruction · four morphologies

Pick up the object and place it in the container. Same instruction across four morphologies.



(B) Cross-Task × Cross-Environment

task-appropriate contact dynamics

Kitchen
Move bowl to
sink



Living room
Fold &
organize



Tabletop
H-R handover



(C) Multi-View Consistency

same episode · three cameras

Main



Wrist-L



Wrist-R

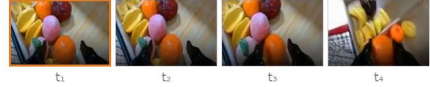


Figure 6: **Generalization across embodiments, tasks, and viewpoints.** (A) **Cross-embodiment:** one instruction drives four morphologies (single-arm, dual-arm, humanoid, dexterous hand); three key frames per cell. (B) **Cross-task × cross-environment:** initial frame (orange border) followed by four generated frames across four tasks. (C) **Multi-view:** main and wrist cameras jointly generated from the same episode as (B, row 1).

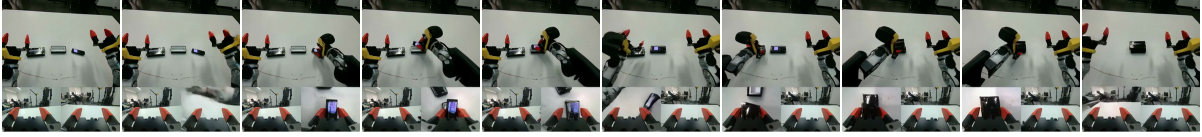
QWEN-ROBOTWORLD more consistently preserves language-grounded execution (correct object/action correspondence and cleaner goal completion) while maintaining coherent multi-view trajectories. The two baselines show different failure patterns. **LVP** more often produces incomplete task execution, while **Cosmos2.5-14B** tends to exhibit weaker alignment between the instruction and the generated manipulation outcomes in more complex cases.

To validate this behavior under a benchmark setting, we evaluate zero-shot performance on **RoboTwin-IF** (*Instruction Following*), a newly proposed benchmark built on the RoboTwin simulator with many newly constructed complex tasks. Notably, although QWEN-ROBOTWORLD mixes only a small amount of open-source RoboTwin data during training, it still shows strong zero-shot performance on RoboTwin-IF together with stable multi-view consistency across synchronized camera streams. These results suggest that the model’s gains are not limited to a few qualitative examples, but generalize to more challenging unseen embodied tasks. Overall, QWEN-ROBOTWORLD demonstrates stronger zero-shot robustness than prior baselines by better aligning instruction following, action realism, and cross-view coherence in a unified generation framework.

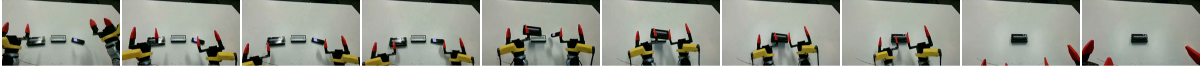
Task 01

Assemble a small electronic device by inserting a black rectangular component into a matching black housing.

Ours (multi-view)



LVP (single-view)



Cosmos (single-view)



Task 02

Wipe yellowish liquid stains from a light table using a gray rag until clean and dry.

Ours (multi-view)



LVP (single-view)



Cosmos (single-view)



Task 03

Place banana, watermelon slice, and avocado onto color-matched plates from the cutting board.

Ours (single-view)



LVP (single-view)



Cosmos (single-view)



Task 04

Clean the whiteboard using a blue eraser and remove the red wavy line/markings.

Ours (single-view)



LVP (single-view)



Cosmos (single-view)

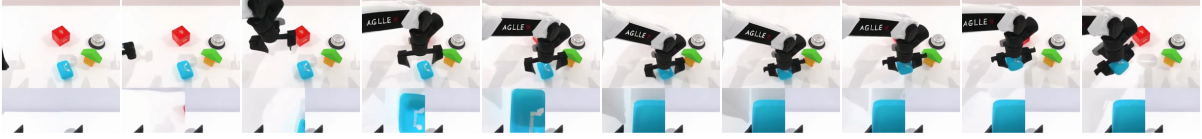


Figure 7: **Zero-shot qualitative comparison on language–action alignment and multi-view coherence.** Side-by-side grids under identical conditioning (same initial frame(s), prompt, and camera layout), comparing QWEN-ROBOTWORLD against LVP and Cosmos2.5-14B.

Figure 8 provides representative RoboTwin-IF zero-shot cases as qualitative evidence for this benchmark result. Each task is visualized with ten uniformly sampled frames anchored by the first and last frame, making intermediate progress and final completion directly visible. Across these newly constructed complex tasks, QWEN-ROBOTWORLD preserves coherent execution and cross-view consistency, which is consistent with the quantitative RoboTwin-IF finding.

Basic Manipulation

Left arm picks up the blue tissue box from the table.



Right arm picks up the Rubik's cube and shakes it once.



Relative Positioning

Left arm picks up the mouse and places it at the upper-left area of the table.



Right arm picks up the wooden box and places it on top of the phone.



Bimanual Collaboration

Left arm picks up the wooden box, handovers to right arm, and right arm places it on the soap.

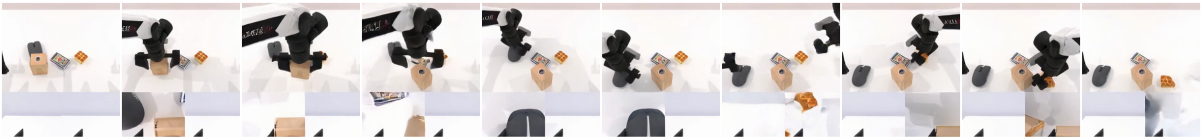


Left arm picks up the Rubik's cube while right arm picks up the bread.



Long-Horizon / Complex Instructions

Place wooden box in front, mouse to left, then bread to right in a three-step sequence.



Place yellow box at center, bell to right, then mouse to left in sequence.



Figure 8: **RoboTwin-IF zero-shot qualitative cases.** The benchmark is built on the RoboTwin simulator with newly constructed complex tasks.

5.3 Cross-Domain Generalization

Beyond manipulation-centric evaluation, we assess the model's generalization to supplementary task families beyond the core manipulation domain. Figure 9 shows human-to-robot transfer across eight target embodiments, where the model preserves task intent from human demonstrations while adapting motion to embodiment-specific kinematic constraints. Figure 10 covers mobility scenarios, including autonomous driving episodes from Bench2Drive, NVIDIA PhysicalAI-AD, Sekai, and Waymo, and egocentric indoor navigation episodes from VLNVerse. Together, these results indicate that the learned language-conditioned transition model generalizes beyond a single embodiment or scenario family.



Figure 9: **Human-to-robot transfer.** Across eight target embodiments, each row compares a human demonstration (left) with the synthesized robot execution (right) for the same task, using five uniformly sampled frames per video. The generated trajectories preserve task intent while adapting motion to embodiment-specific kinematic constraints.

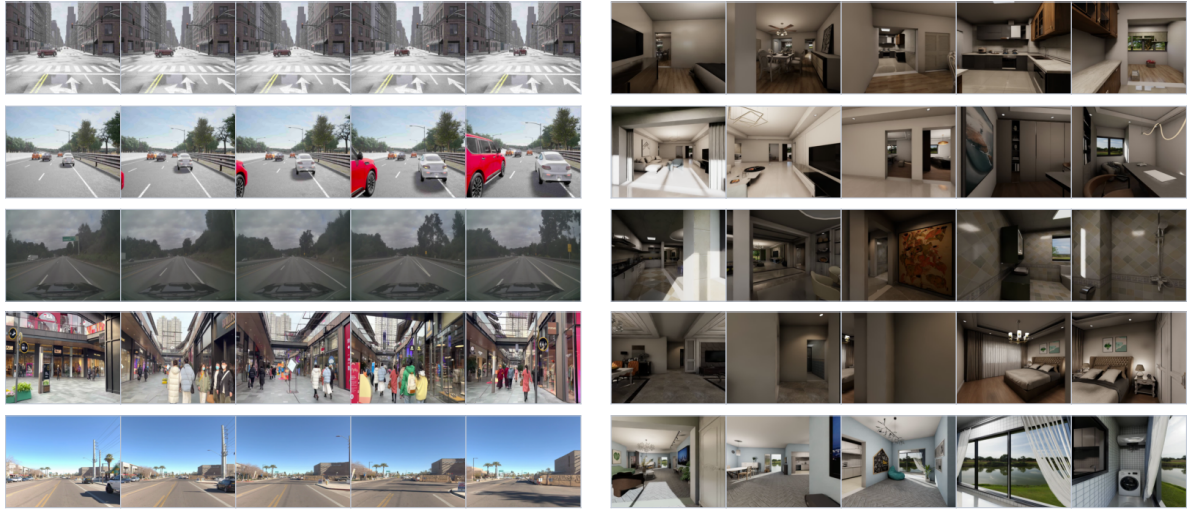


Figure 10: **Mobility generation.** Paired columns with five rows: **(left) Autonomous driving** episodes from Bench2Drive, NVIDIA PhysicalAI-AD, Sekai, and Waymo; **(right) Egocentric indoor navigation** from VLNVerse with language-guided first-person traversal. Each episode uses five uniformly sampled frames.

6 Conclusion

In this report, we present QWEN-ROBOTWORLD, a language-conditioned world model framework for embodied intelligence that unifies robotic manipulation, autonomous driving, indoor navigation, and human-to-robot transfer under a shared natural language action interface. To realize this objective, we develop a three-part system: a double-stream MMDiT architecture with MLLM action encoding for semantically precise and physically grounded generation, the Embodied World Knowledge (EWK) dataset with large-scale cross-embodiment action-language alignment, and a general+expert progressive curriculum that couples broad visual priors with embodied specialization. This design enables one

common backbone that can be adapted toward three representative embodied world model applications—synthetic data generation, policy evaluation, and action planning. Across both benchmark evaluations and zero-shot analyses, QWEN-ROBOTWORLD demonstrates strong, consistent performance and robust multi-view instruction-following generalization. We hope this work provides a practical foundation for building embodied world models that are not only perceptually strong, but also functionally useful for downstream robotic learning and control.

Authors

Jie Zhang*, Xiaoyue Chen*, Anzhe Chen, Chenxu Lv, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, Jiazhao Zhang, Jingren Zhou, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Pei Lin, Qihang Peng, Shengming Yin, Tianhe Wu, Tianyi Yan, Xiao Xu, Yan Shu, Yanran Zhang, Ye Wang, Yi Wang, Yilei Chen, Yixian Xu, Yiyang Huang, Yuxiang Chen, Zekai Zhang, Zhendong Wang, Zixing Lei, Zhixuan Liang, Zihao Liu, Zikai Zhou, Xiong-Hui Chen[†], Chenfei Wu[†]

*Equal contribution.

[†]Corresponding author.

References

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.
- AgiBot-World-Contributors. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025.
- Johan Bjorck, Fernando Castaneda, Linxi Fan, Dieter Fox, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. RT-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems (RSS)*, 2023.
- Build AI. Egocentric-10k. Hugging Face Datasets, 2025. URL <https://huggingface.co/datasets/builddotai/Egocentric-10K>.
- Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control, 2025a. URL <https://arxiv.org/abs/2512.15840>.
- Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025b.
- Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow: Towards a world omniscient world model through embodied interaction, 2025. URL <https://arxiv.org/abs/2509.22642>.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, 2018.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.
- Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. URL <https://arxiv.org/abs/2507.12898>.
- Fourier ActionNet Team and Yao Mu. Actionnet: A dataset for dexterous bimanual manipulation. 2025.
- Galaxea AI. Galaxea open-world dataset and G0 dual-system VLA model. *arXiv preprint arXiv:2509.00576*, 2025.
- Zelin Gao, Qiuyu Wang, Yanhong Zeng, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

-
- Google DeepMind. Veo 3. <https://deepmind.google/technologies/veo/veo-3/>, 2025. URL <https://deepmind.google/technologies/veo/veo-3/>.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. RoomTour3D: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- Byeongho Heo, Song Park, Dongyoon Han, and Sangdoon Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.
- Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RL-Bench: The robot learning benchmark. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.
- Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Li. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS Datasets and Benchmarks Track*, 2024.
- Alexander Khazatsky, Karl Pertsch, Suraj Nair, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.
- Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5:341–353, 2023.
- Kuaishou Technology. Kling: A progressive framework for video generation. <https://klingai.com>, 2024. URL <https://klingai.com>.
- Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2025. URL <https://arxiv.org/abs/2502.20694>.
- Xuanlin Li, Kyle Hsu, Jiayuan Liu, Ken Goldberg, and Sergey Levine. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning (CoRL)*, 2024.
- Lightricks. LTX-Video: Realtime video latent diffusion. <https://github.com/Lightricks/LTX-Video>, 2025. URL <https://github.com/Lightricks/LTX-Video>.
- Sihao Lin, Zerui Li, Xunyi Zhao, Gengze Zhou, Liuyi Wang, Rong Wei, Rui Tang, Juncheng Li, Hanqing Wang, Jiangmiao Pang, Anton van den Hengel, Jiajun Liu, and Qi Wu. VLNVerse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation. *arXiv preprint arXiv:2512.19021*, 2025.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.
- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- NVIDIA. NVIDIA Isaac Sim. <https://developer.nvidia.com/isaac-sim>, 2022.
- NVIDIA. PBench: A physical ai benchmark for world models, 2025a. URL <https://huggingface.co/datasets/nvidia/PBench>.
- NVIDIA. Nvidia physicalai autonomous driving dataset. 2025b. <https://developer.nvidia.com/physicalai>.

-
- OpenAI. Sora: Creating video from text. <https://openai.com/sora>, 2024. URL <https://openai.com/sora>.
- OpenLoong Baihu Team. OpenLoongData-v1.0. <https://www.openloong.org.cn/en/datasets/baihu>, 2025.
- Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *International Conference on Learning Representations (ICLR)*, 2025.
- 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 *ICML*, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.
- Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6), 2017.
- Sekai Team. Sekai: Real-world egocentric walking videos for world model training. 2025. <https://huggingface.co/datasets/sekai>.
- Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. RoboScape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. Gigaworld-0: World models as data engine to empower embodied ai, 2025. URL <https://arxiv.org/abs/2511.19861>.
- Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. InternData-A1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.
- 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.
- 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.
- Wan Team. Wan: Open and advanced large-scale video generative models. <https://wanxai.com>, 2025. URL <https://wanxai.com>.
- Waymo Team. Waymo open dataset: End-to-end driving. 2024. <https://waymo.com/open/>.
- Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, et al. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems (RSS)*, 2025a.
- Shihan Wu et al. RoboCOIN: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025b.

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>.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models, 2025. URL <https://arxiv.org/abs/2505.09694>.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4D embodied world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Joel Zhou, Jordan Juravsky, Sanja Fidler, Umar Bhatt, and Nima Fazeli. DreamGen: Unlocking generalization in robot learning through neural trajectories. *arXiv preprint arXiv:2505.12705*, 2025.