UATL: UNCERTAINTY-AWARE TEMPORAL LEARNING FOR OPEN-LOOP VISION-LANGUAGE-ACTION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models map image observations and language instructions directly to robot actions, but the need for a full forward pass for each action prediction hinders the action rates required for control. Action chunk addresses this by predicting and executing multiple actions open-loop from a single policy query, reducing policy query frequency and increasing VLA action throughput. However, existing action chunk VLA methods based on deterministic continuous regression treat all timesteps and action dimensions within a chunk uniformly, capturing neither differences in predictive confidence nor temporal variation between consecutive actions. Consequently, local prediction errors accumulate during open-loop execution, degrading policy stability in fine-grained manipulation and long-horizon sequential tasks. To address this issue, we propose an Uncertainty-Aware Temporal Learning (UATL) framework for open-loop vision-language-action control. Specifically, we introduce Heteroscedastic Action chunk Learning (HAL) to model continuous action prediction as a conditional Laplace distribution and jointly learn the action mean together with per-timestep, per-dimension scale parameters via negative log-likelihood, thereby explicitly characterizing aleatoric uncertainty in action prediction. Furthermore, we design a second-order Uncertainty-Weighted Temporal Consistency Regularizer (UW-TCR) that uses the reciprocal of the predicted scale as a per-timestep, per-dimension regularization weight, suppressing irregular jitter during open-loop execution while preserving the rapid action changes required at contact and direction reversals. Finally, evaluations on LIBERO, LIBERO-Plus, SimplerEnv, and CALVIN show improved performance in fine-grained manipulation, distribution-shift robustness, and long-horizon control, with negligible additional inference overhead.

1 INTRODUCTION

Vision-language-action (VLA) models transfer knowledge from large-scale vision-language pretraining to map visual observations and natural-language instructions to robot actions, providing a unified framework for open-vocabulary instruction understanding, cross-task knowledge transfer, and end-to-end manipulation (Ma et al., 2026; Zhang et al., 2025; Shao et al., 2025; Xu et al., 2025; Zhong et al., 2025). RT-1 (Brohan et al., 2022) demonstrates the value of large-scale multitask robot data, while RT-2 (Brohan et al., 2023) jointly trains on discretized robot actions and Internet-scale vision-language data to improve semantic reasoning and zero-shot generalization. Open X-Embodiment (O'Neill et al., 2024) subsequently consolidates real-world data from diverse robots and tasks, supporting cross-embodiment pretraining. Models such as OpenVLA (Kim et al., 2025b), Octo (Team et al., 2024), and π_0 (Black et al., 2024a) have further advanced generalist robot policies through open-source VLA development, cross-embodiment learning, and generative continuous-action modeling. Despite these advances, the inference cost of large multimodal backbones remains a challenge for high-frequency, low-latency control (Yue et al., 2024; Wen et al., 2025).

Recent work has sought to address this challenge through more efficient action representations and decoding strategies. FAST (Pertsch et al., 2025) compresses action sequences to reduce autoregressive token counts; PD-VLA (Song et al., 2025) employs parallel fixed-point iterations to mit-

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.

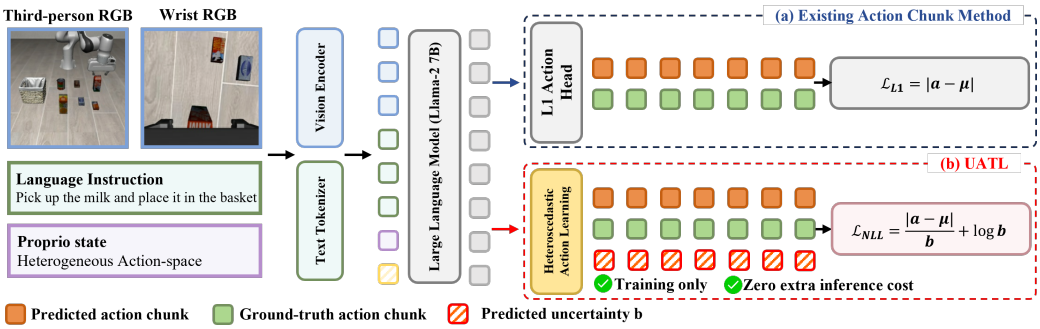

Figure 1: *Existing Action Chunk **vs.** UATL.* (a) An existing action chunk method predicts an action chunk with an L1 head and is trained by the per-step loss against ground-truth actions. (b) UATL keeps the same VLA backbone and the same open-loop chunk execution, and changes only training and adds no extra inference cost.

igate the bottleneck of sequential decoding; and EdgeVLA (Budzianowski et al., 2025) eliminates autoregressive dependencies in action prediction. OpenVLA-OFT (Kim et al., 2025a) combines continuous-action prediction, parallel decoding, and action chunk to generate actions for multiple control steps in a single policy query. By reducing the frequency of expensive policy queries, action chunk increases action throughput.

As we know, action chunk improves execution efficiency by reducing the feedback frequency. As shown in Figure 1(a), under fixed-chunk execution the policy predicts future actions solely from the observation at the beginning of each chunk and cannot promptly adapt to state changes during execution. Perception noise or execution deviations may therefore accumulate within the chunk, causing trajectory drift or even task failure. Existing methods mainly focus on action generation speed and average prediction accuracy, while paying limited attention to variations in prediction reliability across time steps. To reduce this error accumulation and improve task success rates while preserving execution efficiency, we improve only the training process and keep the original policy query frequency and execution protocol unchanged. Our key idea is to model the reliability of the prediction in time steps and action dimensions and adapt the temporal constraints accordingly, as shown in Figure 1(b). We propose **Uncertainty-Aware Temporal Learning (UATL)**, an action chunk learning framework for VLA control. First, **Heteroscedastic Action chunk Learning (HAL)** with conditional Laplace distributions to estimate aleatoric uncertainty. Second, a **second-order Uncertainty-Weighted Temporal Consistency Regularizer (UW-TCR)** adjusts temporal constraints using the estimated uncertainty. Smooths actions in high-confidence regions to reduce jitter, while allowing rapid changes in uncertain regions when needed.

Our main contributions are summarized as follows:

- **UATL:** We propose UATL, an uncertainty-aware VLA framework that jointly models the reliability of action predictions and intra-chunk temporal structure without modifying the model backbone, action chunk protocol, or inference pipeline. UATL mitigates error accumulation during open-loop execution while preserving the efficiency benefits of action chunk.
- **HAL:** We extend deterministic action regression to conditional Laplace distribution modeling, enabling the model to jointly predict action means and uncertainty for each time step and action dimension.
- **UW-TCR:** We propose a second-order uncertainty-weighted temporal consistency regularizer constraint that uses detached uncertainty estimates to adaptively adjust the constraint strength across time steps and action dimensions, suppressing nonphysical oscillations while preserving rapid action changes required by the task.
- **Evaluation across complementary benchmarks.** Experiments on LIBERO, LIBERO-Plus, SimplerEnv, and CALVIN assess standard manipulation, distribution-shift robustness, cross-platform evaluation, and long-horizon control. UATL achieves strong performance with negligible additional inference overhead.

2 Related Work

2.1 Uncertainty Modeling in Robot Policy Learning

Predictive uncertainty is commonly divided into aleatoric and epistemic uncertainty (Loquercio et al., 2020; Senanayake, 2024). Aleatoric uncertainty arises from sensor noise, variation among demonstrations, and ambiguity in robot actions. A heteroscedastic likelihood can model it with input-dependent variance or scale estimates in a single forward pass (Seitzer et al., 2022). Epistemic uncertainty reflects limited data coverage or model capacity; estimating it commonly requires repeated stochastic predictions or model ensembles, increasing latency and memory use (Gawlikowski et al., 2023). Existing VLA methods often use uncertainty during inference. V-GPS (Nakamoto et al., 2024) reranks candidate actions with a learned value function, while MG-Select (Jang et al., 2026) filters candidates using KL divergence from a reference distribution. PDF (Zang et al., 2026) combines uncertainty-aware data augmentation with delayed feedback to correct overconfident predictions. Such approaches can improve reliability but may require candidate generation, additional scoring, or online correction. UATL instead learns per-element scale estimates through HAL during policy training and uses them as reliability proxies for UW-TCR. It obtains action predictions and scales in one policy query without additional backbone evaluations.

2.2 Temporal Consistency in Action Chunk

Smooth trajectories are a standard prior in robot control. Classical trajectory optimization penalizes higher-order temporal derivatives, such as acceleration and jerk, to limit abrupt changes (Fang et al., 2026). Action chunk jointly predicts future controls, but element-wise regression does not explicitly constrain their temporal relationships. Independently generated neighboring chunks can also be discontinuous at their boundaries. Several methods address temporal continuity during inference. ACT (Zhao et al., 2023) ensembles overlapping action predictions; CogACT (Li et al., 2024a) adjusts ensemble weights according to prediction similarity; and RTC (Black et al., 2026) asynchronously generates and patches subsequent chunks to improve their transitions. Diffusion Policy (Chi et al., 2025) models temporal dependencies within a chunk, but uses iterative denoising. Inference-time aggregation primarily addresses relationships across policy queries, whereas deterministic parallel-regression policies still lack an explicit training-time constraint on within-chunk action dynamics. UW-TCR addresses this gap by applying a scale-weighted second-order temporal penalty to predicted action locations. It preserves the original policy-query and execution protocols and adds only marginal action-head computation.

3 Method

3.1 Overview

UATL keeps the backbone, the policy query frequency, and the open-loop execution protocol of an action chunk VLA, and changes only the action head and the training objective, as shown in Figure 2, including two components:HAL and UW-TCR.

3.2 Heteroscedastic Action Learning

While using an L_1 objective applies a uniform error metric across all time steps and action dimensions, the difficulty of predicting robot actions varies across phases such as free motion, contact, and grasping. Deterministic regression struggles to adjust the strength of supervision in the presence of non-uniform noise. To address this, we employ heteroscedastic Laplace modeling to jointly predict actions and element-wise uncertainties, thereby providing a basis for reliability-weighted learning and subsequent temporal constraints, as detailed below:

First, to achieve element-wise reliability modeling, while retaining the core residual MLP structure, the output dimension of the linear projection at the end of the action head is expanded from D to $2D$. This enables the action head to jointly predict action location parameters and log-scales, generating a log-scale output in parallel with the original location output:

$$[\boldsymbol{\mu}_h, \tilde{\mathbf{s}}_h] = g_\psi(\bar{\mathbf{z}}_h), \tag{1}$$

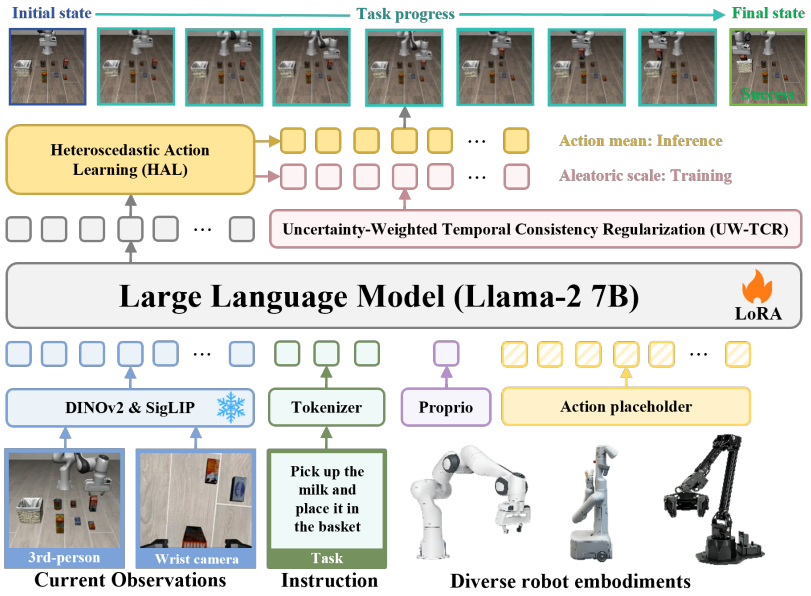

Figure 2: Overview of UATL. HAL predicts action locations and scales, while UW-TCR uses the scales to regularize within-chunk temporal dynamics during training.

where g_ψ is the action head parameterized by ψ, $\bar{\mathbf{z}}_h$ is the action feature at chunk step h, and $\boldsymbol{\mu}_h, \tilde{\mathbf{s}}_h \in \mathbb{R}^D$ are the location and unconstrained log-scale vectors.

Next, to ensure numerical stability during the optimization of the exponential mapping and negative log-likelihood, element-wise truncation is applied to the log-scale. This truncation stabilizes the log-scale by preventing excessively small values from causing an abnormal increase in error weights and gradients, and by preventing excessively large values from overly suppressing the prediction error:

$$\mathbf{s}_h = \mathrm{clip}\left(\tilde{\mathbf{s}}_h, s_{\min}, s_{\max}\right), \qquad \mathbf{b}_h = \exp(\mathbf{s}_h). \tag{2}$$

Here, clipping and exponentiation act element-wise; $s_{\min}$ and $s_{\max}$ are the lower and upper log-scale bounds, and $\mathbf{b}_h \in \mathbb{R}^D_{>0}$ is the resulting Laplace scale vector. Conditioned on the policy input, each expert action element follows:

$$a^*_{h,d} \sim \mathrm{Laplace}(\mu_{h,d}, b_{h,d}), \tag{3}$$

where $a^*_{h,d}$ is the expert action in dimension d at chunk step h, and $b_{h,d}$ is the Laplace scale parameter characterizing the aleatoric uncertainty of the corresponding action prediction.

Finally, given the multimodal inputs, the action chunks are modeled using an element-wise factorized Laplace likelihood. Averaging the negative log-likelihood over a training batch of B samples yields the heteroscedastic action chunk learning objective:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{BHD} \sum_{n=1}^{B} \sum_{h=1}^{H} \sum_{d=1}^{D} \left[\left| a^\star_{n,h,d} - \mu_{n,h,d} \right| \exp(-s_{n,h,d}) + s_{n,h,d} \right]. \tag{4}$$

where B is the batch size; n, h, and d index the sample, chunk step, and action dimension, respectively. The inverse scale adaptively weights the prediction error, while the log-scale term discourages arbitrarily large scales. We use the learned scales as input-dependent proxies for aleatoric uncertainty and their inverses as proxies for prediction reliability.

3.3 Uncertainty-Weighted Temporal Consistency Regularization

During open-loop execution, uncorrected local oscillations may cause end-effector trajectory drift. Uniform smoothing is unsuitable because stable motion requires stronger local consistency, whereas contact establishment and direction changes require rapid action transitions. We therefore introduce

a second-order uncertainty-weighted temporal consistency regularizer whose strength adapts to the estimated prediction reliability at each time step and action dimension, as detailed below.

First, we regularize the second-order finite differences of predicted action means to capture intra-chunk temporal structure. Unlike first-order regularization, which favors locally constant actions, this constraint preserves gradual action changes while suppressing abrupt changes in local curvature. For sample n, time step h, and action dimension d, the second-order difference is defined as:

$$\Delta^2 \mu_{n,h,d} = \mu_{n,h,d} - 2\mu_{n,h-1,d} + \mu_{n,h-2,d}. \tag{5}$$

where a larger $|\Delta^2 \mu_{n,h,d}|$ indicates strong local temporal variation in the action sequence at the corresponding position; constraining this value can reduce abrupt changes and high-frequency oscillations within action segments.

Next, employing reliability-adaptive temporal constraints, element-wise reliability weights are constructed using the inverse scales predicted by the heteroscedastic action head:

$$r_{n,h,d} = \mathrm{sg}(\exp(-s_{n,h,d})) = \mathrm{sg}\left(b_{n,h,d}^{-1}\right), \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Smaller scales yield stronger temporal constraints, while larger scales relax them to preserve necessary rapid action changes.

Finally, since gripper states are discrete, temporal constraints are applied only to the continuous action dimensions $\mathcal{D}_c = \{1, \ldots, D-1\}$. The second-order temporal consistency loss is defined as:

$$\mathcal{L}_{\mathrm{TCR}} = \frac{1}{B(H-2)|\mathcal{D}_c|} \sum_{n=1}^{B} \sum_{h=3}^{H} \sum_{d \in \mathcal{D}_c} r_{n,h,d} \left(\Delta^2 \mu_{n,h,d}\right)^2. \tag{7}$$

Here, $|\mathcal{D}_c|$ denotes the number of continuous action dimensions subject to regularization; this objective focuses on suppressing local temporal abruptness at smaller prediction scales while adaptively relaxing constraints at larger scales, thereby avoiding the excessive suppression of necessary action variations caused by uniform smoothing.

3.4 Training Objective and Inference

Heteroscedastic action chunk learning head jointly predicts action locations and log-scales and is trained with $\mathcal{L}_{\mathrm{NLL}}$ and reliability-weighted $\mathcal{L}_{\mathrm{TCR}}$. Inference preserves the original chunk and open-loop protocols, generating deterministic action chunks from predicted locations in one forward pass without sampling or online smoothing. The overall training objective is:

$$\mathcal{L}_{\mathrm{UATL}} = \mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{TCR}} \mathcal{L}_{\mathrm{TCR}}. \tag{8}$$

where λ_{TCR} controls the temporal regularization strength and is set to 0.05 in the main experiments.

4 Experiments

4.1 Simulation Benchmarks

We evaluate UATL on four complementary simulation benchmarks: LIBERO (Liu et al., 2023) for language-conditioned multitask manipulation, LIBERO-Plus (Fei et al., 2025) for robustness to systematic perturbations, SimplerEnv (Li et al., 2024d) for cross-platform visual transfer of policies trained on real-world data, and CALVIN (Mees et al., 2022) for long-horizon instruction following in an unseen environment.

4.2 Training Details

All models are trained on 4× NVIDIA A6000 GPUs, each with 48 GB of memory. We initialize our model with the pretrained weights of OpenVLA-7B and fine-tune it using LoRA with rank $r = 32$. We optimize the model using AdamW with an initial learning rate of 5×10^{-4}. The per-GPU batch size is 4, and the gradient accumulation factor is 1. The model is trained for 300k steps on the mixed training dataset, with checkpoints saved every 10k steps.

Table 1: Comparison on the LIBERO benchmark. Bold and underlined values indicate the best and second-best results, respectively.

Method	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
Diffusion Policy (Chi et al., 2025)	78.3	92.5	68.3	50.5	72.4
Octo (Team et al., 2024)	78.9	85.7	84.6	51.1	75.1
OpenVLA (Kim et al., 2025b)	84.7	88.4	79.2	53.7	76.5
π_0 (Black et al., 2024a)	96.8	98.8	95.8	85.2	94.2
π_0-FAST (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
$\pi_{0.5}$ (Intelligence et al., 2025)	<u>98.4</u>	98.4	95.8	94.0	96.7
UniVLA (Bu et al., 2025)	96.5	96.8	95.6	92.0	95.2
SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1
CoT-VLA (Zhao et al., 2025)	87.5	91.6	87.6	69.0	81.1
Dita (Hou et al., 2025)	97.4	94.8	93.2	83.6	92.3
PD-VLA (Song et al., 2025)	95.5	96.7	94.9	91.7	94.7
STAR (Li et al., 2025)	95.5	98.3	95.0	88.5	94.3
JALA-dino (Luo et al., 2026)	96.2	98.2	<u>97.4</u>	<u>95.8</u>	96.9
DreamVLA (Zhang et al., 2026a)	97.5	94.0	89.5	89.5	92.6
ThinkAct (Huang et al., 2026a)	88.3	91.4	87.1	70.9	84.4
UD-VLA (Chen et al., 2026)	94.1	95.7	91.2	89.6	92.7
PixelVLA (Liang et al., 2026)	88.5	90.0	85.8	82.6	86.7
MemoryVLA (Shi et al., 2026)	<u>98.4</u>	98.4	96.4	93.4	96.7
CronusVLA (Li et al., 2026)	97.3	<u>99.6</u>	96.9	94.0	<u>97.0</u>
SemanticVLA (Ni et al., 2026)	98.0	<u>98.6</u>	96.8	94.4	<u>97.0</u>
Mantis (Yang et al., 2026)	**98.8**	99.2	94.4	94.2	96.7
UATL (Ours)	<u>98.4</u>	**99.8**	**98.2**	**96.8**	**98.3**

Table 2: Success rates (%) under seven perturbation types on LIBERO-Plus. Bold and underlined values indicate the best and second-best results, respectively.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
OpenVLA (Kim et al., 2025b)	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
π_0 (Black et al., 2024a)	13.8	6.0	58.8	85.0	81.4	<u>79.0</u>	68.9	53.6
π_0-FAST (Pertsch et al., 2025)	<u>65.1</u>	21.6	61.0	73.2	73.2	74.4	68.8	61.6
WorldVLA (Cen et al., 2025)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
NORA (Hung et al., 2025)	2.2	37.0	65.1	45.7	58.6	12.8	62.1	39.0
UniVLA (Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
FastWAM (Yuan et al., 2026b)	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
ForeWAM (Huang et al., 2026b)	62.5	37.4	73.0	74.1	55.1	58.8	**70.4**	61.6
CofactVLA (Zhang et al., 2026b)	44.7	**49.7**	71.8	85.6	<u>83.6</u>	78.0	<u>70.2</u>	69.1
Evo-Depth (Lin et al., 2026)	47.2	<u>49.2</u>	<u>78.9</u>	<u>88.1</u>	76.4	77.6	69.6	<u>69.6</u>
UATL (Ours)	**72.2**	23.6	**79.3**	**88.4**	**88.8**	**84.3**	61.7	**71.2**

4.3 Main Results

Results on LIBERO. As shown in Table 1, UATL achieves the highest average success rate of 98.3% across the four LIBERO suites. It ranks first on Object, Goal, and Long, with success rates of 99.8%, 98.2%, and 96.8%, respectively, and ties for second on Spatial with 98.4%. These results demonstrate consistently strong performance across the four suites, including long-horizon tasks.

Results on LIBERO-Plus. As shown in Table 2, UATL achieves the highest aggregate success rate of 71.2% across seven perturbation types. It ranks first under Camera, Language, Light, Background, and Noise perturbations, reaching 72.2%, 79.3%, 88.4%, 88.8%, and 84.3%, respectively. These results demonstrate strong overall robustness, while Robot and Layout perturbations remain challenging, with success rates of 23.6% and 61.7%, respectively.

Results on SimplerEnv. Table 3 reports results on Google Robot. Under Visual Matching, UATL achieves the highest average success rate of 69.3% and ranks first on Pick Coke Can with 100.0%. Under Variant Aggregation, its average success rate decreases to 54.3%, with Open/Close Drawer showing the largest task-level decline, from 50.0% to 25.0%. On WidowX, Table 4 shows that UATL achieves the highest average success rate of 63.5%. It ranks first on Eggplant in Basket with 95.8% and second on Stack Block with 41.7%, while achieving 58.3% on both Spoon on Towel and Carrot

Table 3: Success rates (%) on Google Robot tasks in SimplerEnv. Bold and underlined values indicate the best and second-best results, respectively.

Method	Visual Matching (VM)				Variant Aggregation (VA)			
	Pick Coke Can	Move Near	Open/Close Drawer	Avg.	Pick Coke Can	Move Near	Open/Close Drawer	Avg.
Octo-Base (Team et al., 2024)	17.0	4.2	22.7	14.6	0.6	3.1	1.1	1.6
RT-1 (Brohan et al., 2022)	<u>85.7</u>	44.2	**73.0**	67.6	**89.8**	50.0	32.3	57.4
RT-1-X (Li et al., 2024d)	56.7	31.7	<u>59.7</u>	49.4	49.0	32.3	29.4	36.9
RT-2-X (Li et al., 2024d)	78.7	**77.9**	25.0	60.5	82.3	**79.2**	35.5	<u>65.7</u>
OpenVLA (Kim et al., 2025b)	16.3	46.2	35.6	32.7	54.5	47.7	17.7	40.0
π_0 (Black et al., 2024a)	72.7	65.3	38.3	58.8	75.2	63.7	25.6	54.8
π_0-FAST (Pertsch et al., 2025)	75.3	67.5	42.9	61.9	77.6	68.2	31.3	59.0
TraceVLA (Zheng et al., 2025)	28.0	53.7	57.0	46.2	60.0	56.4	31.0	49.1
Magma (Yang et al., 2025b)	75.0	53.0	58.9	62.3	68.6	<u>78.5</u>	**59.0**	**68.7**
RoboVLMs (Li et al., 2024b)	77.3	61.7	43.5	60.8	75.6	60.0	10.6	48.7
Dita (Hou et al., 2025)	83.7	<u>76.0</u>	46.3	<u>68.7</u>	<u>85.5</u>	73.0	<u>37.5</u>	65.3
PixelVLA-π_0 (Liang et al., 2026)	80.7	67.7	41.3	63.3	76.8	62.0	30.8	56.5
UATL (Ours)	**100.0**	58.0	50.0	**69.3**	71.0	67.0	25.0	54.3

Table 4: Success rates (%) on WidowX tasks in SimplerEnv. Bold and underlined values indicate the best and second-best results, respectively.

Method	Spoon on Towel	Carrot on Plate	Stack Block	Eggplant in Basket	Avg.
RT-1-X (Li et al., 2024d)	0.0	4.2	0.0	0.0	1.1
Octo-Base (Team et al., 2024)	15.8	12.5	0.0	41.7	17.5
Octo-Small (Team et al., 2024)	41.7	8.2	0.0	56.7	26.7
RoboVLMs (Li et al., 2024b)	45.8	20.8	4.2	79.2	37.5
CogACT (Li et al., 2024a)	**71.7**	50.8	15.0	67.5	51.3
OpenVLA (Kim et al., 2025b)	4.2	0.0	0.0	12.5	4.2
π_0 (Black et al., 2024a)	<u>63.3</u>	<u>58.8</u>	21.3	79.2	55.7
π_0-FAST (Pertsch et al., 2025)	29.1	21.9	10.8	66.6	32.1
TraceVLA (Zheng et al., 2025)	12.5	16.6	16.6	65.0	27.7
Magma (Yang et al., 2025b)	37.5	29.2	20.8	91.7	44.8
ReFineVLA (Van Vo et al., 2025)	38.1	33.3	23.8	<u>95.2</u>	47.7
UniVLA (Bu et al., 2025)	52.8	55.6	2.8	80.6	48.0
ThinkAct (Huang et al., 2026a)	58.3	37.5	8.7	70.8	43.8
DreamVLA (Zhang et al., 2026a)	45.8	45.8	25.0	87.5	51.0
UD-VLA (Chen et al., 2026)	58.3	**62.5**	**54.1**	75.0	<u>62.5</u>
UATL (Ours)	58.3	58.3	<u>41.7</u>	**95.8**	**63.5**

on Plate. These results demonstrate strong performance under Visual Matching and on WidowX, while leaving room for improvement under Variant Aggregation.

Results on CALVIN. As shown in Table 5, UATL achieves the highest average sequence length of 4.05 under the $\mathrm{ABC} \rightarrow \mathrm{D}$ protocol and ranks first at all five chain lengths. Its success rates for completing one to five consecutive tasks are 96.0%, 89.6%, 81.0%, 73.8%, and 65.0%, respectively. In particular, UATL successfully completes the full five-task chain in 65.0% of evaluations, demonstrating strong long-horizon instruction following in the held-out environment.

Overall, UATL achieves the best aggregate performance on LIBERO, LIBERO-Plus, and CALVIN, as well as the highest average success rates on SimplerEnv Google Robot under Visual Matching and on WidowX. By leveraging HAL for aleatoric uncertainty estimation and UW-TCR to impose adaptive temporal constraints, UATL reduces action-chunk error accumulation and delivers strong multitask, robustness, and long-horizon control capabilities across environments and robot platforms. Performance under Variant Aggregation and Robot and Layout perturbations indicates room for further generalization.

Table 5: Results on CALVIN under the ABC→D protocol. Bold and underlined values indicate the best and second-best results, respectively.

Method	Setting	Success Rate for Task Length↑					Avg. Len.
		1/5	2/5	3/5	4/5	5/5	
RoboFlamingo (Li et al., 2024c)	ABC→D	82.4	61.9	46.6	33.1	23.5	2.48
SuSIE (Black et al., 2024b)	ABC→D	87.0	69.0	49.0	38.0	26.0	2.69
Deer-VLA (Yue et al., 2024)	ABC→D	84.8	72.3	54.9	44.6	33.5	2.90
GR-1 (Wu et al., 2024)	ABC→D	85.4	71.2	59.6	49.7	40.1	3.06
3DDA (Ke et al., 2024)	ABC→D	92.2	78.7	63.9	51.2	41.2	3.27
RoboUniView (Liu et al., 2024)	ABC→D	94.2	84.2	73.4	62.2	50.7	3.65
RoboDual (Bu et al., 2024)	ABC→D	94.4	82.7	72.1	62.4	54.4	3.66
OpenVLA (Kim et al., 2025b)	ABC→D	91.3	77.8	62.0	52.1	43.5	3.27
π_0 (Black et al., 2024a)	ABC→D	94.3	87.0	77.9	68.5	59.4	3.87
$\pi_{0.5}$ (Intelligence et al., 2025)	ABC→D	94.8	87.4	78.2	71.7	<u>64.3</u>	3.97
AR-VRM (Yang et al., 2025a)	ABC→D	90.1	75.9	64.2	53.1	46.1	3.29
PD-VLA (Song et al., 2025)	ABC→D	94.1	80.0	68.3	61.4	50.5	3.54
DISCO (Hao et al., 2025)	ABC→D	94.7	82.9	71.0	58.8	49.4	3.57
Dita (Hou et al., 2025)	ABC→D	94.5	82.5	72.8	61.3	50.0	3.61
Seer (Tian et al., 2025)	ABC→D	94.4	87.2	<u>79.9</u>	<u>72.2</u>	<u>64.3</u>	<u>3.98</u>
RoboBERT (Wang et al., 2025)	ABC→D	95.3	85.7	75.4	66.3	56.2	3.79
UniVLA (Bu et al., 2025)	ABC→D	95.5	85.8	75.4	66.9	56.5	3.80
AWP (Yuan et al., 2026a)	ABC→D	92.0	79.6	68.6	63.0	48.0	3.54
DREAM (Sheng et al., 2026)	ABC→D	94.5	85.6	75.8	66.4	58.8	3.81
OmniStream (Yan et al., 2026)	ABC→D	93.1	84.8	76.8	70.3	63.4	3.89
ReconVLA (Song et al., 2026)	ABC→D	<u>95.6</u>	<u>87.6</u>	76.9	69.3	64.1	3.95
UATL (Ours)	ABC→D	**96.0**	**89.6**	**81.0**	**73.8**	**65.0**	**4.05**

Table 6: Component ablation and inference efficiency on LIBERO.

HAL	UW-TCR	Throughput (Hz) ↑	Latency (Sec) ↓	Spatial	Object	Goal	Long	Avg.
×	×	**54.1**	**0.148**	93.2	99.0	96.0	94.8	95.8
✓	×	53.2	0.151	94.4	**99.8**	97.0	95.2	96.6
✓	✓	53.4	0.150	**98.4**	**99.8**	**98.2**	**96.8**	**98.3**

Table 7: Sensitivity to the temporal-consistency weight λ_{TCR} on LIBERO.

λ_{TCR}	Spatial	Object	Goal	Long	Ave.
0.01	94.2	99.6	**98.2**	96.0	97.0
0.05	**98.4**	**99.8**	**98.2**	**96.8**	**98.3**
0.10	94.8	99.0	97.0	96.6	96.9

4.4 Ablation Studies

Components ablation. As shown in Table 6, HAL improves the average success rate from 95.8% to 96.6%. Adding UW-TCR further raises it to 98.3%, yielding an overall gain of 2.5 percentage points. The largest improvement occurs on Spatial, increasing from 93.2% to 98.4%, while Goal and Long improve to 98.2% and 96.8%, respectively. These results demonstrate the complementary benefits of HAL and UW-TCR. Meanwhile, UATL introduces negligible inference overhead, reducing throughput from 54.1 to 53.4 Hz and increasing latency from 0.148 to 0.150 s, without additional policy queries or changes to the execution protocol.

Sensitivity to the temporal-regularization weight. Table 7 evaluates the sensitivity to λ_{TCR} on LIBERO. Increasing λ_{TCR} from 0.01 to 0.05 improves the average success rate from 97.0% to 98.3%, with the largest improvement observed on Spatial, from 94.2% to 98.4%. Object and Long also improve, while Goal remains unchanged. Further increasing the weight to 0.10 reduces

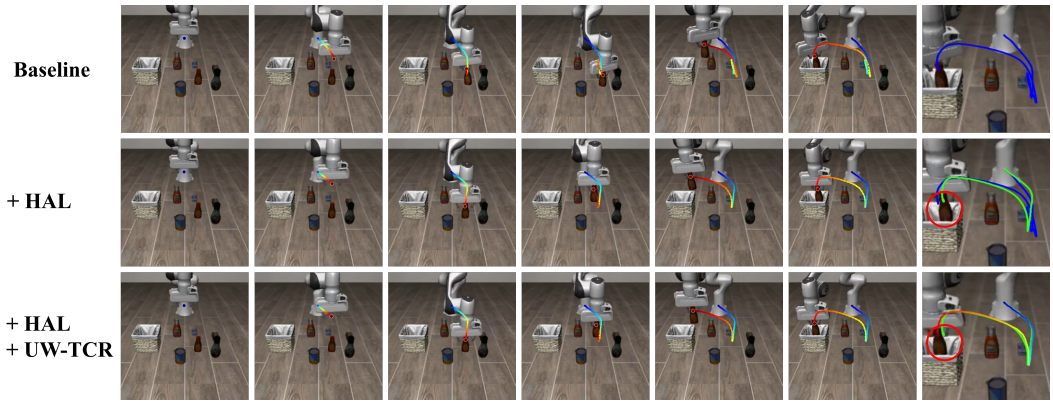

Figure 3: Task execution process and end-effector trajectories on a LIBERO task. The first six columns show the temporal execution process, while the last column presents superimposed trajectories of adjacent variants. Red circles highlight the regions with the most pronounced trajectory differences.

performance across all four suites, lowering the average success rate to 96.9%. We therefore set $\lambda_{\mathrm{TCR}} = 0.05$ by default, as it achieves the best overall performance among the evaluated settings.

4.5 Qualitative Analysis of Task Executions

Figure 3 shows end-effector trajectories on the pick-and-place task. The baseline, HAL only, and the full model all complete the task. The full model uses both HAL and UW-TCR. Rows add one component at a time. The first six columns show the trajectory over time. The color goes from blue at the start to red at the end. The last column zooms into the same workspace. The second row overlays HAL on the baseline, and the third row overlays the full model on HAL.

Introducing HAL produces the most pronounced trajectory change near the target region. Although the baseline reaches the target vicinity, its terminal trajectory is less well aligned with the effective placement area. HAL noticeably corrects the trajectory within the red-circled region, guiding the end effector more accurately into the target.

With the further introduction of UW-TCR, the complete model preserves the overall route while refining the local transition and terminal positioning near the target, resulting in a more coherent and stable process. In summary, HAL primarily improves terminal target alignment, whereas UW-TCR further enhances how smoothly and reliably the target is approached.

5 Conclusion

In this work, we presented UATL, an uncertainty-aware training framework for action chunk VLA policies. HAL models each action element with a conditional Laplace distribution, while UW-TCR uses the predicted scales to adaptively regularize second-order temporal differences within each action chunk. Together, these components account for nonuniform prediction reliability and within-chunk temporal structure without changing the policy-query or execution protocol, introducing only marginal inference overhead. Experiments on LIBERO, LIBERO-Plus, SimplerEnv, and CALVIN demonstrate strong performance across multitask manipulation, distribution shifts, cross-platform evaluation, and long-horizon control. Ablation studies further confirm the complementary contributions of HAL and UW-TCR. These results establish uncertainty-aware temporal modeling as an effective approach to improving the execution reliability of open-loop action chunk VLA policies.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024a.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 33431–33452, 2024b.

Kevin Black, Manuel Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *Advances in Neural Information Processing Systems*, 38:33383–33407, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Qingwen Bu, Hongyang Li, Li Chen, Jisong Cai, Jia Zeng, Heming Cui, Maoqing Yao, and Yu Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*, 2024.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Paweł Budzianowski, Wesley Maa, Matthew Freed, Jingxiang Mo, Winston Hsiao, Aaron Xie, Tomasz Młoduchowski, Viraj Tipnis, and Benjamin Bolte. Edgevla: Efficient vision-language-action models. *arXiv preprint arXiv:2507.14049*, 2025.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Barrett Tang, Donglin Wang, and Haoang Li. Unified diffusion vla: Vision-language-action model via joint discrete denosing diffusion process. In *International Conference on Learning Representations*, volume 2026, pp. 139291–139311, 2026.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Kai Fang, Hailong Pei, and Xuemin Chi. Smoother action chunking flow policy via prior-corrected orthogonal trust-region guidance. *arXiv preprint arXiv:2605.24433*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks: J. gawlikowski et al. *Artificial intelligence review*, 56(Suppl 1):1513–1589, 2023.

Ce Hao, Kelvin Lin, Zhiwei Xue, Siyuan Luo, and Harold Soh. Disco: Language-guided manipulation with diffusion policies and constrained inpainting. *IEEE Robotics and Automation Letters*, 2025.

Zhi Hou, Tianyi Zhang, Yuwen Xiong, Haonan Duan, Hengjun Pu, Ronglei Tong, Chengyang Zhao, Xizhou Zhu, Yu Qiao, Jifeng Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7686–7697. IEEE, 2025.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, 38:82782–82802, 2026a.

Jiakai Huang, Zhongbo Wu, Zheng Zhang, Zihan Wang, Shan You, and Tao Huang. Foresight without seeing: Latent futures for world action models. *arXiv preprint arXiv:2608.11605*, 2026b.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Suhyeok Jang, Dongyoung Kim, Changyeon Kim, Youngsuk Kim, and Jinwoo Shin. Verifier-free test-time sampling for vision-language-action models. In *International Conference on Learning Representations*, volume 2026, pp. 139997–140017, 2026.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025a.

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 2025b.

Hao Li, Qi Lv, Rui Shao, Xiang Deng, Yinchuan Li, Jianye Hao, and Liqiang Nie. STAR: Learning diverse robot skill abstractions through rotation-augmented vector quantization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35562–35580. PMLR, 13–19 Jul 2025.

Hao Li, Shuai Yang, Yilun Chen, Xinyi Chen, Xiaoda Yang, Yang Tian, Hanqing Wang, Tai Wang, Dahua Lin, Feng Zhao, et al. Towards efficient and robust manipulation via multi-frame vision-language-action modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18388–18396, 2026.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 3(7):8, 2024b.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *International Conference on Learning Representations*, volume 2024, pp. 26703–26721, 2024c.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024d.

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*, volume 2026, pp. 110464–110481, 2026.

Tao Lin, Yuxin Du, Jiting Liu, Nuobei Zhu, Yunhe Li, Yuqian Fu, Yinxinyu Chen, Hongyi Cai, Zewei Ye, Bing Cheng, et al. Evo-depth: A lightweight depth-enhanced vision-language-action model. *arXiv preprint arXiv:2605.14950*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024.

Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable vla pretraining in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35047–35058, June 2026.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. *arXiv preprint arXiv:2410.13816*, 2024.

Fei Ni, Zhuo Chen, Yifu Yuan, Zibin Dong, Xianze Yao, Shan Luo, Jianye Hao, Jiankang Deng, and Stefanos Zafeiriou. Semanticvla: Towards semantic reasoning over action memorization via synergistic explicit trace and latent action planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12237–12247, 2026.

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

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. *arXiv preprint arXiv:2501.09747*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Ransalu Senanayake. The role of predictive uncertainty and diversity in embodied ai and robot learning. *arXiv preprint arXiv:2405.03164*, 2024.

Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. Large vlm-based vision-language-action models for robotic manipulation: A survey. *arXiv preprint arXiv:2508.13073*, 2025.

Kai Sheng, Liuyi Wang, Zongtao He, Xiao Lin, Chengju Liu, and Qijun Chen. Dream: Dynamic routing of experts via attention-based mixture for vision-language-action modeling. *Knowledge-Based Systems*, pp. 115585, 2026.

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*, volume 2026, pp. 18567–18602, 2026.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Zhijun Li, Donglin Wang, Lujia Wang, et al. Pd-vla: Accelerating vision-language-action model integrated with action chunking via parallel decoding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13162–13169. IEEE, 2025.

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, pp. 18549–18557, 2026.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 92033–92052, 2025.

Tuan Van Vo, Tan Quang Nguyen, Khang Minh Nguyen, Duy Ho Minh Nguyen, and Minh Nhat Vu. Refinevla: reasoning-aware teacher-guided transfer fine-tuning. *arXiv preprint arXiv:2505.19080*, 2025.

Sicheng Wang, Sheng Liu, Weiheng Wang, Jianhua Shan, and Bin Fang. Robobert: An end-to-end multimodal robotic manipulation model. *arXiv preprint arXiv:2502.07837*, 2025.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 10(4):3988–3995, 2025.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Chao Xu, Suyu Zhang, Yang Liu, Baigui Sun, Weihong Chen, Bo Xu, Qi Liu, Juncheng Wang, Shujun Wang, Shan Luo, et al. An anatomy of vision-language-action models: From modules to milestones and challenges. *arXiv preprint arXiv:2512.11362*, 2025.

Yibin Yan, Jilan Xu, Shangzhe Di, Haoning Wu, and Weidi Xie. Omnistream: Mastering perception, reconstruction and action in continuous streams. *arXiv preprint arXiv:2603.12265*, 2026.

Dejie Yang, Zijing Zhao, and Yang Liu. Ar-vrm: Imitating human motions for visual robot manipulation with analogical reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6818–6827. IEEE, 2025a.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14203–14214. IEEE, 2025b.

Yi Yang, Xueqi Li, Yiyang Chen, Jin Song, Yihan Wang, Zipeng Xiao, Jiadi Su, You Qiaoben, Pengfei Liu, and Zhijie Deng. Mantis: A versatile vision-language-action model with disentangled visual foresight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42505–42515, 2026.

Ge Yuan, Qiyuan Qiao, Jing Zhang, and Dong Xu. Adaworldpolicy: World-model-driven diffusion policy with online adaptive learning for robotic manipulation. *arXiv preprint arXiv:2602.20057*, 2026a.

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*, 2026b.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Zehua Zang, Xi Wang, Fuchun Sun, Xiao Xu, Lixiang Liu, Jiahuan Zhou, and Jiangmeng Li. Test-time perturbation tuning with delayed feedback for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, June 2026.

Dapeng Zhang, Jing Sun, Chenghui Hu, Xiaoyan Wu, Zhenlong Yuan, Rui Zhou, Fei Shen, and Qingguo Zhou. Pure vision language action (vla) models: A comprehensive survey. *arXiv preprint arXiv:2509.19012*, 2025.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026a.

Yan Zhang, Yinan Wu, Haoran Duan, and Jungong Han. Cofactvla: Deconfounding vision-language-action models via counterfactual intervention. *arXiv preprint arXiv:2608.04396*, 2026b.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713. IEEE, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

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*, volume 2025, pp. 54277–54296, 2025.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.