LEARNING SKILLS FROM HISTORICAL ACTION TRAJECTORIES: ACTION EXPERIENCE DICTIONARY FOR WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

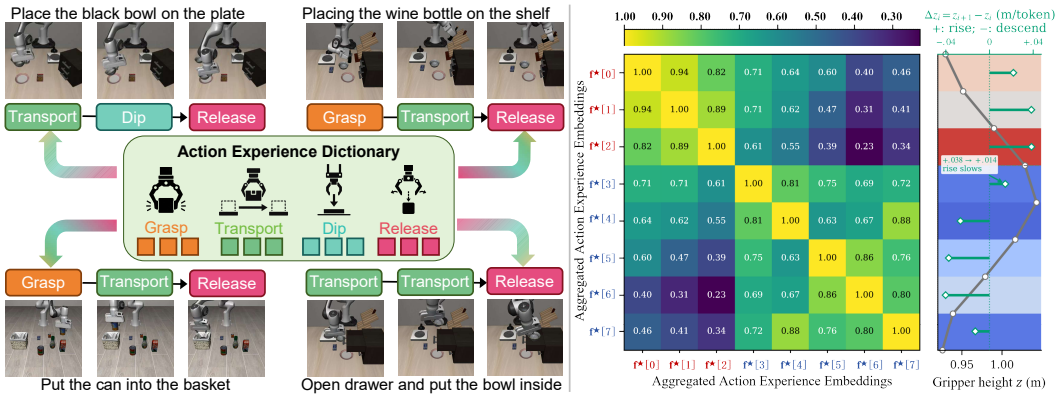
World Action Models (WAMs) couple visual dynamics prediction with action generation, yet they do not explicitly support the reuse of action experience across manipulation tasks. Furthermore, existing WAMs struggle to capture underlying cross-task semantic relationships that could guide target action prediction, as redundant background elements interfere with the extraction of key visual information. To address these challenges, we develop a novel Action Experience Dictionary (AED) that encodes historical physical action trajectories into shared action embeddings to support skill reuse and model cross-task relationships. Specifically, we first aggregate historical actions to align with visual observations and retrieve action embeddings from the AED using a pretrained action tokenizer. Subsequently, we visually condition the pooled embeddings through cross-attention and prepend them to noisy action tokens, providing interaction context and action intent for prediction. To model action-related motion and reduce reliance on irrelevant background cues, we introduce a motion-aware transition loss that supervises visual feature change prediction over random temporal intervals. Experiments on simulation benchmarks and in real-world cross-embodiment settings verify the effectiveness of our AED. The anonymous project website is available at AED.

1 INTRODUCTION

Rapid advances in large-scale foundation models (Touvron et al., 2023; Yang et al., 2025; Team Wan et al., 2025; DeepSeek-AI team, 2024) have spurred interest in transferring visual and linguistic knowledge to physical robots. Vision-Language-Action (VLA) models (Kim et al., 2025; Black et al., 2025b; Qu et al., 2025) adapt pretrained vision-language representations to map observations and instructions to actions, but their direct policy formulations often leave environmental dynamics implicit. To exploit dynamics in video data, World Action Models (WAMs) (Bi et al., 2026; Li et al., 2026; Yuan et al., 2026) couple action generation with future visual prediction, providing supervision on how environments evolve. Recent WAM advances span unified video-action architectures and efficient control pipelines.

However, existing WAMs (Chen et al., 2026a; Yang et al., 2026) typically focus on learning task-specific behaviors, leaving the potential of cross-task semantic relationships to guide manipulation skill learning underexplored. In particular, such underlying relationships among manipulation tasks can help robots draw on action experience relevant to the target task, thereby improving manipulation performance. As illustrated in Fig. 1(a), a robot learning to place a can into a basket can build on experience in grasping, transporting, and releasing objects acquired from other pick-and-place tasks (*e.g.*, placing a wine bottle on a shelf or opening a drawer and placing a bowl inside). By adapting these behaviors to the basket's position and opening, the robot can learn the target manipulation task more effectively. Similarly, action experience gained from placing a can into a basket can also facilitate the learning of related actions in tabletop pick-and-place tasks. This practical example demonstrates that different tasks share reusable action experience despite their distinct goals and visual contexts, as depicted in Fig. 1(b). Nevertheless, simply retaining historical actions is insufficient to make this knowledge reusable, since similar motions can serve different purposes depending on the objects being manipulated and their spatial relationships. Moreover, redundant background elements may hinder the extraction of action-relevant visual information. These challenges lead us

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Example of reusable action skills (b) Visualization of Reusable Action Experience (c)

Figure 1: **(a)** Example of reusable action skills. **(b)** Visualization of reusable action experience in the action experience dictionary (AED). $\mathbf{f}^\star[0]$–$\mathbf{f}^\star[2]$ indicate the `transporting` action pattern, while $\mathbf{f}^\star[3]$–$\mathbf{f}^\star[7]$ encode the `dipping` action pattern in the AED. The visualization of the cosine similarities among action embeddings in the AED shows that learning to place the bowl on the plate involves substantial reuse of both the `transporting` and `dipping` action patterns. **(c)** Gray and green lines show gripper height and its relative changes, respectively, during training to place the bowl on the plate, linking action patterns (e.g., `transporting` and `dipping`) to physical height.

to the central question: *How can WAMs (Cai et al., 2026) use past action trajectories to model underlying relationships among tasks and facilitate skill learning for the target manipulation task?*

To address the above challenges, we propose a novel learnable Action Experience Dictionary (AED) that encodes historical manipulation trajectories as shared action embeddings for WAMs. **First**, we aggregate past physical actions over intervals aligned with visual observations and use a pretrained action tokenizer to retrieve task-relevant action embeddings from the AED. **Second**, the retrieved embeddings are pooled into compact temporal representations and conditioned on compressed historical visual features through cross-attention, enabling the resulting representations to capture both interaction context and action intent. Then, we prepend these visually conditioned action embeddings to the noisy action tokens, enabling the action expert to exploit inter-task relationships when predicting subsequent action chunks. **Third**, we introduce a motion-aware transition loss that supervises the prediction of visual feature changes over randomly sampled temporal intervals using the visual features at the start of each interval and the corresponding AED-conditioned hidden states. This loss encourages the retrieved action embeddings in AED to capture action-related motion while reducing reliance on irrelevant background. **Finally**, we evaluate the effectiveness of the proposed model by comparing it with baselines in simulation on LIBERO, RoboTwin, and LIBERO-Plus, as well as in real-world cross-embodiment experiments. The main contributions are listed below:

- We propose a novel learnable Action Experience Dictionary (AED) that encodes historical action experience into visually conditioned action embeddings, enabling our model to leverage underlying relationships among manipulation tasks to guide action prediction.
- We incorporate action-relevant visual information into the action embeddings in the AED to obtain visually conditioned action embeddings, which combine task-related visual semantics with action experience to capture both task context and action intent.
- We introduce a motion-aware transition loss that supervises visual feature change prediction over random temporal intervals, encouraging learned action embeddings in the AED to capture action-related motion and reducing reliance on irrelevant background cues.

2 Related Work

Vision Language Action (VLA): Recent VLAs increasingly explore structured action representations and visual dynamics for transferable robot control. FAST (Pertsch et al., 2025) introduces frequency-space action tokenization for efficient VLA training, while UniVLA (Bu et al., 2025) and ViPRA (Routray et al., 2026) learn task- or motion-centric latent actions from heterogeneous videos to support transferable control. VLM2VLA (Hancock et al., 2026) represents robot actions in a language-compatible form to preserve pretrained VLM capabilities. DeFI (Zhang et al., 2026)

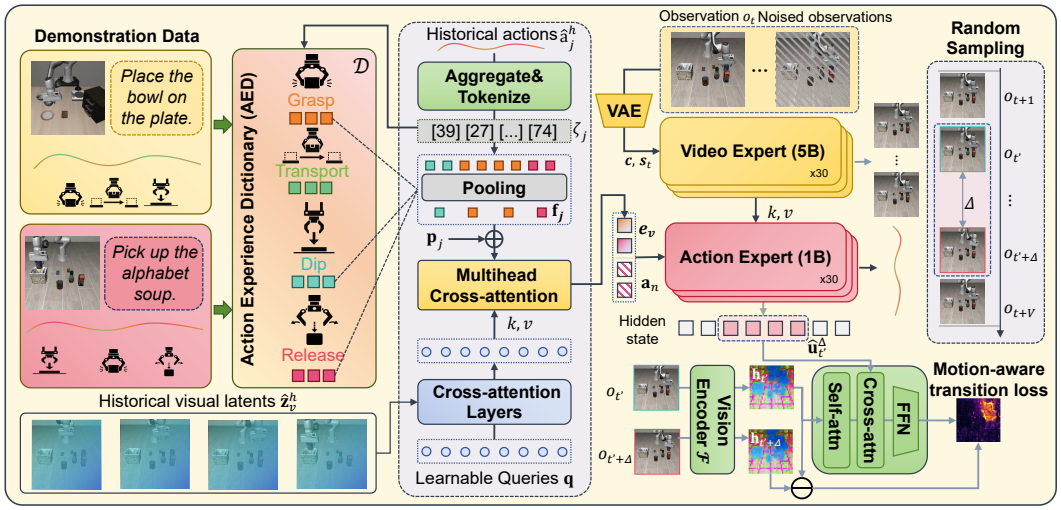


Figure 2: Overview of the proposed AED. After defining learnable action experience dictionary (AED) shared across tasks, we utilize the visually conditioned action embeddings to encode the task-relevant information and employ a motion-aware transition loss to encode action-relevant motion.

decouples forward visual dynamics and inverse action learning. However, cross-task reuse of historical action experience remains underexplored. Unlike UniVLA and ViPRA, which primarily learn transferable latent action spaces, our method explicitly visually conditions the retrieved action embeddings from the AED, and supervises visual feature change prediction over random temporal intervals to capture action-related motion and facilitate skill patterns reuse across tasks.

World Action Models (WAMs): World action models exploit visual prediction to capture physical and temporal structure for robot control. UniPi (Du et al., 2023) performs planning through text-conditioned video generation and action extraction, while GR-2 (Cheang et al., 2024) and DreamGen (Jang et al., 2025) leverage video generative priors for manipulation. More recent WAMs couple visual dynamics and action generation more directly: UWM (Zhu et al., 2025) jointly models video and action diffusion, Motus (Bi et al., 2026) integrates understanding, video, and action experts, while LingBot-VA (Li et al., 2026) and (Yuan et al., 2026) shows that video co-training can benefit control without test-time video generation. Unlike existing WAMs that primarily learn task-specific behaviors through visual dynamics supervision, our method explicitly encodes historical action trajectories into shared action embeddings from AED to model underlying cross-task relationships, while using motion-aware transition supervision to emphasize action-related visual changes.

3 Methodology

3.1 Preliminaries

In embodied manipulation tasks (Li et al., 2026; Driess et al., 2023), robots aim to autonomously plan and make decisions based on task instructions $\mathbf{c}$ and visual observations $\mathbf{o}_t$ at time t. Following FastWAM (Yuan et al., 2026), we adopt flow matching to train control policies with both video and action experts. Let $\mathbf{x}_e$ ($e \in \{v, a\}$) denote the clean target, where $\mathbf{x}_v$ represents future video and $\mathbf{x}_a$ represents an action chunk. Given a Gaussian noise sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a flow time $\tau \in (0, 1)$, we construct $\mathbf{x}_e^\tau = (1-\tau)\mathbf{x}_e + \tau\boldsymbol{\epsilon}$. Accordingly, the flow-matching objective $\mathcal{L}_{\text{FM}}$ is defined as:

$$\mathcal{L}_{\text{FM}} = \sum_{e \in \{v,a\}} \lambda_e \mathcal{L}_e; \quad \mathcal{L}_e = \mathbb{E}_{\mathbf{x}_e, \boldsymbol{\epsilon}, \tau} \left[\left\| \pi_\theta(\mathbf{x}_e^\tau \mid \mathbf{o}_t, \mathbf{c}, \mathbf{s}_t) - (\boldsymbol{\epsilon} - \mathbf{x}_e) \right\|_2^2 \right], \tag{1}$$

where $\pi_\theta(\cdot)$ denotes the policy parameterized by θ, $\mathbf{s}_t$ is the robot's proprioceptive state at time t, and $\lambda_e = 1.0$ is the balancing factor. For joint training, we use $\pi_\theta(\cdot)$ to predict future video latents $\mathbf{z}_{t+1:t+V}$ corresponding to V frames and an action chunk $\mathbf{a}_{t:t+H-1}$ of horizon H to be executed.

3.2 Action Experience Dictionary (AED)

Generally, underlying relationships among manipulation tasks can provide valuable guidance for learning related skills. For example, when learning to place a cup on a shelf, a robot can draw

on experience acquired from tabletop pick-and-place, such as grasping, transporting, and releasing objects, while adapting these behaviors to the target task. Conversely, experience gained from shelf placement can also benefit other tasks that involve these shared skills, enabling mutual knowledge transfer across related tasks. However, even during joint training, existing WAMs (Yuan et al., 2026; Li et al., 2026; Kim et al., 2026) often learn task-specific behaviors without explicitly leveraging inter-task relationships, overlooking the potential of cross-task semantic connections to facilitate the learning of manipulation skills. This limitation motivates us to investigate ***how to leverage reusable action experience shared across related tasks to improve manipulation performance.***

To address the above limitation, we develop a novel learnable action experience dictionary (AED) to learn manipulation skills from historical action trajectories, as depicted in Fig. 2. Specifically, the proposed AED encodes historical physical action trajectories as a sequence of action embeddings to capture latent relationships across skills. Subsequently, for each training batch, we retrieve action embeddings that are highly relevant to the target task from the AED. These action embeddings are aggregated and then conditioned on the corresponding temporal visual embeddings through transformer blocks with cross-attention. Finally, we prepend the resulting action embeddings to the noisy action tokens within the action expert for prediction. This prepending strategy enables the model to exploit inter-task relationships for subsequent skill learning. To further encode action-relevant visual information into AED, we propose a motion-aware transition (MT) loss predicting temporal visual feature changes from the sampled observation and AED-guidance action hidden states.

▷ **Construction of Action Experience Dictionary:** Let $\mathcal{D} \in \mathbb{R}^{N \times d_a}$ denote a learnable Action Experience Dictionary (AED) containing N randomly initialized action embeddings, each of dimension d_a. Notably, an action skill, such as grasping an object placed on a tabletop or inside a container, may comprise multiple action patterns with distinct motion characteristics, including approach direction, gripper height, displacement, and speed. Accordingly, each action embedding in the AED represents a specific action pattern, whereas a set of related action embeddings jointly characterizes the action skill. During training, the proposed AED is shared across all manipulation tasks to facilitate the reuse of action experience, thereby leveraging cross-task relationships to benefit the target task. To encode $\mathcal{D}$, we use a pretrained action tokenizer Φ to convert historical physical action trajectories $\mathbf{a}^h \in \mathbb{R}^{H \times d_f}$ into a sequence of indices for retrieving task-relevant action embeddings from $\mathcal{D}$, where d_f denotes the number of degrees of freedom. However, directly tokenizing these trajectories, each consisting of H actions, results in a temporal frequency mismatch with the V visual observations and incurs additional computational costs. To tackle this issue, we aggregate historical trajectories to obtain $\widehat{\mathbf{a}}^h \in \mathbb{R}^{V \times d_f}$, and define the j-th aggregated action $\widehat{\mathbf{a}}^h[j] \in \mathbb{R}^{d_f}$ as:

$$\widehat{\mathbf{a}}^h[j] = \sum_{r=1}^{k} \mathbf{a}^h[(j-1)k+r] \odot \mathbf{m}[(j-1)k+r] + \mathbf{a}^h[jk] \odot (\mathbf{1} - \mathbf{m}[jk]), \quad \forall j = 1, \cdots, V, \quad (2)$$

where $k = \frac{H}{V}$ is the number of historical actions between two consecutive observations, and $\mathbf{m} \in \{0,1\}^{H \times d_f}$ denotes a binary mask whose entries are set to 1 for the arm control dimensions and 0 for the gripper dimensions. Here, $\mathbf{m}[(j-1)k+r]$ and $\mathbf{m}[jk]$ represent the $((j-1)k+r)$-th and jk-th rows of $\mathbf{m}$, respectively. The same indexing convention applies to $\mathbf{a}[(j-1)k+r]$ and $\mathbf{a}[jk]$. In Eq. (2), we accumulate the arm commands to summarize the motion executed over the interval of observations while retaining the final gripper command to preserve the gripper's terminal state.

Subsequently, we adopt Φ to map the j-th aggregated action $\widehat{\mathbf{a}}^h[j]$ to a sequence of indices $\zeta_j = \Phi(\widehat{\mathbf{a}}^h[j]) \in \mathbb{R}^M$, where M denotes the number of retrieved action embeddings. Each index in ζ_j is then used to retrieve the corresponding action embedding from $\mathcal{D}$. These embeddings are averaged to obtain a compact action representation $\mathbf{f}_j \in \mathbb{R}^d$ for the j-th ($j = 1, \ldots, V$) aggregated action $\widehat{\mathbf{a}}^h_j$:

$$\mathbf{f}_j = \mathbf{p}_j + \frac{1}{M} \sum_{l=1}^{M} \mathcal{D}(\zeta_j[l]), \quad (3)$$

where $\zeta_j[l] \in \mathbb{R}$ denotes the l-th index of ζ_j, and $\mathcal{D}(\zeta_j[l]) \in \mathbb{R}^d$ represents the $\zeta_j[l]$-th row of $\mathcal{D}$. $\mathbf{p}_j \in \mathbb{R}^{d_a}$ denotes the temporal positional encoding for the j-th action trajectory. Afterwards, we stack $\{\mathbf{f}_j\}_{j=1}^{V}$ in temporal order to obtain aggregated action embeddings $\mathbf{f}^\star \in \mathbb{R}^{V \times d_a}$:

$$\mathbf{f}^\star = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_V]. \quad (4)$$

▷ **Visually Conditioned Action Embeddings:** The action embeddings $\mathbf{f}^\star$ obtained in Eq. (4) encode only the action patterns themselves (e.g., grasping and moving patterns), without capturing semantic information about task-relevant objects or their relationships with the target manipulation task. This may result in an incomplete understanding of the task context and action intent. To this end, we condition the action embeddings $\mathbf{f}^\star$ on visual information. Specifically, the historical visual latent embeddings comprise R spatiotemporal patch tokens $\mathbf{z}_v^h \in \mathbb{R}^{R\times d_v}$, where d_v denotes the dimension of visual embeddings. Since many of these tokens correspond to background content or content irrelevant to the interaction, directly feeding them into the action expert introduces redundancy and causes the input sequence length to grow with the observation horizon. Therefore, we compress $\mathbf{z}_v^h$ into a fixed number of latent visual tokens $\widehat{\mathbf{z}}_v^h \in \mathbb{R}^{L\times d_a}$ using L learnable queries $\mathbf{q} \in \mathbb{R}^{L\times d_a}$:

$$\widehat{\mathbf{z}}_v^h = [\mathcal{A}_1 \oplus \mathcal{A}_2 \oplus \cdots \oplus \mathcal{A}_\psi]\,\mathbf{w}_o,\ \mathcal{A}_i = \sigma\Big(\frac{\mathbf{q}\mathbf{w}_q\left(\mathbf{z}_v^h\mathbf{w}_k\right)^\top}{\sqrt{d_a/\psi}}\Big)(\mathbf{z}_v^h\mathbf{w}_v),\ \forall i = 1,\ldots,\psi, \tag{5}$$

where $\oplus$ denotes concatenation along the feature dimension, ψ is the number of cross-attention heads, and $\mathcal{A}_i \in \mathbb{R}^{L\times(d_a/\psi)}$ represents the i-th ($i = 1,\ldots,\psi$) attention head. $\sigma(\cdot)$ is the softmax function. For each attention head, $\mathbf{w}_q \in \mathbb{R}^{d_a\times(d_a/\psi)}$, $\mathbf{w}_k \in \mathbb{R}^{d_v\times(d_a/\psi)}$, and $\mathbf{w}_v \in \mathbb{R}^{d_v\times(d_a/\psi)}$ denote the linear projection matrices for the query, key, and value. Furthermore, $\mathbf{w}_o \in \mathbb{R}^{d_a\times d_a}$ indicates the output projection matrix used to fuse the features from all attention heads.

To incorporate the task-relevant visual context encoded in $\widehat{\mathbf{z}}_v^h$ into $\mathbf{f}^\star$ while preserving their action semantics, we fuse them using a dictionary encoder $\mathcal{E}$, implemented as a one-layer Transformer encoder with cross-attention. Here, $\mathbf{f}^\star$ serves as the query, while $\widehat{\mathbf{z}}_v^h$ provides the keys and values. Using cross-attention, $\mathcal{E}$ generates visually conditioned action embeddings $\mathbf{e}_v \in \mathbb{R}^{V\times d_a}$, which are then concatenated with the noisy action tokens $\mathbf{a}_n \in \mathbb{R}^{H\times d_a}$ to obtain $\mathbf{a}^\star \in \mathbb{R}^{(H+V)\times d_a}$:

$$\mathbf{a}^\star = [\mathbf{e}_v; \mathbf{a}_n], \quad \mathbf{e}_v = \mathcal{E}(\mathbf{f}^\star, \widehat{\mathbf{z}}_v^h), \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation along the token dimension. After concatenation, we feed $\mathbf{a}^\star$ to the action expert for action chunk prediction. Using the formulation in Eq. (6), we integrate the task-relevant action experience retrieved from $\mathcal{D}$ and the associated visual information into the historical context, thereby enriching the context available for subsequent action prediction.

▷ **Motion-Aware Transition Loss:** Although visual conditioning incorporates historical scene information into the retrieved action embeddings for action prediction via Eq. (6), the historical observations contain action-relevant objects and background content, and the flow-matching objective does not explicitly distinguish visual cues associated with action-dependent motion from incidental scene appearance. Thus, the resulting action embeddings in the AED $\mathcal{D}$ may retain background correlations that are less useful when reusing action experience across tasks. To encourage using action-relevant visual information, as illustrated in Fig. 2, we develop a motion-aware transition (MT) loss that predicts temporal visual changes from the starting state and AED-conditioned action hidden states. Relatively stable background components can partially cancel in the feature difference, providing a supervision signal highlighting observable changes. Conditioning this prediction on action hidden states encourages the representations to capture object motion associated with the actions, helping reduce reliance on irrelevant background cues during action chunk prediction.

During training at time t, we randomly sample a future time step $t' \in \{t+1,\ldots,t+V-1\}$ and use a visual encoder $\mathcal{F}$ (e.g., Lingbot-vision (Fu et al., 2026) or DINOv3 (Siméoni et al., 2025)) to extract latent features $\mathbf{h}_{t'} = \mathcal{F}(\mathbf{o}_{t'}) \in \mathbb{R}^{B\times d_z}$ from the future observation $\mathbf{o}_{t'}$. Here B is the number of latent features and d_z is the dimensionality of the latent features. We then sample a time window $\Delta \sim \mathrm{Unif}\{1,2,\ldots,t+V-t'\}$, where Unif is the discrete uniform distribution. After extracting the final-layer hidden states $\mathbf{u}_{t'}^\Delta \in \mathbb{R}^{\Delta\times d_a}$ from the action expert over the interval from t' to $t'+\Delta$, we propose the MT loss $\mathcal{L}_{\mathrm{MT}}$ that emphasizes visual objects whose motion is associated with the actions, thereby reducing the influence of irrelevant background cues on action chunk prediction:

$$\mathcal{L}_{\mathrm{MT}} = \frac{1}{Bd_z}\left\|\Delta\widehat{\mathbf{h}}_{t'} - \Delta\mathbf{h}_{t'}\right\|_F^2,\ \Delta\widehat{\mathbf{h}}_{t'} = \mathcal{G}(\mathbf{h}_{t'}, \mathbf{u}_{t'}^\Delta),\ \Delta\mathbf{h}_{t'} = \mathbf{h}_{t'+\Delta} - \mathbf{h}_{t'}, \tag{7}$$

where $\Delta\widehat{\mathbf{h}}_{t'} \in \mathbb{R}^{B\times d_z}$ is the predicted visual feature change from time t' to $t'+\Delta$. It is predicted using a three-layer predictor $\mathcal{G}$ with cross-attention between $\mathbf{u}_{t'}^\Delta$ and $\mathbf{h}_{t'}$. Here, cross-attention uses $\mathbf{h}_{t'}$ as the query and $\mathbf{u}_{t'}^\Delta$ as the keys and values. $\Delta\mathbf{h}_{t'} \in \mathbb{R}^{B\times d_z}$ is the ground-truth change in

visual features from time t' to $t'+\Delta$, and $\mathbf{h}_{t'+\Delta} = \mathcal{F}(\mathbf{o}_{t'+\Delta})$ is the latent representation of the future observation $\mathbf{o}_{t'+\Delta}$. Since the encoding of hidden state $\mathbf{u}_{t'}^{\Delta}$ incorporates the relevant action embeddings from $\mathcal{D}$, optimizing Eq. (7) encourages $\mathcal{D}$ to capture action-related visual changes, focus on action-relevant objects, and reduce its reliance on irrelevant background during action prediction.

Theorem 1 (Temporal Composition Error Bound) *Let $V \geq 3$ and assume consistent endpoint features from the same trajectory, with interval selection independent of the trajectory and auxiliary randomness. For any fixed parameters θ with finite transition risk, the sampled loss in Eq. (7) is an unbiased estimator of the sampling-weighted risk $\mathcal{R}_{\mathrm{MT}}(\theta)$ defined below. The corresponding composition risk satisfies*

$$\mathcal{C}_{\mathrm{MT}}(\theta) \leq (V-1)(3V-4)\mathcal{R}_{\mathrm{MT}}(\theta). \tag{8}$$

Thus, vanishing expected MT loss implies vanishing mean-square composition error over the future intervals.

Proof. Let X contain a trajectory and the auxiliary randomness defining its interval predictions, and let $\mathcal{I} = \{(a,b) : t+1 \leq a < b \leq t+V\}$. For $(a,b) \in \mathcal{I}$, define $\Delta\widehat{\mathbf{h}}_{a:b} = \mathcal{G}_\theta(\mathbf{h}_a, \mathbf{u}_a^{b-a})$, $\mathbf{r}_{a:b} = \Delta\widehat{\mathbf{h}}_{a:b} - (\mathbf{h}_b - \mathbf{h}_a)$, and $\ell_{a:b}(X;\theta) = \|\mathbf{r}_{a:b}\|_F^2/(Bd_z)$. The sampling distribution satisfies $p_{a:b} = 1/[(V-1)(t+V-a)] \geq 1/(V-1)^2$. Define $\bar{\ell}_{\mathrm{MT}}(X;\theta) = \sum_{(a,b)\in\mathcal{I}} p_{a:b}\ell_{a:b}(X;\theta)$ and $\mathcal{R}_{\mathrm{MT}}(\theta) = \mathbb{E}_X[\bar{\ell}_{\mathrm{MT}}(X;\theta)]$. For the sampled interval $I = (t', t'+\Delta)$, $\mathbb{E}_I[\ell_I(X;\theta)|X] = \bar{\ell}_{\mathrm{MT}}(X;\theta)$, and taking expectation over X proves unbiasedness. For $t+1 \leq a < b < c \leq t+V$, define $\mathbf{r}_{a:b:c} = \Delta\widehat{\mathbf{h}}_{a:b} + \Delta\widehat{\mathbf{h}}_{b:c} - \Delta\widehat{\mathbf{h}}_{a:c}$. Since the true feature differences telescope, $\mathbf{r}_{a:b:c} = \mathbf{r}_{a:b} + \mathbf{r}_{b:c} - \mathbf{r}_{a:c}$. Let $\kappa_{a:b:c} = p_{a:b}^{-1} + p_{b:c}^{-1} + p_{a:c}^{-1}$. Weighted Cauchy–Schwarz gives

$$\frac{\|\mathbf{r}_{a:b:c}\|_F^2}{Bd_z} \leq \kappa_{a:b:c}\bar{\ell}_{\mathrm{MT}}(X;\theta) \leq (V-1)(3V-4)\bar{\ell}_{\mathrm{MT}}(X;\theta). \tag{9}$$

Indeed, $\kappa_{a:b:c} = (V-1)[3V-2(a-t)-(b-t)] \leq (V-1)(3V-4)$. Finally, define $\mathcal{C}_{\mathrm{MT}}(\theta) = \mathbb{E}_X\left[\max_{t+1\leq a<b<c\leq t+V} \|\mathbf{r}_{a:b:c}\|_F^2/(Bd_z)\right]$. Taking the maximum in Eq. (9) and then expectation over X proves Eq. (8). Theorem 1 shows that the MT loss with randomly sampled intervals controls the temporal composition error of visual feature change predictions, providing a theoretical basis for consistent supervision across temporal scales. Lower compounding errors enable the model to better ignore extraneous disturbances that accumulate over long temporal horizons, such as changes in task-irrelevant objects arising from viewpoint shifts, manipulator motion, and incidental scene dynamics during task execution. Since these predictions depend on AED-conditioned action hidden states, AED learns the action experience by focusing on action-relevant visual information across different interaction stages and temporal scales through backpropagation.

3.3 Training and Inference

Training: We jointly train the video and action branches with the flow-matching objective $\mathcal{L}_{\mathrm{FM}}$ and the motion-aware transition loss $\mathcal{L}_{\mathrm{MT}}$. Therefore, the overall optimization $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_m \mathcal{L}_{\mathrm{MT}}, \tag{10}$$

where $\lambda_m = 0.01$ denotes the balancing weight. For $\mathcal{L}_{\mathrm{FM}}$, we set $\lambda_e = 1$ ($e \in \{v, a\}$) in Eq. (1).

Inference: Video and action latents are initialized with Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoised using flow velocities predicted from the observation $\mathbf{o}_t$, proprioceptive state $\mathbf{s}_t$, task instruction $\mathbf{c}$, and AED embeddings $\mathbf{f}^\star$, with K Euler steps of the flow-matching ordinary differential equation (ODE) yielding an action chunk of length H. Deployment requires neither decoding the video latents into pixel-space frames nor evaluating the motion-aware transition predictor.

4 Experiments

4.1 Implementation Details

Following Fast-WAM (Yuan et al., 2026), the video expert (5B) is initialized from Wan2.2 (Team Wan et al., 2025), retaining its video DiT, text encoder, and video VAE. The action expert (1B) adopts the same architectural design as the video branch, with its hidden dimension reduced to $d_a = 1024$. The action tokenizer follows the design of Pertsch et al. (2025). All trainable parameters are optimized using AdamW for 10 epochs on LIBERO with 8 NVIDIA A100 GPUs and for 5 epochs on RoboTwin 2.0 with 32 NVIDIA H100 GPUs. We report success rates on various benchmarks, including LIBERO (Liu et al., 2023), RoboTwin 2.0 (Chen et al., 2026b), and LIBERO Plus (Fei et al., 2025). Physical experiments are conducted on two robotic platforms, Spirit AI MOZ1 and ROKAE AR5. Additional implementation details and evaluation settings are provided in the appendix.

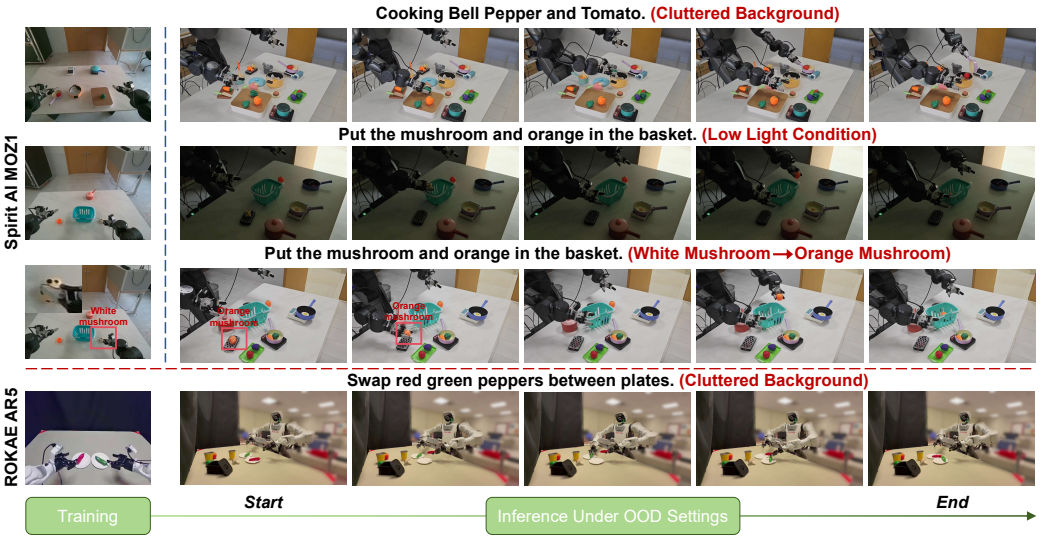
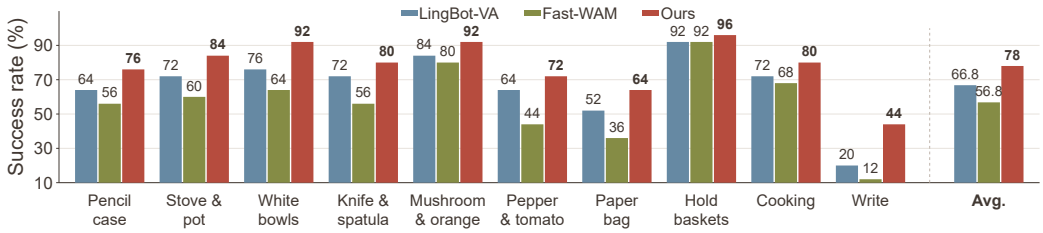
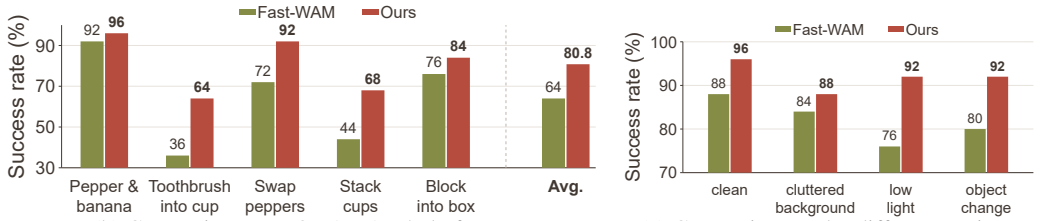


Figure 3: Visualization of manipulation tasks performed by our model in OOD settings.



(a) Comparison on the Spirit AI MOZ1 platform.



(b) Comparison on ROKAE AR5 platform. (c) Comparison under different settings.

Figure 4: Results on real-world manipulation tasks across robotic embodiments under OOD settings.

4.2 Main Comparison Results

Out-of-Distribution (OOD) Performance: Since real-world environments involve various types of perturbations, we evaluate our model under OOD settings using the Spirit AI MOZ1 and ROKAE AR5 platforms. As shown in Fig. 3, we consider three types of OOD conditions: cluttered backgrounds, low light conditions, and unseen objects during training. The visualization of manipulation tasks performed by our model demonstrates its robustness under these OOD conditions. Additionally, as shown in Fig. 4(a)(b), we compare the success rates of our model with those of state-of-the-art baselines (e.g., LingBot-VA and Fast-WAM) across different robotic embodiments under OOD settings. As shown in Fig. 4(c), we further report the success rates on a representative real-world manipulation task, i.e., "Put the mushroom and orange in the basket", under different OOD conditions. Our model consistently achieves higher success rates than the existing methods across different embodiments and OOD conditions, demonstrating the effectiveness of the proposed AED.

Benchmark Comparison: Tab. 1 shows that our method achieves the highest reported average success rates on both RoboTwin 2.0 (92.8%) and LIBERO (98.8%). On **RoboTwin 2.0**, our method achieves success rates of 93.2% under the Clean setting and 92.3% under the Random setting, yielding an average of 92.8%. These results outperform the strongest reported baseline for each metric by 0.3, 0.5, and 0.6 percentage points, respectively. On **LIBERO**, our method outperforms the strongest baseline, Fast-WAM (Yuan et al., 2026), by 1.2 percentage points on average, with improvements of 0.8%, 1.4%, and 2.6% on Spatial, Goal, and Long, respectively, while matching its

Table 1: Success rate (%) on LIBERO and RoboTwin 2.0. LIBERO averages 50 rollouts per task over ten tasks per suite, and RoboTwin 2.0 averages 100 trials per task over 50 tasks in the 'Clean" and 'Random" environments. "PT" indicates whether robotic policy pretraining is used.

Method	PT	RoboTwin 2.0			LIBERO				
		Clean	Random	Avg.	Spatial	Object	Goal	Long	Avg.
OpenVLA (Kim et al., 2025)	✓	–	–	–	84.7	88.4	79.2	53.7	76.5
π_0 (Black et al., 2025b)	✓	65.9	58.4	62.2	96.8	**98.8**	95.8	85.2	94.1
$\pi_{0.5}$ (Black et al., 2025a)	✓	82.7	76.8	79.8	98.8	98.2	**98.0**	92.4	96.9
Motus (Bi et al., 2026)	✓	88.7	87.0	87.8	96.8	**99.8**	96.6	97.6	97.7
Fast-WAM (Yuan et al., 2026)	✗	91.9	**91.8**	91.8	98.2	**100.0**	97.0	95.2	97.6
LingBot-VA (Li et al., 2026)	✓	**92.9**	91.6	**92.2**	98.5	99.6	97.2	**98.5**	**98.5**
AED (Ours)	✗	**93.2**	**92.3**	**92.8**	**99.0**	**100.0**	**98.4**	**97.8**	**98.8**

Table 2: Success rates (%) under different perturbation types on LIBERO-Plus. "PT" indicates whether robotic policy pretraining is used. All success rates are computed over 10,030 trials.

Method	PT	Cam.	Robot.	Lang.	Light.	Back.	Noise.	Layout.	Avg.
OpenVLA (Kim et al., 2025)	✓	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
π_0 (Black et al., 2025b)	✓	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$ (Black et al., 2025a)	✓	**75.4**	**77.5**	**85.6**	**96.9**	**94.6**	**89.7**	**85.7**	**85.7**
Fast-WAM (Yuan et al., 2026)	✗	48.4	**75.7**	66.8	96.6	68.5	70.1	76.5	70.8
AED (Ours)	✗	**93.4**	72.9	**72.3**	**99.6**	**83.1**	**98.4**	**85.3**	**86.2**

Table 3: Results on LIBERO-10.

Variant	MT	VC	AED	Avg.
Ours	✓	✓	✓	**97.4**
– MT		✓	✓	96.8
– VC			✓	95.6
– AED				94.8

Table 4: Inference cost on LIBERO.

Steps	Latency (ms)	Memory (GB)
1	167.3 ± 1.6	24.1
4	284.6 ± 1.6	24.1
10	521.2 ± 3.0	24.1
15	708.5 ± 8.0	24.1

Table 5: Latency (ms) comparisons.

Method	LIBERO Goal	MOZ1
$\pi_{0.5}$	220.0	494.
Motus†	2131.6	2163.7
Fast-WAM	497.1	513.8
Ours	518.7	520.3

100.0% success rate on Object. Our method ranks first on Spatial and Goal and ties for first on Object. As shown in Tab. 2, our method achieves an overall success rate of 86.2% on **LIBERO-Plus**, outperforming both Fast-WAM and $\pi_{0.5}$ without using embodied policy pretraining. It ranks first under the Camera, Light, and Noise perturbations and outperforms Fast-WAM in six of the seven categories, with Robot being the only exception. These results indicate that our robustness gains are concentrated in variations involving viewpoint, lighting, and sensor noise.

4.3 Ablation Study and Efficiency Analysis

Ablation Study: To evaluate the effectiveness of each component, we conduct ablation studies on the proposed motion-aware transition (MT) loss, visually conditioned action embeddings (VC), and learnable action experience dictionary (AED). As presented in Tab. 3, our full model achieves the highest success rate of **97.4%** on LIBERO-10. Removing the MT loss reduces the success rate to 96.8%, showing that MT loss supervision improves action prediction. Removing VC further degrades the performance, while further removing the AED results in a 2.6% lower success rate than that of the full model. These ablation studies demonstrate the effectiveness of our model in reusing action experience across tasks to facilitate the learning of target manipulation tasks.

Efficiency Analysis: As shown in Tab. 4, we measure inference latency and peak GPU memory usage on LIBERO-10, with future frame decoding disabled and the transition predictor removed during inference. Increasing the number of denoising steps from 1 to 15 increases the latency from 167.3 to 708.5 ms, while peak memory usage remains constant at 24.1 GB. Tab. 5 further compares the inference latency of our model on LIBERO Goal and a real-world robotic platform (e.g., Spirit AI MOZ1) against SOTA WAMs (Yuan et al., 2026) and VLA models Black et al. (2025a) using a single NVIDIA A100 GPU, demonstrating comparable efficiency and strong potential for both simulated and real-world deployment. While maintaining efficiency comparable to that of the baselines, our model achieves significant performance improvements (see Tabs. 1–2 and Fig. 4).

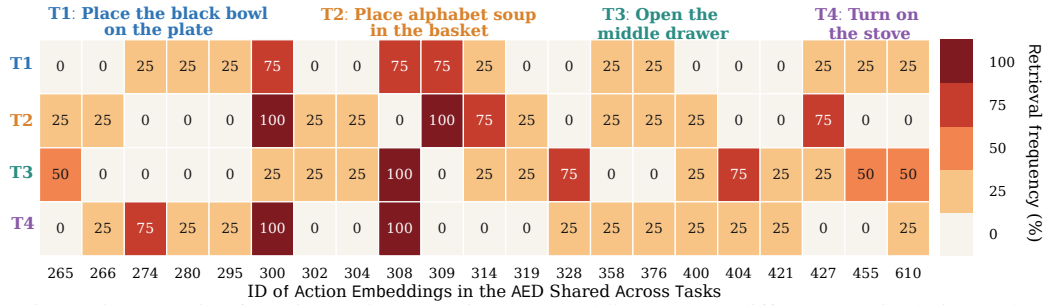
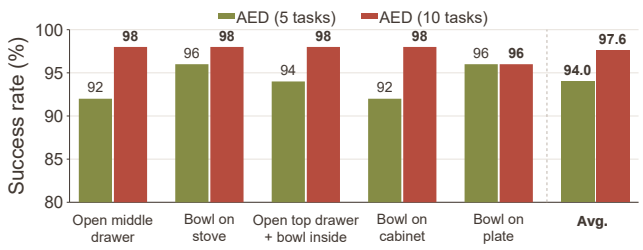
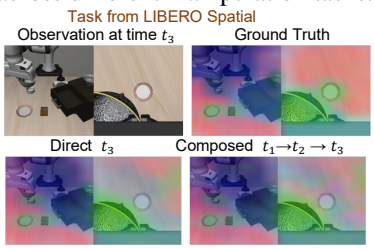


Figure 5: Analysis of reusing action experience embeddings across different manipulation tasks.



(a) Quantitative Analysis of AED.



(b) Direct versus composed transitions.

Figure 6: **(a)** Quantitative performance evaluation of the proposed AED on LIBERO Goal. **(b)** Visualization of direct and composed transition predictions under supervision from the MT loss.

4.4 Analysis of Action Experience Dictionary (AED)

To analyze how different manipulation tasks reuse action embeddings (i.e., skill patterns) shared across tasks, Fig. 5 visualizes the reuse frequencies of shared embeddings across four randomly selected LIBERO tasks. We observe that many action embeddings are frequently reused across these tasks, indicating that the proposed AED can encode action-relevant skill patterns shared across tasks and leverage them to improve the performance of target manipulation tasks during training. A high frequency of reusing the same action embeddings across different tasks suggests stronger cross-task relationships. The proposed model captures such inter-task relationships through shared action embeddings in the AED and leverages them to facilitate future action chunk prediction.

To quantitatively evaluate the efficacy of reusable action experience across tasks, as shown in Fig. 6(a), we randomly select five LIBERO Goal tasks as target tasks and treat the remaining five as auxiliary source tasks. Despite having different goals, these tasks share reusable skill patterns, such as grasping, transporting, and placing objects. We train one model on five target tasks and another on the same target tasks plus five auxiliary source tasks, evaluating both on the same target tasks. In Fig. 6(a), training the proposed AED on all ten tasks consistently yields higher success rates. Such improvement suggests positive transfer from the additional related tasks, i.e., shared skill patterns learned from the five auxiliary source tasks benefit the learning of the five target tasks.

4.5 Analysis of Motion-Aware Transition (MT) Loss

To qualitatively evaluate the MT loss, we visualize its compositional generalization ability in Fig. 6(b). "Direct t_3" predicts the feature change from t_1 to t_3 directly, while "Composed $t_1 \rightarrow t_2 \rightarrow t_3$" predicts it through two consecutive transitions. The two methods produce similar visualization results and identify action-relevant objects (e.g., grippers and bowls), indicating effective transition composition. With the guidance of MT loss, the policy learns to suppress background interference. Additional analyses of sampling strategies, action aggregation, visual encoders, prefix designs, action experience embeddings, and learnable visual queries are provided in the appendix.

5 Conclusion

In this paper, we propose a novel Action Experience Dictionary (AED) for learning skills from historical trajectories. AED encodes trajectories into shared action embeddings and conditions them on visual context to capture task-relevant interactions. We further propose a motion-aware transition loss that predicts visual feature changes over random temporal intervals, encouraging action-centric representations. Experiments across simulation benchmarks and real-world cross-embodiment evaluations demonstrate improved manipulation performance over baseline approaches.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript to assist with literature organization, experimental record organization, table formatting, language revision, and discussion of possible refinements to the methodology and evaluation protocol. AI-assisted suggestions were carefully reviewed by the authors, and all equations, implementation details, experimental procedures, citations, reported results, and interpretations were independently checked. The authors remain fully responsible for the final methodology, experiments, results, code, and manuscript.

ETHICS STATEMENT

This work studies robotic manipulation in both simulated environments and real-world settings. The real-world experiments were conducted within bounded workspaces using predefined operating constraints and appropriate human supervision. The study does not intentionally collect personal or sensitive information; visual observations are limited to the robot and its surrounding workspace. Public benchmark datasets and pretrained components are used according to their respective licenses, with the relevant sources identified in the manuscript and supplementary materials.

REPRODUCIBILITY STATEMENT

The main paper describes the proposed Action Experience Dictionary, visual conditioning mechanism, motion-aware transition objective, and evaluation protocol. The supplementary material provides detailed model configurations, optimization settings, preprocessing procedures, inference parameters, per-task benchmark results, data-collection information, robot-platform details, ablation studies, and additional qualitative analyses. Anonymous code and robotic manipulation demonstrations are provided through the supplementary project page. Together, these materials document the settings and procedures required to reproduce the reported experiments.

REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

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 Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

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 2025b. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html.

Jisong Cai, Long Ling, Shiwei Chu, Zhongshan Liu, Jiayue Kang, Zhixuan Liang, Wenjie Xu, Yinan Mao, Weinan Zhang, Xiaokang Yang, Ru Ying, Ran Zheng, and Yao Mu. Aha-wam: Asynchronous horizon-adaptive world-action modeling with observation-guided context routing, 2026. URL https://arxiv.org/abs/2606.09811.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation, 2024. URL https://arxiv.org/abs/2410.06158.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. Lawam: Latent world action models for efficient dynamics-aware robot policies, 2026a. URL https://arxiv.org/abs/2606.15768.

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 *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=itonej9GIV.

DeepSeek-AI team. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8469–8488, 2023.

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. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: In-depth robustness analysis of vision-language-action models, 2025. URL https://arxiv.org/abs/2510.13626.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception, 2026. URL https://arxiv.org/abs/2607.05247.

Asher Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7a0f8055c838df8e62329a76c7c6403d-Abstract-Conference.html.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories, 2025. URL https://arxiv.org/abs/2505.12705.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of

Proceedings of Machine Learning Research, pp. 2679–2713. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

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://proceedings.iclr.cc/paper_files/paper/2026/hash/748becc400a57c0e31cfe6a2e7951467-Abstract-Conference.html.

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*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.016.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.012. URL https://www.roboticsproceedings.org/rss21/p012.html.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, Jiayuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.15830.

Sandeep Kumar Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. Vipra: Video prediction for robot actions. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/707e34efcabd2b9375f7a64019600aa8-Abstract-Conference.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui

Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Fan Yang, Yuting Su, Xiaobo Wang, Yuncheng You, Fugui Fan, Yuting Wu, Minghui Wu, Chenxu Zhao, JiaHong Ning, and Peiguang Jing. Lila-wam: Lightweight latent reasoning world-action model for robotic manipulation, 2026. URL https://arxiv.org/abs/2608.03701.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666.

Wenyao Zhang, Bozhou Zhang, Zekun Qi, Wenjun Zeng, Xin Jin, and Li Zhang. Disentangled robot learning via separate forward and inverse dynamics pretraining. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/793bfa8f8c8db6e33a7ecf410ae573a8-Abstract-Conference.html.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.015. URL https://www.roboticsproceedings.org/rss21/p015.html.

A Reproducibility and Robotic Manipulation Demos

The real-world and simulation demos and reproducible code are available on the anonymous project: https://anonymous.4open.science/w/supplementary_materials-4586.

B Limitation

Our method has two main limitations. Firstly, a frozen visual model needs to be loaded, which may slow down the optimization speed during training, although this can be solved by caching visual features in advance. Secondly, the finite history window limits access to longer-term dependencies. We will address this issue in our future work.

C Hyperparameters and Implementation Details

C.1 Model and Representation Configuration

Backbone and trainable modules. We initialize the video expert from Wan2.2-TI2V-5B from pretrained Wan2.2 weights, with the ActionDiT action expert initialized by linear interpolation of the parameters from the video expert. Both experts contain 30 layers, with hidden dimensions of 3,072 and 1,024, respectively. We train both experts, the Action Experience Dictionary (AED), the visual memory and action–visual fusion modules, and the transition predictor. The video VAE and the LingBot-vision encoder remain frozen. We employ delta pose to control the robot.

Action Experience Dictionary. The AED contains 2,048 valid entries of dimension 1,024, plus a separate padding (PAD) entry. FAST+ token indices retrieve dictionary embeddings, which are aggregated with temporal positional encoding as described in the main paper.

Visual conditioning. Historical visual features are extracted by the VAE and compressed into 128 embeddings using two memory-compression layers. A single action–visual fusion block combines the visual memory with the action history to produce eight historical prefix tokens for the action expert. The frozen LingBot encoder provides the visual targets for transition supervision.

C.2 Optimization and Training Configuration

We use AdamW with a learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay of 0.01, and a gradient-clipping threshold of 1. The schedule includes 10% warmup updates followed by cosine decay to 10^{-6} at the end of the planned run. Training uses 8 NVIDIA A100 GPUs and 32 NVIDIA H100 GPUs with eight samples per GPU. We use BF16 precision and DeepSpeed ZeRO-1. The full training schedule comprises 10 epochs for LIBERO and 5 epochs for RoboTwin. The video and action flow-matching losses each have a weight of 1. The MT loss weight is 0.01.

C.3 Observation and Action Preprocessing

For single-arm manipulation, the input includes a third-person view and a wrist-camera view. For dual-arm manipulation, we use three views: a head-mounted view and one wrist-camera view for each arm. Each image is resized to 224×224, and the views are concatenated horizontally.

Single-arm actions are represented by seven-dimensional vectors. Actions are normalized using min–max scaling. The history contains 32 action steps, which are aggregated into eight consecutive groups of four steps before FAST+ tokenization.

C.4 Inference and Deployment

We use NVIDIA A100 GPUs for inference. The policy predicts a chunk of 32 future actions using 10 flow-matching ordinary differential equation (ODE) integration steps. We execute the first 10 predicted actions before replanning, at a control frequency of 30 Hz.

Algorithm 1 Pipeline of the Proposed AED-WAM

Input: Training batch $\mathcal{B}$ with instruction $\mathbf{c}$, observations $\mathbf{o}$, proprioceptive states $\mathbf{s}$, historical actions $\mathbf{a}^h$, and flow-matching targets $(\mathbf{x}_v, \mathbf{x}_a)$. At inference, use $(\mathbf{c}, \mathbf{o}_t, \mathbf{s}_t, \mathbf{a}^h, \mathbf{z}_v^h)$, where $\mathbf{z}_v^h = \mathrm{VAE}(\mathbf{o}^h)$.
Output: Action chunk $\widehat{\mathbf{a}}_{t:t+H-1}$.

▷ **Training**

1: Set $k = H/V$; compute $\widehat{\mathbf{a}}^h[j]$ by Eq. (2), tokenize $\zeta_j = \Phi(\widehat{\mathbf{a}}^h[j])$, and retrieve the temporally ordered action embeddings $\mathbf{f}^\star = [\mathbf{p}_j + M^{-1}\sum_l \mathcal{D}(\zeta_j[l])]_{j=1}^V$.
2: Compress historical visual latents $\mathbf{z}_v^h$ to $\widehat{\mathbf{z}}_v^h$ and obtain visually conditioned action embeddings $\mathbf{e}_v = \mathcal{E}(\mathbf{f}^\star, \widehat{\mathbf{z}}_v^h)$; prepend them to noisy action tokens (Eq. (6)).
3: Sample τ and $\boldsymbol{\epsilon}_e$; set $\mathbf{x}_e^\tau = (1-\tau)\mathbf{x}_e + \tau\boldsymbol{\epsilon}_e$ for $e \in \{v, a\}$ and compute $\mathcal{L}_{\mathrm{FM}}$ (Eq. (1)).
4: Sample (t', Δ); predict $\Delta\widehat{\mathbf{h}}_{t'} = \mathcal{G}(\mathbf{h}_{t'}, \mathbf{u}_{t'}^{\Delta})$ and compute $\mathcal{L}_{\mathrm{MT}}$ against $\Delta\mathbf{h}_{t'} = \mathbf{h}_{t'+\Delta} - \mathbf{h}_{t'}$ (Eq. (7)).
5: Update trainable parameters by $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + 0.01\,\mathcal{L}_{\mathrm{MT}}$; keep the video VAE and target encoder $\mathcal{F}$ frozen.

▷ **Inference**

6: Receive $(\mathbf{c}, \mathbf{o}_t, \mathbf{s}_t, \mathbf{a}^h, \mathbf{z}_v^h)$ and repeat Steps 1–2 to construct $\mathbf{e}_v$.
7: Initialize the video and action latents with Gaussian noise at $\tau_0 = 1$; integrate the joint flow-matching ODE backward to $\tau_K = 0$ with K Euler steps, conditioned on $(\mathbf{o}_t, \mathbf{s}_t, \mathbf{c}, \mathbf{e}_v)$. Only the resulting action chunk is executed.
8: Decode $\widehat{\mathbf{a}}_{t:t+H-1}$, execute the first 10 actions at 30 Hz, append the new history and latents, and repeat Step 6.

D Per-Task Benchmark Results

D.1 LIBERO

Table 6 reports the 42K checkpoint's success rates on standard LIBERO, with 1,976 successes over 2,000 trials (98.80%). Task IDs match the S0–S9, O0–O9, G0–G9, and L0–L9 definitions in the task protocol.

Table 6: Per-task success rates on all 40 standard LIBERO tasks, grouped by suite. The same 42,000-step checkpoint is evaluated with seed 3407, 50 trials per task, and replanning after ten executed actions. SR denotes success rate in percent. Each suite average covers 500 trials; the overall average covers 2,000 trials.

ID	Task instruction	Successes	SR (%)
LIBERO-Spatial			
S0	pick up the black bowl between the plate and the ramekin and place it on the plate	50/50	100.00
S1	pick up the black bowl next to the ramekin and place it on the plate	50/50	100.00
S2	pick up the black bowl from table center and place it on the plate	50/50	100.00
S3	pick up the black bowl on the cookie box and place it on the plate	48/50	96.00
S4	pick up the black bowl in the top drawer of the wooden cabinet and place it on the plate	49/50	98.00
S5	pick up the black bowl on the ramekin and place it on the plate	49/50	98.00
S6	pick up the black bowl next to the cookie box and place it on the plate	50/50	100.00
S7	pick up the black bowl on the stove and place it on the plate	49/50	98.00
S8	pick up the black bowl next to the plate and place it on the plate	50/50	100.00
S9	pick up the black bowl on the wooden cabinet and place it on the plate	50/50	100.00
Suite average		**495/500**	**99.00**
LIBERO-Object			
O0	pick up the alphabet soup and place it in the basket	50/50	100.00
O1	pick up the cream cheese and place it in the basket	50/50	100.00
O2	pick up the salad dressing and place it in the basket	50/50	100.00
O3	pick up the bbq sauce and place it in the basket	50/50	100.00
O4	pick up the ketchup and place it in the basket	50/50	100.00
O5	pick up the tomato sauce and place it in the basket	50/50	100.00
O6	pick up the butter and place it in the basket	50/50	100.00

Continued on next page

Table 6: LIBERO per-task results (continued).

ID	Task instruction	Successes	SR (%)
O7	pick up the milk and place it in the basket	50/50	100.00
O8	pick up the chocolate pudding and place it in the basket	50/50	100.00
O9	pick up the orange juice and place it in the basket	50/50	100.00
Suite average		**500/500**	**100.00**
LIBERO-Goal			
G0	open the middle drawer of the cabinet	49/50	98.00
G1	put the bowl on the stove	48/50	96.00
G2	put the wine bottle on top of the cabinet	48/50	96.00
G3	open the top drawer and put the bowl inside	50/50	100.00
G4	put the bowl on top of the cabinet	49/50	98.00
G5	push the plate to the front of the stove	50/50	100.00
G6	put the cream cheese in the bowl	49/50	98.00
G7	turn on the stove	50/50	100.00
G8	put the bowl on the plate	50/50	100.00
G9	put the wine bottle on the rack	49/50	98.00
Suite average		**492/500**	**98.40**
LIBERO-Long (LIBERO-10)			
L0	put both the alphabet soup and the tomato sauce in the basket	49/50	98.00
L1	put both the cream cheese box and the butter in the basket	50/50	100.00
L2	turn on the stove and put the moka pot on it	49/50	98.00
L3	put the black bowl in the bottom drawer of the cabinet and close it	47/50	94.00
L4	put the white mug on the left plate and put the yellow and white mug on the right plate	48/50	96.00
L5	pick up the book and place it in the back compartment of the caddy	50/50	100.00
L6	put the white mug on the plate and put the chocolate pudding to the right of the plate	48/50	96.00
L7	put both the alphabet soup and the cream cheese box in the basket	50/50	100.00
L8	put both moka pots on the stove	49/50	98.00
L9	put the yellow and white mug in the microwave and close it	49/50	98.00
Suite average		**489/500**	**97.80**
Overall average		**1976/2000**	**98.80**

D.2 RoboTwin 2.0

Table 7 compares our method with Fast-WAM (Yuan et al., 2026) on all 50 RoboTwin 2.0 tasks. Our evaluation uses a single checkpoint and 100 trials per task in each of the Clean and Random settings, yielding 10,000 trials in total. Our method succeeds in 4,659/5,000 Clean trials and 4,617/5,000 Random trials, corresponding to 93.18% and 92.34%, respectively, and an overall success rate of 92.76%.

Table 7: Per-task success rate (%) on RoboTwin 2.0. Fast-WAM entries are reported baseline results; ours use 100 trials per task and setting with unseen instructions. Avg. averages Clean and Random, and the final row averages all 50 tasks. Bold indicates the better result between the two methods for each metric, including ties.

	Fast-WAM			Ours		
Task	Clean	Random	Avg.	Clean	Random	Avg.
adjust_bottle	**100.00**	**100.00**	**100.00**	**100.00**	99.00	99.50
beat_block_hammer	**99.00**	97.00	98.00	98.00	**99.00**	**98.50**
blocks_ranking_rgb	**100.00**	**100.00**	**100.00**	**100.00**	99.00	99.50
blocks_ranking_size	**94.00**	**98.00**	**96.00**	92.00	96.00	94.00
click_alarmclock	**100.00**	**100.00**	**100.00**	**100.00**	99.00	99.50
click_bell	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
dump_bin_bigbin	97.00	96.00	96.50	**98.00**	**98.00**	**98.00**
grab_roller	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Continued on next page

Table 7: RoboTwin 2.0 per-task results (continued).

	Fast-WAM			Ours		
Task	Clean	Random	Avg.	Clean	Random	Avg.
handover_block	**95.00**	81.00	88.00	94.00	**84.00**	**89.00**
handover_mic	**99.00**	**100.00**	**99.50**	**99.00**	99.00	99.00
hanging_mug	**58.00**	**62.00**	**60.00**	56.00	58.00	57.00
lift_pot	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
move_can_pot	90.00	88.00	89.00	**99.00**	**98.00**	**98.50**
move_pillbottle_pad	**100.00**	99.00	99.50	**100.00**	**100.00**	**100.00**
move_playingcard_away	**100.00**	**100.00**	**100.00**	99.00	**100.00**	99.50
move_stapler_pad	77.00	64.00	70.50	**86.00**	**76.00**	**81.00**
open_laptop	**98.00**	**100.00**	**99.00**	96.00	**100.00**	98.00
open_microwave	62.00	45.00	53.50	**86.00**	**68.00**	**77.00**
pick_diverse_bottles	80.00	85.00	82.50	**82.00**	**87.00**	**84.50**
pick_dual_bottles	**100.00**	**96.00**	**98.00**	**100.00**	**96.00**	**98.00**
place_a2b_left	95.00	93.00	94.00	**100.00**	**95.00**	**97.50**
place_a2b_right	93.00	**99.00**	96.00	**96.00**	97.00	**96.50**
place_bread_basket	91.00	93.00	92.00	**92.00**	**97.00**	**94.50**
place_bread_skillet	90.00	**93.00**	91.50	**92.00**	**93.00**	**92.50**
place_burger_fries	96.00	**99.00**	97.50	**99.00**	97.00	**98.00**
place_can_basket	71.00	**69.00**	70.00	**81.00**	65.00	**73.00**
place_cans_plasticbox	**99.00**	**96.00**	**97.50**	98.00	94.00	96.00
place_container_plate	96.00	**100.00**	98.00	**99.00**	99.00	**99.00**
place_dual_shoes	**94.00**	88.00	**91.00**	89.00	**91.00**	90.00
place_empty_cup	**100.00**	**100.00**	**100.00**	99.00	**100.00**	99.50
place_fan	**96.00**	**96.00**	**96.00**	94.00	94.00	94.00
place_mouse_pad	83.00	89.00	86.00	**95.00**	**90.00**	**92.50**
place_object_basket	**89.00**	**88.00**	**88.50**	**89.00**	85.00	87.00
place_object_scale	90.00	**97.00**	93.50	**93.00**	96.00	**94.50**
place_object_stand	90.00	**94.00**	92.00	**93.00**	**94.00**	**93.50**
place_phone_stand	97.00	**99.00**	98.00	**99.00**	98.00	**98.50**
place_shoe	**96.00**	**99.00**	**97.50**	94.00	**99.00**	96.50
press_stapler	90.00	**97.00**	93.50	**93.00**	95.00	**94.00**
put_bottles_dustbin	**95.00**	90.00	**92.50**	89.00	**95.00**	92.00
put_object_cabinet	**94.00**	**89.00**	**91.50**	88.00	**89.00**	88.50
rotate_qrcode	93.00	89.00	91.00	**95.00**	**90.00**	**92.50**
scan_object	89.00	92.00	90.50	**93.00**	**94.00**	**93.50**
shake_bottle	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
shake_bottle_horizontally	**100.00**	**100.00**	**100.00**	**100.00**	99.00	99.50
stack_blocks_three	**95.00**	**97.00**	**96.00**	77.00	70.00	73.50
stack_blocks_two	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
stack_bowls_three	80.00	81.00	80.50	**87.00**	**82.00**	**84.50**
stack_bowls_two	92.00	**98.00**	**95.00**	**93.00**	96.00	94.50
stamp_seal	**90.00**	**94.00**	**92.00**	87.00	89.00	88.00
turn_switch	61.00	59.00	60.00	**70.00**	**78.00**	**74.00**
Average	91.88	91.78	91.83	**93.18**	**92.34**	**92.76**

E Training Data, Data Collection, and Robot Platforms

E.1 Training Data Overview

Figure 7 illustrates ten real-world manipulation tasks on the Spirit AI MOZ1 platform. We use a VR device to collect these real-world data.

E.2 Robot Platforms

Figure 8 shows the Spirit AI MOZ1 and ROKAE AR5-5_0.7 platforms used in our real-world experiments, together with the deployment pipeline. The green, red, and orange markers indicate the head-mounted camera, wrist-mounted RealSense cameras, and grippers, respectively. The camera



Figure 7: Overview of ten real-world manipulation tasks on Spirit AI MOZ1. Each row shows six frames from the high camera, covering an episode from start to finish.

views provide observations of the workspace and the regions near the grippers. During deployment, camera images and a natural-language task instruction are passed to the model running on a host computer, which predicts actions and sends them to the robot for execution.

F Analysis of Shared Action Experience

F.1 Cross-Task Similarity of Action Content Embeddings

Manipulation tasks with distinct goals and visual contexts nevertheless share reusable action experience. For example, placing a can into a basket can draw on grasping, transporting, and releasing behaviors also used in other pick-and-place tasks. We examine this underlying relationship among

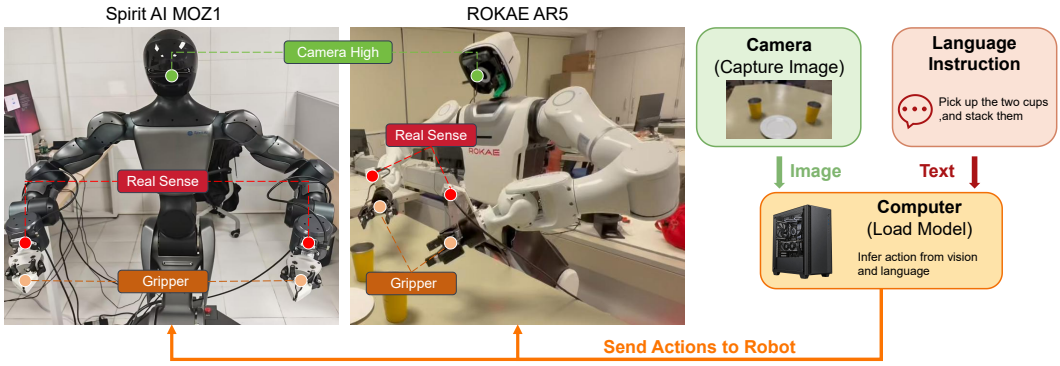


Figure 8: Robot platforms and deployment pipeline. Left and center: Spirit AI MOZ1 and ROKAE AR5-5_0.7 with camera and gripper annotations. Right: camera images and language instructions are processed by the model on a host computer, and predicted actions are sent to the robot.

manipulation tasks from two complementary perspectives: the similarity structure of action embeddings and the reuse of Action Experience Dictionary (AED) entries across task windows.

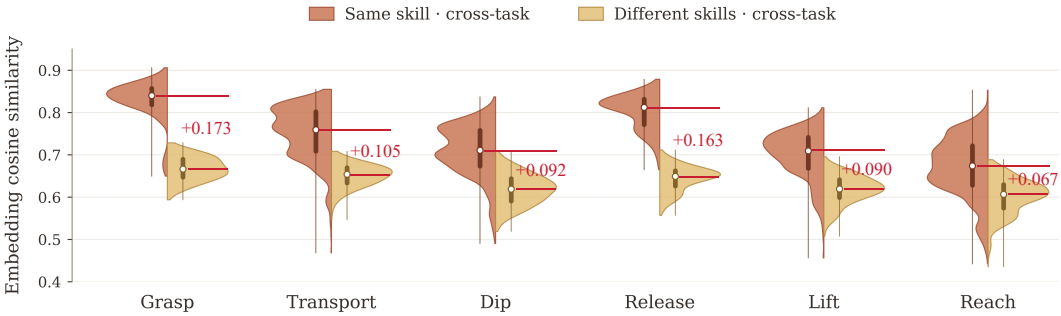


Figure 9: **Cross-task similarity of action content embeddings.** Orange and yellow compare the same skill and different skills across tasks, respectively. Each distribution sample is a weighted average cosine-similarity score for a cross-task pair; white points mark the medians of these task-pair scores. The annotated gaps for grasp, transport, dip, release, lift, and reach are $+0.173$, $+0.105$, $+0.092$, $+0.163$, $+0.090$, and $+0.067$, respectively.

Representation and statistical unit. For each four-action interval, its FAST+ tokens undergo AED lookup and projection, followed by mean pooling over the retrieved valid tokens to produce a 1024-dimensional *action content embedding*. This is the aggregated action representation before adding positional encoding or applying visual cross-attention, rather than a single AED vocabulary vector or a visually conditioned action embedding. Each sample in the violin distributions is a weighted average of cosine similarities for a cross-task pair, not an individual action vector or a single vector-pair similarity. The white points summarize these task-pair scores by their medians.

Cross-task organization of action content. Figure 9 shows higher median task-pair similarity for the same skill than for different skills across all six categories. Both comparison groups are cross-task. The largest annotated gaps occur for grasping (0.173) and releasing (0.163), while transporting also exhibits a positive gap (0.105). These behaviors directly match the reusable action experience in our motivation; dipping, lifting, and reaching extend the pattern beyond pick-and-place.

The overlapping distributions indicate graded relationships among skills rather than completely disjoint categories. Because this analysis precedes positional encoding and visual cross-attention, it locates the observed cross-task structure in the pooled action content itself. This supports the intended role of historical action trajectories as a source of reusable action experience across tasks.

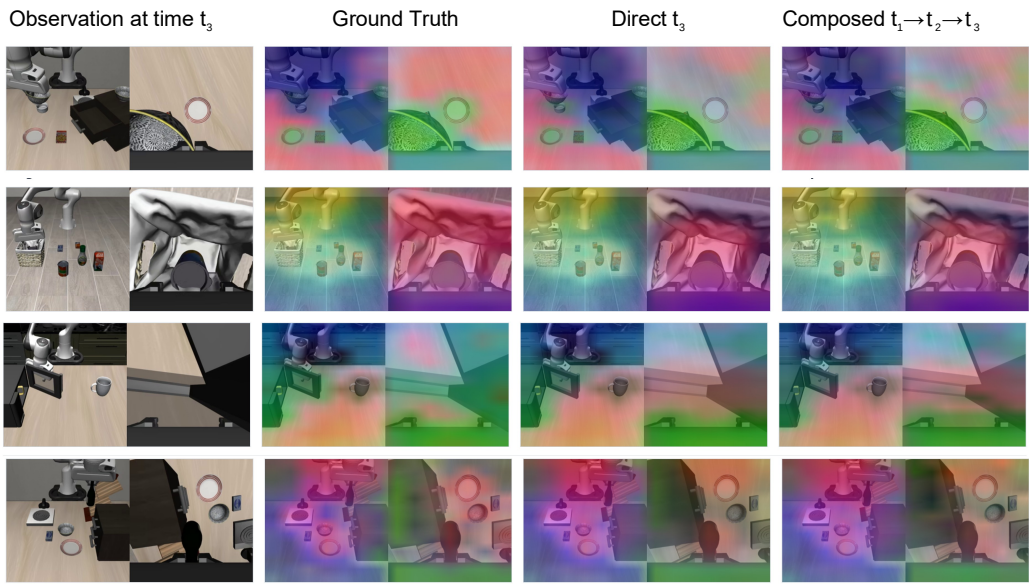


Figure 10: **Additional transition-prediction examples.** Each row shows, from left to right, the observation at t_3, a PCA visualization of its ground-truth features, the endpoint features estimated by direct $t_2 \to t_3$ prediction, and those estimated by composing $t_1 \to t_2$ and $t_2 \to t_3$ predictions. The two prediction routes produce similar spatial feature patterns across the illustrated scenes.

Connection to the proposed mechanism. The pretrained action tokenizer supplies indices into the learnable AED. Repeated IDs establish access to the same dictionary entries, while the similarity distributions characterize the pooled, projected action content embeddings. An individual token ID need not denote an entire high-level skill. These complementary views connect shared dictionary access with the organization of action experience across tasks.

F.2 Shared AED Entries across Manipulation Tasks

Retrieval frequency. For each selected task window, the eight original four-action intervals are paired into four eight-action statistical groups. This grouping retains the original tokenization; it does not re-tokenize eight-action segments. Let $\mathcal{G}_{t,g}$ be the set of token IDs occurring in group g of the selected window for task t. Then

$$\text{Frequency}(t,k) = \frac{1}{4}\sum_{g=1}^{4} \mathbf{1}\{k \in \mathcal{G}_{t,g}\} \times 100\%. \tag{11}$$

Thus, 75% means presence in three of the four groups, irrespective of repetitions within a group. Retrieval is restricted to the selected window and is not a whole-task token frequency.

Shared dictionary access. In Fig. 5, ID 300 occurs in all four task windows, with frequency of 75%, 100%, 25%, and 100% for T1–T4. ID 309 has frequency of 75% and 100% in the two pick-and-place windows (T1 and T2); ID 308 has frequency of 75%, 100%, and 100% in T1, T3, and T4. These overlapping, non-identical sets reveal shared AED access across different objects and goals, spanning both the pick-and-place examples in our motivation and drawer and stove interactions.

As emphasized in the Introduction, similar motions can serve different purposes depending on objects and their spatial relationships. Shared AED entries supply reusable action experience, while subsequent visual conditioning supplies interaction context and action intent for the target task. These results indicate combining common action content with task-relevant visual information to exploit underlying relationships among manipulation tasks.

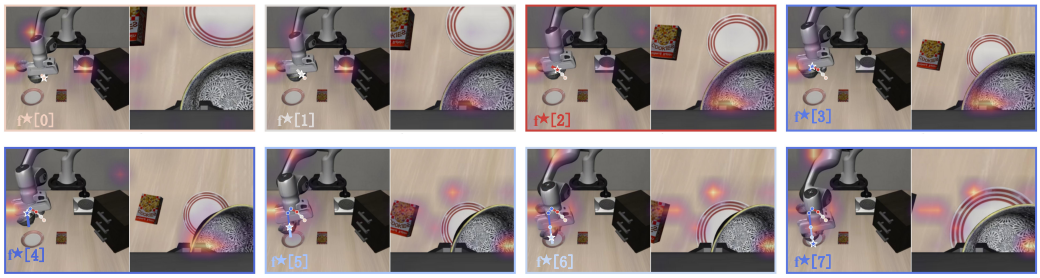


Figure 11: **Visual context associated with historical action embeddings.** Panels are ordered from left to right and top to bottom, corresponding to $\mathbf{f}^\star[0]$–$\mathbf{f}^\star[7]$. Each panel pairs two camera views with overlaid visual responses and trajectory markers. The responses vary with the interaction stage and include regions around the robot gripper and nearby objects.

G Additional Qualitative Results

G.1 Transition Prediction

Figure 10 examines whether the learned transitions describe visual changes consistently across temporal intervals. For three time steps $t_1 < t_2 < t_3$, the direct prediction adds the predicted change over $t_1 \to t_3$ to the features at t_1, whereas the composed prediction adds the changes over $t_1 \to t_2$ and $t_2 \to t_3$. Both therefore estimate the same endpoint features at t_3. The ground-truth column visualizes features extracted from the observation at t_3.

Across the four examples, the direct and composed predictions exhibit similar spatial organization and broadly reproduce the ground-truth feature patterns around the robot, scene objects, and surrounding surfaces. Agreement with the ground truth indicates that the two routes capture meaningful endpoint structure, while agreement between the routes supports temporal composition consistency: subdividing an interval yields a compatible estimate of the resulting visual state. This observation is consistent with the main-paper analysis linking the motion-aware transition (MT) objective to composition error, and supports learning action-related visual transitions across temporal scales.

G.2 Visual Conditioning

Figure 11 illustrates how historical action embeddings are associated with visual regions over a manipulation sequence. The eight panels correspond to $\mathbf{f}^\star[0]$ through $\mathbf{f}^\star[7]$ in temporal order, with two camera views in each panel. The overlaid responses show the visual regions associated with each action embedding, and the corresponding visualized trajectories are also provided.

The response patterns change as the gripper moves through the scene. In the earlier panels, visible responses occur near the robot and objects adjacent to the gripper; in later panels, responses also appear around the gripper–object interaction and the plate region as the trajectory approaches it. This stage-dependent association is consistent with visual conditioning supplying the object and spatial context needed to interpret historical motions. Such context matters because a reusable motion alone does not specify which object it acts on or how that object relates to the current goal.

Together with the shared action-content analysis in Appendix F.1, these examples support complementary roles for the two components: AED provides reusable action experience, and visual conditioning associates that experience with the observed interaction context. The localized responses are also consistent with the intended emphasis of MT supervision on action-relevant visual information.

H Ablation Studies

H.1 Experimental Protocol

The ablation uses seed 0, 32 predicted actions per call, ten executed actions before replanning, with 50 trials per task across ten LIBERO-10 tasks. SR denotes success rate in percent. All six ablation plots use 97.4% as the shared full-model reference.

H.2 Action Aggregation and Temporal Sampling

Action aggregation. In Fig. 12, Agg→Tok aggregates each four-action interval before FAST+ tokenization, Tok→Agg jointly tokenizes the four actions before pooling their embeddings, and MLP encodes the aggregated continuous actions with a multilayer perceptron. The Agg→Tok (Ours) outperforms the other methods.

Temporal sampling. Figure 13 compares Start→Span, which samples a start first and then a valid span, Span→Start, which reverses this order, a fixed temporal window, and the configuration without interval sampling. The 96.8% result jointly removes random sampling and the motion-aware transition (MT) objective.

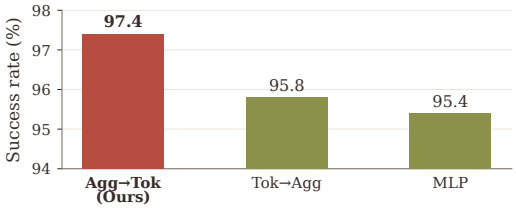


Figure 12: Action aggregation.

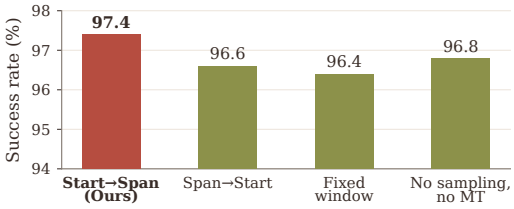


Figure 13: Temporal sampling.

H.3 Sampling Window Distribution

To clarify the Start→Span strategy, Fig. 14 visualizes the probability of sampling each temporal window for motion-aware transition supervision. Let $r = t' - t$ denote the start offset in visual intervals. The sampler first draws r uniformly from $\{1, \ldots, V-1\}$, then draws Δ uniformly from $\{1, \ldots, V-r\}$. Consequently,

$$P(r, \Delta) = \frac{1}{(V-1)(V-r)}, \qquad 1 \le r \le V-1, \quad 1 \le \Delta \le V-r, \tag{12}$$

with zero probability outside this support. Summing over valid start offsets gives

$$P(\Delta) = \frac{1}{V-1} \sum_{r=1}^{V-\Delta} \frac{1}{V-r}, \qquad \Delta \in \{1, \ldots, V-1\}. \tag{13}$$

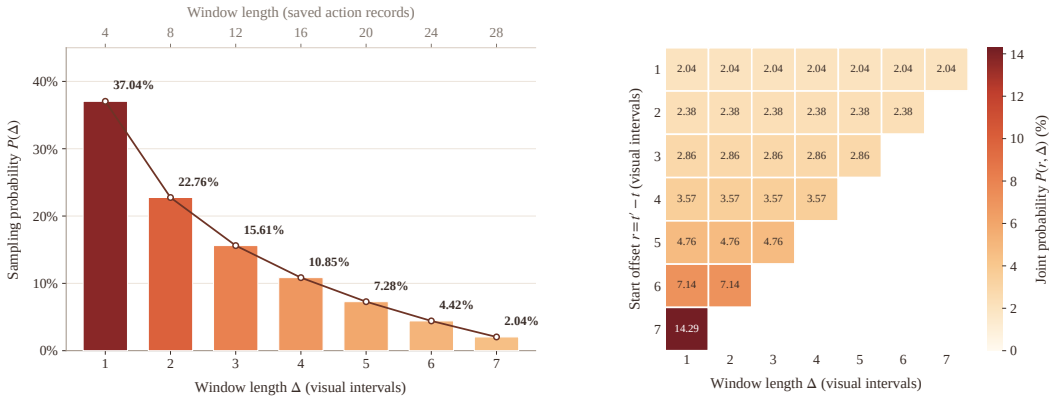


Figure 14: **Temporal-window sampling probabilities for Start→Span**, with $V = 8$. Left: marginal length distribution $P(\Delta)$; the upper axis counts saved action records (four per visual interval). Right: joint distribution $P(r, \Delta)$ over start offsets and lengths; blank cells are invalid windows. All values are percentages.

Uniform draws at each stage do not give uniform window lengths: for $V = 8$, $P(\Delta = 1) = 37.04\%$ and $P(\Delta = 7) = 2.04\%$. Short windows are valid at more start offsets, including late starts with few remaining choices. The resulting MT supervision emphasizes short-term visual transitions while retaining positive probability for every valid longer interval.

H.4 Visual Encoder and Query Count

Pretrained visual encoder. Figure 15 compares LingBot-Vision and DINOv3 as alternative pretrained encoders for visual feature extraction.

Visual-memory query count. Figure 16 varies the number of learnable visual-memory queries across 32, 64, 128, and 256, with 128 as the proposed setting.

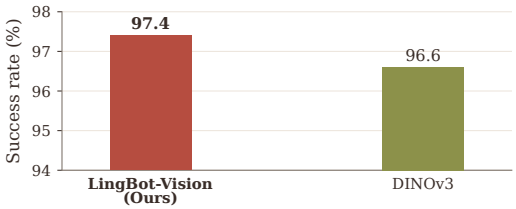


Figure 15: Pretrained visual encoder.

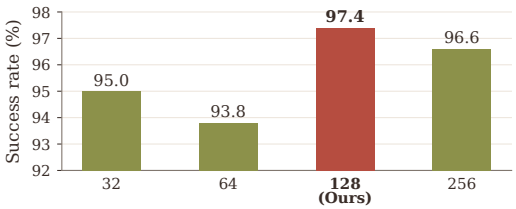


Figure 16: Visual-memory query count.

H.5 Prefix Content and Injection

Prefix content. In Fig. 17, Vision-only constructs the history prefix from historical visual features, whereas Action-only uses historical action embeddings without VC. The 95.6% Action-only setting also removes MT; the 94.8% setting jointly removes AED, VC, and MT.

Prefix injection. In Fig. 18, Prefix prepends the history tokens to the action sequence, whereas Cross-attn conditions the action expert through separate residual cross-attention modules.

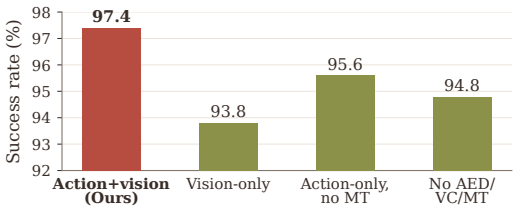


Figure 17: Prefix content.

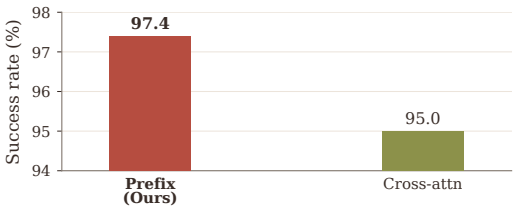


Figure 18: Prefix injection.