



Xiaomi-Robotics-U0: Unified Embodied Synthesis with World Foundation Model

Xiaomi Robotics¹

Abstract

Recent foundation image and video generation models offer strong generalization and controllability, but their direct application to embodied scenarios is limited by requirements for multi-view consistency, geometric coherence, and robot embodiment constraints. Existing methods typically adapt foundation models with limited robot data, often sacrificing visual knowledge acquired during large-scale pre-training.

We present **Xiaomi-Robotics-U0**, a 38-billion-parameter multimodal autoregressive model for unified embodied synthesis. It treats embodied generation as an extension of foundation image and video generation and jointly optimizes text-to-image generation, image editing, embodied scene generation, embodied transfer, and embodied video generation. This unified framework preserves the generalization of the pre-trained world foundation model while adapting it to embodied settings. **Xiaomi-Robotics-U0** is the first model to support high-quality multi-view scene generation across multiple robot embodiments and to introduce structured, controllable embodied transfer for fine-grained editing while preserving multi-view consistency and interaction dynamics. It achieves state-of-the-art results on single-step and sequential generation tasks, outperforming GPT-Image-2.0 in human evaluations of embodied scene generation and transfer, ranking first on World Arena for embodied video generation, and improving the out-of-distribution success rate of $\pi_{0.5}$ from 36.9% to 63.2% on challenging real-world manipulation tasks. These results show that foundation world models can serve both as embodied world models and scalable data engines for embodied intelligence. Code and checkpoints are available at <https://robotics.xiaomi.com/xiaomi-robotics-u0.html>.

1 Introduction

Recent foundation image and video generation models [4, 13, 16, 27, 42, 45, 55, 60, 64] have made remarkable progress in semantic understanding, controllable generation, and visual reasoning through training with data on the Internet. Large-scale generative models are now capable of synthesizing highly realistic images and videos from various multimodal inputs, demonstrating impressive generalization far beyond the distribution of their training data. Such capabilities make foundation generative models an attractive starting point for embodied intelligence [34, 50, 66, 70], where robots are required to reason about complex environments and imagine future interactions before acting.

However, embodied generation [30, 32, 36] introduces challenges that differ fundamentally from conventional image and video synthesis. Unlike natural image generation, embodied scenarios require strict multi-view consistency, accurate geometric and physical coherence across cameras, explicit robot embodiment constraints,

¹See Contributions section for full author list. Please send correspondence to mi-robotics@xiaomi.com.



Figure 1 Embodied and general capabilities of **Xiaomi-Robotics-UO**. The rectangle corresponds to the initial observations for the same embodiment, the pairwise transfer sample, and the keyframes within a video for the embodied capabilities. All frames are referenced and generated images.

and temporally consistent interaction dynamics. The generated observations must remain compatible with robot kinematics, camera calibration, and downstream manipulation policies rather than merely appearing visually realistic. Consequently, directly applying existing foundation image or video generation models to embodied scenarios often leads to inconsistent geometry, implausible robot states, and poor compatibility with robot control.

Recent embodied world models [1, 50, 72] attempt to bridge this gap by continually adapting pre-trained foundation models using robot trajectories [5, 15, 24, 57, 58, 69] or egocentric manipulation videos [2, 21]. While these approaches have demonstrated encouraging progress, their adaptation paradigm remains fundamentally different from that of foundation model pre-training. Instead of preserving the diverse image and video generation tasks that originally endow foundation models with strong semantic understanding and controllable generation, post-training is typically performed solely on robot-specific datasets. Since embodied datasets are considerably smaller, less diverse and highly repetitive compared with internet-scale visual corpora [18, 38, 48], such adaptation inevitably weakens the generalization capability inherited from foundation models and limits the diversity, controllability, and scalability of the embodied generation.

In this work, we present **Xiaomi-Robotics-UO**, a Unified Embodied Synthesis Model that unifies foundation generation and embodied generation within a single training paradigm. As shown in Figure 1, rather than specializing a foundation model into a robot-specific generator, **Xiaomi-Robotics-UO** reformulates embodied synthesis as a natural extension of foundation image and video generation. Starting from a world foundation model, **Xiaomi-Robotics-UO** is continually trained on both general-domain and embodied datasets under a unified autoregressive objective, jointly learning text-to-image generation, image editing, embodied scene generation, embodied transfer, and embodied video generation. Figure 2 exhibits the above-mentioned capabilities of **Xiaomi-Robotics-UO**. This unified formulation preserves the rich semantic knowledge and controllable generation capability of foundation models while effectively adapting them to robot-centric multi-view reasoning and interaction modeling.

Beyond single-step generation, **Xiaomi-Robotics-UO** further extends the embodied synthesis to sequential world modeling. By jointly learning interleaved embodied sequences and manipulation videos across multiple temporal resolutions, the model captures both long-horizon task progression and fine-grained interaction

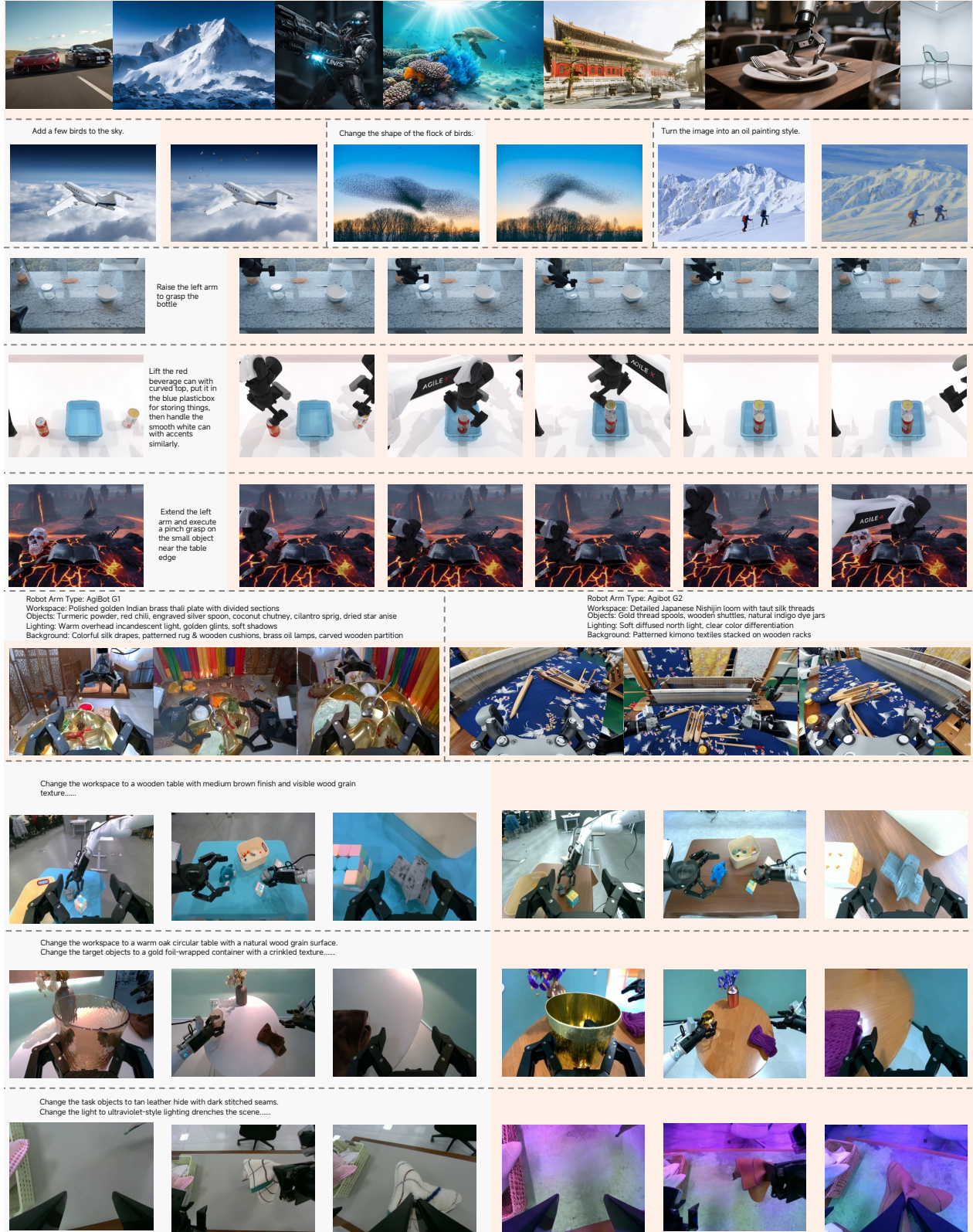


Figure 2 Overview of all tasks in Xiaomi-Robotics-UO, Gray: inputs; Orange: outputs.

dynamics. More importantly, this enables embodied generation to evolve from static scene synthesis into a scalable trajectory generation engine: generated embodied scenes can be directly rolled out into temporally coherent manipulation videos, providing abundant synthetic trajectories for downstream policy learning. We believe that this capability represents an important step toward agentic embodied data generation, where foundation world models continuously generate increasingly diverse robot experiences through autonomous interaction rollout.

Extensive experiments demonstrate that **Xiaomi-Robotics-U0** establishes a new state of the art across both single-step and sequential embodied generation tasks. It is the first unified embodied synthesis model capable of high-quality multi-view embodied scene generation, structured controllable embodied transfer, and zero-shot embodied video generation within a single autoregressive framework. **Xiaomi-Robotics-U0** outperforms GPT-Image-2.0 in human evaluation on embodied scene generation and embodied transfer, achieves state-of-the-art performance on the World Arena benchmark, and significantly improves the out-of-distribution robustness of downstream robot policies using generated data. These results suggest that world foundation models can serve not only as embodied world models, but also as scalable data engines for embodied intelligence.

Our contributions are summarized as follows:

- We introduce **Xiaomi-Robotics-U0**, the first unified embodied synthesis model that seamlessly integrates foundation image generation, image editing, embodied scene generation, embodied transfer, and embodied video generation within a single autoregressive framework. By jointly optimizing general-domain and embodied generation tasks under a unified continual training paradigm, **Xiaomi-Robotics-U0** effectively transfers the generalization capability of foundation world models to embodied generation while preserving their original visual generation ability.
- We propose the first unified framework for **multi-view embodied generation**, including embodied scene generation and controllable embodied transfer. By jointly training with large-scale text-to-image and image editing data, **Xiaomi-Robotics-U0** achieves state-of-the-art performance with superior multi-view consistency, fine-grained controllability, and instruction following, outperforming GPT-Image-2.0 in human evaluations. Furthermore, we introduce a structured control formulation that disentangles workspace, background, foreground irrelevant objects, target objects, and lighting into independent control dimensions, enabling scalable embodied video augmentation while preserving geometric consistency and interaction dynamics. The generated data substantially improves the robustness of downstream robot policies on challenging real-world manipulation tasks.
- We extend embodied generation from static scene synthesis to **sequential embodied world modeling** through interleaved subtask-subgoal learning and multi-FPS embodied video generation. **Xiaomi-Robotics-U0** achieves state-of-the-art performance on the World Arena benchmark and demonstrates strong zero-shot multi-view embodied video generation capability. By enabling coherent long-horizon rollout from synthesized embodied scenes, **Xiaomi-Robotics-U0** transforms foundation world models into scalable embodied data engines, providing a practical path toward agentic generation of diverse embodied interaction trajectories for future policy learning.

2 Xiaomi-Robotics-U0

2.1 Problem Formulation

Our goal is to develop a unified embodied world model that supports diverse generation and prediction tasks under a single autoregressive framework. Given a multi-modal context

$$\mathcal{C} = \{x_1, x_2, \dots, x_n\}, \quad (1)$$

where each element can be a text token, an image token, or a robot-related control token, the model learns the joint distribution over the subsequent multi-modal sequence

$$P(\mathcal{Y}|\mathcal{C}) = \prod_{t=1}^T P(y_t|y_{<t}, \mathcal{C}), \quad (2)$$

using standard next-token prediction (NTP).

To support embodied intelligence, **Xiaomi-Robotics-U0** is trained on several complementary data formats with different temporal structures. Specifically, we categorize all training samples into two groups.

Single-step samples describe a one-shot mapping between inputs and outputs without explicit temporal evolution. These samples include: (i) Text-to-Image (T2I), where the model generates an image conditioned on textual descriptions; (ii) Any-to-Image (X2I), where one to three reference images together with text are used for image editing and generation; (iii) Embodied Scene Generation, where the model generates the initial multi-view robot observations conditioned on robot embodiment and scene descriptions; and (iv) Embodied Transfer, where the model predicts transferred multi-view observations given the current embodied observations and a target scene description.

Sequential samples model long-horizon interactions and future evolution. We consider two representative formats: (i) image-text interleaved subtask-subgoal sequences, where each subtask description is followed by the corresponding multi-view observations after completion; and (ii) embodied manipulation videos collected at multiple frame rates, enabling the model to capture both sparse task-level transitions and dense interaction dynamics.

Despite their different semantics, all tasks are represented as unified multi-modal token sequences and optimized under the same autoregressive objective, allowing knowledge to be naturally shared across image generation, scene understanding, embodied reasoning, and future prediction.

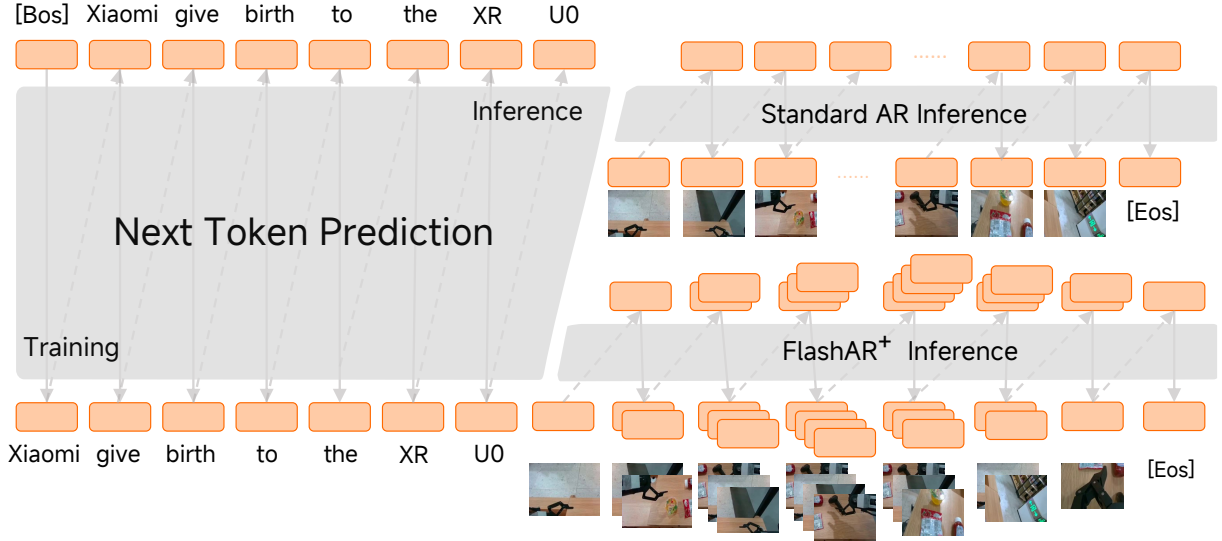


Figure 3 Overview structure of **Xiaomi-Robotics-U0**. The model is trained end-to-end at scale with a unified next-token prediction objective. An efficient inference extension is provided along.

2.2 Model Architecture

Figure 3 illustrates the overall architecture of **Xiaomi-Robotics-U0**. We adopt the open-sourced EMU3.5 [16] as the initialization model, which is built upon the Qwen-3-32B decoder-only Transformer and has demonstrated strong capabilities in image generation and image-text interleaved modeling. Images are first tokenized by the IBQ Tokenizer [51] with a spatial compression ratio of 16×16 . By extending the original Qwen vocabulary with the IBQ codebook, all modalities are represented within a unified discrete vocabulary and modeled using next-token prediction.

Compared with conventional embodied world models that introduce task-specific prediction heads, **Xiaomi-Robotics-U0** formulates all generation tasks—including image generation, embodied scene synthesis, scene transfer, sequential subtask prediction, and video generation—as a single autoregressive sequence

modeling problem. This unified formulation enables efficient multi-task co-training while preserving the strong visual generation capabilities inherited from the large-scale pre-trained model.

To improve inference efficiency, we further adopt an inference-accelerating adaptation inspired by FlashAR [73]. Specifically, an additional vertical prediction head is introduced to decode image tokens in an anti-diagonal order, allowing multiple visual tokens to be generated simultaneously. Combined with optimized KV-cache management and continuous batching provided by vLLM [26], **Xiaomi-Robotics-U0** achieves up to $82.9\times$ faster image generation at 1024×1024 resolution compared with the original NTP pattern, while preserving most generation quality.

2.3 Dataset

2.3.1 Dataset Curation

Data Collection. To endow **Xiaomi-Robotics-U0** with broad visual understanding and generation capabilities, we assemble a diverse training corpus spanning six distinct domains, each contributing complementary visual and semantic priors.

- **General image-text** data provides foundational visual-semantic alignment. We include ShareGPT4V [9] and similar large-scale vision-language benchmarks to establish a strong baseline for text-conditioned generation.
- **Embodied manipulation** data bridges the gap between passive visual understanding and active physical interaction. This domain combines both real-world and simulated sources. On the real side, we leverage AgiBotWorld-Beta [8], Open X-Embodiment [15], and a proprietary multi-view robot manipulation dataset collected in house, which we refer to as MiBot. On the simulation side, we incorporate data from RoboTwin2.0 [11], GenieSim [69], and InternData-A1 [57], enabling coverage of scenarios that are difficult or dangerous to collect physically.
- **Autonomous driving** data captures structured outdoor scenes with complex spatial layouts and dynamic objects. We source this from Cosmos-Drive-Dreams [47], which provides high-quality driving scenarios with diverse weather, lighting, and traffic conditions.
- **Egocentric** data, drawn from EgoWalk [2], supplies first-person perspective recordings that reflect natural human viewpoint and motion patterns, enriching the model’s understanding of human-centric visual experience.
- **3D reconstruction** data offers dense multi-view correspondences and accurate geometry. We utilize ScanNet++ v2 [68] and DL3DV-10K [31], both of which provide high-quality indoor and outdoor 3D scene captures.
- **Game** data, sourced from Open-P2P [71], introduces stylized and procedurally generated environments that expand the diversity of visual distributions beyond real-world captures.

Data Filtering. Raw data collected at scale inevitably contains artifacts that can impair generation quality. Inspired by Abot Physical World [14], we implement an autonomous data filtering pipeline including motion blur detection, image quality detection, and gripper-state alignment.

To sum up, we collect 9.5 million samples for single-step generation, which corresponds to 56.4 billion tokens. For sequential generation, we select 2.6 million video clips, which corresponds to 49.6 billion tokens.

2.3.2 Dataset Labeling

Unified Annotation Pipeline. Most raw samples lack fine-grained text annotations and only carry coarse task-level descriptions, which cannot support precise text-driven embodied generation. We thus construct a unified annotation pipeline to standardize supervision signals across heterogeneous data.

Unified Annotation Pipeline. We leverage Qwen3-VL-235B [3], a powerful vision-language model (VLM), to generate dense, structured annotations for generic frames, multi-source video sequences, and robotic manipulation trajectories. The pipeline consists of four decoupled branches with distinct annotation objectives:

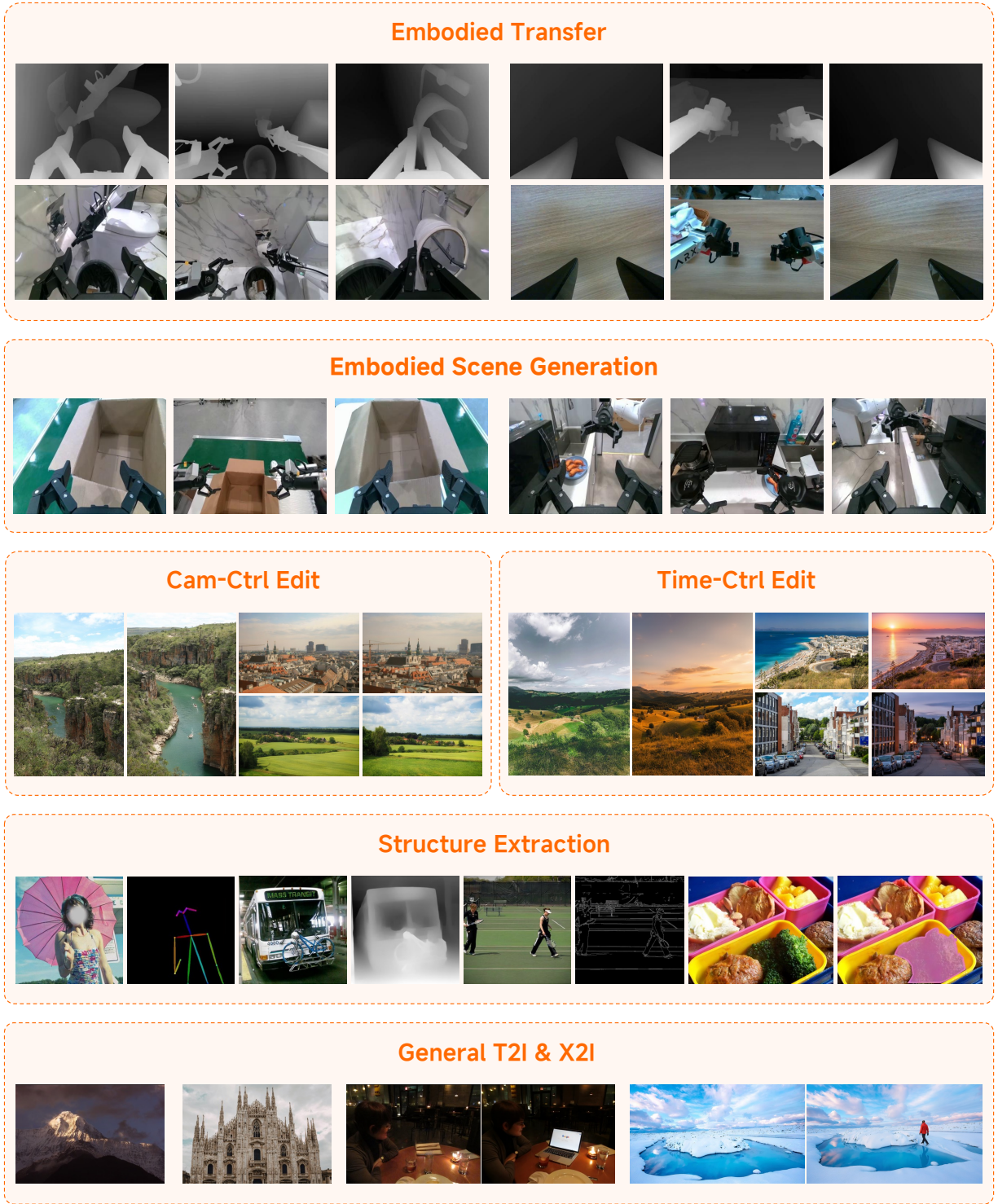


Figure 4 The illustration of the constructed single-step generation dataset. Except for the embodied transfer and the embodied scene generation data, we additionally introduce Text-to-Image(T2I) and Any-to-Image(X2I) data for co-training. Concretely speaking, we add camera-control and time-control data to improve the spatial and temporal understanding capability. For structure and geometry perception, we construct skeleton, depth, edge, and semantic mask datasets with advanced computer vision tools. To maintain the ability of the pre-trained model, we introduce general T2I and X2I datasets for co-training.

- **General Scene Annotation** All video data spanning general visual corpora and embodied robot sequences are processed without domain restrictions. Each video clip undergoes temporal downsampling to select high-clarity keyframes. The VLM produces holistic captions describing overall scene layouts, object interactions, and cross-object spatial relationships for every sampled frame.
- **Structured Embodied Scene Annotation** Annotation is only performed on embodied manipulation data. Decomposed subtask outputs from trajectory processing guide scene partitioning into five orthogonal semantic dimensions: workspace, task objects, irrelevant objects, lighting, and background. Separate descriptions are generated for each dimension according to task relevance, supporting compositional reasoning and fine-grained controllable generation during inference.
- **Geometric Modality Extraction** Temporally consistent inverse depth maps are extracted for all frames and videos via Video Depth Anything [10] to supply geometric conditioning signals for cross-domain embodied transfer tasks. Brighter pixels correspond to closer surfaces and darker pixels to distant regions within this depth representation, allowing the model to synthesize photorealistic RGB outputs conditioned on geometric priors.
- **Trajectory-level Subtask Decomposition** As illustrated in Figure 5, HDBSCAN clustering [37] segments robot trajectories using end-effector poses and gripper signals to identify motion and grasp transition boundaries. Keyframes with dense physical interactions are sampled at these boundaries and sent to VLM to generate temporally aligned subtask plans with frame boundaries. Transition frames with rich physical interactions are processed with scene annotation to produce training samples for embodied scene generation and embodied transfer tasks.

Task-Oriented Data Organization. To support the three training objectives of *Xiaomi-Robotics-U0*, we reorganize all annotated data into task-specific formats:

- **Embodied Transfer.** Each training sample consists of a system instruction, a scene description, and multi-view depth maps as input, with multi-view RGB images as the supervision target. This task trains the model to hallucinate photorealistic appearance from geometric cues.
- **Scene Generation.** Each sample pairs a system instruction with a scene description, supervised by multi-view RGB images. This task learns text-to-multi-view-scene synthesis without geometric priors.
- **Video Generation.** Each sample comprises a system instruction, a task instruction, and an initial observation frame, supervised by a sequence of 15–25 RGB frames. For datasets that provide robot action labels, a sequence of robot mask images rendered from the robot action and URDF is additionally appended after the initial observation frame, serving as action guidance. This task learns to predict future visual states conditioned on task semantics and initial context.

Data Balancing. Our training corpus contains heterogeneous multi-domain data, which easily leads to domain bias and overfitting toward dominant data distributions. We adopt sample-level reweighting conditioned on data domains and generation task types (multi-view generation, scene editing, single-view generation), with the core target of maintaining full coverage of robotic manipulation scenarios.

Specifically, we first eliminate invalid embodied episodes, including sequences with corrupted visuals, misaligned timestamps, missing proprioceptive or action signals, failed execution records, and static trajectories lacking robot-object interactions. All remaining samples are categorized via hierarchical annotations derived from subtask and structured scene labeling. Grouping criteria cover task goals, scene contexts, target objects, spatial relations, subtask phases, and atomic manipulation skills.

Such fine-grained grouping supports more precise balancing than simple dataset-level weighting: redundant samples sharing identical embodiment-task-object combinations are downsampled, whereas clips with novel visuals, object layouts, manipulation skills or temporal stages are prioritized. Higher sampling weights are assigned to long-tail categories, including rare objects, seldom-used manipulation skills, and core interaction phases (grasping, releasing, placing, and object state transitions). The resulting training distribution eliminates redundant repeated demonstrations while retaining diverse visual scenes and comprehensive manipulation skill coverage.

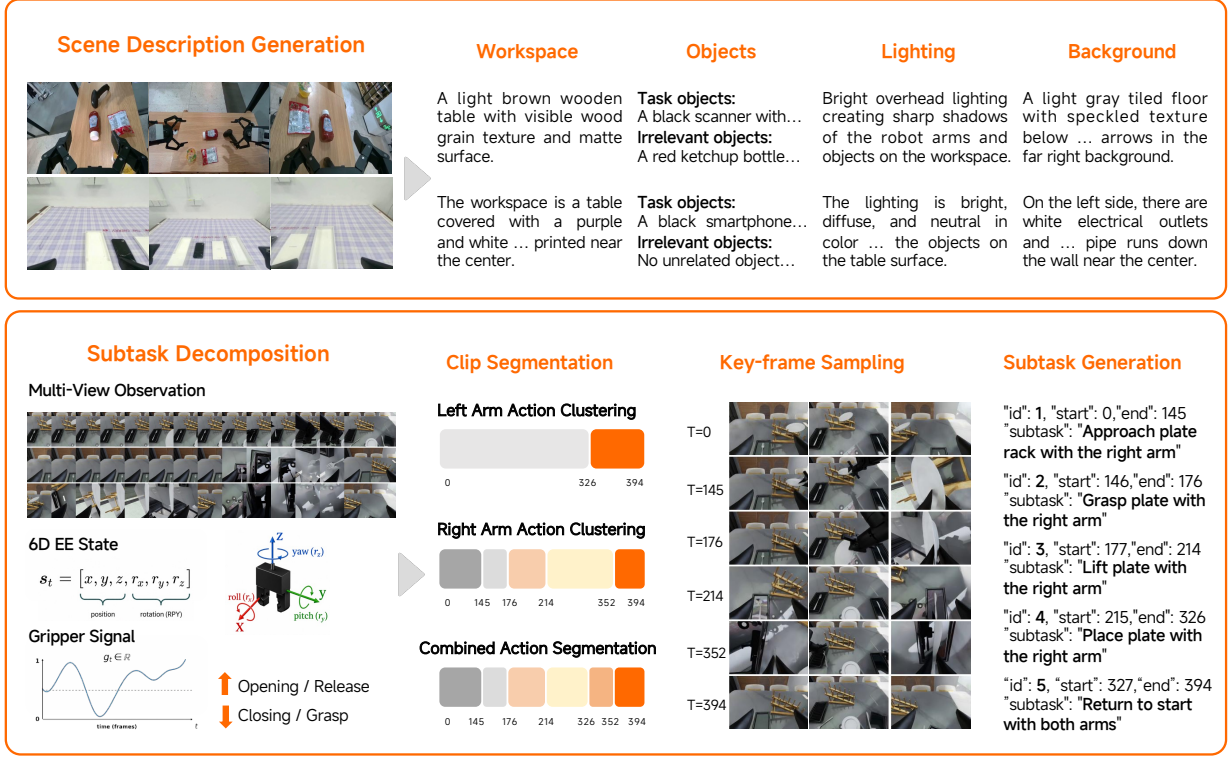


Figure 5 Overview of the unified data labeling pipeline. The upper part illustrates scene description generation, where each sample is annotated with structured semantic attributes. The lower part shows trajectory-level subtask decomposition, including action-based clip segmentation, key-frame sampling, and subtask generation. This pipeline provides consistent scene, geometry, and action-level supervision across heterogeneous robotic datasets.

2.4 Training

To efficiently adapt the pre-trained image generation model into a unified embodied world model while preserving its original visual generation capability, we adapt two continued training setups: single-step training and sequential training.

Single-step training. This setup focuses on aligning the model with diverse one-step generation tasks while maintaining the original image generation capability. We jointly train four complementary datasets, including Text-to-Image (T2I), Any-to-Image (X2I), Embodied Scene Generation, and Embodied Transfer.

The T2I and X2I datasets continuously reinforce the general visual generation and image editing capabilities inherited from the pre-trained model, preventing catastrophic forgetting during embodied adaptation. Meanwhile, Embodied Scene Generation teaches the model to synthesize physically plausible initial robot observations conditioned on robot embodiments and scene descriptions, while Embodied Transfer enables cross-scene multi-view generation that preserves robot interaction states and camera geometry. Co-training these tasks effectively aligns the visual generation prior with embodied scene understanding under a unified autoregressive objective.

Sequential training. The sequential training extends the model to long-horizon embodied reasoning and interaction modeling.

We first introduce image-text interleaved subtask-subgoal sequences, where each language-described subtask is paired with the corresponding multi-view observations after execution. This data format teaches the model to reason over task decomposition while predicting future embodied observations, further aligning the sequence modeling capability inherited from the base model.

To enhance dynamic interaction modeling, we additionally incorporate large-scale embodied manipulation videos collected from InternData-A1 [57], GenieSim [69], AgiBotWorld-Beta [8], RoboCOIN [62], and RoboMIND [61]. Unlike conventional low-frame-rate trajectory datasets, we construct multi-frame-rate video sequences with FPS values of 1, 3, and 5. Sparse sequences encourage long-horizon planning and future prediction, whereas denser sequences enable the model to capture fine-grained manipulation dynamics and continuous interaction processes. Training across multiple temporal resolutions significantly improves the model’s ability to model both high-level task progression and low-level embodied dynamics within a unified autoregressive framework.

2.5 Inference

To further accelerate inference, we introduce a FlashAR⁺ extension to **Xiaomi-Robotics-U0**. The extension keeps the conditional modeling of reference images and other input signals unchanged, while applying parallel autoregressive generation only to the final target image region. This design is intended to preserve the multi-source conditioning capability of **Xiaomi-Robotics-U0** and reduce the serial decoding cost during X2I inference.

2.5.1 FlashAR⁺ Extension

FlashAR baseline. FlashAR [73] proposes a post-training and inference acceleration paradigm for autoregressive text-to-image generation. Its core idea is to preserve the generation capability of the original model while introducing extra vertical prediction, so that image tokens can be generated in parallel along anti-diagonal groups. This reduces the number of serial decoding steps in the image region. Inspired by FlashAR, we perform FlashAR⁺ post-training on top of **Xiaomi-Robotics-U0**, extending it from single-image T2I generation to broader X2I scenarios, and use the resulting model for X2I inference acceleration in **Xiaomi-Robotics-U0**. Figure 6a demonstrates the decoding steps of the FlashAR⁺ extension.

In **Xiaomi-Robotics-U0**, X2I covers tasks such as T2I, I2I, and multi-reference image generation. The input conditions may include text instructions and one or more reference images.

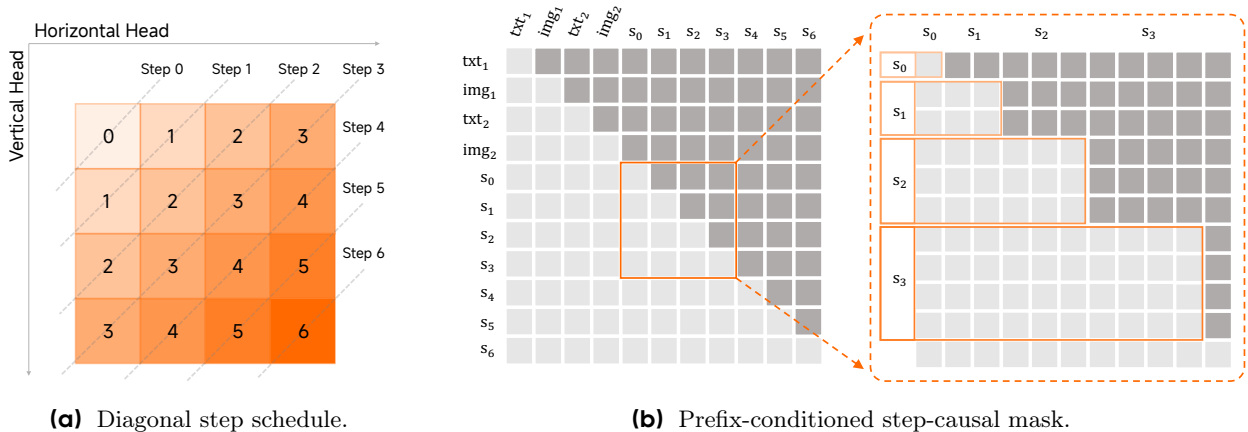


Figure 6 Xiaomi-Robotics-U0 FlashAR⁺ extension for X2I generation. Target image tokens are grouped by diagonal steps, while all input conditions are kept in the prefix and the target image is generated with a step-causal attention mask.

FlashAR⁺ mask and attention. Compared with standard autoregressive generation, **Xiaomi-Robotics-U0** FlashAR⁺ Extension introduces the key changes summarized in Table 1.

During training, the FlashAR⁺ extension identifies image blocks during data parsing and defines the target image. All preceding content, including text instructions, reference images, control conditions, and the target image header, is treated as the conditional prefix. Parallel autoregressive losses are computed only over the target image region, while reference images serve only as a conditional context. This prevents reference images

Table 1 Comparison between standard autoregressive generation and the **Xiaomi-Robotics-U0 FlashAR⁺** Extension.

Aspect	Standard AR	Xiaomi-Robotics-U0 FlashAR⁺ Extension	Effect
Target image generation	Serial token-by-token generation in raster order	Parallel generation by anti-diagonal steps in the target image	Reduces serial decoding steps
Target image attention	Standard one-dimensional causal attention	The conditional prefix is visible; the target image uses two-dimensional step-causal attention internally	Prevents information leakage
Prediction heads	A single AR prediction head	Horizontal head, vertical head, and gated fusion	Supports anti-diagonal generation
Directional auxiliary loss	Standard supervision	Introduces H/V loss constraints	Constrains both directional heads

from being incorrectly supervised as generation targets and ensures that the model learns to generate the target image from multi-source conditions.

The attention constraint is also restricted to the target image region, as shown in figure 6b. The conditional prefix preserves standard autoregressive visibility. Target image tokens can attend to the entire conditional prefix, but within the target image they can only attend to tokens from earlier anti-diagonal steps. For a target grid position (r, c) , its parallel step is defined as $r + c$; tokens within the same step are mutually invisible, which prevents information leakage during parallel generation. Reference images, text tokens, and control tokens are not involved in the anti-diagonal mask construction inside the target image, preserving the semantic integrity of X2I conditions.

This constraint can be expressed as a visibility function within the target image. For a target-image query token $q = (r_q, c_q)$ and key token $k = (r_k, c_k)$,

$$M(q, k) = 1 [s(r_k, c_k) < s(r_q, c_q)].$$

FlashAR⁺ loss. **Xiaomi-Robotics-U0 FlashAR⁺** Extension introduces a key adjustment to directional auxiliary supervision. We match the horizontal auxiliary loss weight to the vertical auxiliary loss weight:

$$\mathcal{L} = \mathcal{L}_{\text{fuse}} + 0.05\mathcal{L}_h + 0.05\mathcal{L}_v + 0.2\mathcal{L}_{\text{distill}}.$$

Here, $\mathcal{L}_{\text{fuse}}$ is the target-image prediction loss after H/V gated fusion, $\mathcal{L}_h$ and $\mathcal{L}_v$ supervise the horizontal and vertical heads respectively, and $\mathcal{L}_{\text{distill}}$ is the AR distillation term. For a target position (r, c) , let $z_h(r, c)$ and $z_v(r, c)$ denote the logits from the horizontal and vertical heads, and let $g(r, c) \in [0, 1]$ denote the gate value. The fused logits are

$$z_{\text{fuse}}(r, c) = g(r, c)z_h(r, c) + (1 - g(r, c))z_v(r, c).$$

The corresponding fused loss is

$$\mathcal{L}_{\text{fuse}} = - \sum_{(r, c) \in \mathbf{Y}} \log \text{softmax}(z_{\text{fuse}}(r, c))_{y_{r, c}}.$$

The directional auxiliary losses $\mathcal{L}_h$ and $\mathcal{L}_v$ use the same target tokens to supervise z_h and z_v , respectively. The H head also needs to adapt to the target-region distribution. If $\mathcal{L}_h = 0$, the H head may fail to converge



Figure 7 Qualitative and efficiency comparison between AR, FlashAR⁺, and FlashAR⁺ with vLLM. FlashAR⁺ preserves high visual quality on Embodied Transfer and Embodied Scene Generation while substantially reducing generation latency.

sufficiently, and the H/V gate may collapse toward the H branch in later training stages. When the gate collapses to an under-trained H head, the fused prediction degenerates into an unstable single-direction prediction, and further training can lead to degraded generation quality. Adding $0.05\mathcal{L}_h$ explicitly supervises the H head and stabilizes H/V fusion.

Xiaomi-Robotics-U0 FlashAR⁺ inference. During inference, Xiaomi-Robotics-U0 first encodes the text instruction and reference images as a conditional prefix, followed by the target image header. The model then performs anti-diagonal generation only on the target image grid: positions are generated step by step in increasing order of $r + c$, and all positions within the same step are sampled in parallel. Boundary positions use the available single-direction prediction, while interior positions use H/V gated fusion.

This procedure preserves the constraints imposed by multi-source conditions while extending image-region parallel decoding to Xiaomi-Robotics-U0 FlashAR Extension inference. As a result, the model maintains Any-to-Image conditional generation capability while reducing the serial generation cost in the target image region.

2.5.2 vLLM Integration

To further improve the image-generation efficiency of Xiaomi-Robotics-U0, we integrate vLLM on top of FlashAR⁺. The implementation keeps the FlashAR⁺ decoding rule unchanged: vLLM handles the conditional prefix, batched execution, and paged KV-cache management, while a diagonal scheduler emits the visual tokens of each anti-diagonal step and synchronizes the paired conditional and unconditional requests used for

classifier-free guidance.

As shown in Figure 7, the accelerated path preserves high generation quality on embodied tasks. On T2I generation at 1024×1024 resolution with a single H20 GPU, standard AR takes 450.77 seconds per image. FlashAR⁺ reduces this latency to 16.56 seconds per image, achieving a $27.22\times$ speedup over AR, and FlashAR⁺ with vLLM further reduces the latency to 5.44 seconds per image with `max_num_seq=28`, achieving an additional $3.04\times$ speedup over FlashAR⁺.

3 Experiments

Our model unifies embodied transfer, embodied scene generation, and embodied video generation while retaining general image editing and generation capabilities. This section presents benchmark and qualitative results for our embodied transfer and scene generation tasks, which outperform GPT-Image-2.0 in human evaluation. Specifically, our embodied transfer method substantially enhances robot trajectory diversity while preserving geometric consistency, boosting the success rate of $\pi_{0.5}$ from 36.9% to 63.2%. We further validate the model’s general image generation capacity via evaluations on the Geneval and ImageEdit benchmarks, confirming that our method retains dominant general image generation performance. For embodied video generation, our model achieves state-of-the-art results on the World Arena benchmark. It also demonstrates strong generalization to self-synthesized initial scenes, enabling a closed-loop data engine for infinite video generation conditioned on arbitrary scene descriptions and robotic instructions.

3.1 Embodied Transfer

Benchmark and protocol. To evaluate embodied transfer, we construct a benchmark of 300 samples, each consisting of a textual scene description together with a set of multi-view depth maps; the model is required to synthesize multi-view-consistent RGB images of the described robot manipulation scene. Each sample additionally provides the multi-view RGB reference images from the original dataset. These references do not fully correspond to the (edited) scene description, but they capture the structural information beyond depth and are used to compute the benchmark metrics. Since the depth maps are drawn from the same domain as the training set, we deliberately edit a subset of the scene descriptions so that they call for content not present in the training data, thereby probing the model’s general-purpose generation ability rather than mere reconstruction. According to how much the edited description deviates from the original scene, we split the benchmark into 150 Easy samples (descriptions close to the underlying scene) and 150 Hard samples (descriptions requiring substantial novel content). We compare against GPT-Image-2, a strong image generation model.

Metrics. We assess generation quality from three complementary aspects, all computed against the reference views. For **depth consistency**, we extract monocular depth from both the generated image and the reference with a shared estimator and report SI-RMSE (scale-invariant RMSE, $\downarrow$), δ_1 (the fraction of pixels with relative depth error below 1.25, $\uparrow$), and AbsRel (absolute relative error, $\downarrow$); these measure how faithfully the generated image respects the 3D geometry imposed by the input depth. For **structural fidelity**, we compute Canny edge F1, Precision, and Recall ($\uparrow$), capturing whether the generated layout preserves the fine-grained contours of the scene. For **semantic grounding**, we run open-vocabulary detection and segmentation and report Segmentation mIoU and Recall ($\uparrow$), reflecting whether the described objects are correctly synthesized and localized. As summarized in Table 2, **Xiaomi-Robotics-U0** outperforms GPT-Image-2 by a large margin across all metrics on both the Easy and Hard splits, indicating markedly better adherence to the input geometry and stronger semantic alignment with the scene description.

Qualitative results. Figure 8 presents qualitative examples of embodied transfer. For each example we display the scene description, the original multi-view RGB images from the source dataset, and the images generated by **Xiaomi-Robotics-U0**. Our pipeline first estimates the multi-view depth maps from the original images, then edits the scene description—with the edited spans highlighted—to call for content that departs from the original scene; the extracted depth maps and the edited description are jointly fed to the model for generation. Across diverse scenes, and given a scene description in either natural-language or structured form, **Xiaomi-Robotics-U0** faithfully follows both the edited textual description and the multi-view depth

Table 2 Objective metrics on the embodied transfer benchmark. Values in parentheses denote the change of **Xiaomi-Robotics-U0** relative to **GPT-Image-2** on the same split (signed toward the actual value; a lower value is better for $\downarrow$ metrics and a higher value for $\uparrow$ metrics).

Metric	GPT-Image-2		Xiaomi-Robotics-U0 (Ours)	
	Easy	Hard	Easy	Hard
Depth SI-RMSE ($\downarrow$)	0.4007	0.3444	0.1407 (-0.260)	0.1435 (-0.201)
Depth δ_1 ($\uparrow$)	0.5307	0.5340	0.8017 (+0.271)	0.7846 (+0.251)
Depth AbsRel ($\downarrow$)	0.4350	0.4045	0.1647 (-0.270)	0.1702 (-0.234)
Canny F1 ($\uparrow$)	0.0519	0.0507	0.2066 (+0.155)	0.1815 (+0.131)
Canny Precision ($\uparrow$)	0.0560	0.0536	0.2037 (+0.148)	0.1770 (+0.123)
Canny Recall ($\uparrow$)	0.0572	0.0582	0.2290 (+0.172)	0.2176 (+0.159)
Segmentation mIoU ($\uparrow$)	0.4105	0.4098	0.7621 (+0.352)	0.7436 (+0.334)
Segmentation Recall ($\uparrow$)	0.6208	0.6524	0.8136 (+0.193)	0.8032 (+0.151)

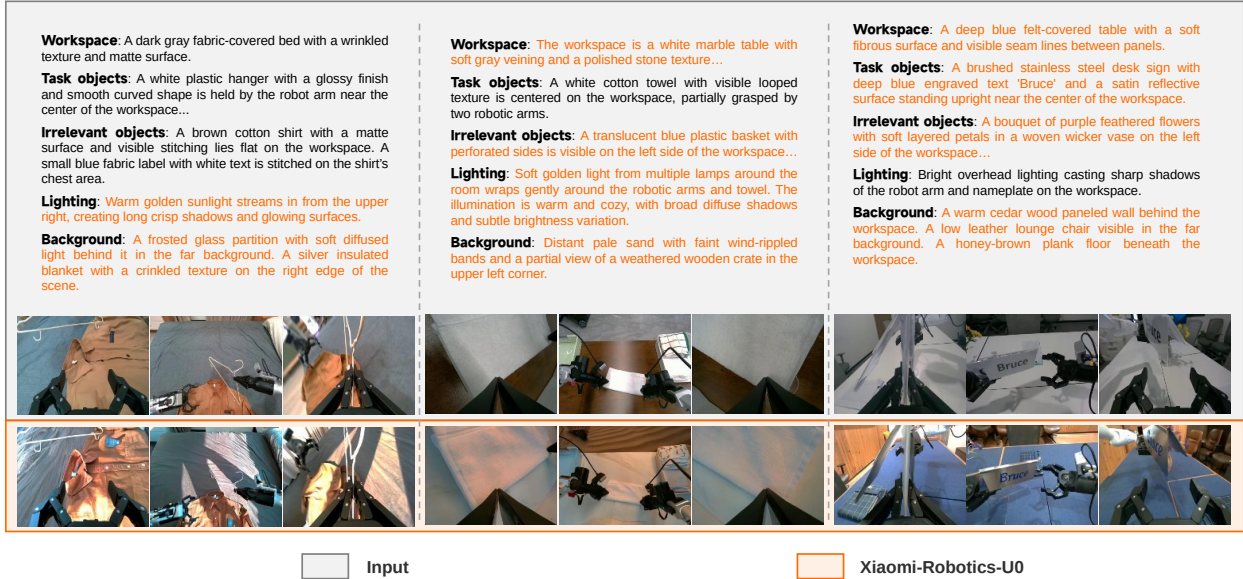


Figure 8 Qualitative results on embodied transfer. For each example we show the scene description, the original multi-view RGB images from which the depth conditions are extracted, and the images generated by **Xiaomi-Robotics-U0**. Starting from the original images, we estimate their multi-view depth maps and edit the scene description (edited text highlighted); the depth maps together with the edited description are then fed to **Xiaomi-Robotics-U0**, which synthesizes multi-view-consistent RGB images that adhere to both the edited text and the input geometry.

conditions, correctly rendering the described objects, lighting, and background while respecting the spatial geometry of each view. Crucially, the generated views remain mutually consistent in object identity, appearance, and arrangement, demonstrating that our unified model produces coherent multi-view scenes rather than independently plausible but conflicting images.

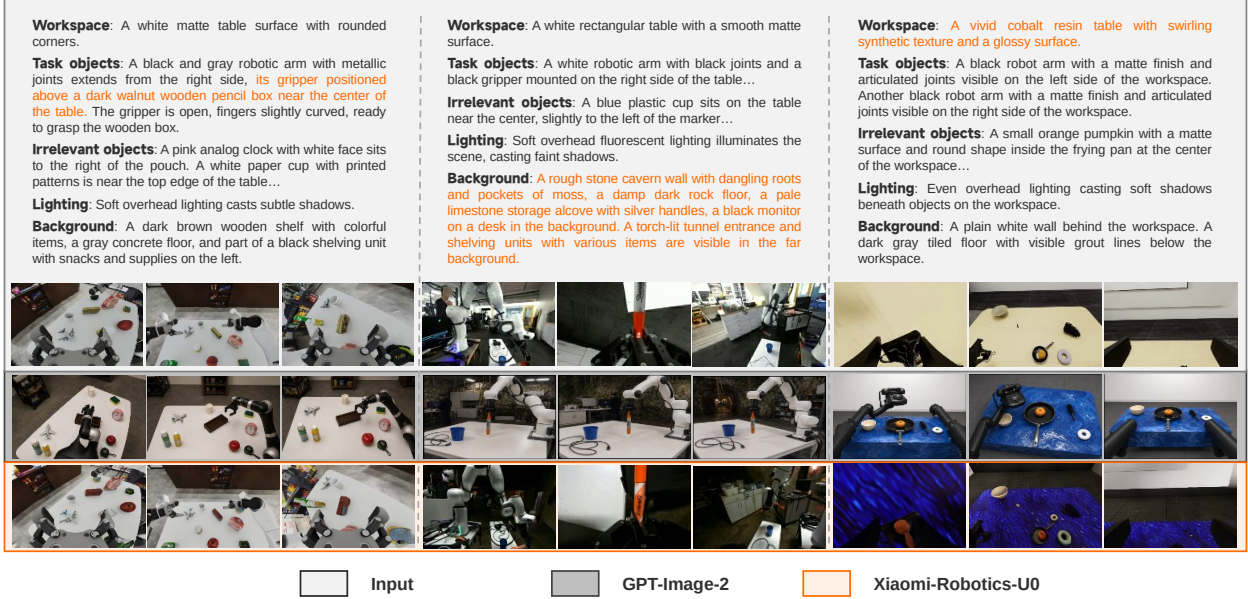


Figure 9 Qualitative comparison between GPT-Image-2 and Xiaomi-Robotics-U0 on the embodied transfer benchmark. GPT-Image-2 generates plausible, well-aligned images but violates the input depth and produces multi-view-inconsistent results, whereas Xiaomi-Robotics-U0 adheres to the geometry of every view and keeps the views mutually consistent.

We further compare Xiaomi-Robotics-U0 with GPT-Image-2 in Figure 9, on three samples from the source datasets, each showing the scene description, the input multi-view depth maps, and the two models’ outputs. GPT-Image-2 produces visually appealing images with strong text-image alignment: the described objects, materials, lighting, and background are all rendered plausibly, and in isolation each view looks convincing. However, it exhibits a poor understanding of the input geometry and of cross-view relationships. Its outputs frequently violate the depth conditions—objects are placed at inconsistent distances, surfaces are warped, and the overall spatial layout drifts away from the imposed geometry—which is reflected in its high SI-RMSE and low δ_1 in Table 2. Moreover, the views generated for the same scene are largely independent: object identity, pose, and arrangement fail to remain consistent across viewpoints, so the three images do not correspond to a single coherent 3D scene. In contrast, Xiaomi-Robotics-U0 faithfully respects the depth of every view while keeping the multi-view outputs mutually consistent, indicating that it not only matches the strong text-following ability of GPT-Image-2 but also genuinely grounds generation in the input geometry.

Zero-shot data augmentation. Beyond benchmark evaluation, we leverage the structured text-to-image capability of Xiaomi-Robotics-U0 to augment the demonstration trajectories collected for four real-world tasks (pack box, fold towel, pack phone, and pack earphone), aiming to endow downstream policies with better generalization to novel scenes. For each trajectory, we prompt a VLM (Qwen3-VL) to generate, conditioned on the current depth maps and the original scene description, 1000 diverse alternatives for every dimension except the task objects, encouraging maximal diversity across background, workspace, and lighting. We then randomly combine these per-dimension descriptions and feed them to Xiaomi-Robotics-U0 for generation. As shown in Figure 10 and Figure 11, the resulting scenes exhibit rich diversity, and none of these compositions appear in the training set, demonstrating that Xiaomi-Robotics-U0 can substantially expand the visual coverage of the collected trajectories in a zero-shot manner.

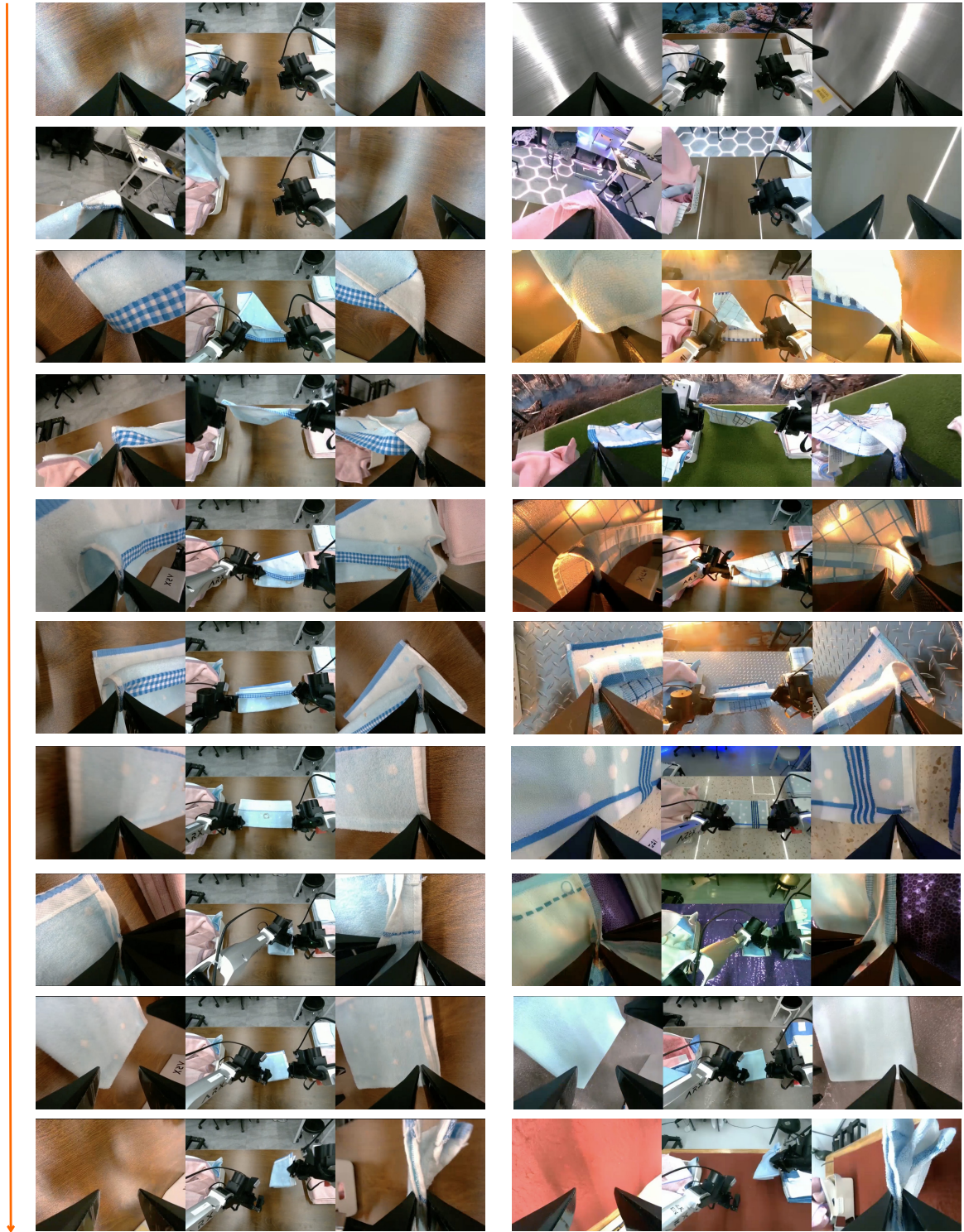


Figure 10 The zero-shot augmented keyframes of the multi-view manipulation trajectories with the embodied transfer capability of *Xiaomi-Robotics-U0*. We generate 1000 descriptions for the background, workspace, and lighting condition with Qwen3-VL and replace the original condition randomly. The left panel is the original keyframes, and the right corresponds to the transferred.

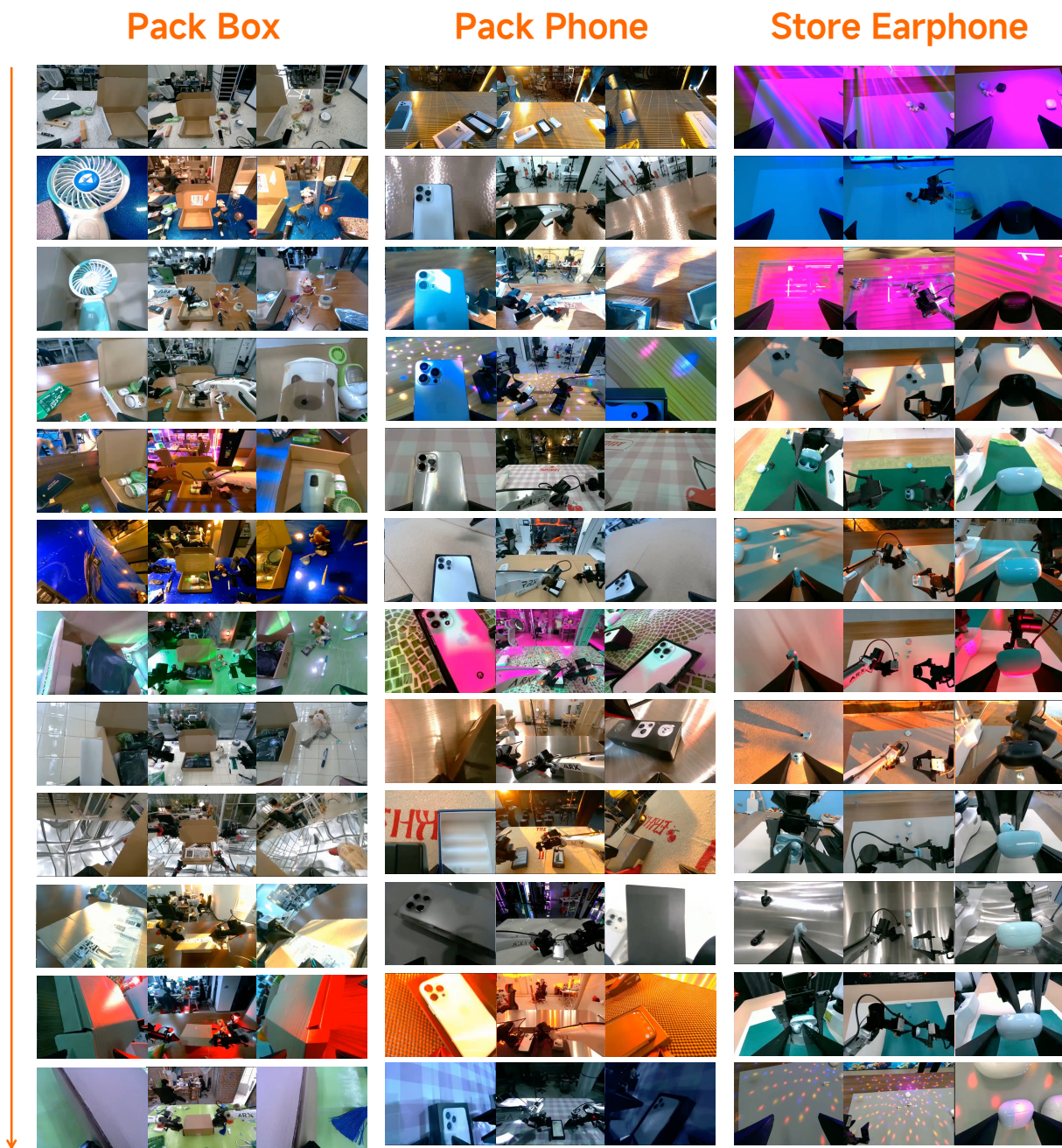


Figure 11 More results of the zero-shot augmented keyframes of the multi-view manipulation trajectories with the embodied transfer capability of Xiaomi-Robotics-U0.

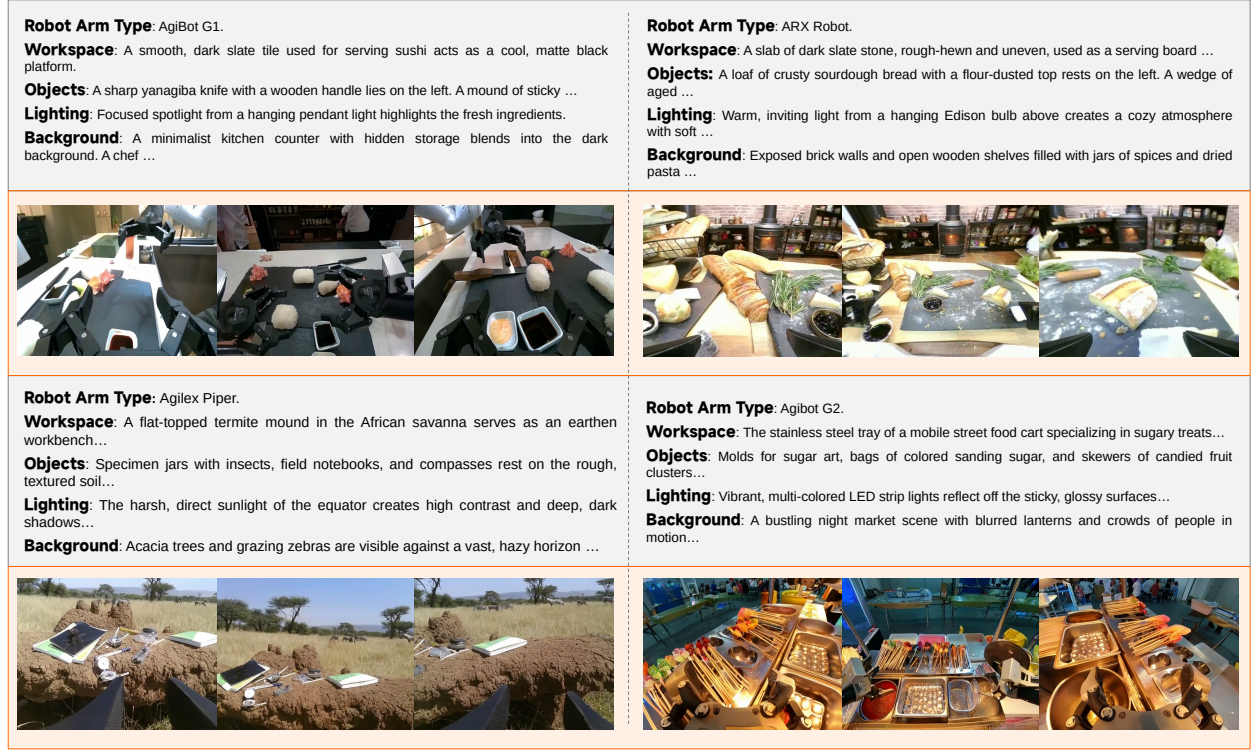


Figure 12 Qualitative results on embodied scene generation.

3.2 Embodied Scene Generation

Benchmark and Protocol. To evaluate embodied scene generation, we construct a dedicated benchmark using LLM-generated scene prompts. The benchmark contains 400 samples split equally into two difficulty levels: 200 Easy samples for conventional indoor manipulation scenes, and 200 Hard samples for unconstrained open-domain scenes. Each prompt specifies a robot arm type and a structured scene description covering workspace layout, object configuration, lighting, and background. We compare **Xiaomi-Robotics-U0** against GPT-Image-2 as the baseline. To ensure a fair comparison, multi-view reference images corresponding to each target robot arm type are provided to GPT-Image-2 as visual context.

Human-Based Pairwise Evaluation. Quantifying cross-view geometric consistency remains an open challenge, as no automated metric reliably evaluates global 3D coherence across viewpoints. We thus adopt a human pairwise preference protocol for quantitative assessment. Annotators are presented with paired outputs from **Xiaomi-Robotics-U0** and GPT-Image-2 in randomly shuffled left-right orders to mitigate positional preference bias. Annotators select the superior result based on two criteria: (1) multi-view geometric consistency, measuring unified 3D layouts with invariant object positions, scales, and occlusion across views; (2) textual instruction following, measuring alignment with the workspace, objects, lighting, and background in prompts.

Figure 14 reports the overall win rate. **Xiaomi-Robotics-U0** outperforms GPT-Image-2 in most comparisons, achieving clear advantages in multi-view consistency while maintaining comparable instruction-following performance. Qualitative comparisons in Figure 13 show that GPT-Image-2 produces high-quality images but suffers from severe cross-view geometric inconsistencies, with conflicting object positions, scales, and layouts across cameras. By contrast, **Xiaomi-Robotics-U0** generates geometrically coherent multi-view observations with consistent spatial layouts, while retaining photorealistic quality. This verifies that **Xiaomi-Robotics-U0** simultaneously achieves high visual fidelity and robust cross-view consistency.

Qualitative Results. Figure 12 presents qualitative results covering four robot arm types. **Xiaomi-Robotics-U0**

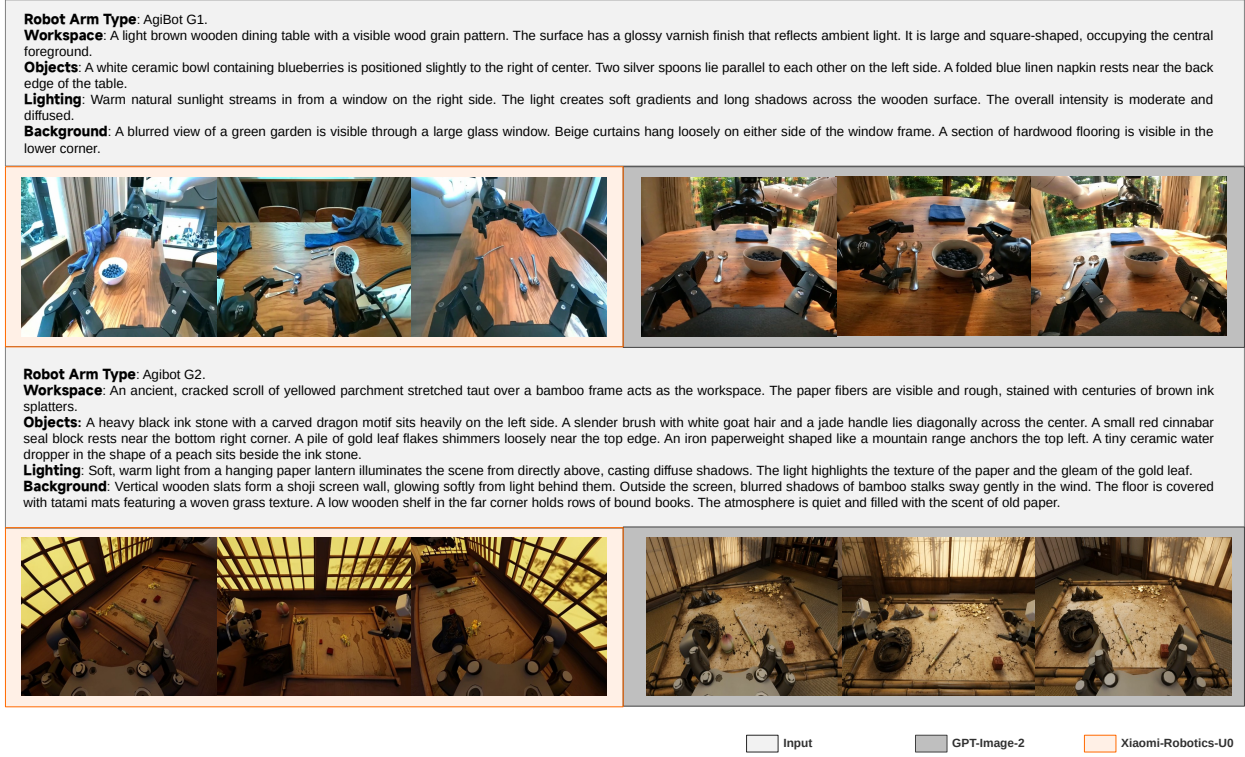


Figure 13 Qualitative comparison between GPT-Image-2 and Xiaomi-Robotics-U0 on the embodied scene generation benchmark.

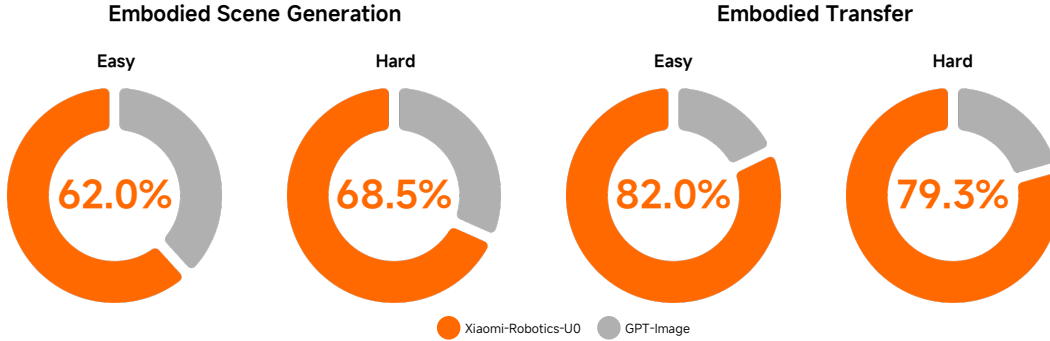


Figure 14 Pairwise evaluation results of embodied transfer and embodied scene generation.

generates photorealistic and geometrically consistent multi-view initial scenes for diverse scene prompts, demonstrating strong open-ended generalization to arbitrary textual descriptions. This shows that our model extends conventional single-view text-to-image synthesis to multi-view robotic observation generation by learning intrinsic geometric relations between arm configurations and camera projection. Notably, among the four evaluated robot arm types, the AGIBOT G2 is trained exclusively on simulated data, yet Xiaomi-Robotics-U0 can still produce photorealistic, real-world-aligned observations for this specific arm type. This validates the model’s promising sim-to-real generalization potential via learned real-world visual priors.

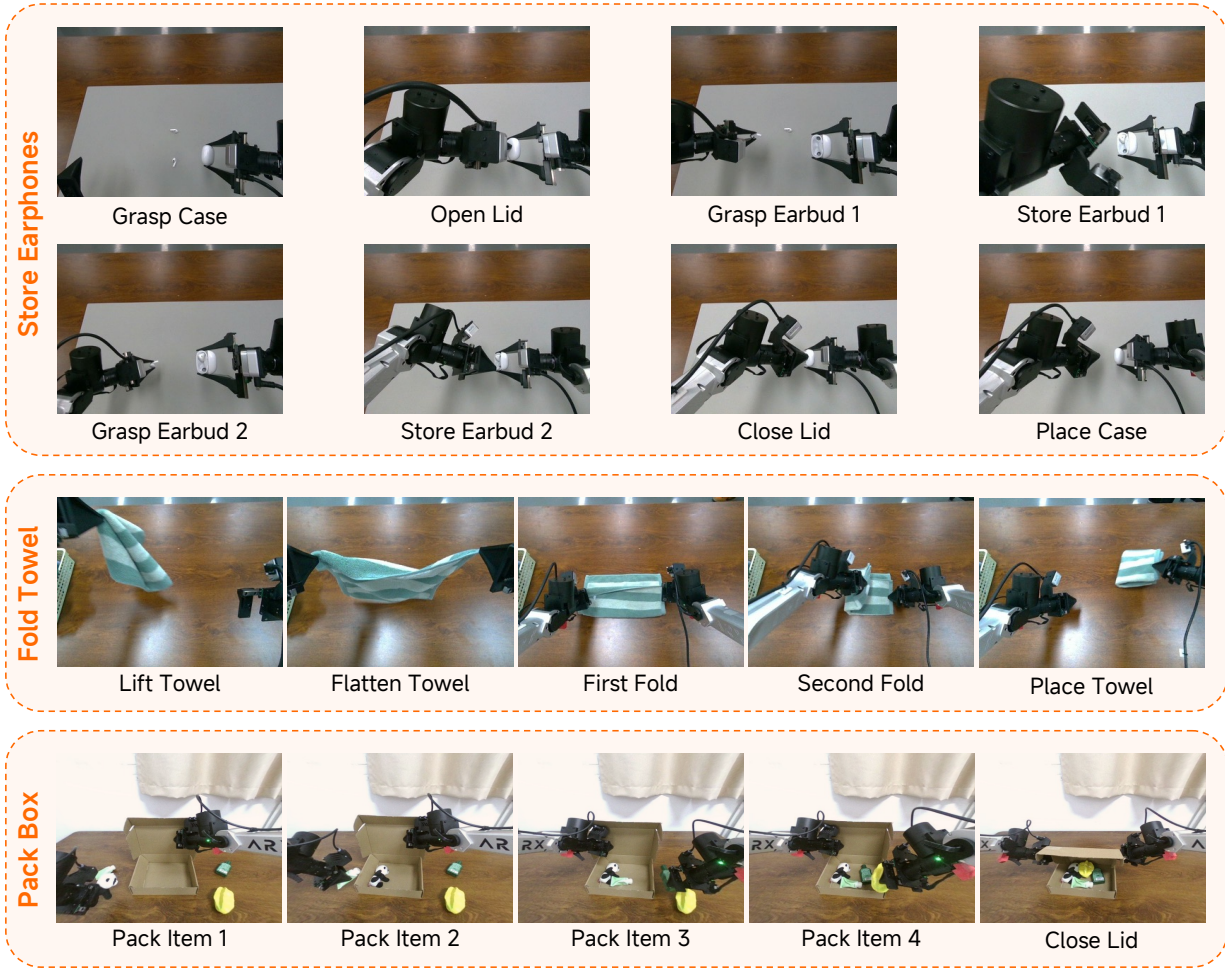


Figure 15 Stage-level progress visualization for each task. Each row corresponds to one task, and each column shows the head view at the completion of a progress milestone. For *Store Earphones*, the eight milestones run from grasping the case, through storing both earbuds, to closing the lid and placing the case. For *Fold Towel*, the milestones cover lifting and flattening the towel, two sequential folds, and placing the folded towel. For *Pack Box*, they cover four pick-and-place steps followed by closing the box lid.

3.3 Real World Experiments

We conduct the real-world evaluation to investigate whether the proposed world model improves downstream policy learning. Specifically, we use the *Xiaomi-Robotics-U0* model to augment the expert demonstrations, and compare the VLA policies trained on the original and augmented data across a suite of real-world tasks.

Task suite. The benchmark consists of three real-world tabletop manipulation tasks: *Store Earphones*, *Fold Towel*, and *Pack Box*, encompassing diverse challenges such as dexterous manipulation, deformable object handling, and multi-object interaction. Each task is defined by a language instruction and an ordered set of progress milestones. Figure 15 visualizes the milestone states for each task, providing a concrete reference for the progress metric defined later in this section.

Policy training and data. For each task in the suite above, we independently post-train a $\pi_{0.5}$ [23] policy starting from the `pi05_base` checkpoint, using its official PyTorch implementation. All training settings are kept identical across compared policies; the only variable is the supervised fine-tuning data mixture. For each task, we collect approximately 40 hours of real-world demonstrations on the robot platform described below, forming the task-specific clean data. To construct augmentation data, we apply the embodied scene generation

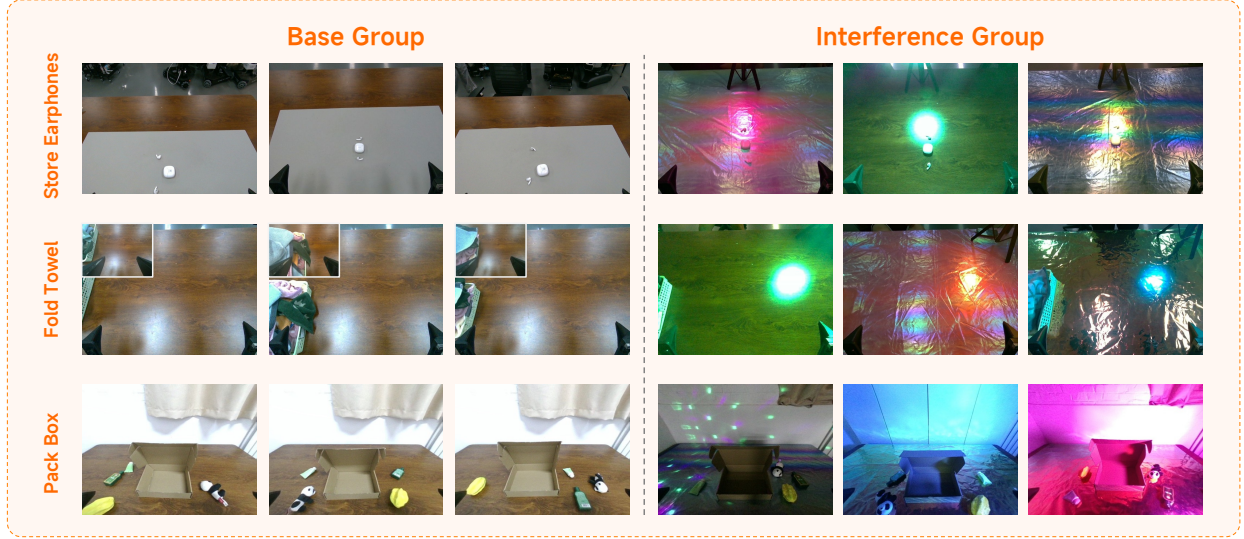


Figure 16 Task layouts for all three tasks. The base group varies the object layout for each task. For *Store Earphones*, different layouts vary the positions of the case and the earbuds. For *Fold Towel*, the towel is initialized in a randomized configuration, which is clearer in the left-wrist view shown as an inset at the top-left corner. For *Pack Box*, a new layout is generated by randomly placing the four objects on the table. The interference group uses the same layout randomization but additionally introduces substantial changes in tablecloth and lighting.

pipeline (Section 3.2) to the clean demonstrations, producing approximately 40 hours of style-transferred episodes per task that vary visual appearance—backgrounds, lighting, and textures—while preserving the original robot states and action labels. We compare two resulting per-task policies: **Original** is post-trained exclusively on the task-specific clean data; **Xiaomi-Robotics-U0-Aug** is post-trained on the mixture of clean data and its corresponding style-transferred augmentation. No inference-time world-model module is used during robot execution; the world model influences the policy solely through the post-training data mixture.

Evaluation schedule. To evaluate robustness beyond the original demonstration distribution while keeping the comparison controlled, we evaluate each policy under two groups of real-world conditions: a base group and an interference group. Each group contains three object-layout variants per task, and both policies are evaluated on an identical condition schedule. In the base group, the three layouts differ only in object placement; the background uses tablecloths seen in the demonstration data, and the scene is evaluated under nominal laboratory illumination. In the interference group, the layouts again vary object placement, but the background is replaced by held-out tablecloths absent from the demonstration data—often with more complex patterns and different textures or surface reflectance. The interference group further introduces held-out illumination, including low light, colored light, and dynamic patterned lighting such as disco-ball effects. The full set of evaluation layouts for every task and group is shown in Figure 16. For every policy, task, test group, and layout, we run three independent real-robot trials. This yields $2 \times 3 \times 3 = 18$ trials (two groups $\times$ three layouts $\times$ three trials) per policy-task pair.

Robot setup. We deploy all trained policies on a bimanual ARX robot platform with two 6-DoF arms, each equipped with a 1-DoF parallel-jaw gripper. At each policy query, the observation consists of the current proprioceptive arm state and three RGB views: one head view from the head camera and two wrist views from the left and right wrist cameras. For *Store Earphones* and *Fold Towel*, these views are captured with Intel RealSense D405 cameras; for *Pack Box*, we use fisheye cameras in the same viewpoint configuration to cover the larger packing workspace. The fisheye images are undistorted before use, both as policy inputs and as inputs to the **Xiaomi-Robotics-U0** augmentation. During real-robot evaluation, the policy is running on a separate inference server and communicates with the robot control process over a WebSocket connection on the local-area network. At 1 Hz, the robot transmits the current arm state and the three RGB views to the server, which returns an action chunk that is executed until the next query. The observation stream,

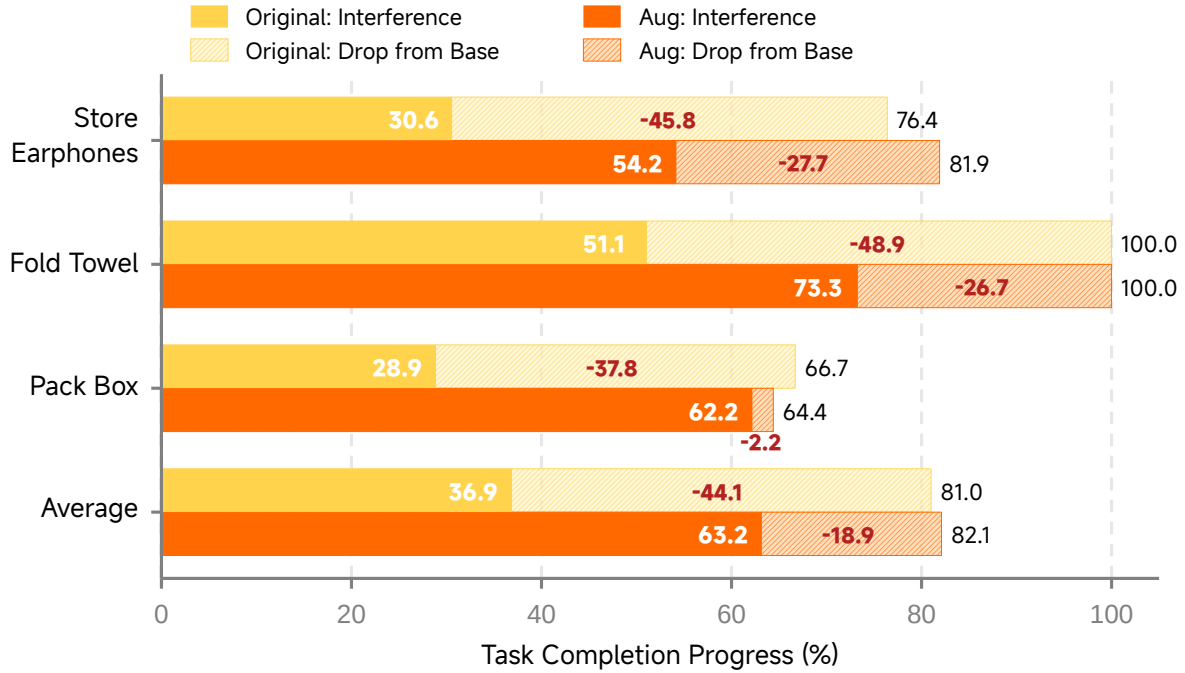


Figure 17 Comparison of task completion progress between Original and Xiaomi-Robotics-U0-Aug, split by evaluation group. Xiaomi-Robotics-U0-Aug is on par with Original under in-distribution conditions (base group) but achieves substantially higher progress under held-out backgrounds and lighting (interference group), leading to a clear overall advantage.

communication stack, and action-chunk execution interface are identical across all compared policies.

Metric. We use task completion progress as our sole evaluation metric. Rather than scoring only full success, progress assigns partial credit based on the ordered milestones a rollout completes, making it better suited to long-horizon manipulation, where a policy may reliably solve early subgoals yet fail at a later stage. Specifically, if a task has K ordered milestones and the rollout completes the first ℓ of them, its progress is $\ell/K \in [0, 1]$, where 1 denotes full success. For task t and group $g \in \{\text{base}, \text{interference}\}$, the reported progress of the tested policy is defined as:

$$\text{Prog}(t, g) = \frac{1}{N} \sum_{i=1}^N \frac{\ell_i}{K_t},$$

where N is the number of rollouts for each task and group (three layouts $\times$ three trials in our setting). A higher progress score for Xiaomi-Robotics-U0-Aug indicates that the world model provides useful post-training variation for improving real-world robustness across both fully and partially completed executions.

Experiment results. Figure 17 reports the task completion progress of both policies. In the base group, where test conditions closely resemble the demonstration distribution, Xiaomi-Robotics-U0-Aug performs comparably to Original, with a slight decrease on some tasks—an expected consequence of sharing training capacity with augmented data that does not add new information for in-distribution scenes. In the interference group, however, Xiaomi-Robotics-U0-Aug consistently outperforms Original across all tasks, indicating that style-transferred augmentation induces visual invariance to novel backgrounds and lighting unseen in the clean demonstrations. Overall, the robustness gains under distribution shift outweigh the marginal in-distribution trade-off, which is especially valuable for real-world deployment, where scene conditions cannot be tightly controlled.

To illustrate this gap qualitatively, Figure 18 presents representative Xiaomi-Robotics-U0-Aug rollouts under interference conditions for all three tasks. Despite challenging perturbations—held-out tablecloths,

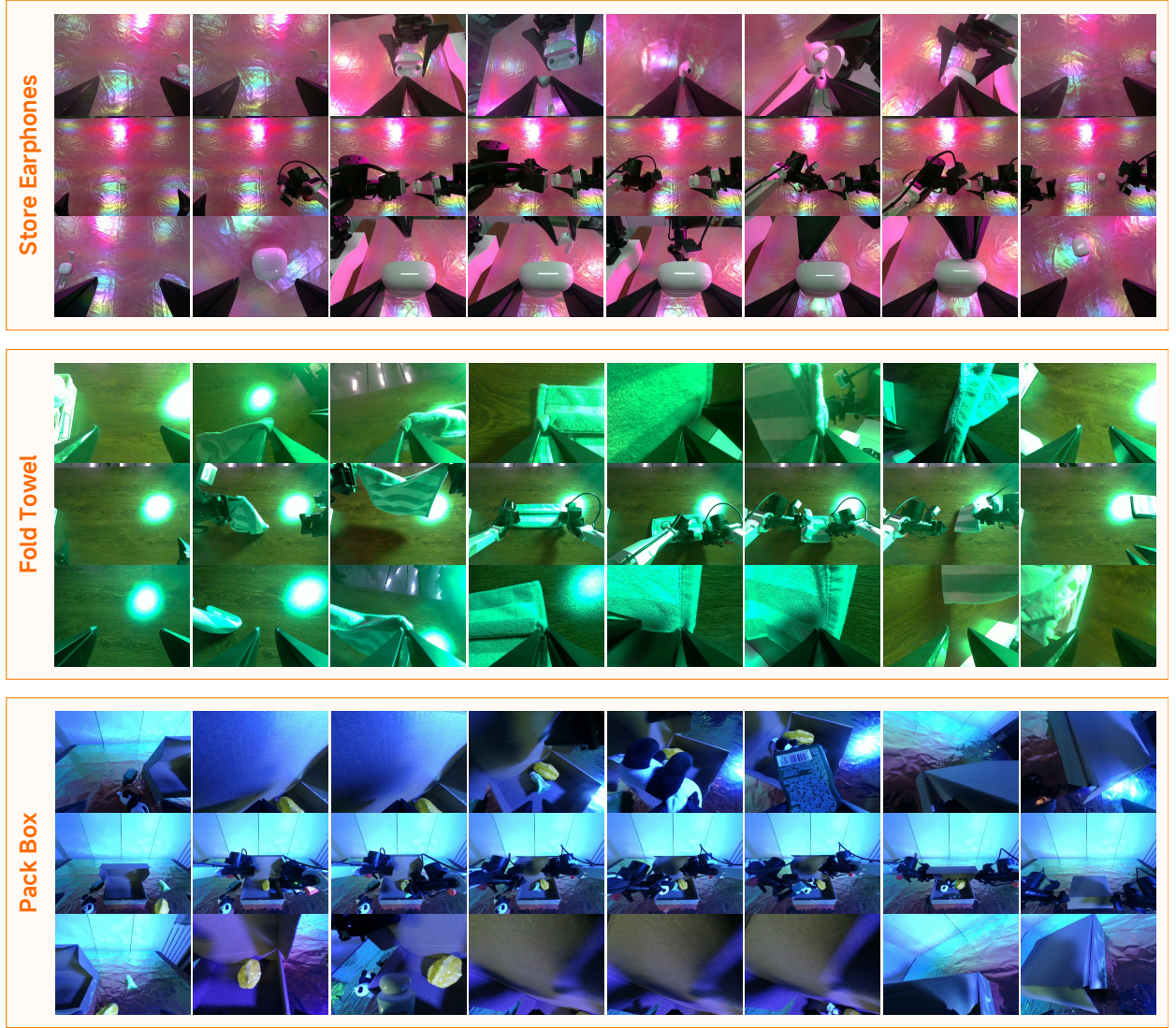


Figure 18 Xiaomi-Robotics-U0-Aug rollouts under interference scenarios for *Store Earphones*, *Fold Towel*, and *Pack Box* (top to bottom). Within each task, the three rows are the head view, left wrist view, and right wrist view, with frames sampled across the trajectory.

colored lighting, and projected patterns—the *Xiaomi-Robotics-U0-Aug* policy still executes the intended manipulation sequences: it stores both earbuds and closes the case in *Store Earphones*, completes the full folding sequence in *Fold Towel*, and packs all four objects before closing the lid in *Pack Box*.

Figure 19 further shows the typical failure modes that arise in our evaluation: (1) failing to grasp the earbud under saturated lighting; (2) misidentifying the orientation of the earphone case; (3) failing to locate the towel edges against a visually similar tablecloth; (4) unfolding an already-folded towel due to an incomplete grasp of the cloth edge; (5)(6) attempting to grasp a projected light spot instead of the target object. Under heavily corrupted visual observations, both policies occasionally exhibit these unexpected behaviors; however, *Xiaomi-Robotics-U0-Aug* is more likely to act correctly despite the perturbations or self-correct after a brief deviation. The Original policy, by contrast, tends to become permanently stuck on visual artifacts or commit unrecoverable errors at a higher rate.

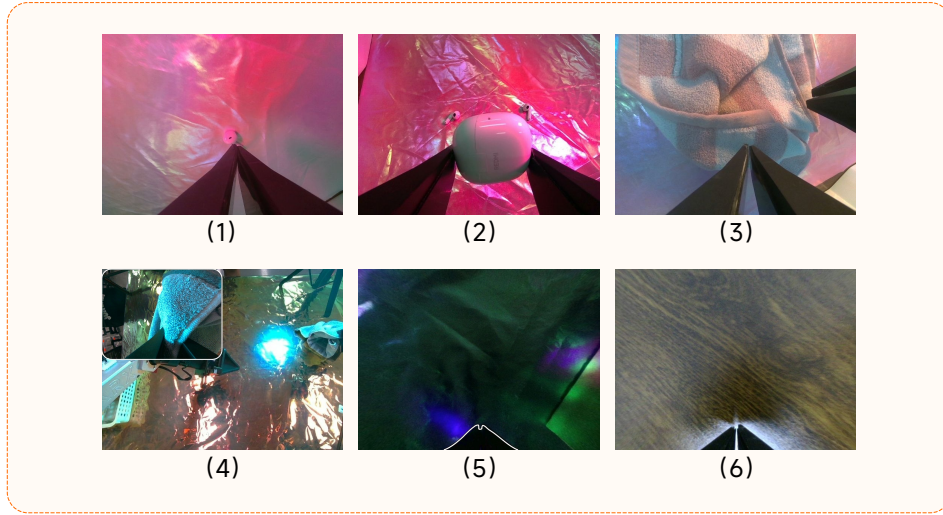


Figure 19 Typical failure cases across tasks: (1)–(2): *Store Earphones*, (3)–(4): *Fold Towel*, (5)–(6): *Pack Box*.

3.4 Video Generation

WorldArena. A key capability of an Embodied World Model is serving as a data engine that generates physically plausible and visually high-quality videos from a single observation and a corresponding instruction. To evaluate this video generation capability for embodiment tasks, we benchmark **Xiaomi-Robotics-U0** on WorldArena [49]. Given an initial frame along with a language instruction and robot action, WorldArena assesses the generated video across 15 metrics spanning six dimensions: visual quality, motion quality, content consistency, physics adherence, 3D accuracy, and controllability. We formulate this task as processing interleaved vision-language inputs and producing multi-frame interleaved outputs with empty language tokens. The initial frame paired with the language instruction serves as input, along with a sequence of robot mask images rendered from the robot action, camera parameters, and URDF as action guidance. Quantitative results are reported in Table 3 and Figure 21, qualitative visualizations are shown in Figure 22 (rows 1–5). **Xiaomi-Robotics-U0** achieves the best overall EWMScore (73.64), ranking first among over 100 submitted models. In terms of controllability, **Xiaomi-Robotics-U0** attains the highest Instruction Following (93.86) and ranks competitively on Semantic Alignment, demonstrating its strong ability to translate language instructions into faithful visual outcomes. For physics adherence, **Xiaomi-Robotics-U0** leads on Interaction Quality (87.30), indicating that it captures physically plausible object interactions. **Xiaomi-Robotics-U0** also achieves the best Perspectivity (98.84) and Motion Smoothness (95.51), reflecting accurate 3D viewpoint geometry and temporally stable motion. Overall, **Xiaomi-Robotics-U0** strikes the best balance across all evaluation dimensions, confirming its effectiveness as a data engine for embodiment tasks.

Embodiment Video Generation. Beyond the simulated WorldArena benchmark, we further leverage the scene generation capability of **Xiaomi-Robotics-U0** to produce novel initial frames and subsequently generate videos conditioned on language instructions. Representative examples are shown in Figure 20. In the first row, the robot arm correctly interacts with the cup—grasping it, lifting it, and placing it stably on the table, demonstrating accurate contact modeling. In the second row, beyond the correct tabletop manipulation, the model also renders dynamic background changes by simulating a person walking toward the table to observe the robot’s operation, indicating awareness of scene-level dynamics beyond the immediate task. In the third row, the model faithfully simulates the physics of a deformable object (a sheet of paper): the paper bends

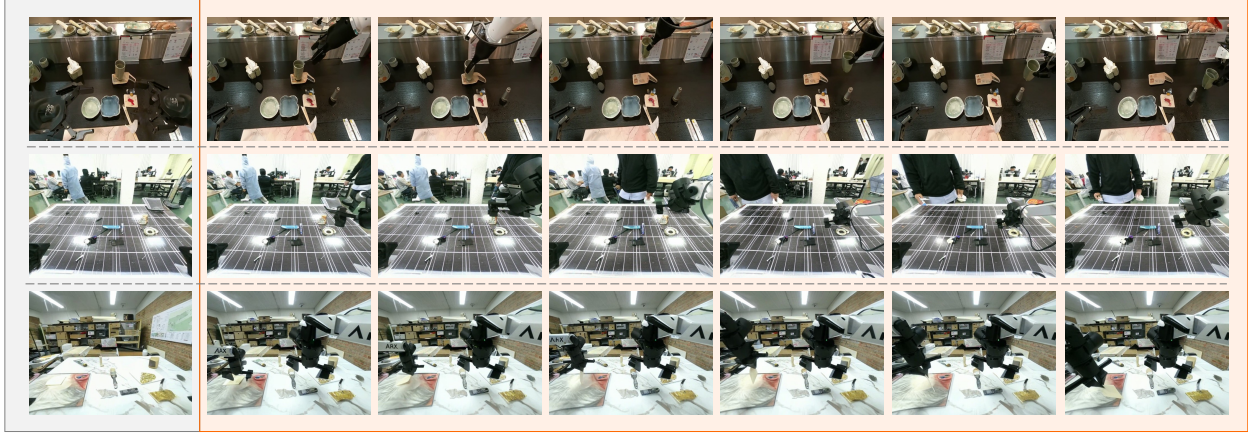


Figure 20 Qualitative video generation results of Xiaomi-Robotics-U0.

realistically when picked up and set down, and exhibits a slight lag during lateral movement due to its low mass, reflecting an understanding of material-dependent physical properties. More visualizations are provided in Figure 22.

Model	EWMScore_P	Rank	Visual Quality	Motion Quality	Content Consistency	Physics Adherence	3D Accuracy	Controllability
UNIS	73.64 🏆	1	61.78	85.08	55.87	64.60	92.04	91.60 🏆
SisyphusWorld	73.06	2	60.18	89.17	58.11	58.28	92.60	85.93
BWM-Fast	72.71	3	63.08	79.67	58.22	62.38	91.78	89.68
SACWM	72.67	4	59.80	85.58	58.86	59.62	93.66	85.41
DexWorldEngine	72.66	5	59.70	79.25	58.24	66.22	91.88	91.04
PAIWorld	72.34	6	63.12	80.49	57.82	61.75	91.58	87.07
FlowWAM-FiveAges	72.00	7	63.54	80.14	57.34	60.54	90.94	86.96
EvoPhysWorld	71.73	8	60.07	84.15	58.04	59.74	89.20	85.62
NoxWM	71.70	9	60.73	87.28	55.22	58.42	87.83	86.68
LeviosaWM	71.64	10	59.28	88.26	55.08	56.34	88.48	88.56

Figure 21 WorldArena leaderboard (July 13, 2026). UNIS is the anonymous evaluation codename for Xiaomi-Robotics-U0.

3.5 Text2Image & Image Edit Benchmarks

We evaluate the autoregressive **Xiaomi-Robotics-U0** image model on general text-to-image generation and image editing benchmarks. The goal of this evaluation is not to claim that **Xiaomi-Robotics-U0** is a dedicated open-domain image generator, but to quantify how much general image generation and editing ability is retained after adapting the model toward embodied scene synthesis. We use GenEval [20] for compositional T2I evaluation and ImgEdit [67] Basic for single-turn image editing.

As shown in Table 4, **Xiaomi-Robotics-U0** preserves strong general T2I and image-editing capabilities after embodied post-training. It performs competitively on most appearance-oriented generation and local editing tasks. Performance drops are largely confined to highly complex scenarios that require multi-constraint

Table 3 WorldArena dimension-level scores and overall EWMScore. The best score in each column is **bolded**.

		Xiaomi-Robotics-U0	SisyphusWorld	BWM-Fast	SACWM	DexWorldEngine
Visual Quality	Image Quality	53.94	45.57	51.22	45.83	51.44
	Aesthetic Quality	40.79	38.38	40.15	37.30	40.96
	JEPA Similarity	90.60	96.58	97.87	96.26	86.71
Motion Quality	Dynamic Degree	73.70	76.27	69.58	70.35	69.16
	Flow Score	86.02	99.82	75.11	95.79	74.31
	Motion Smoothness	95.51	91.42	94.33	90.60	94.27
Content Consistency	Subject Consistency	79.05	82.11	81.42	82.17	81.46
	Background Consistency	86.44	86.52	90.17	87.90	88.85
	Photometric Consistency	2.13	5.69	3.08	6.50	4.42
Physics Adherence	Interaction Quality	87.30	71.98	79.88	72.28	81.04
	Trajectory Accuracy	41.89	44.58	44.89	46.95	51.41
3D Accuracy	Depth Accuracy	85.25	94.85	86.41	94.99	86.19
	Perspectivity	98.84	90.34	97.14	92.34	97.58
Controllability	Instruction Following	93.86	82.68	90.22	81.66	92.62
	Semantic Alignment	89.35	89.18	89.15	89.16	89.45
EWMScore ($\uparrow$)		73.64	73.06	72.71	72.67	72.66

Table 4 Text-to-image and image-editing benchmark comparison with category-level breakdowns. GenEval scores are reported on a 0–1 scale, and ImgEdit scores use a 1–5 judge scale.

Model	Overall	Single object	Two object	Counting	Colors	Position	Attribute binding
Qwen-Image	0.87	0.99	0.92	0.89	0.88	0.76	0.77
Xiaomi-Robotics-U0-AR	0.74	0.99	0.92	0.49	0.88	0.60	0.53

Model	Overall	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action
Qwen-Image	4.27	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69
Xiaomi-Robotics-U0-AR	4.08	4.21	4.15	3.28	4.35	4.21	4.03	4.72	3.27	4.53

composition, precise spatial grounding, or symbolic binding.

Table 4 demonstrates that **Xiaomi-Robotics-U0** retains robust general T2I and image-editing capabilities after embodied post-training. While it slightly lags behind Qwen-Image in highly complex cases—such as multi-constraint composition or precise spatial alignment—**Xiaomi-Robotics-U0** maintains highly comparable performance on common image generation and local editing instructions without catastrophic forgetting.

4 Related Work

4.1 Foundation Generative Models and Embodied World Models

Recent years have witnessed rapid advances in foundation generative models, evolving from image synthesis to unified multimodal world modeling. Large-scale image generation and multimodal generation models, including Stable Diffusion [4], FLUX [27], GPT-Image-2.0 [42], Show-o [64], Janus [60], Janus-Pro [13], Chameleon [55], BAGEL [17], and EMU3.5 [16], have demonstrated remarkable capabilities in semantic understanding, instruction following, compositional reasoning, and controllable visual generation. More recently, foundation video and world models, such as Genie [7], Genie2 [44], Sora [41], MovieGen [46], Wan [59], and Cosmos [1], further extend generation from static images to long-horizon videos and interactive world simulation, enabling coherent prediction of future observations under a unified autoregressive or diffusion framework. Despite their impressive generalization ability, these models are designed for generic visual content and do not explicitly model the geometric constraints, robot embodiments, calibrated camera systems, or physical interaction dynamics required by embodied manipulation.

To bridge this gap, recent embodied world models introduce robot-centric world modeling by adapting foundation models to manipulation trajectories and robot videos. Representative approaches include Dreamer Series [22], World Action Models (WAMs) [28, 56, 65, 70], DreamZero [66], and Qwen-RobotWorld [72],

which learn predictive models of future observations for planning, simulation, or policy learning. While these methods significantly advance embodied prediction, they are typically optimized using robot-only datasets whose scale and diversity remain substantially smaller than those used for foundation generative models. Consequently, continual adaptation often sacrifices part of the semantic richness, controllability, and visual diversity inherited from foundation pre-training. In contrast, our work treats embodied generation as a natural extension of foundation image and video generation. Rather than specializing a foundation model solely on robot trajectories, we jointly optimize general-domain generation and embodied synthesis within a unified autoregressive framework, preserving the original generation capability while acquiring robot-centric geometric reasoning and interaction modeling.

4.2 Controllable and Multi-view Embodied Generation

Recent advances in controllable image generation and image editing have substantially improved semantic manipulation through instruction-guided synthesis and reference-conditioned editing. Representative approaches, including InstructPix2Pix [6], AnyDoor [12], OmniGen [63], GPT-Image-2.0 [42], and FLUX Kontext [27], enable high-quality editing with strong instruction following and semantic consistency. However, these methods are primarily designed for natural images and generally assume single-view observations, making them difficult to directly apply to embodied manipulation, where multiple calibrated cameras, robot embodiments, and interaction states must remain geometrically consistent.

Meanwhile, multi-view generation has attracted increasing attention in image synthesis and 3D content creation. Representative methods, including MVDream [53], SyncDreamer [33], Wonder3D [35], Era3D [29], Zero123++ [52], and related multi-view diffusion models [19, 54], explicitly enforce cross-view consistency to facilitate high-quality 3D reconstruction and asset generation. Nevertheless, these methods mainly focus on object-centric generation and do not explicitly consider robot embodiments, calibrated manipulation cameras, or interaction-aware scene synthesis.

Our work bridges these two research directions by unifying controllable image editing, multi-view embodied scene generation, and embodied transfer within a single autoregressive framework. Furthermore, we introduce a structured control formulation that disentangles workspace layout, background appearance, foreground irrelevant objects, target objects, and lighting into independent control dimensions, enabling significantly richer embodied scene variations while preserving robot geometry, embodiment constraints, and multi-view consistency.

4.3 Embodied Video Generation and Synthetic Data Engines

Video generation and embodied world simulation have recently become increasingly important for scalable robot learning. Foundation video models such as MovieGen [46], Wan [59], Cosmos [1], Genie [7], and Qwen-RobotWorld [72] demonstrate impressive capabilities in long-horizon video synthesis and future observation prediction. Meanwhile, large-scale robot datasets and simulation platforms, including Open X-Embodiment [43], RoboCasa [40], ManiSkill [39], BridgeData V2 [58], and DROID [25], have substantially accelerated the development of data-driven robot learning. More recently, embodied video datasets and synthetic data pipelines, such as RoboMIND [61], RoboCoin [62], GenieSim [69], InternA1 [57], and AgiBotWorld [8], further explore large-scale embodied video generation and robot data augmentation.

Despite these advances, existing approaches typically synthesize isolated images, simulation assets, or short robot videos independently, without jointly modeling controllable scene generation, embodied transfer, and long-horizon trajectory rollout. In contrast, our work unifies embodied scene generation, embodied transfer, interleaved subtask-subgoal prediction, and embodied video generation within a single autoregressive model. By jointly learning embodied videos at multiple temporal resolutions, our model captures both long-horizon task progression and fine-grained interaction dynamics, transforming foundation world models into scalable embodied data engines for downstream policy learning and future agentic embodied systems.

5 Conclusions

In this report, we presented **Xiaomi-Robotics-U0**, a unified embodied synthesis model built upon a world foundation model through lightweight continual training on general-domain and embodied datasets. **Xiaomi-Robotics-U0** unifies text-to-image generation, image editing, multi-view embodied scene generation, embodied transfer, and embodied video generation within a single autoregressive framework, successfully transferring the rich visual knowledge of foundation models into embodied intelligence. To the best of our knowledge, **Xiaomi-Robotics-U0** is the first unified embodied synthesis model supporting both high-quality multi-view embodied generation and sequential embodied video generation, as well as the first to introduce a structured control formulation for scalable embodied data generation. Extensive experiments demonstrate state-of-the-art performance across both single-step and sequential embodied generation tasks, outperforming GPT-Image-2.0 in human evaluation, achieving state-of-the-art results on the World Arena benchmark, and substantially improving the robustness of downstream robot policies through generated data. We hope this work provides a practical step toward scalable agentic data engines for embodied intelligence.

Despite the promising results, **Xiaomi-Robotics-U0** has several limitations. First, embodied transfer still relies on depth estimation as an intermediate representation, which may introduce occasional artifacts and limit fine-grained control over object textures and local appearances. Future work will investigate embodied multi-view image editing directly on raw observations. Second, embodied scene generation and video generation are performed separately, making long-horizon rollout susceptible to accumulated errors; jointly generating scenes and subsequent videos is a promising direction. Finally, the current 32K context window limits long-duration video modeling, and extending the model to longer contexts will enable minute-level embodied interaction generation.

Contributions

*Core contributors, alphabetically listed by first name **External contributor

Project Lead

- Xinghang Li

Data:

- Jun Guo*
- Qiwei Li*
- Hang Lai*
- Long Qian*
- Yueze Wang*
- Nan Sun*
- Hongyu Yan*
- Jingen Qu
- Jiayi Song
- Jiahang Cao
- Hanyue Zhao
- Heyun Wang
- Yunhong Wang
- Jiayi Song
- Caoyu Xia
- Jack Zhao

Training:

- Jun Guo*
- Qiwei Li*
- Long Qian*
- Yueze Wang*
- Jingen Qu
- Jiayi Song

- Nan Sun
- Hongyu Yan

Deployment & Experiments:

- Jiahang Cao*
- Jun Guo*
- Hang Lai*
- Qiwei Li*
- Long Qian*
- Jiayi Song*
- Hongyu Yan*
- Hanyue Zhao*
- Futeng Liu
- Wanli Peng

Inference Speedup:

- Xi Chen*
- Jingen Qu*

Infrastructure:

- Xinghang Li*
- Jingen Qu
- Xi Chen

Senior Lead:

- Jason Li*
- Diyun Xiang
- Hangjun Ye
- Heng Qu
- Huaping Liu**

Acknowledgment

We would like to express our sincere appreciation to the broader team for their tremendous support, including those not listed above: Li Jiang, Zhiyin Lai, Xiaoke Xilinjueluo, Qingyi Li, Qi Liu, Yayun Liu, Jun Xia, Feng Qiu, Donghao Wang, Yan Hou, Dong Wang, Quanyun Zhou, Liangliang He, Jiayin Liu, Kang Zhou, Rui Cai, Shuoxue Bi, Yingchao Zhou, Kun Ma, Yiwei Zhou



Figure 22 Qualitative video generation results of *Xiaomi-Robotics-U0*. Rows 1–5: WorldArena benchmark samples conditioned on language instructions and initial observations. Rows 6–10: videos generated from *Xiaomi-Robotics-U0* produced scene images. Each row shows the initial frame and language instructions followed by predicted future frames.



Figure 23 Qualitative results of embodied transfer. We extract depths and modify descriptions from the original scene to generate the new scene.

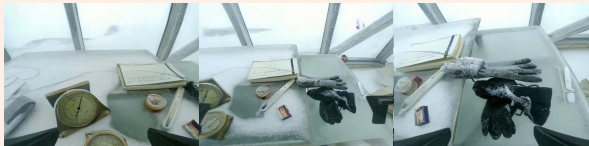
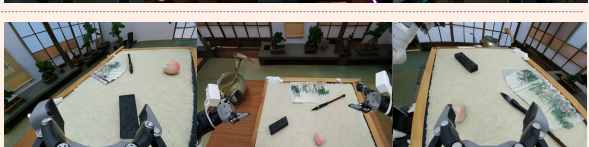
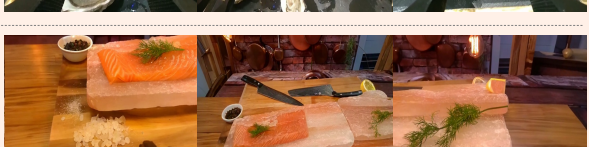
Prompt	Generated Multi-View Observations
Robot Arm Type: AgiBot G1. Workspace: A block of floral foam soaked in water, vibrant green and porous, used for arrangements. The surface is damp, crumbly, and rectangular. Objects: A bundle of long-stemmed red roses with thorns stripped lies on the left. A roll of green floral tape sits upright near the center. A pair of pruning shears with red handles rests on the right edge. A handful of baby's breath, delicate and white, scatters near the front. A spool of gold ribbon, shiny and narrow, unspools toward the back. A plastic vase, clear and cylindrical, stands empty in the far corner. Lighting: Bright, colorful studio lights illuminate the flowers, enhancing their vivid hues. Soft shadows fall behind the stems, creating depth in the arrangement. Background: Walls lined with shelves of dried flowers and ribbons in every color surround the area. Large windows look out onto a bustling flower market street. Buckets of fresh blooms fill the floor space in the background. A chalkboard menu listing flower prices hangs on the back wall.	
Robot Arm Type: AgiBot G1. Workspace: A section of the International Space Station's exterior handrail, coated in white thermal insulation material and micrometeoroid pits. Objects: A specialized torque wrench with a tether clip is magnetically attached near the joint connection. A sample canister sealed with a red cap floats secured by a bungee cord in the center. A fragment of golden multilayer insulation foil flutters loosely near the right clamp. A data slate with a cracked screen is strapped to the rail near the left edge. A gloved tool pouch made of Kevlar hangs vertically at the back. Lighting: Unfiltered solar radiation provides intense, blinding white light with absolute black shadows where the sun is blocked. Background: The curvature of the Earth with swirling blue oceans and white clouds dominates the view below. The infinite black void of space is dotted with bright, unblinking stars. Solar panels with a blue tint extend into the distance on the left side. A robotic arm segment is partially visible against the backdrop of the planet.	
Robot Arm Type: ARX Robot. Workspace: The glass top of a high-altitude meteorological station table, frosted with ice crystals and vibrating slightly from the wind. Objects: A brass barometer with a cracked face sits securely bolted to the near left corner. A stack of wet, handwritten logbooks tied with string rests in the center. A frozen mercury thermometer lies horizontally across the right side. A pair of thick woolen gloves with ice clinging to the fibers sits by the edge. A small tin of matches, half-empty and damp, is placed near the logbooks. A metal compass with a glowing dial rests on top of the gloves. Lighting: Harsh, pale blue emergency LED strips mounted on the ceiling cast sterile, shadowless light that reflects sharply off the icy glass. Background: Thick reinforced windows reveal a blinding white blizzard swirling outside. Steel support beams coated in frost crisscross the room. A red emergency fire extinguisher is mounted on the far wall. The sound of howling wind vibrates through the entire structure.	
Robot Arm Type: ARX Robot. Workspace: An ancient, ink-stained wooden printing press bed, carved with intricate floral patterns and smelling of turpentine and old paper. Objects: A heavy lead type tray filled with reversed letters sits on the left side. A roller covered in thick black ink rests diagonally across the center. A sheet of handmade rice paper with a fresh crimson print lies on the right. A small porcelain bowl of water for cleaning brushes is near the front edge. A wooden mallet with a worn handle leans against the ink roller. Scattered copper shavings from type casting gleam near the back corner. Lighting: Soft, directional daylight streams through a high clerestory window, illuminating dust motes dancing above the ink-stained wood. Background: Tall shelves packed with thousands of wooden type cases line the dimly lit walls. A large hanging pulley system for lifting paper reams dominates the ceiling space. The floor is concrete, stained with decades of spilled ink. Faint sounds of a city street drift in through the open window above.	
Robot Arm Type: AgiBot G2. Workspace: The control console of a retro-futuristic spaceship cockpit from a 1970s sci-fi movie set. Objects: Blinking analog dials, toggle switches with worn labels, and a joystick with a glowing crystal tip. Lighting: Pulsating neon lights in pink and teal embedded in the dashboard, casting a cybernetic glow. Background: A large viewport showing a starfield streaking by at warp speed, with nebulae in shades of purple and gold.	
Robot Arm Type: AgiBot G2. Workspace: A sheet of thick, handmade rice paper, cream-colored and textured with visible fibers, stretched taut over a wooden frame. The surface is delicate and matte. Objects: A traditional calligraphy brush with black ink dripping from the tip lies diagonally across the center. A solid ink stick, carved with a dragon, rests on the left side. A small porcelain water dropper shaped like a peach sits near the right edge. A red seal stamp with gold characters is placed at the front. A folded fan with a landscape painting leans against the back frame. Lighting: Soft, natural light from a paper lantern hanging nearby creates a gentle, warm glow. Shadows are soft and barely perceptible on the textured paper. Background: A traditional Japanese tea room with sliding shoji screens and tatami mat flooring. A scroll painting of a waterfall hangs on the far wall. A bronze incense burner emits thin trails of smoke in the corner. Bonsai trees in shallow pots line the windowsill in the background.	
Robot Arm Type: Agilix Piper. Workspace: A slab of dark slate stone commonly used for serving oysters, wet with seawater and crushed ice. The surface is black, reflective, and extremely cold. Objects: Six fresh oysters on the half shell are arranged in a semi-circle on the left. A wedge of lemon with bright yellow rind sits in the center. A small glass dish of red cocktail sauce is placed on the right edge. A sprig of green parsley garnishes the front corner. A bottle opener with a wooden handle rests on the back edge. Lighting: Bright, cool restaurant spotlighting from above highlights the glistening moisture on the oysters and ice. Reflections of the dining room chandeliers dance on the wet slate. Background: A bustling upscale seafood restaurant with blurred diners in the background. White tablecloths and silverware are visible on adjacent tables. Large aquarium tanks with swimming fish line the far wall. Waiters in black vests move quickly through the scene. Soft jazz music is implied by the elegant atmosphere.	
Robot Arm Type: Agilix Piper. Workspace: A massive block of clear Himalayan salt, pink and crystalline, carved into a flat preparation surface. The texture is grainy and dry, leaving a fine dust on anything it touches. Objects: A fillet of raw salmon, orange and glistening, lies centrally on the salt block. A sprig of fresh dill, bright green and feathery, rests on the fish. A small bowl of black peppercorns, rough and dark, sits on the left corner. A lemon wedge, yellow and juicy, is placed near the right edge. A chef's knife with a Damascus steel blade leans against the back. A cluster of coarse sea salt crystals, white and jagged, scatters near the front. Lighting: Warm, golden light from a hanging Edison bulb enhances the pink hues of the salt and the orange of the fish. Shadows are soft and inviting, creating a cozy culinary atmosphere. Background: Rustic kitchen walls made of exposed brick and hanging copper pots fill the background. Wooden shelves stocked with spice jars and dried herbs line the rear. A stone fireplace glows warmly in the far corner. The floor is made of wide-plank oak, worn smooth by years of use.	

Figure 24 Embodied scene generation results.

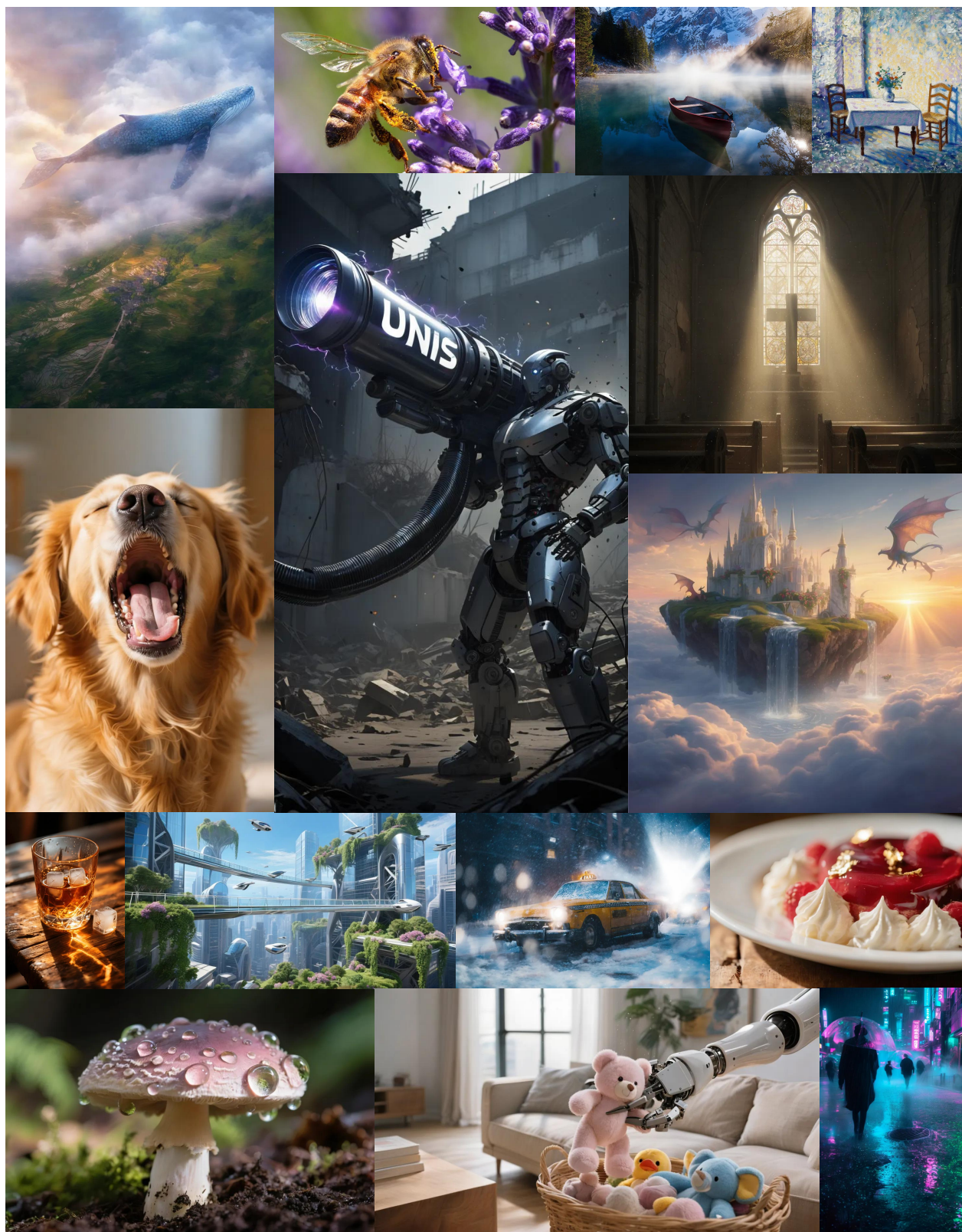


Figure 25 Text-to-image results of Xiaomi-Robotics-U0.

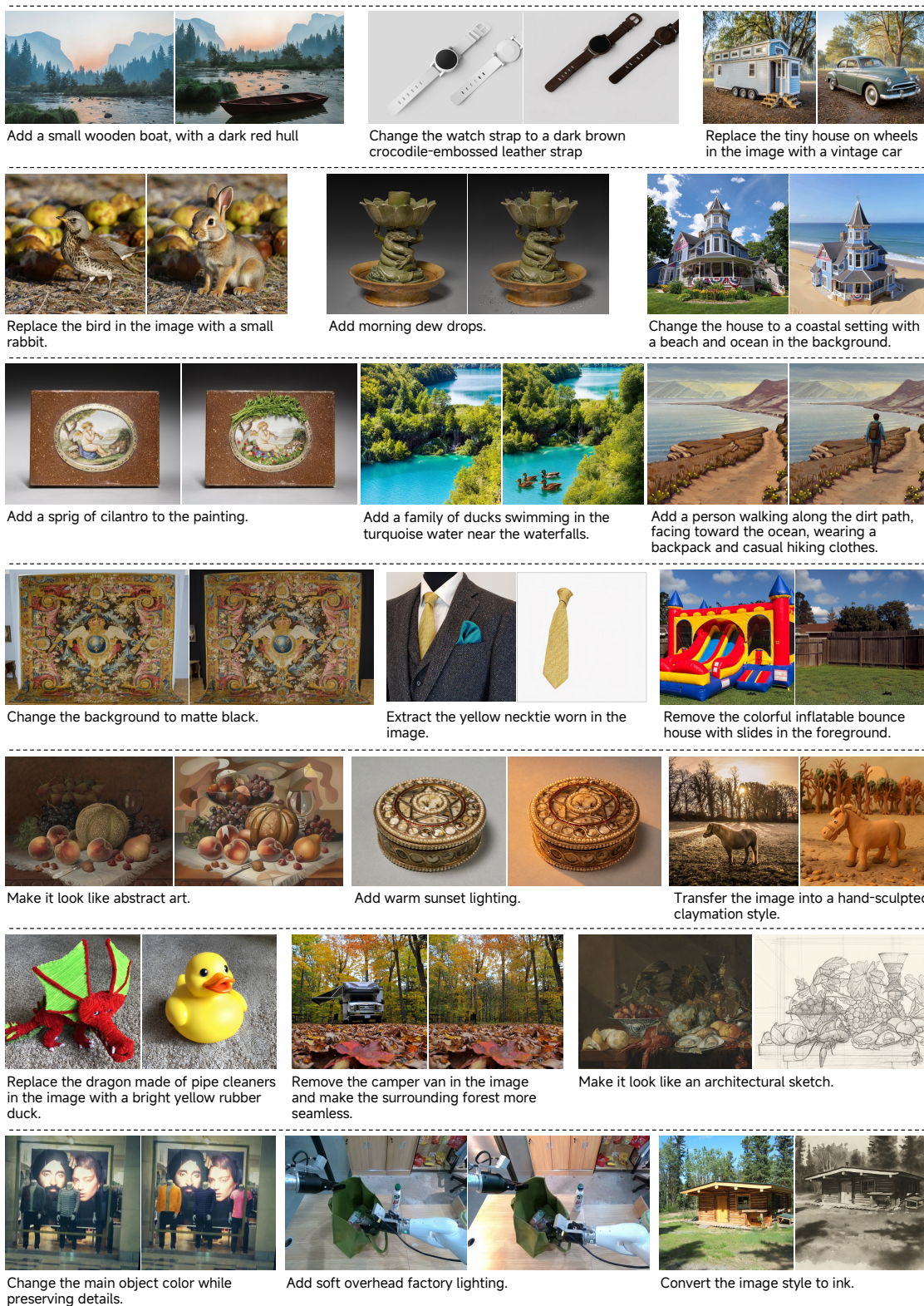


Figure 26 Image editing results of Xiaomi-Robotics-U0.

References

- [1] Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. [arXiv preprint arXiv:2606.02800](#), 2026.
- [2] Timur Akhtyamov, Mohamad Al Mdfaa, Javier Antonio Ramirez Benavides, Arthur Nigmatzyanov, Sergey Bakulin, German Devchich, Denis Fatykhov, Diego Ruiz Salinas, Alexander Mazurov, Kristina Zipa, et al. Egowalk: A multimodal dataset for robot navigation in the wild. [arXiv preprint arXiv:2505.21282](#), 2025.
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. [arXiv preprint arXiv:2511.21631](#), 2025.
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. [arXiv preprint arXiv:2311.15127](#), 2023.
- [5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. [arXiv preprint arXiv:2212.06817](#), 2022.
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 18392–18402, 2023.
- [7] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In [Forty-first International Conference on Machine Learning](#), 2024.
- [8] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. AgiBot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. [arXiv preprint arXiv:2503.06669](#), 2025.
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In [European Conference on Computer Vision](#), pages 370–387. Springer, 2024.
- [10] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In [Proceedings of the Computer Vision and Pattern Recognition Conference](#), pages 22831–22840, 2025.
- [11] Tianxing Chen, Zanzin 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](#), 2025.
- [12] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 6593–6602, 2024.
- [13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. [arXiv preprint arXiv:2501.17811](#), 2025.
- [14] Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, et al. Abot-physicsworld: Interactive world foundation model for robotic manipulation with physics alignment. [arXiv preprint arXiv:2603.23376](#), 2026.
- [15] Open X-Embodiment Collaboration, Abby O’Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. <https://arxiv.org/abs/2310.08864>, 2023.
- [16] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. [arXiv preprint arXiv:2510.26583](#), 2025.

- [17] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025.
- [18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems, 36:27092–27112, 2023.
- [19] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314, 2024.
- [20] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36:52132–52152, 2023.
- [21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18995–19012, 2022.
- [22] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603, 2019.
- [23] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. arXiv preprint arXiv:2504.16054, 2025.
- [24] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianing 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.
- [25] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945, 2024.
- [26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.
- [27] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.
- [28] 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.
- [29] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. Advances in Neural Information Processing Systems, 37:55975–56000, 2024.
- [30] Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. arXiv preprint arXiv:2510.16732, 2025.
- [31] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22160–22169, 2024.
- [32] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. IEEE/ASME Transactions on Mechatronics, 2025.
- [33] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In International conference on learning representations, volume 2024, pages 27676–27697, 2024.

- [34] Qian Long, Yueze Wang, Jiayi Song, Junbo Zhang, Peiyan Li, Wenxuan Wang, Yuqi Wang, Haoyang Li, Shaoxuan Xie, Guocai Yao, et al. Scaling world model for hierarchical manipulation policies. arXiv preprint arXiv:2602.10983, 2026.
- [35] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9970–9980, 2024.
- [36] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from physical simulators and world models. arXiv preprint arXiv:2507.00917, 2025.
- [37] Leland McInnes, John Healy, and S. Astels. hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2:205, 2017. URL <https://api.semanticscholar.org/CorpusID:53231359>.
- [38] Byeon Minwoo, Park Beomhee, Kim Haecheon, Lee Sungjun, Woonhyuk Baek, and Kim Saehoon. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- [39] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. arXiv preprint arXiv:2107.14483, 2021.
- [40] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. arXiv preprint arXiv:2406.02523, 2024.
- [41] OpenAI. Video generation models as world simulators. <https://openai.com/research/video-generation-models-as-world-simulators>, 2024. Technical report.
- [42] OpenAI. Gpt image api. <https://platform.openai.com/docs/guides/image-generation>, 2025. Accessed: 2026-07-03.
- [43] 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.
- [44] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. URL <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/>.
- [45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In International Conference on Learning Representations, volume 2024, pages 1862–1874, 2024.
- [46] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024.
- [47] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. arXiv preprint arXiv:2506.09042, 2025.
- [48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems, 35:25278–25294, 2022.
- [49] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyu Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. arXiv preprint arXiv:2602.08971, 2026.

- [50] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *Advances in Neural Information Processing Systems*, 38:63674–63698, 2026.
- [51] Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16037–16046, 2025.
- [52] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.
- [53] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International conference on learning representations*, volume 2024, pages 39838–39859, 2024.
- [54] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdifffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 175–191. Springer, 2024.
- [55] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL <https://arxiv.org/abs/2405.09818>, 9(8), 2024.
- [56] MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.
- [57] Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 976–985, 2026.
- [58] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- [59] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [60] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.
- [61] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinyu Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.
- [62] 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.
- [63] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2025.
- [64] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pages 28240–28264, 2025.
- [65] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world-action model. *arXiv preprint arXiv:2603.17240*, 2026.
- [66] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shen Yuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

- [67] Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. Advances in Neural Information Processing Systems, 38, 2026.
- [68] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12–22, 2023.
- [69] Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, Zhaobo Liu, Zhen Xiao, Sheng Zhang, Lei Bao, Rui Feng, Zhenquan Pang, Jiayu Li, Qian Wang, and Maoqing Yao. Genie sim 3.0 : A high-fidelity comprehensive simulation platform for humanoid robot, 2026. URL <https://arxiv.org/abs/2601.02078>.
- [70] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? arXiv preprint arXiv:2603.16666, 2026.
- [71] Yuguang Yue, Irakli Salia, Samuel Hunt, Chris Green, Wenzhe Shi, and Jonathan J Hunt. Scaling behavior cloning improves causal reasoning: An open model for real-time video game playing. arXiv preprint arXiv:2601.04575, 2026.
- [72] Jie Zhang, Xiaoyue Chen, Anzhe Chen, Chenxu Lv, Deqing Li, Gengze Zhou, Hang Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-robotworld technical report: Unifying embodied world modeling through language-conditioned video generation. arXiv preprint arXiv:2606.17030, 2026.
- [73] Junkang Zhou, Yefei He, Feng Chen, Weijie Wang, and Bohan Zhuang. Flashar: Efficient post-training acceleration for autoregressive image generation, 2026. URL <https://arxiv.org/abs/2605.09430>.