WorldScape Policy 2.0: Empowering Steerable World Action Modeling with Reasoning-Augmented Memory

Haisheng Su[1,†], **Zongdai Liu**[1], **Xin Jin**[1], **Haoxuan Dou**[1], **Chengming Hu**[1], **Baorun Li**[1], **Zhanwang Liu**[1], **Ruiyan Xu**[1], **Jianjie Fang**[2], **Xin Zhang**[1], **Zhenjie Yang**[3], **Xue Yang**[3], **Chen Gao**[2], **Junchi Yan**[3], **Yong Li**[2], **Wei Wu**[1,‡]

[1]Manifold AI [2]Tsinghua University [3]Shanghai Jiao Tong University

[†]**Project Lead.** [‡]**Corresponding Author.**

Abstract. World Action Models (WAMs) offer a promising paradigm for robotic manipulation by jointly modeling visual state transitions and robot actions. However, existing WAMs are constrained by limited temporal context, coarse episode-level language supervision, and predominantly text-only conditioning, which hinder task-progress tracking and fine-grained language-video-action grounding while limiting visual-context reasoning and cross-embodiment transfer. In this paper, we introduce **WorldScape Policy 2.0**, a controllable WAM with reasoning-augmented long short-term memory. Its causal short-term visual memory supplies recent observations as DiT prefill to preserve local interaction dynamics, while its long short-term event memory organizes historical VLM outputs into global-history, local-active, and event-boundary representations for progress-aware retrieval. The retrieved history augments perception and autoregressively generated planning tokens, yielding an implicit subgoal condition for autonomous planning; semantic forcing further transfers event-level instruction semantics into this latent planning pathway. To establish fine-grained multimodal controllability, we construct **ManipEvent-5M**, an event-grounded embodied pretraining dataset containing nearly 5 million event segments with aligned action trajectories, episode-level task instructions, segment-level subtask captions, goal images, and video demonstrations. These designs provide a unified interface for autonomous planning from high-level instructions and controllable execution from fine-grained text, goal-image, or video-context prompts. Experiments in both simulation and real-world platforms demonstrate superior capabilities in long-horizon autonomous planning, fine-grained instruction following and in-context adaptation.

Date: July 20, 2026
Webpage: https://manifoldai-research.github.io/WorldScape-Policy/

1 Introduction

Robotic learning is shifting from task-specific policies toward generalist embodied models that can interpret task intent, anticipate action consequences, and adapt to new tasks from contextual cues [1, 2]. Vision-Language-Action (VLA) models [3–5] have demonstrated the promise of mapping language and perception to executable robot actions, but they typically emphasize direct policy prediction rather than how the visual world evolves under those actions. World Action Models (WAMs) provide a complementary paradigm by jointly modeling visual state transitions and robot actions. This coupling provides a natural basis for action planning and opens a viable pathway toward context-conditioned adaptation. However, transforming a WAM from a short-horizon predictive model into a controllable long-horizon policy requires addressing three interrelated questions: what historical information should be retained, how atomic actions should be grounded through fine-grained supervision, and how task intent can be expressed beyond language alone.

Despite remarkable progress, current WAMs remain limited along these three dimensions, as illustrated in Fig. 1. **Limited progress-aware memory.** Most existing methods condition world-action modeling on either the current observation [6–9] or a short history window [10, 11]. Such limited context obscures global task progress when different stages share similar visual observations. For example, a tabletop may appear nearly identical before and after an intermediate subgoal, although the correct next action depends on what has already been completed. MemoryWAM [12]

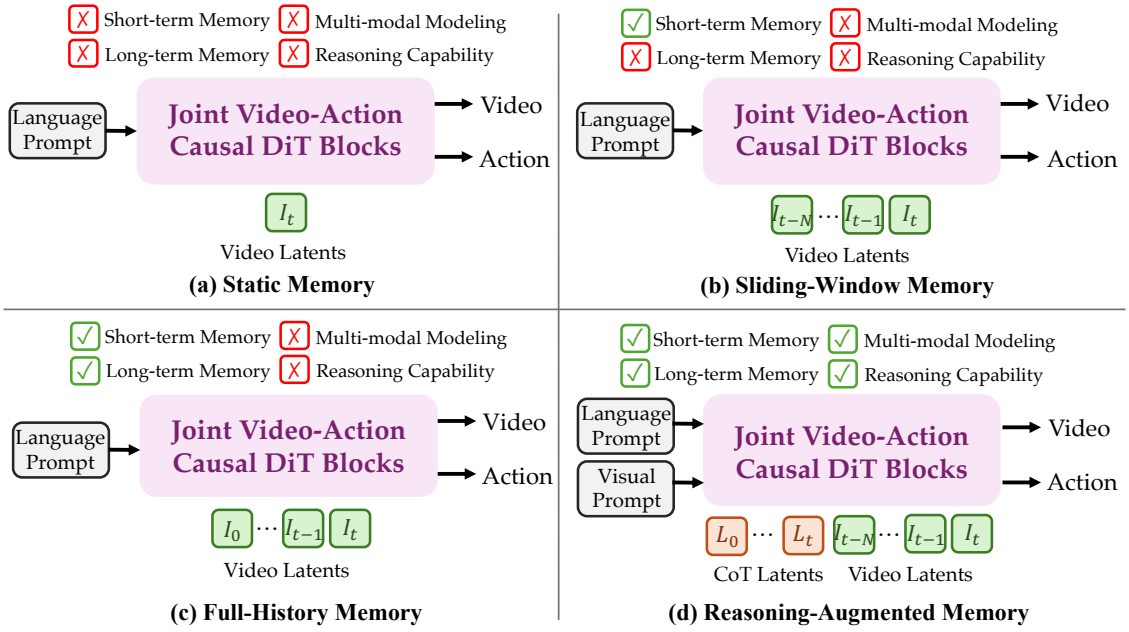

Figure 1: Comparison of different WAM paradigms. Existing models use (a) static observations, (b) short-term visual history, or (c) full-history visual memory without progress-aware reasoning or multimodal control. (d) WorldScape Policy 2.0 introduces multimodal controllability and reasoning-augmented long short-term memory, enabling interactive video-action modeling for long-horizon robotic manipulation.

expands access to the full visual history through an efficient hybrid memory mechanism. However, retrieving past visual evidence alone does not explicitly organize history around task progress or convert it into a latent subgoal condition for autonomous planning.

Coarse language-video-action grounding. In many embodied datasets, different segments of an episode inherit the same task-level instruction, providing weak linguistic supervision for diverse atomic actions. Models can therefore rely on visual shortcuts instead of learning how an event-level intent aligns with a specific state transition and action chunk. Consequently, a model may generate plausible video-action trajectories while remaining insensitive or brittle to fine-grained changes in the requested atomic action.

Restricted interaction interfaces. Existing WAMs are also largely controlled through text, whereas users may express intent more naturally through a goal image, a demonstration video, or an example from another embodiment. Without native visual prompting, a WAM cannot fully exploit target-state evidence, object correspondences, or latent task rules, limiting both visual-context reasoning and cross-embodiment transfer.

Our key insight is that long-horizon controllability requires two complementary forms of temporal context: semantic event memory for reasoning about task progress and frame-level visual memory for preserving local interaction dynamics. Based on this insight, we propose **WorldScape Policy 2.0**, a controllable WAM with reasoning-augmented long short-term memory. Its VLM branch organizes historical chunks into global-history, local-active, and event-boundary memory views, then retrieves relevant evidence to augment perception and autoregressive planning tokens of current observation. In parallel, its causal DiT retains recent observations as short-term visual prefill and optional visual prompts as persistent context. This division enables event memory to infer the next subgoal while visual memory preserves evolving interaction dynamics.

To provide scalable supervision, we construct **ManipEvent-5M**, an event-based multimodal dataset containing nearly 5 million segments aligned with action trajectories, episode-level task instructions, event-level subtask captions, goal images, and video demonstrations. Our three-stage curriculum first establishes fine-grained controllability and causal visual memory through event-grounded pretraining, then introduces the event memory branch which transfers event semantics into autonomous latent planning through semantic forcing, and finally adapts the model to downstream embodiments and various interaction modes. The resulting joint video-action model supports autonomous planning

from high-level instructions, direct control from fine-grained captions, and in-context adaptation from visual prompts.

Our main contributions are summarized as follows:

- We propose **WorldScape Policy 2.0**, a controllable WAM that couples progress-aware long short-term event memory with causal short-term visual memory for implicit subgoal planning in long-horizon manipulation.
- We introduce **ManipEvent-5M**, an event-grounded embodied pretraining dataset containing nearly 5 million segments with aligned action trajectories, hierarchical text instructions, goal images, and video demonstrations, enabling fine-grained multimodal WAM interactions.
- We develop a three-stage learning strategy that transfers event-level semantics from controllable pretraining to autonomous planning through semantic forcing and supports a unified interface for high-level planning, atomic instruction following, and visual-prompted adaptation. Experiments on simulation benchmarks and a dual-arm PiPER platform evaluate these capabilities and the contribution of each design component.

2 Related Work

2.1 World Action Models for Robotic Manipulation

World Action Models (WAMs) leverage visual dynamics for robot action planning, offering a dynamics-centric alternative to reactive observation-to-action policies. Early world models and video-prediction policies use future observations, visual goals, or latent states as intermediate planning representations [13–25]. V-JEPA 2-AC [26] further demonstrates that an action-conditioned latent world model, post-trained from large-scale video representations with limited robot data, can support image-goal planning. Complementary approaches use world models as scalable embodied-data engines, as in GigaWorld-0 [27], or directly adapt pretrained video models into policies, as in Cosmos Policy [28], which represents actions, future states, and values within the video latent space.

Recent WAMs move toward unified video-action modeling, showing that action-centered or causal video-action models can serve as zero-shot robotic policies [6, 7, 9, 11]. LingBot-VA [10] emphasizes causal autoregressive video-action modeling with persistent KV-cache, while MemoryWAM [12] improves memory efficiency with recent frames, event-boundary anchors, and gist tokens. Beyond holistic video or global-latent prediction, OA-WAM introduces persistent object-addressable slots and jointly predicts object-centric world states and flow-matching action chunks [29]. However, prior WAMs often retrofit text-guided video backbones, use video prediction mainly as training-time supervision, or lack explicit task-progress reasoning [6, 9]. WorldScape Policy 2.0 addresses these gaps through native controllable WAM pretraining and reasoning-augmented long short-term memory.

2.2 Long-Term Memory and Long-Horizon Planning

Long-horizon manipulation requires models to remember completed subgoals and decide what remains to be done. End-to-end approaches such as Long-VLA [30] introduce phase-aware modeling of movement and interaction to improve skill chaining without an explicit planner. Other VLAs expose intermediate planning representations: $\pi_{0.5}$ [3] predicts high-level subtasks before action generation, while $\pi_{0.7}$ [4] uses subtask language, metadata, history, and visual subgoals for long-horizon and language-coached execution. VLA-OS [31] systematically studies language, visual, and hierarchical planning representations, highlighting the benefit of visually grounded intermediate plans.

Memory-oriented methods such as MEM [32] and MemoryVLA [33] retrieve and fuse multi-scale embodied or perceptual-cognitive memory for temporally aware action prediction. LoHo-Manip [34] maintains an explicit done-and-remaining subtask plan together with an updated visual trace under receding-horizon control. Dual-system methods, including Goal2Skill [35] and DSWAM [36], similarly decouple high-level planning from low-level execution by passing subgoals, visual traces, or executable instructions to VLA / WAM executors. NovaPlan [37] combines a closed-loop VLM planner with generated video references and geometrically grounded execution for zero-shot long-horizon manipulation. These works highlight the value of memory and subtask reasoning, but typically rely on VLA policies, external planners, or explicit intermediate plans. WorldScape Policy 2.0 instead integrates recent visual dynamics, event-level progress memory, and latent subgoal reasoning within a controllable WAM.

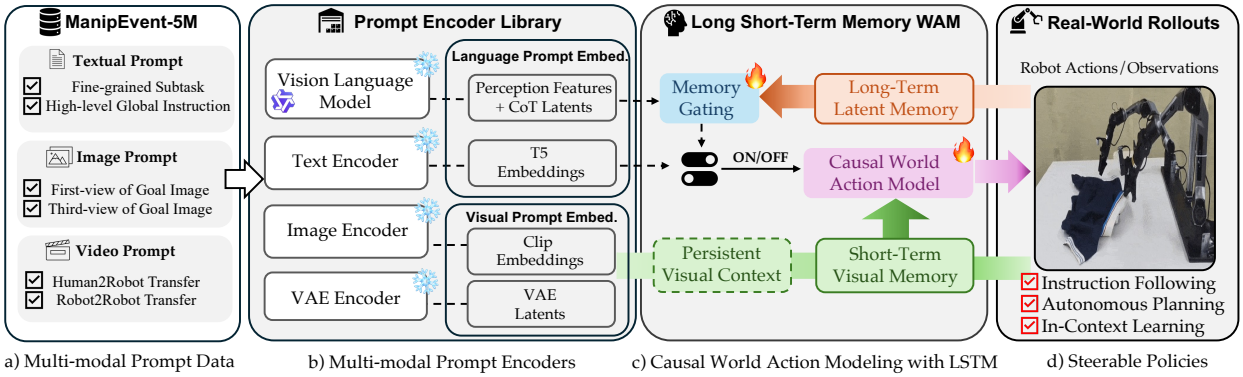

Figure 2: Overview of the WorldScape Policy 2.0 framework. Built on the event-based ManipEvent-5M pretraining data, WorldScape Policy 2.0 encodes observations together with interchangeable text, goal-image, and video-demonstration prompts into a unified causal video-action modeling pipeline. The model couples short-term visual transitions with long-term reasoning-augmented event memory, enabling implicit subgoal planning and joint video-action prediction. This design supports both autonomous planning from coarse task instructions and controllable manipulation from fine-grained subtask captions or visual prompts.

2.3 Fine-Grained Prompts for Steerable Policies

Fine-grained language prompts provide an important control interface for steerable manipulation, since coarse trajectory-level instructions cannot specify diverse atomic actions, object states, execution styles, and intermediate subgoals. At the generalist-model level, Qwen-VLA [38] unifies continuous action and trajectory generation across tasks, environments, and robot embodiments through embodiment-aware prompting, while retaining fine-grained control within the same policy. FineVLA [39] instead isolates language granularity as a supervision axis, aligning goal-level and process-level instructions with robot trajectories to improve factor-level steerability. Related works introduce dense control or grounding signals via pixel-level visual prompts [40], 3D spatial action representations [41], reconstruction-based visual grounding [42], and auxiliary spatial co-training [43].

Visual prompting provides a complementary interface for specifying where and how to act. TraceVLA [44] renders state-action trajectories as visual traces to improve spatiotemporal grounding, whereas CoT-VLA [45] autoregressively predicts future subgoal images as visual reasoning steps before action generation. Human-demonstration-conditioned policies [46] further use a single human video as a task prompt and align human and robot representations for cross-embodiment transfer. Explicit subtask decomposition can also provide fine-grained executable instructions to WAM executors [36]. WorldScape Policy 2.0 unifies these complementary directions through event-grounded WAM pretraining, aligning fine-grained subtask captions, goal images, and video demonstrations with visual state transitions and atomic action trajectories in ManipEvent-5M.

3 Method

3.1 Overview and Problem Formulation

WorldScape Policy 2.0 is a controllable World Action Model designed to support autonomous planning and interactive manipulation within a shared video-action backbone. Let o_t denote the multi-view observation at time t, a_t the corresponding robot action, and z_t the VAE-encoded visual latent of o_t. We distinguish two language interaction modes, $\mu \in \{\text{auto}, \text{fine}\}$, whose language inputs are an episode-level instruction y and an event-level instruction c_t, respectively. The corresponding prompt set is $\mathcal{P}_t^{\mu} = \{p_{\text{lang}}^{\mu}, p_{\text{goal}}, p_{\text{video}}\}$, where $p_{\text{lang}}^{\text{auto}} = y$ and $p_{\text{lang}}^{\text{fine}} = c_t$. Historical context is summarized by $\mathcal{M}_t = (\mathcal{Q}_t, \mathcal{Z}_t^{\text{vis}})$, comprising the event-memory queue $\mathcal{Q}_t$ and the recent visual-latent buffer $\mathcal{Z}_t^{\text{vis}}$. The event queue is activated for autonomous planning during the mid-training stage (see Sec. 3.6) and is empty when event memory is disabled, whereas the visual buffer is used throughout all training stages. The model predicts an

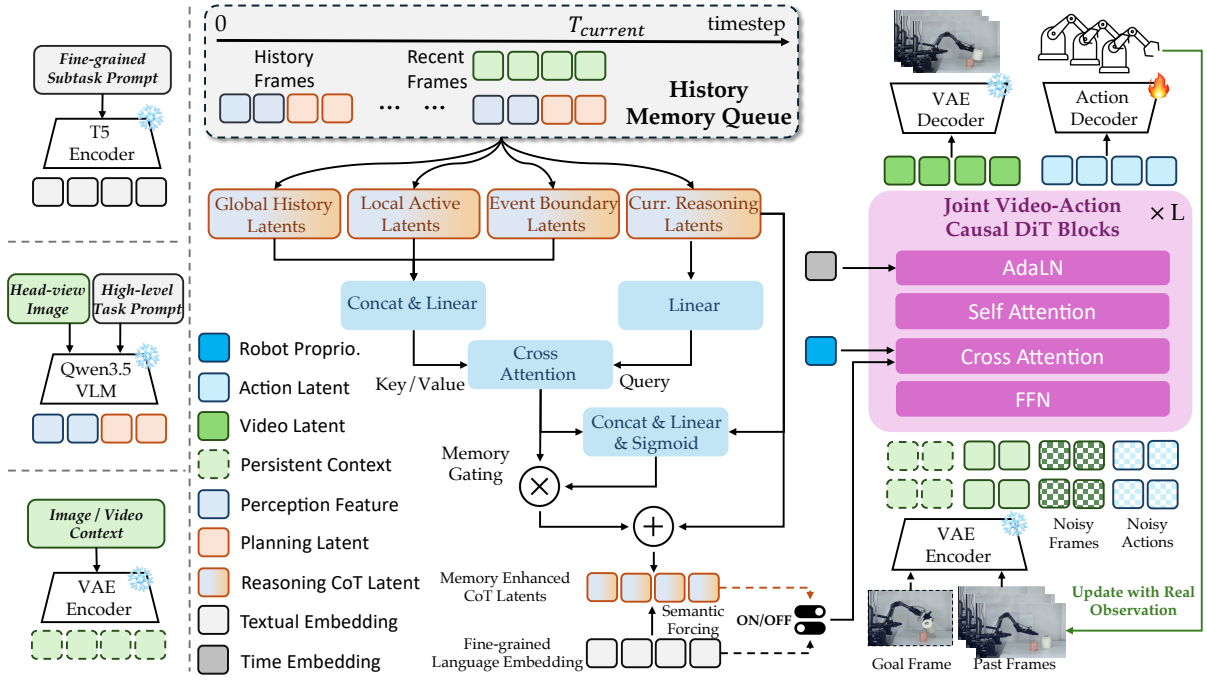

Figure 3: Illustration of Reasoning-Augmented Long Short-Term Memory. Historical VLM outputs form global-history, local-active, and event-boundary memories, which are retrieved and gated to augment the current reasoning tokens; semantic forcing aligns them with subtask semantics. Recent observations and optional visual prompts enter the DiT as causal visual prefill, while memory-enhanced VLM tokens or T5 embeddings condition joint video-action flow prediction according to the interaction mode.

H-step action chunk and H future visual latents as:

$$p_\theta(A_t^{(e)}, z_{t+1:t+H} \mid o_t, a_{<t}, \mathcal{P}_t^{\mu}, \mathcal{M}_t). \tag{1}$$

(1) Unified action representation. For robot embodiment e, actions are expressed relative to the first frame of each action chunk. For a chunk beginning at time t and future offset $\tau \in \{1, \ldots, H\}$, each per-arm action is

$$a_{t,\tau}^{(e)} = \left[\Delta p_{t,\tau}^{(e)}; \phi_6\left(\Delta R_{t,\tau}^{(e)}\right); g_{t+\tau}^{(e)}\right] \in \mathbb{R}^{10}, \quad \Delta p_{t,\tau}^{(e)} = p_{t+\tau}^{(e)} - p_t^{(e)}, \quad \Delta R_{t,\tau}^{(e)} = \left(R_t^{(e)}\right)^\top R_{t+\tau}^{(e)}. \tag{2}$$

Here, $p_s^{(e)} \in \mathbb{R}^3$ and $R_s^{(e)} \in \mathrm{SO}(3)$ are the Cartesian position and orientation at time s, and $\phi_6 : \mathrm{SO}(3) \to \mathbb{R}^6$ converts the relative rotation to its continuous 6D representation. Thus, the position and rotation encode a delta pose relative to the chunk's first frame, whereas $g_{t+\tau}^{(e)} \in \mathbb{R}$ is the absolute one-DoF gripper command. We collect the future controls as $A_t^{(e)} = (a_{t,\tau}^{(e)})_{\tau=1}^{H} \in \mathbb{R}^{H \times d_a}$, where $d_a = 20$ for a dual-arm embodiment formed by concatenating two 10D per-arm vectors.

(2) Embodiment-specific action adapters. Each embodiment has a dedicated action encoder $E_a^{(e)}$ and action decoder $D_a^{(e)}$:

$$\widetilde{A}_\rho^{(e)} = E_a^{(e)}(A_\rho^{(e)}), \qquad \hat{v}_\theta^{A,(e)} = D_a^{(e)}(h_\theta^{A,(e)}), \tag{3}$$

where $E_a^{(e)}$ projects a flow-perturbed raw action chunk into the shared DiT token space and $D_a^{(e)}$ maps the resulting action hidden states back to a velocity field in the raw action space. The causal video-action DiT is shared across embodiments, while these adapters absorb embodiment-specific kinematics, action scales, and control conventions. Importantly, the adapters only provide the input and output interfaces: flow matching and flow integration are both performed on the raw action chunk rather than on $\widetilde{A}_\rho^{(e)}$.

As shown in Fig. 2, language and visual prompts follow distinct conditioning paths. Fine-grained instructions are encoded by T5, high-level instructions are interpreted jointly with the current observation by the VLM reasoning branch, and visual prompts are encoded as persistent VAE prefill. The resulting mode-specific conditions are integrated with event and visual memory before the DiT jointly predicts future video and action chunks.

3.2 Event-Grounded World Action Modeling

Existing WAMs are often trained with coarse episode-level instructions or a single interaction modality, providing weak supervision for the diverse state changes and atomic actions inside long manipulation episodes. WorldScape Policy 2.0 instead grounds WAM learning at the event level, where each segment aligns fine-grained subtask semantics with visual transitions and paired robot actions. This event-level structure supports two language interaction modes and complementary visual prompting. A fine-grained subtask instruction is encoded by T5 and directly controls the WAM, whereas a high-level task instruction is interpreted by the VLM branch to support autonomous subgoal planning. Goal images and video demonstrations provide additional persistent visual context. We formulate the encoded prompt as:

$$s_t = E_{\mathrm{T5}}(c_t), \quad z_t^{\mathrm{goal}} = E_{\mathrm{VAE}}(p_{\mathrm{goal}}), \quad v_t = E_{\mathrm{VAE}}(p_{\mathrm{video}}), \tag{4}$$

where c_t is a fine-grained subtask instruction, s_t is its T5 embedding, and z_t^{goal} and v_t are clean VAE latents of goal images and video demonstrations. The persistent visual condition is selected according to the available visual prompt:

$$p_t^{\mathrm{vis}} = \begin{cases} z_t^{\mathrm{goal}}, & \text{goal-image prompting,} \\ v_t, & \text{video-demonstration prompting,} \\ [z_t^{\mathrm{goal}}; v_t], & \text{both prompts available,} \\ \varnothing, & \text{otherwise.} \end{cases} \tag{5}$$

The language condition is likewise selected by μ, rather than concatenating the T5 and VLM branches into a single input.

Mode-dependent language grounding. In fine-grained instruction following, the segment-level caption c_t explicitly specifies the atomic action or intermediate subgoal. Its T5 embedding s_t is directly injected into the DiT through cross-attention, enabling precise and steerable action execution. In autonomous planning, the robot receives only an episode-level high-level instruction y. The VLM jointly interprets y, the current observation, and historical event memory to infer the active subgoal; the resulting memory-enhanced VLM tokens, rather than a T5 embedding of y, condition the DiT. Event-level captions supervise this latent planning pathway through semantic forcing during training.

Goal-conditioned policy learning. Goal image prompts specify the desired target state visually. WorldScape Policy 2.0 encodes the goal image as z_t^{goal} and retains it in the causal visual prefill, allowing flow-perturbed future chunks to attend to the desired state throughout generation. The model is thus trained to infer an executable action path from the current observation to the visual goal.

Demonstration-conditioned in-context adaptation. Video context prompts introduce cross-embodiment adaptation tasks, where the context may come from human-to-robot demonstrations, robot-to-robot demonstrations, or previous contextual episodes with different embodiments and viewpoints. WorldScape Policy 2.0 represents these demonstrations as v_t and retains them as persistent visual prefill context alongside recent observations. This training mode encourages the model to infer object correspondences, transferable action patterns, and latent task rules from visual demonstrations, directly matching the video-context annotations provided by ManipEvent-5M dataset introduced in Sec. 3.5.

3.3 Reasoning-Augmented Long Short-Term Memory

Long-horizon manipulation requires memory at two complementary representation levels. At the event level, the model must retain semantic evidence about completed subtasks and retrieve history relevant to the current task stage. At the visual level, it must preserve recent frame-level dynamics, such as contact changes and object motion, for temporally coherent video-action prediction. As illustrated in Fig. 3, WorldScape Policy 2.0 therefore combines a *long short-term event memory* in the VLM reasoning branch with a *short-term visual memory* in the causal DiT branch. The former performs progress-aware retrieval over historical chunk representations, whereas the latter directly supplies recent visual latents as causal prefill context. Optional goal images and video demonstrations are further retained as persistent visual context throughout the rollout.

Latent reasoning formulation. For each action chunk under autonomous planning, perception and planning share a single VLM context rather than using separate forward passes. The context combines the current head-view observation o_t^{head} with a task-conditioned planning prompt $\widetilde{y}_t = T_{\mathrm{plan}}(y)$. We instantiate T_{plan} with the following fixed template, where `{task}` is replaced by the episode-level instruction y:

"You are a robot planner. Instructions: {task}. Given the current high-level task instruction and current head-view observation, predict the next atomic action subtask for the next second."

A single VLM prefill jointly encodes $(o_t^{\text{head}}, \widetilde{y}_t)$ and produces the final-layer context hidden states H_t^{ctx} together with their KV cache. The perception tokens are read directly from this VLM output:

$$u_t = H_t^{\text{ctx}} = H_{\text{VLM}}^{L}(o_t^{\text{head}}, \widetilde{y}_t). \tag{6}$$

Starting from the same cached context, the VLM then continues autoregressively and generates K discrete planning tokens using greedy decoding:

$$w_t^k = \arg\max_{w} \; p_{\text{VLM}}(w \mid o_t^{\text{head}}, \widetilde{y}_t, w_t^{<k}), \quad k = 1, \ldots, K. \tag{7}$$

The final-layer hidden state returned online at each decoding step is retained as the corresponding latent planning token:

$$r_t^k = \left[H_{\text{VLM}}^{L}(o_t^{\text{head}}, \widetilde{y}_t, w_t^{1:k})\right]_{\text{pos}(w_t^k)}, \quad k = 1, \ldots, K. \tag{8}$$

Thus, u_t and $r_t^{1:K}$ belong to one continuous VLM inference trajectory: u_t encodes the shared visual-language prefill, while $r_t^{1:K}$ summarizes its autoregressively inferred task continuation. We concatenate the two token groups to form the current reasoning latent

$$q_t = [u_t; r_t^{1:K}], \tag{9}$$

which jointly represents the current state and the inferred next subgoal. Historical observations are processed using the same perception and planning branches, ensuring that every memory entry has the same two-part token structure.

Long short-term event memory construction. The event-memory branch maintains the queue $\mathcal{Q}_t = \{q_1, \ldots, q_{t-1}\}$ of historical chunk-level VLM outputs introduced above. Each q_j contains the perception tokens and latent planning tokens associated with chunk j. To control the cost of full-history retrieval, we compress the perception tokens of every historical chunk into a fixed number of learned gist tokens through attention pooling, while retaining its planning tokens. Concatenating the resulting tokens gives a compact full-history bank H_t. We then explicitly construct three complementary memory views. **Global-History latents** summarize the complete trajectory, **Local-Active latents** represent the currently active event, and **Event-Boundary latents** sparsely preserve major semantic transitions:

$$\begin{aligned} m_t^{\text{gh}} &= P_{\text{gh}}\big(\text{Expand}\left(F_{\text{gh}}([e_y; \text{Mean}(H_t)])\right)\big), \\ m_t^{\text{la}} &= P_{\text{la}}([q_j \mid j \in \mathcal{I}_t^{\text{la}}]), \quad \mathcal{I}_t^{\text{la}} = \{\max(1, t - S_e), \ldots, t-1\}, \\ m_t^{\text{eb}} &= P_{\text{eb}}([q_j \mid j \in \mathcal{B}_t]), \end{aligned} \tag{10}$$

Here, e_y is the high-level instruction embedding, Mean averages over the token dimension of the compact history bank H_t, F_{gh} fuses task intent with pooled history, and Expand produces a fixed number of global-history slots. The learned projections P_{gh}, P_{la}, and P_{eb} map the three memory views into a shared feature space. The notation $[q_j \mid j \in \mathcal{S}]$ denotes temporally ordered token concatenation over index set $\mathcal{S}$; $\mathcal{I}_t^{\text{la}}$ contains the latest S_e completed chunks, where S_e is the local-active window length, and $\mathcal{B}_t$ is the event-boundary index set defined below. The recent local-active and selected boundary chunks are retained as full-token anchors, preserving fine-grained perception and planning evidence.

Event-boundary chunks are selected directly from latent changes in the historical VLM outputs. Let $\bar{q}_j = \text{Mean}(q_j)$ denote the token-averaged representation of chunk j. We compute

$$d_j = 1 - \cos(\bar{q}_j, \bar{q}_{j-1}), \qquad \mathcal{B}_t = \text{TopK}_{\Delta}(\{d_j\}_{j=2}^{t-1}; S_b), \tag{11}$$

where d_j measures the semantic change between consecutive chunks. TopK_{Δ} greedily selects at most S_b highest-change indices while enforcing a minimum temporal separation Δ between selected indices. The resulting event-boundary latents sparsely retain states at which a subtask likely starts, terminates, or changes, without requiring explicit online boundary annotations. As with local-active memory, the selected chunks are stored as full-token anchor frames, allowing the model to revisit detailed perception and planning evidence at these critical transitions rather than only their pooled boundary scores. Consequently, event memory combines long-term task-progress context from global-history and event-boundary latents with short-term semantic context from local-active latents. The anchor-frame implementation preserves local detail within this semantic memory organization, which remains distinct from the frame-level visual memory described below.

Event-memory retrieval and gated fusion. Following the implementation, the three specialized memory views are concatenated with the compact full-history bank, ensuring that selection does not discard non-boundary evidence. Let

$$B_t = \mathrm{cat}_{\mathrm{tok}}\left(m_t^{\mathrm{gh}}, m_t^{\mathrm{la}}, m_t^{\mathrm{eb}}, H_t\right) \tag{12}$$

denote the concatenated memory tokens. For $q_t \in \mathbb{R}^{N_q \times d_{\mathrm{VLM}}}$ and $B_t \in \mathbb{R}^{N_m \times d_{\mathrm{VLM}}}$, we form independent query, key, and value projections and retrieve memory by

$$\begin{aligned} Q_t &= P_q(q_t), \qquad K_t = P_k(B_t), \qquad V_t = P_v(B_t), \\ \mathcal{A}_t^{\mathrm{mem}} &= \mathrm{softmax}_{\mathrm{mem}}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}} + M_t\right), \\ \widetilde{m}_t &= P_o\left(\mathcal{A}_t^{\mathrm{mem}} V_t\right) \in \mathbb{R}^{N_q \times d_{\mathrm{VLM}}}, \end{aligned} \tag{13}$$

where $Q_t \in \mathbb{R}^{N_q \times d_k}$, $K_t \in \mathbb{R}^{N_m \times d_k}$, $V_t \in \mathbb{R}^{N_m \times d_v}$, $M_t \in \mathbb{R}^{N_q \times N_m}$ masks padded history and boundary slots, and $P_o : \mathbb{R}^{d_v} \to \mathbb{R}^{d_{\mathrm{VLM}}}$ maps the attended values back to the VLM hidden dimension. Because historical evidence is not equally useful at every step, a learned gate controls how much retrieved memory is injected into each current token:

$$\gamma_t = \sigma\left(W_g\,[q_t; \widetilde{m}_t] + b_g\right), \qquad \hat{q}_t = q_t + \alpha\,\gamma_t \odot \widetilde{m}_t, \tag{14}$$

where σ is the sigmoid function, $W_g : \mathbb{R}^{2d_{\mathrm{VLM}}} \to \mathbb{R}$ and b_g are learned gate parameters applied to each token, $[\cdot;\cdot]$ denotes feature-wise concatenation, and $\odot$ denotes element-wise multiplication with broadcasting over the feature dimension. Thus, $\gamma_t \in \mathbb{R}^{N_q \times 1}$ is a token-wise memory-importance gate, α is a residual scaling factor, and $\hat{q}_t$ is the memory-enhanced reasoning latent. This retrieval mechanism allows the model to selectively use task-level progress, recent event context, sparse transition evidence, and the compressed full history.

Visual short-term memory and persistent context. In parallel to event-memory retrieval, the causal DiT models recent frame-level dynamics through visual prefilling. We define the recent visual-latent buffer as $\mathcal{Z}_t^{\mathrm{vis}} = z_{t-S_v:t}^{\mathrm{obs}}$, containing the clean VAE latents of the current observation and the most recent S_v visual chunks. These latents precede the flow-interpolated future video latents and embodiment-encoded action tokens in the DiT sequence:

$$h_t^{\mathrm{DiT},(e)} = [p_t^{\mathrm{vis}}; \mathcal{Z}_t^{\mathrm{vis}}; z_{t+1:t+H}^{\rho}; E_a^{(e)}(A_{\rho,t}^{(e)})], \tag{15}$$

where $\mathcal{Z}_t^{\mathrm{vis}}$ is updated online after each executed action chunk and $E_a^{(e)}$ is selected according to the current embodiment. As summarized in Fig. 4, a causal visual-attention mask allows future video-action tokens to attend to persistent visual prompts and available clean history while preventing information leakage from future chunks. This pathway supplies local motion and interaction continuity through DiT self-attention, rather than first compressing recent frames into the VLM event memory.

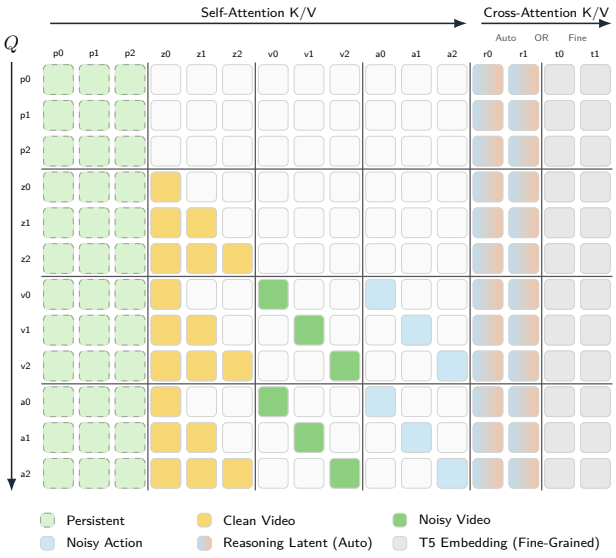

Figure 4: Attention mask of WorldScape Policy 2.0.

The optional prompt latents p_t^{vis} remain fixed across rollout steps, unlike the sliding recent-frame buffer, allowing the model to continuously reference the desired goal state or demonstrated trajectory. Both recent and persistent visual latents enter the DiT through causal self-attention, while the cross-attention condition is selected by interaction mode: T5 embeddings are used for fine-grained instruction following, whereas memory-enhanced VLM tokens are used for autonomous planning. This separation preserves the respective roles of semantic event retrieval, local visual dynamics, and multimodal task conditioning.

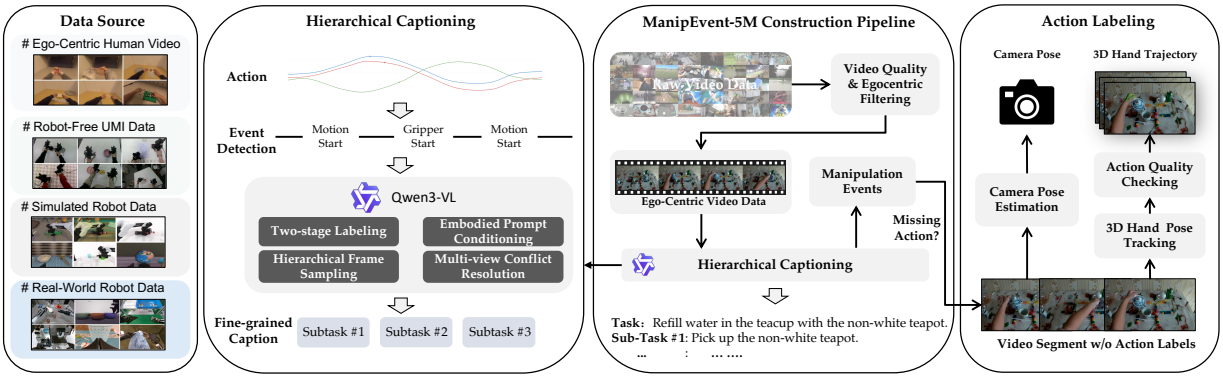

Figure 5: Construction pipeline of ManipEvent-5M. We aggregate heterogeneous manipulation sources, including human-arm ego videos, robot-free UMI data, simulated trajectories, and real-robot demonstrations, and convert them into event-level training samples. Each trajectory is segmented into ordered atomic subgoals and annotated with paired action trajectories, episode-level global instructions, segment-level fine-grained captions, goal-image prompts, and video-context prompts. This event-based multimodal schema supports the event-grounded WAM interface in Fig. 2, enabling fine-grained language grounding, goal-conditioned policy learning, demonstration-conditioned in-context learning, and long-horizon memory-aware mid-training.

3.4 Implicit Subgoal Latent Planning

WorldScape Policy 2.0 uses event-level subtask captions in two different ways. In fine-grained instruction following, the T5 embedding $s_t = E_{\mathrm{T5}}(c_t)$ acts as the direct language condition. In autonomous planning, the caption is used only as training-time semantic supervision for the VLM planning branch, which produces $\hat{q}_t$ from perception, autoregressive planning tokens, and retrieved historical memory.

Subgoal semantic forcing. During training, s_t provides a semantic target for the autonomous-planning latent. Because s_t and $\hat{q}_t$ are token sequences of potentially different lengths, we first obtain fixed-dimensional summaries $\bar{s}_t = \mathrm{Pool}(s_t)$ and $\bar{q}_t = \mathrm{Pool}(\hat{q}_t)$. We keep the T5 target encoder fixed and apply stop-gradient to its normalized representation, while a trainable projector ϕ_q maps the reasoning summary into the T5 semantic space:

$$\begin{aligned} \widetilde{s}_t &= \mathrm{sg}\left(\frac{\bar{s}_t}{\max(\|\bar{s}_t\|_2, \varepsilon)}\right), \\ \widetilde{q}_t &= \frac{\phi_q(\bar{q}_t)}{\max(\|\phi_q(\bar{q}_t)\|_2, \varepsilon)}, \\ \mathcal{L}_{\mathrm{sem}} &= 1 - \widetilde{s}_t^{\top} \widetilde{q}_t, \end{aligned} \tag{16}$$

where sg denotes stop-gradient and $\varepsilon > 0$ prevents division by zero. Fixing the semantic target prevents the alignment objective from drifting toward a jointly collapsed solution, while ϕ_q learns to align the reasoning latent with explicit subtask semantics. During autonomous inference, the fine-grained caption and its T5 embedding are absent; the model receives only the high-level instruction and relies on $\hat{q}_t$ to infer the active subgoal from the current observation and event memory.

Mode-dependent video-action prediction. The language-side cross-attention condition is selected according to the requested interaction mode:

$$\ell_t = \begin{cases} s_t, & \text{fine-grained instruction following,} \\ \hat{q}_t, & \text{autonomous planning,} \end{cases} \tag{17}$$

where s_t provides explicit atomic-action control, while $\hat{q}_t$ provides an implicitly inferred and progress-aware subgoal condition. Let ρ denote the flow timestep embedded through AdaLN. The shared causal DiT produces action and video hidden states, after which the embodiment-specific action decoder and shared video output head predict their respective velocity fields:

$$\begin{aligned} (h_\theta^{A,(e)}, h_\theta^{z}) &= F_\theta(h_t^{\mathrm{DiT},(e)}, \ell_t, \rho), \\ \hat{v}_\theta^{A,(e)} &= D_a^{(e)}(h_\theta^{A,(e)}), \qquad \hat{v}_\theta^{z} = D_z(h_\theta^{z}), \end{aligned} \tag{18}$$

Table 1: Statistics of WorldScape Policy 2.0 pretraining data in ManipEvent-5M. The number of segments is reported for datasets with event-level decomposition. "Avg. Seg. / Episode" is the number of event segments divided by the number of episodes, and "Single-Seg. Ratio" is the proportion of episodes without further event decomposition.

Data Type	Dataset Source	Frames (M)	Duration (H)	Episodes (K)	Segments (M)	Avg. Seg. / Episode	Single-Seg. Ratio (%)	Training Ratio (%)
Real-World Robot Data	Self-Collected PiPER Dataset	54.00	506.60	35.82	0.56	15.6	0.5	45.0
Public Robot Data	AgiBot World [47]	285.56	2644.08	160.15	1.16	7.2	0.2	32.53
	RoboMIND [48]	6.56	60.75	10.27	–	–	100.0	0.75
	RoboCOIN [49]	32.00	302.94	44.27	–	–	100.0	3.65
	DROID [50]	27.00	500.82	92.23	–	–	100.0	3.07
Robot-Free UMI Data	Self-Collected UMI Dataset	9.50	90.60	26.02	0.06	2.3	30.8	8.0
Simulated Robot Data	RoboTwin 2.0 [51]	8.16	42.16	35.73	0.145	4.1	26.8	3.94
	LIBERO [52]	0.12	3.86	1.71	0.005	2.9	0.12	0.06
Ego-Centric Human Data	EgoDex [53]	89.24	831.00	338.23	2.96	8.8	18.8	3.0
Total		**512.14**	**4982.81**	**744.43**	**4.89**	**6.6**	–	–

where F_θ denotes the shared causal video-action DiT, D_z is the video output head, and visual prompts, when available, are already included in $h_t^{\mathrm{DiT},(e)}$ as persistent prefill context. At inference, the video field is integrated in VAE-latent space, whereas $\hat{v}_\theta^{\mathrm{A},(e)}$ is integrated directly in the raw action space from $\rho = 1$ (noise) to $\rho = 0$ (data). At each flow step, the current raw action chunk is re-encoded by $E_a^{(e)}$ before the next velocity evaluation. In autonomous planning, the model repeatedly updates event and visual memory, infers $\hat{q}_t$, predicts the next chunk, and rolls forward with new observations. In fine-grained instruction following, the externally specified s_t directly steers each atomic action without requiring the model to infer the subgoal from the high-level instruction alone.

3.5 ManipEvent-5M: Event-Based Multimodal Dataset

To facilitate native pretraining of WorldScape Policy 2.0, we construct **ManipEvent-5M**, the event-based multimodal dataset illustrated in Fig. 5. It aggregates human-arm egocentric videos, robot-free UMI demonstrations, simulated trajectories, and real-robot data from self-collected and public sources. Each trajectory is decomposed into an ordered sequence of events:

$$\tau = \{(o_{t_b^j:t_e^j}, a_{t_b^j:t_e^j}, y, c_j, p^j_{\mathrm{goal}}, p^j_{\mathrm{video}})\}_{j=1}^{N_\tau}, \tag{19}$$

where (t_b^j, t_e^j) is the temporal boundary of event j, y is the episode-level instruction, c_j is its fine-grained event caption, and p^j_{goal} and p^j_{video} are optional visual prompts. This schema separates global task intent from event-level subgoal semantics while preserving their alignment with observations and actions.

Action canonicalization and human-hand retargeting. Robot trajectories are converted to the Cartesian end-effector representation defined in Sec. 3.1. For human data, estimated hand poses are retargeted to robot gripper poses and commands before pretraining, yielding action-aligned human demonstrations under the same position–rotation–gripper semantics. This canonical layout enables joint pretraining across human and robot data, while the embodiment-specific action encoders and decoders preserve distinct kinematic mappings and control conventions.

Text-prompt construction. We convert each trajectory into an episode-level task description and temporally aligned subtask captions using a four-stage hierarchical annotation procedure with Qwen3-VL [54], as illustrated in Fig. 6.

(1) Two-stage labeling. We first establish temporal structure from robot-state signals and then perform open-vocabulary semantic relabeling with the VLM. For temporal segmentation, embodiment-normalized translational and rotational velocities are computed from end-effector or hand trajectories and smoothed. Motion-state transitions, gripper-state changes, and episode endpoints form candidate boundaries. Short segments are merged with adjacent events, and an adaptive granularity constraint prevents both semantically overloaded segments and fragments caused by sensor jitter. Each resulting event is assigned a coarse primitive prior, such as move, grasp / release, or idle, based on motion and

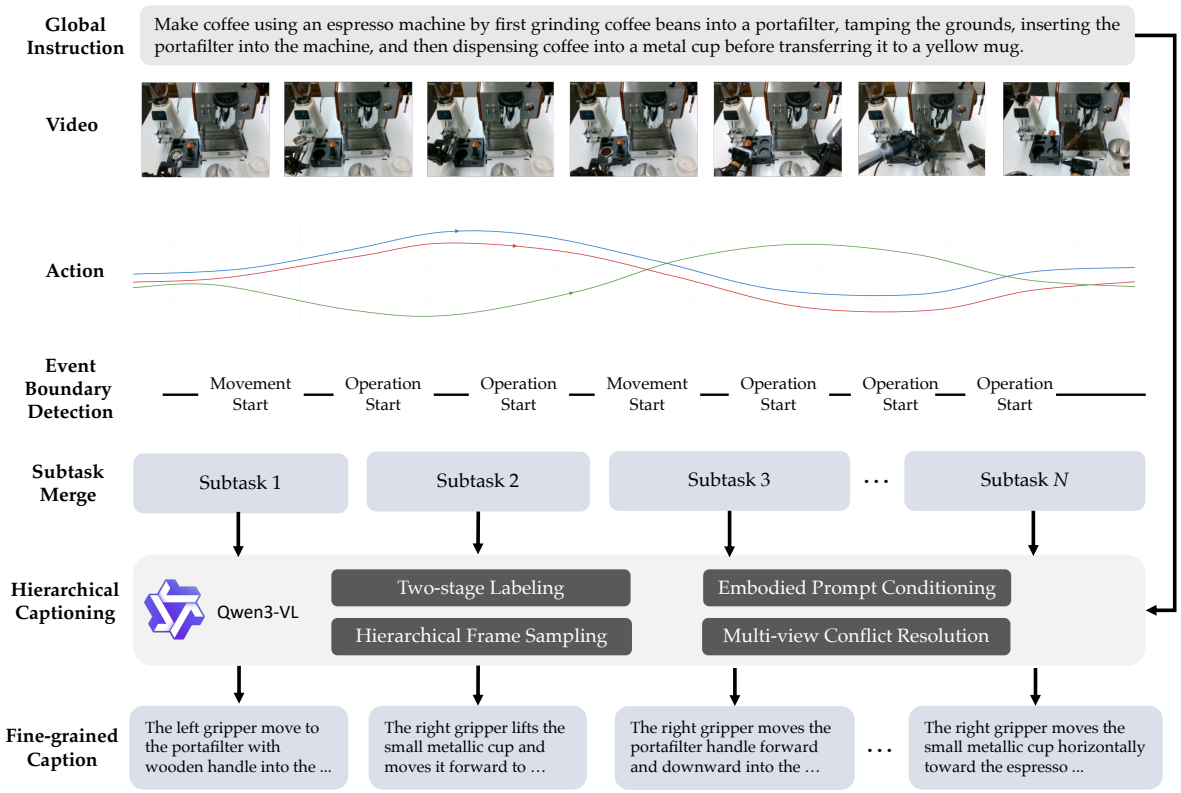

Figure 6: Examples of hierarchical captioning in ManipEvent-5M. The pipeline combines two-stage temporal-semantic labeling, hierarchical episode / event frame sampling, embodied prompt conditioning, and multi-view conflict resolution. Representative instances map segmented events to fine-grained captions, complementing the construction pipeline in Fig. 5.

gripper signals. Given these boundaries, the VLM determines the semantic content of the trajectory and each event. This two-stage design decouples *when* an action changes from *what* the action means, reducing missed steps, ordering errors, and temporal misalignment that arise when directly captioning a long video.

(2) Hierarchical frame sampling. Captioning is performed at both episode and event levels. At the episode level, we sample up to 16 representative timestamps from the full trajectory and provide their images together with the ordered event list, temporal boundaries, primitive-action priors, and execution outcome. Qwen3-VL returns a structured `high_level` description that summarizes the task actually completed, i.e., a hindsight task description rather than a verbatim copy of the original command. This makes the label robust to missing instructions, execution deviations, and recovery attempts. At the event level, we uniformly sample up to 12 timestamps within each segment and collect their available camera views. Conditioned on the episode-level `high_level`, event boundaries, primitive prior, and local visual evidence, the VLM generates a single-action `step_text`. Separating global summarization from local captioning binds every instruction to a specific event while avoiding omissions and ordering errors in long action sequences.

(3) Embodied prompt conditioning. We translate the annotation guidelines into structured prompt constraints rather than requesting an unconstrained video summary. The prompt supplies the temporal metadata and primitive-action prior, and encourages each caption to identify the action order, acting end effector, target object, initial and terminal states, contact mode, motion direction, object interaction, body motion, and, when present, failure or recovery behavior. These constraints increase semantic density while retaining a natural-language action sentence; for example, the caption specifies which gripper interacts with which object and toward which target, rather than producing a generic description such as "move the cup."

(4) Multi-view conflict resolution. For event-level annotation, multi-view observations are presented with a local-first organization, `LEFT_GRIPPER` → `RIGHT_GRIPPER` → `HEAD`. The head camera provides global scene and task context, whereas wrist / gripper cameras offer more reliable evidence about contact, grasp state, and local target identity. The

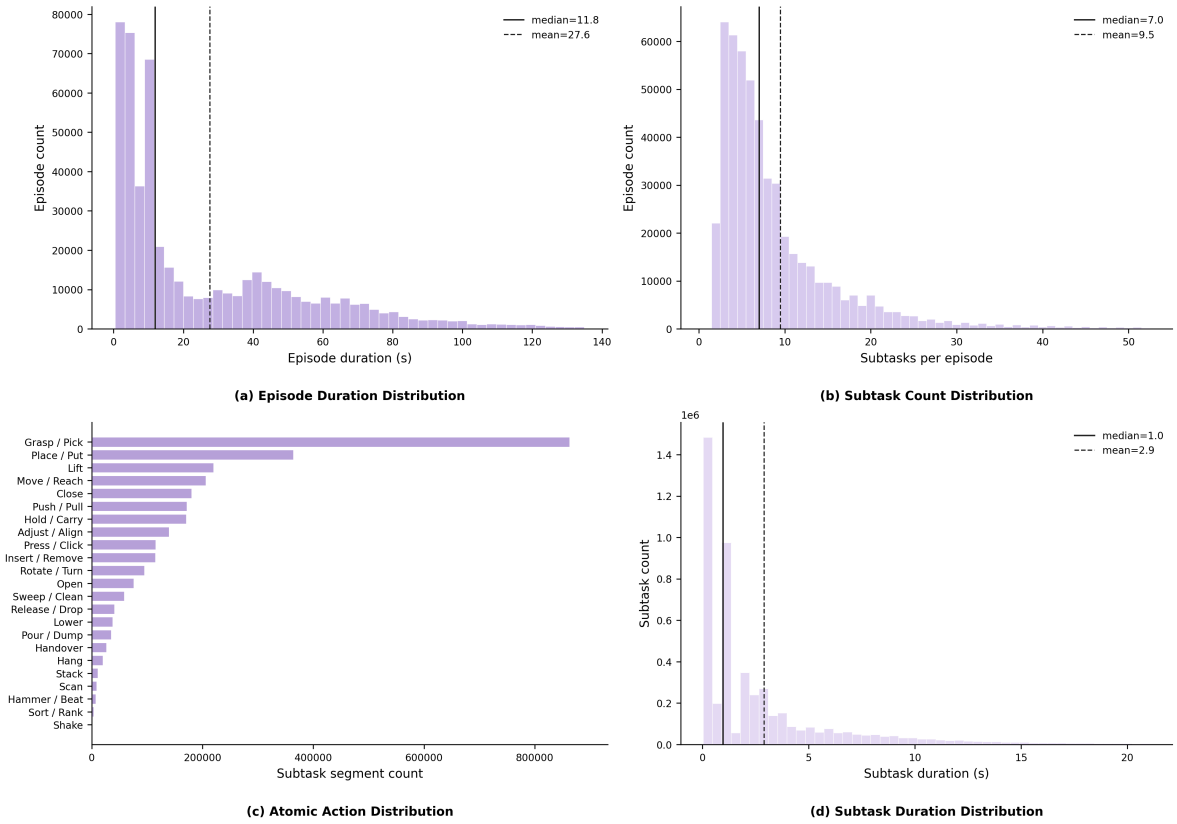

Figure 7: Distribution statistics of ManipEvent-5M. The figure shows the detailed distribution of episode duration, subtask count, atomic actions, and subtask duration across diverse sources from real-world / simulated robot data, robot-free data, and ego-centric human data in the constructed ManipEvent-5M dataset.

prompt therefore instructs the VLM to prioritize gripper-view evidence when it conflicts with the head view, mitigating target-object confusion in cluttered scenes. Finally, outputs are parsed and validated as structured records; invalid episode captions trigger a stricter retry or rule-based fallback, while a failed event caption is replaced locally without discarding valid labels from the remaining trajectory.

As summarized in Table 1 and Fig. 7, ManipEvent-5M combines large-scale visual coverage with event-level subtask decomposition across heterogeneous data sources. Beyond fine-grained text annotations, we construct goal-image and video-demonstration prompts to support visual goal specification and cross-embodiment in-context adaptation.

Goal-image prompt construction. We construct two complementary types of goal-image prompts. A *first-view goal image* is directly extracted from the terminal observation of each segmented event, depicting the desired post-condition from the same camera configuration and embodiment as the corresponding action trajectory. It therefore provides an unambiguous visual target for learning how the current state should transition after an atomic action. A *third-view goal image* is sampled from a task-matched human demonstration at the completion of the corresponding event. Such images retain the human demonstrator and show the target object configuration from an external viewpoint. They require the policy to abstract away the demonstrator and viewpoint, infer the intended outcome, and reproduce the demonstrated state using the robot embodiment.

Video-prompt construction. Video prompts are designed primarily for human-to-robot and robot-to-robot task transfer. For human-to-robot transfer, we pair self-collected real-robot trajectories with UMI demonstrations collected for the same tasks. Pairs are matched by the global task, ordered event sequence, and terminal outcome, ensuring that the human and robot demonstrations express consistent action procedures and reducing ambiguity about which skill should be transferred. For robot-to-robot transfer, we collect or retrieve trajectories of the same semantic task executed by different robot embodiments and align them using their ordered event descriptions and goal states. The resulting

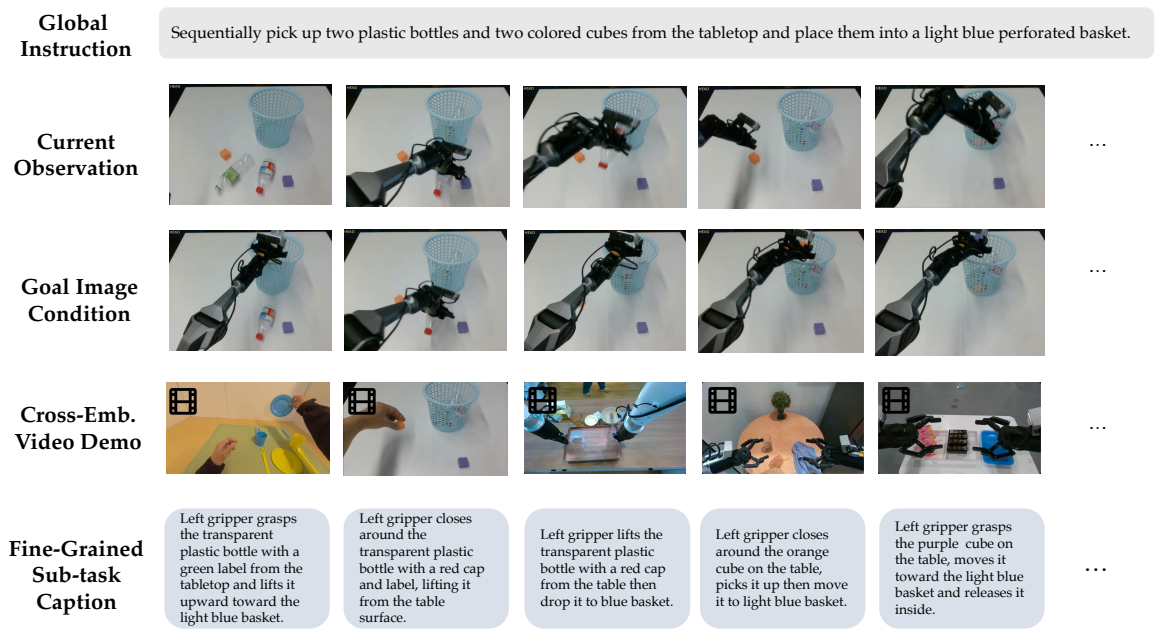

Figure 8: Multimodal prompt example from ManipEvent-5M: `desktop cleaning`. The sample pairs a global task instruction and current observations with goal-image conditioning, event-level fine-grained subtask captions, and a cross-embodiment video demonstration for placing bottles and colored cubes into a basket.

context-target pairs encourage the model to identify embodiment-invariant object correspondences, action structure, and task intent while adapting execution to the target robot.

Figs. 8 and 9 show representative multimodal prompt samples for `desktop cleaning` and long-horizon `coffee making`, respectively. Together, this event-based multimodal organization converts long demonstrations into ordered subgoal segments and aligns language instructions, visual goals, demonstration context, state transitions, and robot actions within a common training interface.

3.6 Training Objectives and Three-Stage Curriculum

WorldScape Policy 2.0 is optimized with a joint video-action objective and, when autonomous planning is enabled, the semantic-forcing objective is further adopted:

$$\mathcal{L} = \mathcal{L}_{act} + \lambda_w \mathcal{L}_{world} + \lambda_s \mathcal{L}_{sem}. \tag{20}$$

Here, $\mathcal{L}_{act}$ and $\mathcal{L}_{world}$ are flow-matching objectives for action and future visual-latent generation, respectively, and $\mathcal{L}_{sem}$ follows Eq. 16; λ_s is 0.001 when semantic forcing is active and zero otherwise. Multimodal conditioning and memory gating are trained end to end through these objectives without separate prompt or memory labels. For embodiment e, let $A_0^{(e)}$ denote a clean raw action chunk in $\mathbb{R}^{H\times d_a}$, with $d_a = 20$ in the dual-arm setting, and let $A_1^{(e)} \sim \mathcal{N}(0, I)$ be action noise of the same shape. Likewise, z_0 and $z_1 \sim \mathcal{N}(0, I)$ denote the clean and noisy future visual latents. At $\rho \sim \mathcal{U}(0, 1)$, we form $A_\rho^{(e)} = (1-\rho)A_0^{(e)} + \rho A_1^{(e)}$ and $z_\rho = (1-\rho)z_0 + \rho z_1$, with raw-space targets $v^{A,(e)} = A_1^{(e)} - A_0^{(e)}$ and $v^z = z_1 - z_0$. Writing $\Xi^{(e)} = \{A_0^{(e)}, A_1^{(e)}, z_0, z_1\}$, the two flow-matching losses are:

$$\begin{aligned} \mathcal{L}_{act} &= \mathbb{E}_{e,\rho,\Xi^{(e)}}\left[\|\hat{v}_\theta^{A,(e)}(A_\rho^{(e)}, z_\rho, \rho) - v^{A,(e)}\|_2^2\right], \\ \mathcal{L}_{world} &= \mathbb{E}_{e,\rho,\Xi^{(e)}}\left[\|\hat{v}_\theta^{z}(A_\rho^{(e)}, z_\rho, \rho) - v^{z}\|_2^2\right]. \end{aligned} \tag{21}$$

Although $E_a^{(e)}$ projects $A_\rho^{(e)}$ into DiT tokens for joint video-action attention, the prediction $\hat{v}_\theta^{A,(e)}$ and target $v^{A,(e)}$ remain in raw action space. Thus, $\mathcal{L}_{act}$ does not impose flow matching on action embeddings.

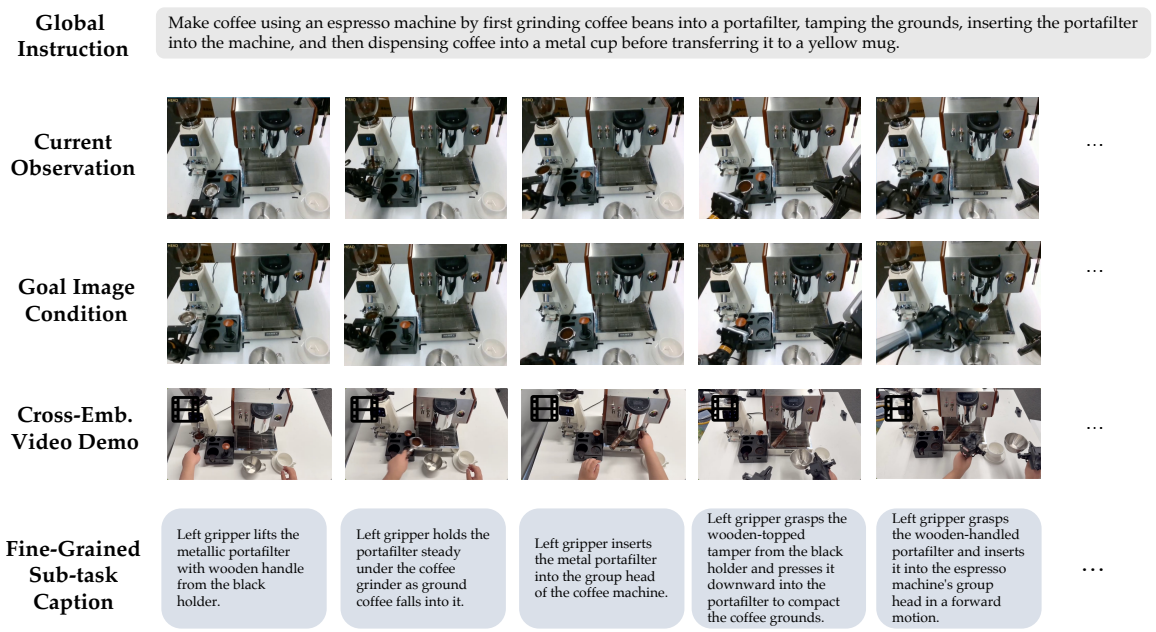

Figure 9: Multimodal prompt example from ManipEvent-5M: `coffee making.` The sample illustrates how a multi-stage task is represented by a global instruction, current observations, a goal-image condition, fine-grained subtask captions, and a cross-embodiment video demonstration.

Training follows a three-stage curriculum that progressively acquires multimodal controllability, autonomous planning, and downstream interactive manipulation.

Stage 1: event-grounded multimodal WAM pretraining. We first pretrain the causal video-action backbone on ManipEvent-5M using event-level samples paired with fine-grained captions, goal images, video demonstrations, visual transitions, and robot actions. Different prompt modalities are sampled as interchangeable control conditions: fine-grained captions are injected through T5 cross-attention, while goal images and video demonstrations are encoded as persistent visual prefill. The short-term causal visual memory is already active at this stage: recent observation chunks are dynamically maintained as clean visual prefill, enabling the DiT to model local temporal continuity through causal self-attention. Stage 1 does not yet introduce the VLM-based event memory or autonomous latent subgoal reasoning. Its objective is:

$$\mathcal{L}^{(1)} = \mathcal{L}_{act} + \lambda_w \mathcal{L}_{world}. \tag{22}$$

This stage establishes fine-grained language-video-action grounding, short-horizon visual dynamics, and controllable video-action generation before event-level memory reasoning is introduced.

Stage 2: memory-aware mid-training. Starting from the Stage-1 WAM and retaining its short-term causal visual memory, we introduce the VLM planning branch and long short-term event memory for autonomous planning. This stage uses temporally extended robot trajectories, for which episode-level instructions provide global task intent and event-level captions identify the active atomic subgoals. Given only the high-level instruction as input to the planning branch, the model constructs memory-enhanced VLM tokens $\hat{q}_t$ by retrieving global-history, local-active, and event-boundary latents together with the compact history bank; local-active and boundary chunks retain full-token anchor representations for detailed retrieval. The corresponding fine-grained T5 embedding provides a training-time semantic target through $\mathcal{L}_{sem}$, but is not directly used as the autonomous-planning condition. The Stage-2 objective is:

$$\mathcal{L}^{(2)} = \mathcal{L}_{act} + \lambda_w \mathcal{L}_{world} + \lambda_s \mathcal{L}_{sem}. \tag{23}$$

By aligning latent subgoal reasoning with the fine-grained semantic space learned in Stage 1, semantic forcing transfers the pretrained language understanding and controllable video-action generation capabilities to the autonomous planning pathway. The model consequently learns to track task progress, infer the active subgoal, and select the next action without requiring an explicit fine-grained instruction at inference time.

Stage 3: downstream interactive post-training. Finally, we post-train the model on downstream robot tasks using the interaction mode required by each task. Long-horizon manipulation tasks activate high-level instructions, reasoning-augmented event memory, and online visual history for autonomous planning. Fine-grained instruction-following tasks directly condition the WAM on T5 subtask embeddings. In-context adaptation tasks retain goal images or video demonstrations as persistent visual prefill and combine them with online visual memory and, when required, latent reasoning. This stage adapts the shared pretrained model to task-specific embodiments and action distributions while preserving a unified interface for autonomous planning, explicit instruction following, and visual-prompted adaptation.

4 Experiments

4.1 Benchmark Setup and Evaluation Protocol

We evaluate WorldScape Policy 2.0 through simulation benchmarks, controlled ablations, and real-world deployment. For the main RoboTwin 2.0 comparison [51], all methods are fine-tuned for 50K steps on the full clean-plus-randomized training data to match the setting used by the compared baselines, and are evaluated over 100 trials per task across 50 tasks under both clean and randomized conditions. Across these settings, we assess task success, subgoal completion, instruction following, progress consistency, and in-context adaptation.

Controlled ablations on RoboTwin 2.0 isolate the contributions of memory components, staged training and semantic forcing. We further evaluate the model on an AgileX Robotics platform across four main capabilities: long-horizon autonomous planning, memory-dependent reasoning, cross-embodiment ICL (In-Context Learning), and sequential fine-grained control. For each real-world task, we perform 20 trials and report the average success rate.

4.2 Implementation Details

Model configurations. For event annotation, we use Qwen3-VL-32B [54] to generate and verify segment-level subtask captions, temporal boundaries, and prompt metadata. The reasoning-augmented LSTM uses a lightweight Qwen3-VL-4B model [54] as the latent reasoning backbone. To reduce memory cost, this module receives only the head-view observation image, resized to 320×160. For each action chunk, a single VLM prefill jointly encodes this image and the task-conditioned planning prompt, after which the model autoregressively generates $K = 4$ planning tokens from the same cached context.

The WAM input concatenates three camera views into a single visual canvas. The video DiT is initialized from the Wan2.2-5B text-to-video pretrained model [55], and the video and action DiTs share the same backbone. Video-action interaction uses joint bidirectional attention between video and action tokens, while attention to clean tokens and visual-history tokens remains causal to preserve autoregressive prediction. During multi-embodiment pretraining, each embodiment uses its own action encoder–decoder pair $(E_a^{(e)}, D_a^{(e)})$, while the video-action backbone is shared. The short-term visual memory retains up to 4 recent chunks. The long-term memory keeps 8 history chunks by default, and this length can be adjusted according to task horizon and memory demand.

Training setup. The action interface follows Sec. 3.1: each arm contributes a chunk-relative 3D delta position, a continuous 6D relative-rotation representation, and an absolute one-DoF gripper command, giving a 20D raw action for dual-arm control. As specified in Eq. 21, action flow matching is optimized directly in this raw space; the action encoder and decoder serve only as embodiment-specific interfaces to the shared DiT. Event-grounded pretraining uses batch size 768 and learning rate 5×10^{-4}; post-training uses batch size 128 and learning rate 6×10^{-5} for 50K steps.

4.3 Simulation Benchmark Results

We evaluate manipulation robustness across all 50 tasks of RoboTwin 2.0 [51] under both clean and randomized settings. The benchmark covers diverse bimanual skills and introduces visual and physical perturbations in the randomized setting, measuring robustness under the same clean-plus-randomized training protocol used for all methods in Table 2.

As shown in Table 2, WorldScape Policy 2.0 achieves the best overall average success rate of 94.3%, with 94.3% under clean conditions and 94.2% under randomization. It outperforms the representative VLA baseline $\pi_{0.5}$ [3] by

Table 2: Evaluation on the RoboTwin 2.0 benchmark. Results are averaged across all 50 tasks under clean and randomized evaluation settings; all methods use clean-plus-randomized training data for a consistent comparison.

Model	Clean	Randomized	Average
# Vision-Language-Action (VLA) Models			
π_0 [56]	65.9%	58.4%	62.2%
X-VLA [57]	72.9%	72.8%	72.9%
$\pi_{0.5}$ [3]	82.7%	76.8%	79.8%
Abot-M0 [58]	81.2%	80.4%	80.8%
LingBot-VLA [59]	86.5%	85.3%	85.9%
HoloBrain-0-QW [60]	91.9%	92.3%	92.1%
# World Action Models			
Motus [7]	88.7%	87.0%	87.9%
GigaWorld-Policy [9]	85.6%	85.3%	85.5%
LingBot-VA [10]	92.9%	91.6%	92.3%
Fast-WAM [6]	91.9%	91.8%	91.9%
Abot-M0.5 [58]	94.0%	**94.2%**	94.1%
LingBot-VA 2.0 [61]	93.8%	93.4%	93.6%
WorldScape Policy 1.0 [62]	93.2%	91.7%	92.5%
WorldScape Policy 2.0	**94.3%**	**94.2%**	**94.3%**

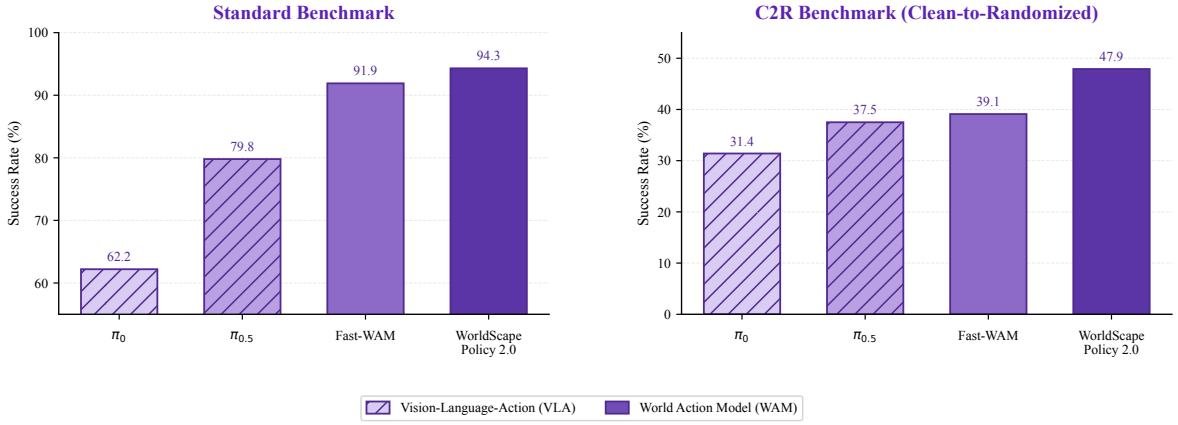

Figure 10: Comparison on the RoboTwin 2.0 standard and C2R benchmarks. The standard benchmark uses clean-plus-randomized training data, whereas C2R trains policies only on clean demonstrations and reports the average success rate across clean and randomized evaluation settings. Hatched bars denote VLA methods, and solid bars denote WAM methods.

11.6%, 17.4%, and 14.5% on clean, randomized, and average success, respectively. Compared with Fast-WAM [6], a representative WAM baseline, it still improves clean, randomized, and average success by 2.4% each. While compared with LingBot-VA 2.0 [61], a recent natively pretrained WAM, our method further improves clean, randomized, and average success by 0.5%, 0.8%, and 0.7%, respectively, further validating the effectiveness of our event-grounded pretraining strategy. The negligible gap between clean and randomized performance further demonstrates the robustness and strong generalization capability of WorldScape Policy 2.0 across diverse bimanual manipulation tasks.

Figure 10 further compares the standard benchmark with the stricter Clean-to-Randomized (C2R) setting. Unlike the standard protocol, C2R uses only clean demonstrations for training and evaluates the resulting policy under both clean and randomized conditions; the reported score averages success across these two evaluation settings, providing a balanced measure of in-domain performance and OOD generalization without exposure to randomized training data. Under this setting, WorldScape Policy 2.0 achieves the highest average success rate of 47.9%, outperforming π_0, $\pi_{0.5}$, and Fast-WAM by 16.5%, 10.4%, and 8.8%, respectively.

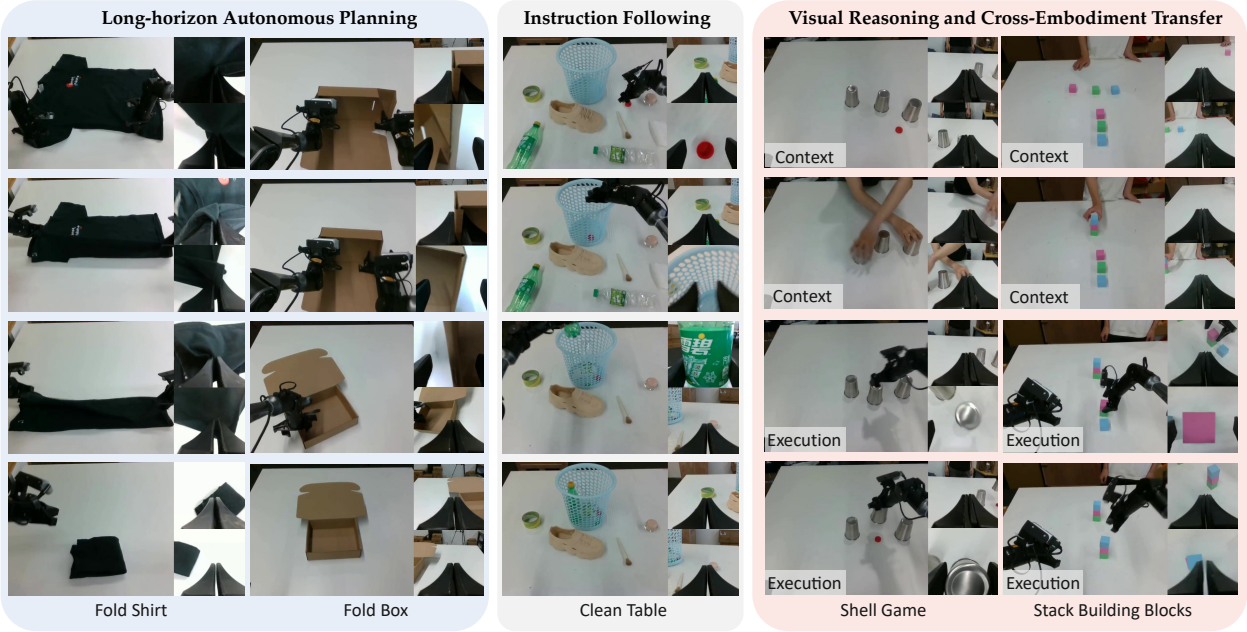

Figure 11: Illustration of real-world tasks on the dual-arm PiPER platform. The benchmark covers long-horizon autonomous planning through `shirt and box folding`, fine-grained instruction following through `table cleaning`, and visual-context reasoning and cross-embodiment transfer through the `shell game` and demonstration-conditioned `block stacking`, respectively.

Table 3: Real-robot manipulation benchmark across diverse capabilities. The evaluation reports episode-level task success rate for autonomous planning, sequential instruction following, and in-context reasoning / skill transfer with text, goal-image, and video-demonstration prompts.

Capability	Tasks	Prompt / Context	$\pi_{0.5}$	DreamZero	WorldScape Policy 2.0
Long-Horizon Autonomous Planning	`Folding Clothes`	Global Instruction	60%	45%	**75%**
	`Folding Box`		65%	55%	**75%**
Memory-Dependent Visual Reasoning	`Shell Game`	Video Demo	30%	50%	**75%**
Cross-Embodiment Skill Transfer	`Stacking Blocks`	Goal Image	10%	20%	**60%**
		Video Demo	20%	20%	**70%**
Sequential Fine-Grained Control	`Clean Table` (Full Sequence)	Sequential Subtask Captions	70%	60%	**80%**

4.4 Real-World Evaluation Results

We evaluate five real-world tasks on the dual-arm PiPER platform, as illustrated in Fig. 11. Table 3 covers comprehensive evaluation across four core capabilities: ❶ ***Long-horizon autonomous planning*** evaluates the success rates of `clothes and box folding` tasks respectively under a global instruction. ❷ ***Memory-dependent visual reasoning*** uses a video-conditioned `shell game` task for evaluation, requiring the robot to track history and infer a hidden object state. ❸ ***Cross-embodiment skill transfer*** evaluates `block stacking` policy conditioned on either a goal image or a demonstration video. ❹ ***Sequential fine-grained controllability*** evaluates the full `table-cleaning` sequence driven by successive subtask captions. Table 4 further isolates text grounding through six atomic `table-cleaning` instructions under reset scenes. As discussed in the Sec. 1, standard VLA policies such as $\pi_{0.5}$ [3] rely on static single-frame observations as input and therefore lack the temporal context required by the `shell game` task and demonstration-conditioned `block stacking` task. For a fair and task-feasible comparison, we temporally extend the input of $\pi_{0.5}$ in these evaluations, allowing it to access the observation history necessary to attempt the tasks.

Table 4: Fine-grained text-instruction following on the real-world `table cleaning` task. In-domain object categories are included in task-specific post-training, whereas out-of-domain categories are held out and introduced only at evaluation. Each row independently evaluates one atomic instruction under a reset scene. We report instruction-level success rate over 20 trials.

Task	Domain Split	Fine-Grained Text Instruction	$\pi_{0.5}$	DreamZero	WorldScape Policy 2.0
`Clean Table`	In-domain	Pick up the *white tissue* and place it into the *basket*.	70%	75%	**80%**
		Pick up the *paper cup* and place it into the *basket*.	**90%**	75%	**90%**
		Pick up the *plastic bottle* and place it into the *basket*.	**90%**	70%	**90%**
	Out-of-domain	Pick up the *black pen* and place it into the *basket*.	50%	40%	**70%**
		Pick up the *green tape* and place it into the *basket*.	55%	35%	**60%**
		Pick up the *beige shoe* and place it into the *basket*.	25%	15%	**50%**
In-Domain Average			83.3%	73.3%	**86.7%**
Out-of-Domain Average			43.3%	30.0%	**60.0%**
Overall Average			63.3%	51.7%	**73.3%**

Table 5: Component analysis on RoboTwin 2.0. STM denotes short-term visual memory, LTM denotes long-term event memory, and LSR denotes latent subgoal reasoning.

STM	LTM	LSR	Clean	Randomized	Average
✗	✗	✗	64.60%	17.22%	40.91%
✓	✗	✗	66.92%	22.42%	44.67%
✓	✓	✗	68.49%	24.01%	46.25%
✓	✓	✓	**69.74%**	**26.03%**	**47.89%**

Across these real-world tasks, WorldScape Policy 2.0 outperforms both the representative VLA method $\pi_{0.5}$ [3] and the WAM method DreamZero [11] in overall episode success rate. As shown in Table 3, it achieves 75% success on both long-horizon `folding` tasks and 80% on sequential `table cleaning`, consistently exceeding both baselines. The advantage is even more pronounced on visual-context tasks, where it reaches 75% on the `shell game` and 60–70% on cross-embodiment `block stacking`, demonstrating strong memory-dependent reasoning and visual-prompt transfer capabilities. The fine-grained instruction-following results in Table 4 further validate the benefit of the proposed event-grounded pretraining. WorldScape Policy 2.0 achieves the best in-domain average success rate (86.7%) and a substantially higher overall average (73.3%) than $\pi_{0.5}$ (63.3%) and DreamZero (51.7%). The improvement is especially clear under held-out object categories: WorldScape Policy 2.0 reaches 60.0% OOD (Out-of-Domain) success rate, compared with 43.3% for $\pi_{0.5}$ and 30.0% for DreamZero. These results indicate that fine-grained event-level pretraining improves atomic instruction grounding and compositional generalization, demonstrating the effectiveness and necessity of the proposed framework for controllable manipulation.

4.5 Ablation Study

Unless otherwise specified, all ablative experiments are conducted on the RoboTwin 2.0 benchmark and follow the official clean-only training setting: each variant is trained only on the clean data split, rather than the full clean-plus-randomized data used for the main comparison in Table 2. Consequently, the absolute ablation scores should not be directly compared with the headline results in Table 2.

❶ ***Memory and latent reasoning components.*** Table 5 progressively adds short-term visual memory (STM), long-term event memory (LTM), and latent subgoal reasoning (LSR). Compared with the no-memory baseline, STM improves clean, randomized, and average success by 2.32%, 5.20%, and 3.76%, respectively. Adding LTM further raises the three metrics to 68.49%, 24.01%, and 46.25%, showing that event-level progress memory complements recent visual context in general. Finally, with the help of implicit subgoal reasoning, WorldScape Policy 2.0 achieves the best results as expected. Overall, the complete model consistently outperforms the no-memory baseline under both clean and randomized settings, strongly validating the importance of integrating short-term visual memory, long-term event memory, and latent subgoal planning for robust manipulation.

Table 6: Ablation of Stage-1 event-grounded pretraining and Stage-2 memory-aware mid-training on RoboTwin 2.0. SF denotes semantic forcing. All variants use the same Stage-3 downstream post-training protocol.

Stage-1	Stage-2	SF	Clean	Randomized	Average
✗	✗	✗	65.67%	20.71%	43.19%
✓	✗	✗	67.90%	25.36%	46.63%
✓	✓	✗	68.95%	25.64%	47.30%
✓	✓	✓	**69.74%**	**26.03%**	**47.89%**

❷ ***Training stages and semantic forcing.*** Table 6 separates the effects of Stage-1 event-grounded pretraining, Stage-2 memory-aware mid-training, and semantic forcing (SF), while keeping Stage-3 downstream post-training fixed. The four rows compare training from scratch, Stage 1 alone, Stages 1–2 without SF, and the complete curriculum with SF. This design measures whether multimodal event-grounded pretraining provides a useful initialization, whether memory mid-training adds progress-aware prediction, and whether semantic forcing transfers explicit event semantics into latent subgoal reasoning. It can be observed that Stage-1 pretraining yields the largest improvement, particularly under randomization, while Stage-2 mid-training and semantic forcing provide further consistent gains through progress-aware memory and semantically aligned latent reasoning. The monotonic improvement across stages validates the effectiveness and complementarity of the proposed multi-stage pretraining framework for robust manipulation.

5 Conclusion

We presented WorldScape Policy 2.0, a controllable World Action Model for long-horizon robotic manipulation. Its reasoning-augmented long short-term memory couples causal short-term visual context with event-level progress memory and latent subgoal reasoning, enabling a single video-action model to support autonomous planning from high-level instructions, fine-grained text control, goal-image conditioning, and demonstration-based in-context adaptation. To train these capabilities, we introduced ManipEvent-5M, an event-grounded multimodal dataset with nearly five million segments, and a three-stage curriculum that transfers explicit event semantics into autonomous reasoning through semantic forcing. Evaluations on both simulation and real-world platforms demonstrate the complementary roles of visual memory, event memory, and latent reasoning, substantially outperforming representative VLA and WAM baselines and validating the effectiveness of event-level pretraining. These results advance WAMs from passive future prediction toward memory-grounded and multimodally controllable manipulation.

References

[1] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[2] Haisheng Su, Feixiang Song, Cong Ma, Wei Wu, and Junchi Yan. Robosense: Large-scale dataset and benchmark for egocentric robot perception and navigation in crowded and unstructured environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27446–27455, 2025.

[3] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

[4] Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, et al. $\pi_{0.7}$: A steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

[5] Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10678–10688, 2026.

[6] 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.

[7] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 35101–35113, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bi_Motus_A_Unified_Latent_Action_World_Model_CVPR_2026_paper.html.

[8] Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

[9] 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.

[10] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Proceedings of Robotics: Science and Systems*, 2026. URL https://roboticsconference.org/program/papers/16/.

[11] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan 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.

[12] Sizhe Yang, Juncheng Mu, Tianming Wei, Chenhao Lu, Xiaofan Li, Linning Xu, Zhengrong Xue, Zhecheng Yuan, Dahua Lin, Jiangmiao Pang, and Huazhe Xu. MemoryWAM: Efficient world action modeling with persistent memory. *arXiv preprint arXiv:2606.20562*, 2026.

[13] Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 79115–79135. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

[14] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Yue Liao, Zhengkai Jiang, Yue Hu, Peng Gao, Hongsheng Li, Maoqing Yao, and Guanghui Ren. Enerverse: Envisioning embodied future space for robotics manipulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/360052c2c6d0c8ec24c476d43236ab25-Abstract-Conference.html.

[15] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24328–24346. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25g.html.

[16] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 5170–5194. PMLR, 2025. URL https://proceedings.mlr.press/v305/jang25a.html.

[17] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[18] Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

[19] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 3936–3951. PMLR, 2025. URL https://proceedings.mlr.press/v305/bharadhwaj25a.html.

[20] Haisheng Su, Kunchang Li, Jinyuan Feng, Dongliang Wang, Weihao Gan, Wei Wu, and Yu Qiao. Tsi: temporal saliency integration for video action recognition. *arXiv preprint arXiv:2106.01088*, 2021.

[21] Yong-Lu Li, Hongwei Fan, Zuoyu Qiu, Yiming Dou, Liang Xu, Hao-Shu Fang, Peiyang Guo, Haisheng Su, Dongliang Wang, Wei Wu, et al. Discovering a variety of objects in spatio-temporal human-object interactions. *arXiv preprint arXiv:2211.07501*, 2022.

[22] Haisheng Su, Jing Su, Dongliang Wang, Weihao Gan, Wei Wu, Mengmeng Wang, Junjie Yan, and Yu Qiao. Collaborative distillation in the parameter and spectrum domains for video action recognition. *arXiv preprint arXiv:2009.06902*, 2020.

[23] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61885–61896. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24f.html.

[24] Haisheng Su, Wei Wu, Feixiang Song, Junjie Zhang, Zhenjie Yang, and Junchi Yan. Drivemamba: Task-centric scalable state space model for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2602.13301*, 2026.

[25] Haisheng Su, Wei Wu, Zhenjie Yang, and Isabel Guan. Egofsd: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end self-driving. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2026.

[26] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

[28] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10006732.

[29] Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. OA-WAM: Object-addressable world action model for robust robot manipulation, 2026. URL https://arxiv.org/abs/2605.06481.

[30] Yiguo Fan, Shuanghao Bai, Xinyang Tong, Pengxiang Ding, Yuyang Zhu, Hongchao Lu, Fengqi Dai, Wei Zhao, Yang Liu, Siteng Huang, Zhaoxin Fan, Badong Chen, and Donglin Wang. Long-VLA: Unleashing long-horizon capability of vision language action model for robot manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 2018–2037. PMLR, 2025. URL https://proceedings.mlr.press/v305/fan25a.html.

[31] Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, and Lin Shao. VLA-OS: Structuring and dissecting planning representations and paradigms in vision-language-action models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c7af751b5e0a407c62ac023e3cb381f9-Abstract-Conference.html.

[32] Marcel Torne, Karl Pertsch, Homer Walke, Kyle Vedder, Suraj Nair, Brian Ichter, et al. MEM: Multi-scale embodied memory for vision language action models, 2026. URL https://arxiv.org/abs/2603.03596.

[33] Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=54U3XHf7qq.

[34] Isabella Liu, An-Chieh Cheng, Rui Yan, Geng Chen, Ri-Zhao Qiu, Xueyan Zou, Sha Yi, Hongxu Yin, Xiaolong Wang, and Sifei Liu. Long-horizon manipulation via trace-conditioned VLA planning, 2026. URL https://arxiv.org/abs/2604.21924.

[35] Zhen Liu, Xinyu Ning, Zhe Hu, Xinxin Xie, Weize Li, Zhipeng Tang, Chongyu Wang, Zejun Yang, Hanlin Wang, Yitong Liu, and Zhongzhu Pu. Goal2Skill: Long-horizon manipulation with adaptive planning and reflection, 2026. URL https://arxiv.org/abs/2604.13942.

[36] Jian Zhu, Jianjun Zhang, Taiyi Su, Tianbin Liu, Zhangyuan Wang, Kai Xie, Zitai Huang, Chong Ma, Youzhang He, Tianjian Wang, et al. DSWAM: A dual-system world action foundation model for fine-grained robot manipulation, 2026. URL https://arxiv.org/abs/2607.04927.

[37] Jiahui Fu, Junyu Nan, Lingfeng Sun, Hongyu Li, Jianing Qian, Jennifer L. Barry, Kris Kitani, and George Konidaris. NovaPlan: Zero-shot long-horizon manipulation via closed-loop video language planning, 2026. URL https://arxiv.org/abs/2602.20119.

[38] Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-VLA: Unifying vision-language-action modeling across tasks, environments, and robot embodiments, 2026. URL https://arxiv.org/abs/2605.30280.

[39] Xintong Hu, Xuhong Huang, Jinyu Zhang, Yutong Yao, Yuchong Sun, Qiuyue Wang, Mingsheng Li, Sicheng Xie, Yitao Liu, Junhao Chen, et al. FineVLA: Fine-grained instruction alignment for steerable vision-language-action policies, 2026. URL https://arxiv.org/abs/2605.27284.

[40] Wenqi Liang, Gan Sun, Yao He, Jiahua Dong, Suyan Dai, Ivan Laptev, Salman Khan, and Yang Cong. PixelVLA: Advancing pixel-level understanding in vision-language-action model. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7M6ryCABIc.

[41] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.011.

[42] Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18549–18557, 2026. doi: 10.1609/aaai.v40i22.38921. URL https://ojs.aaai.org/index.php/AAAI/article/view/38921.

[43] Ruisen Tu, Arth Shukla, Sohyun Yoo, Xuanlin Li, Junxi Li, Jianwen Xie, Hao Su, and Zhuowen Tu. SG-VLA: Learning spatially-grounded vision-language-action models for mobile manipulation, 2026. URL https://arxiv.org/abs/2603.22760.

[44] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=b1CVu9l5GO.

[45] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, June 2025. doi: 10.1109/CVPR52734.2025.00166. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_CoT-VLA_Visual_Chain-of-Thought_Reasoning_for_Vision-Language-Action_Models_CVPR_2025_paper.html.

[46] Xiang Zhu, Yichen Liu, Hezhong Li, and Jianyu Chen. Learning generalizable robot policy with human demonstration video as a prompt, 2025. URL https://arxiv.org/abs/2505.20795.

[47] AgiBot-World Team. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3549–3556. IEEE, 2025. doi: 10.1109/IROS60139.2025.11247088. URL https://ieeexplore.ieee.org/document/11247088/.

[48] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems (RSS) 2025*. Robotics: Science and Systems Foundation, 2025. URL https://www.roboticsproceedings.org/rss21/p152.pdf.

[49] Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, Zhaoye Long, et al. RoboCOIN: An open-sourced bimanual robotic data collection for integrated manipulation, 2025. URL https://arxiv.org/abs/2511.17441.

[50] 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. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

[51] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62192.

[52] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 44776–44791, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

[53] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. URL https://mlanthology.org/iclr/2026/hoque2026iclr-egodex/.

[54] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[56] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

[57] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Tai Wang, Ya-Qin Zhang, Jingjing Liu, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kt51kZH4aG.

[58] Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, et al. Abot-m0. 5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026.

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

[60] Xuewu Lin, Tianwei Lin, Yun Du, Hongyu Xie, Yiwei Jin, Jiawei Li, Shijie Wu, Qingze Wang, Mengdi Li, Mengao Zhao, et al. Holobrain-0 technical report. *arXiv preprint arXiv:2602.12062*, 2026.

[61] Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, et al. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026.

[62] Manifold AI. Worldscape policy: Generalizable robotic learning via a foundation world model, 2026.