ACTION-CONSISTENT ATTENTION WITH HISTORY FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative policies based on flow matching have achieved remarkable success in robotics by formulating action-chunk prediction as a conditional probability path over a sequence of time-steps. However, when predicting the conditional velocity field, existing methods typically treat the action chunk as a flattened vector sequence, overlooking two fundamental semantic structures: the temporal evolution of each individual joint across consecutive steps, and the inter-joint coordination required within each time-step for physically coherent control. In this paper, we propose **AC-HFlow** (*Action-Consistent History Flow Matching*), a novel paradigm that reduces multi-step inference to a few integration steps, with as few as one, without sacrificing performance, while capturing coherent action semantics. It employs Action-Consistent Attention (ACA) to explicitly reason over the intra-chunk joint–time grid, capturing the partial order of joint evolution across timestamps and the directed coupling among joints within the same step. Meanwhile, pooled historical action statistics serve as an informative source for flow matching, which mitigates the overhead introduced by random-noise initialization. Extensive experiments demonstrate that AC-HFlow delivers superior action consistency, strong generalization capability under visual and scene shifts, and effective few-step inference in which a single integration step already attains competitive success.

1 INTRODUCTION

Visuomotor imitation learning has progressed rapidly by casting control as conditional generative modeling over short horizons of actions, with diffusion (Chi et al., 2023; Ho et al., 2020; Song et al., 2021), flow matching (Lipman et al., 2023; Liu et al., 2023b; Tong et al., 2024), DiT-based backbones (Peebles & Xie, 2023), and action chunking (Zhao et al., 2023) as common building blocks. However, two coupled questions remain: how to represent the action semantics *inside* a chunk, and how to generate that chunk efficiently without relying on a long trajectory from unstructured random noise.

This representational question arises because an action chunk is not merely a sequence of exchangeable vector tokens, but a matrix whose two axes have distinct physical meanings. Along the time axis, each joint must form a coherent trajectory; within a timestep, multiple joints must cooperate to realize one control command. A standard action transformer first embeds the complete action vector at each timestep and attends over the resulting horizon tokens. While it can learn both dependencies indirectly, it never exposes an individual joint–time element to a relation operator specialized for these axes. Consequently, the conditional velocity field predicted by the flow network lacks an explicit inductive bias for preserving joint-wise temporal dynamics and within-step mechanical coupling. The problem is therefore not a lack of global context, but a lack of an explicit route for *inter-step* and *inter-joint* reasoning within the vector field (Figure 1).

The second question concerns the starting point of the probability path. Conventional flow matching initializes each action chunk from independent Gaussian noise, leaving the model to reconstruct task-relevant motion through numerical integration. Yet closed-loop execution already provides a recent action history that encodes the robot's motion tendency and joint coordination. Leveraging this history to shape the source distribution can make the initialization depend on recent motion

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.

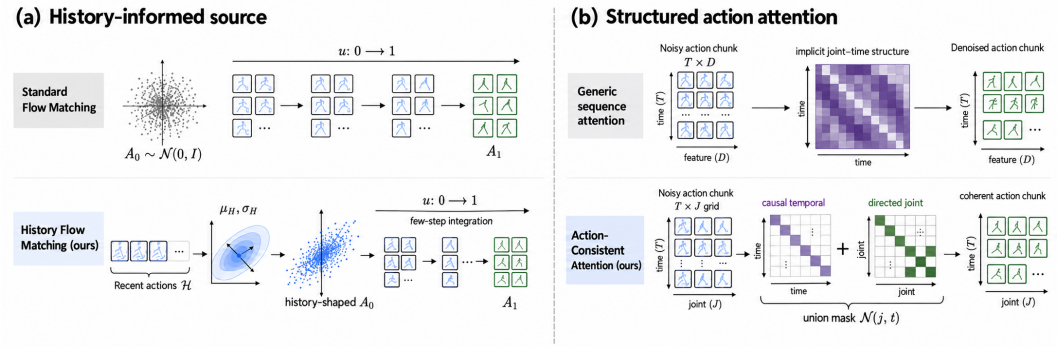


Figure 1: Design contrast between AC-HFlow and prior flow/diffusion policies, which start from isotropic random noise and treat the action chunk as a flat sequence of horizon tokens. Demonstrations may be recorded in joint-angle or end-effector command space, and on our real robot (Figure 5) the deployed actions are joint-angle changes, whereas LIBERO training uses OSC end-effector commands.

statistics, thereby shortening the transport distance and complementing structured attention with a history-shaped flow endpoint (Jia et al., 2026; Ma et al., 2025).

We introduce **AC-HFlow** (*Action-Consistent History Flow Matching*), a unified framework that addresses both questions. AC-HFlow re-parameterizes each timestep feature as J learned joint factors and performs relational attention on the resulting $T \times J$ grid. For a query at joint j and timestep t, the visibility mask admits keys from the same joint at timesteps up to t (causal temporal evolution), or from the same timestep at joint indices up to j (directed joint factorization). This union-masked attention handles both relations in one efficient call, imposing a partial order on the velocity field without removing the backbone's global temporal context. In parallel, we aggregate and normalize recent actions into a history pool whose empirical statistics are used to construct the stochastic flow source directly in action space. Unlike action-to-action approaches (Jia et al., 2026) that require an additional feature-extraction process, this action-space construction introduces no separate history encoder and provides a history-shaped stochastic endpoint for efficient generation with very few integration steps. Extensive experiments on LIBERO (Liu et al., 2023a) and LIBERO-Plus (Fei et al., 2025) validate these claims: AC-HFlow improves action consistency, generalizes robustly under visual and scene shifts, and attains competitive success with very few—even a single—integration step.

Contributions.

- We introduce **ACA**, a single union-masked attention layer that explicitly models causal inter-step and same-step inter-joint relations inside a conditional flow velocity network.
- We use pooled action history to construct a history-shaped flow source that supports effective few-step inference, including single-step inference, without a separate history encoder or distillation procedure.
- Extensive simulation and real-robot experiments validate AC-HFlow's action consistency, few-step efficiency, and robustness, while further studies investigate how grid positional encoding affects generalization under visual and scene shifts.

2 Related Work

Generative visuomotor policies. Diffusion Policy (Chi et al., 2023) casts closed-loop control as conditional denoising of action sequences with DDPM (Ho et al., 2020) or DDIM (Song et al., 2021) samplers, and flow matching (Lipman et al., 2023; Liu et al., 2023b; Tong et al., 2024) replaces discrete steps with continuous probability paths in policies such as VITA (Gao et al., 2025), FlowPolicy (Zhang et al., 2025), Streaming Flow Policy (Jiang et al., 2025), and π_0 (Black et al., 2024; 2025). DiT-style backbones (Peebles & Xie, 2023; Reuss et al., 2024) and generalist VLA

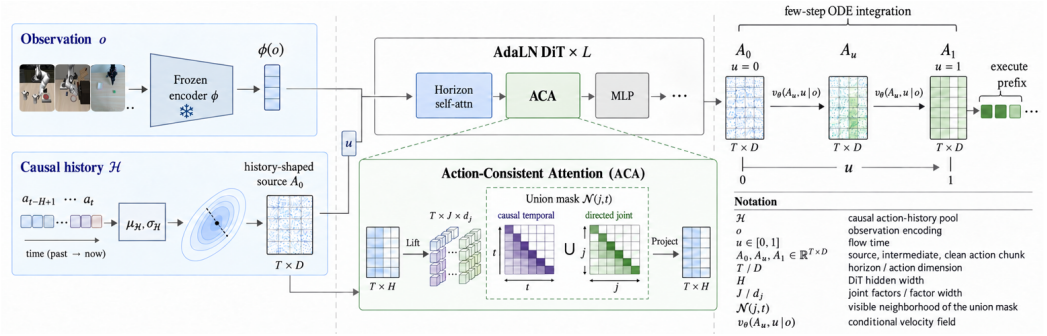


Figure 2: Overview of the AC-HFlow policy. A frozen per-step observation encoding conditions the flow velocity network, which lifts the action chunk onto a joint–time grid and applies union-masked relational attention inside each AdaLN DiT block; pooled recent actions construct the history-shaped flow source $\boldsymbol{A}_0$ used for few-step integration. Index j on the grid is a *joint factor* that factorizes one component of the action vector at each timestep; see Figure 1 for how this notion of joint relates to joint-angle vs. end-effector action parameterizations.

policies (Team et al., 2024; Kim et al., 2025; Wen et al., 2025; Reuss et al., 2025) further scale visual-language conditioning. Our work remains in *action-space* flow matching, addressing two aspects these approaches leave implicit: how the velocity network should preserve joint–time action semantics, and how recent actions can define a more informative source distribution for efficient generation.

Action chunking and temporal structure. Action chunking (Zhao et al., 2023), implicit behavioral cloning (Florence et al., 2022), bimanual chunk attention (Lee et al., 2025), and frequency-domain action tokenization (Pertsch et al., 2025) show that action coordinates are not exchangeable, yet most chunked policies rely on generic temporal attention (Vaswani et al., 2017), often with rotary embeddings (Su et al., 2024). We instead expose a joint–time product grid inside the flow velocity network, explicitly modeling the evolution of each joint factor across timesteps and the correlations among joint factors within the same timestep.

Action-to-action, history priors, and fast sampling. Action-to-action flow matching (Jia et al., 2026) and causal diffusion policies (Ma et al., 2025) show that recent motion provides a stronger starting point than unstructured random noise, while consistency training and distillation (Song et al., 2023; Prasad et al., 2024; Wang et al., 2025; Zhang et al., 2025) reduce inference steps through modified objectives or teacher–student transfer. AC-HFlow shares this history-based motivation but constructs its flow source from pooled, normalized historical action statistics directly in action space—with no separate history encoder and no distillation stage—while ACA structures the velocity field along the joint and temporal axes.

3 METHOD

We formulate action prediction as an observation-conditioned probability flow over a finite action horizon. Our central observation is that a chunk carries two distinct forms of structure: each joint factor evolves across the prediction horizon, while different joints must coordinate within every action. Throughout the paper, we use *action consistency* to refer to two concrete properties of the generated chunk: (i) within-chunk joint coordination—how coherently the different action dimensions co-move inside a single action, and (ii) within-chunk temporal evolution—how smoothly each action dimension varies across the consecutive steps of the chunk. These are precisely the two relations encoded by ACA below (directed inter-joint coupling within a step and causal inter-step evolution along each joint factor), and they are probed by the chunk diagnostics in Appendix A.7; the task-success results in Table 1 further confirm their behavioral benefit. AC-HFlow makes these relations explicit in the velocity network, and History Flow Matching complements this representation with a source distribution informed by recent execution (Figure 2).

3.1 PRELIMINARIES

Action chunks. At each replan, the policy predicts a chunk of future actions

$$\boldsymbol{A} = \left(\boldsymbol{a}_{t+1}, \ldots, \boldsymbol{a}_{t+T}\right)^{\top} \in \mathbb{R}^{T \times D}, \tag{1}$$

where T is the horizon and D the action dimension, and executes a prefix before observing the environment and replanning. Unlike independent action vectors, $\boldsymbol{A}$ describes a coordinated multi-step motion: each control dimension must remain temporally coherent, and all dimensions at a given step must jointly realize the intended behavior.

Conditional flow matching. Let $\boldsymbol{A}_0 \sim \mathcal{N}(0, I)$ be a random-noise chunk, $\boldsymbol{A}_1 \sim p_{\mathrm{data}}(\boldsymbol{A} \mid \boldsymbol{o})$ a demonstrated chunk conditioned on observation encoding $\boldsymbol{o}$, and $u \sim \mathcal{U}(0, 1)$ the flow time. Following (Lipman et al., 2023; Liu et al., 2023b), we use the linear conditional path

$$\boldsymbol{A}_u = (1-u)\boldsymbol{A}_0 + u\boldsymbol{A}_1, \qquad \dot{\boldsymbol{A}}_u = \boldsymbol{A}_1 - \boldsymbol{A}_0. \tag{2}$$

The conditional velocity network $\boldsymbol{v}_\theta(\boldsymbol{A}_u, u \mid \boldsymbol{o})$ is trained to regress this path velocity:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\boldsymbol{A}_0, \boldsymbol{A}_1, u, \boldsymbol{o}} \left[\left\| \boldsymbol{v}_\theta(\boldsymbol{A}_u, u \mid \boldsymbol{o}) - (\boldsymbol{A}_1 - \boldsymbol{A}_0) \right\|_2^2 \right]. \tag{3}$$

At inference, a sample is obtained by integrating the learned conditional field from noise to the action distribution,

$$\boldsymbol{A}_1 = \boldsymbol{A}_0 + \int_0^1 \boldsymbol{v}_\theta(\boldsymbol{A}_u, u \mid \boldsymbol{o})\, \mathrm{d}u. \tag{4}$$

Modeling this field is demanding for action chunks because it must transport a high-dimensional joint distribution while preserving long-range temporal evolution and within-step coordination; AC-HFlow builds both relations *directly into the velocity network* as a structural inductive bias.

3.2 ACTION-CONSISTENT ATTENTION

The conditional velocity is parameterized by L AdaLN-conditioned DiT blocks (Peebles & Xie, 2023) in which the visual observation and flow time modulate every block. Within each block, ACA is applied after horizon self-attention and before the feed-forward network, with a separate zero-initialized residual gate; the original horizon attention retains global context while ACA adds explicit action-specific relations without changing the conditioning interface or flow-matching objective.

We regard the lifted action chunk as a signal on the Cartesian product $\mathcal{G}_T \times \mathcal{G}_J$, where $\mathcal{G}_T$ indexes prediction time and $\mathcal{G}_J$ indexes learned joint factors (slots that factorize one component of the per-timestep action vector; see Figure 1). Formally, ACA imposes a *partial order* on the velocity field's receptive field: information can flow forward in time for the same joint and forward along the joint ordering within the same timestep, decoupling temporal dynamics from spatial coordination and preventing spurious correlations between future actions and unrelated joints during flow integration.

ACA starts from $\boldsymbol{X} \in \mathbb{R}^{B \times T \times H}$ and lifts it to control features

$$\boldsymbol{Z} = \mathrm{reshape}\big(\boldsymbol{X} W_{\mathrm{j}}\big) \in \mathbb{R}^{B \times T \times J \times d_j}, \tag{5}$$

with Jd_j independent of the DiT width H. A single masked attention layer on this grid refines the conditional velocity through structured message passing: for a query token (j_q, t_q), keys in its visible neighborhood $\mathcal{N}(j_q, t_q)$ are aggregated with standard attention, and invisible keys are masked out.

Inter-step relation. For a fixed joint factor j, attention is causal over the length-T sequence $\boldsymbol{Z}_{:,j,:} \in \mathbb{R}^{B \times T \times d_j}$: keys at timesteps $t_k \le t_q$ are visible, encoding dependencies among timesteps of the same factor without introducing information from later chunk positions.

Inter-joint relation. For a fixed timestep t, attention is directed along joint indices $j_k \le j_q$ in $\boldsymbol{Z}_{t,:,:} \in \mathbb{R}^{B \times J \times d_j}$; the order follows the learned slot layout and provides a consistent directed factorization among same-step joint factors.

Union-masked relational attention. ACA flattens control–time tokens into a length-JT sequence, with control as the outer index, and runs a single masked attention whose visible neighborhood is

$$\mathcal{N}(j_q, t_q) = \{(j_q, t) : t \le t_q\} \cup \{(j, t_q) : j \le j_q\}. \tag{6}$$

A key is thus visible exactly when it represents an earlier state of the same joint factor or an earlier joint factor at the same timestep. Compared with dense bidirectional attention over the JT flattened tokens, which would consider all $(JT)^2$ query–key pairs, the union mask retains only $JT(T+J)/2$ visible pairs, so attention-map memory scales as $O(JT(T+J))$; equivalently, the induced graph is the union of causal temporal chains and directed within-step joint cliques. Because every ACA layer follows an ordinary horizon-level DiT attention layer, this relational sparsity does not sever global information flow. Writing $\widetilde{\boldsymbol{Z}} = \mathrm{LN}\big(\boldsymbol{Z} + \mathrm{Attn}(Q, K, V; \mathcal{N})\big)$, we reshape and project back to width H before applying the block-level AdaLN gate:

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + \alpha_{\mathrm{A}}(\phi(\boldsymbol{o}), \psi(u)) \odot \big(\mathrm{reshape}(\widetilde{\boldsymbol{Z}})\, W_{\mathrm{out}}\big). \tag{7}$$

Positional encoding on the joint–time grid. Each lifted token must encode temporal index t and joint-slot index j. Following (Su et al., 2024), we write $\mathrm{RoPE}(\boldsymbol{x}; p, b)$ for rotary position embedding of a d-dimensional vector $\boldsymbol{x}$ with index p and base b.

Default ACA. We flatten the lifted grid with control index outermost, add a learned slot embedding $\boldsymbol{e}_j$ to each token, and apply standard RoPE (Su et al., 2024) to queries and keys on the flattened sequence:

$$r(j, t) = jT + t, \qquad \widetilde{\boldsymbol{Z}}_{j,t} = \boldsymbol{Z}_{j,t} + \boldsymbol{e}_j. \tag{8}$$

GRoPE. We name the sequential grid encoding **GRoPE** (*Grid RoPE*), since it embeds the joint–time grid by sequential rotations rather than by learned slots. GRoPE replaces equation 8 and flattened RoPE with sequential grid encoding on the full d_j-dimensional joint feature of each lifted token:

$$\mathrm{GRoPE}(\boldsymbol{z}_{j,t}; t, j) = \mathrm{RoPE}\big(\mathrm{RoPE}(\boldsymbol{z}_{j,t}; t); j\big), \tag{9}$$

with the same rotary base b as in the RoPE definition, for both the temporal and the joint rotation; we denote the resulting variant **AC-HFlow+GRoPE**. AC-HFlow+GRoPE otherwise matches AC-HFlow (union mask and History Flow Matching). Unlike the learned slot embedding of default AC-HFlow, which is fit to the joint layout observed in training, GRoPE expresses grid coordinates as relative rotations and introduces no learned positional parameters; it therefore avoids memorizing task-specific coordinate statistics and should transfer more readily when the observation distribution shifts. As reported in Table 3, this variant preserves AC-HFlow's in-distribution performance while attaining the highest average success under LIBERO-Plus visual and scene shifts among all evaluated policies.

Ablation variants. Restricting visibility to temporal edges, inter-joint edges, or their union (equation 6) yields the structure ablations in Table 2. All variants keep the outer DiT self-attention and MLP unchanged, isolating the effect of relational factorization.

3.3 History Flow Matching

While ACA refines the *shape* of the transport path, History Flow Matching controls its *starting location*: instead of isotropic noise, the source endpoint is drawn from a pool of recent normalized action chunks. Let $\mathcal{H} = \{\boldsymbol{h}_m\}_{m=1}^{M}$ denote the valid historical actions, with per-control empirical mean $\boldsymbol{\mu}_{\mathcal{H}}$ and standard deviation $\boldsymbol{\sigma}_{\mathcal{H}}$ (lower-bounded by ϵ_{std}). For independent $\boldsymbol{\epsilon}_h, \boldsymbol{\epsilon}_n \sim \mathcal{N}(0, I)$, we form

$$\widehat{\boldsymbol{A}}_0 = p\big(\boldsymbol{\mu}_{\mathcal{H}} + \boldsymbol{\sigma}_{\mathcal{H}} \odot \boldsymbol{\epsilon}_h\big) + (1-p)\boldsymbol{\epsilon}_n, \tag{10}$$

where p controls the history contribution, and then standardize the source,

$$\boldsymbol{A}_0 = \frac{\widehat{\boldsymbol{A}}_0 - p\boldsymbol{\mu}_{\mathcal{H}}}{\sqrt{p^2\boldsymbol{\sigma}_{\mathcal{H}}^2 + (1-p)^2}}, \tag{11}$$

so that the source marginal remains centered and standardized. Training samples p around 0.9 via $p = 0.85 + 0.10\xi$ with $\xi \sim \mathrm{Beta}(5, 5)$; evaluation fixes p=0.9. This construction supplies the source

Table 1: Success (%) on seven LIBERO tasks at each method's operating point (30 trials per evaluation). Higher is better; best value per column in bold; Avg. denotes the unweighted average over the seven tasks. AC-HFlow+GRoPE results at five and ten steps appear in Appendix A.5 (Table 7).

Policy	Steps	drawer	stove	micro.	pudd.	soup	moka	s+moka	Avg.
AC-HFlow	3	**100.0**	**100.0**	90.0	**60.0**	**73.3**	53.3	76.7	**79.0**
AC-HFlow	1	100.0	100.0	90.0	56.7	60.0	56.7	83.3	78.1
AC-HFlow+GRoPE	3	100.0	100.0	90.0	53.3	63.3	53.3	**86.7**	78.1
AC-HFlow+GRoPE	1	100.0	100.0	**93.3**	46.7	60.0	40.0	83.3	74.8
DDIM-UNet	90	96.7	100.0	80.0	40.0	70.0	**66.7**	63.3	73.8
DDPM-DiT	90	93.3	100.0	76.7	50.0	46.7	66.7	83.3	73.8
DDPM-UNet	90	93.3	100.0	63.3	53.3	56.7	66.7	83.3	73.8
VITA	10	100.0	100.0	70.0	53.3	33.3	50.0	80.0	69.5
A2A-Noise	10	86.7	100.0	83.3	0.0	0.0	56.7	60.0	55.2
A2A	10	66.7	100.0	70.0	0.0	0.0	33.3	26.7	42.4
FM-DiT	10	80.0	66.7	20.0	0.0	0.0	0.0	3.3	24.3

endpoint of the flow path equation 2 directly in normalized action space, needs no separate history encoder, and falls back to isotropic Gaussian noise when no valid history anchor exists. The history pool is strictly causal, and the component-wise variants are retained only for ablation; configuration and buffer details are given in Appendix A.1.

4 Experiments

We study three questions: (1) Does explicitly modeling the joint–time grid improve over a timestep-only FM-DiT (Lipman et al., 2023; Peebles & Xie, 2023)? (2) What is contributed by inter-step edges, inter-joint edges, and their union? (3) What does the history-shaped source contribute, and does the complete policy remain effective with few integration steps and under distribution shift?

4.1 Setup

Tasks and control. We evaluate on LIBERO manipulation (Liu et al., 2023a) with the official OSC end-effector interface (7D actions, comprising Cartesian translations, rotations, and gripper opening), 128×128 egocentric RGB, and five no-op warmup steps before control begins. The real-robot experiments (Figure 5) instead use joint-position (joint-angle) actions on the SO-101 arm, and the same ACA operates over joint factors in both parameterizations. Each reported cell uses 30 evaluation trials per run.

Shared training configuration. All compared policies are trained for 90 epochs with 500 warmup steps and batch size 64 under the same AdamW/EMA recipe, unless their original family requires a different sampler, and are evaluated under the same protocol so that the reported differences reflect the compared architectures and source construction rather than training-schedule or checkpoint-selection variations. The optimizer settings for AC-HFlow are given in Appendix A.1.

Protocols. **AC-HFlow** denotes the complete policy with union-masked Action-Consistent Attention and History Flow Matching enabled; **AC-HFlow+GRoPE** is the same policy with the sequential grid encoding GRoPE in place of flattened RoPE and learned slot embeddings (equation 9). We compare against flow-based policies VITA (Gao et al., 2025) and A2A (Jia et al., 2026), diffusion policies DDPM (Ho et al., 2020) and DDIM (Song et al., 2021) with UNet or DiT (Peebles & Xie, 2023) backbones, and a timestep-only FM-DiT flow-matching baseline (Lipman et al., 2023; Peebles & Xie, 2023). Network sizes, training schedules, integration step counts, and the LIBERO-Plus shared-ID protocol are detailed in Appendix A.1.

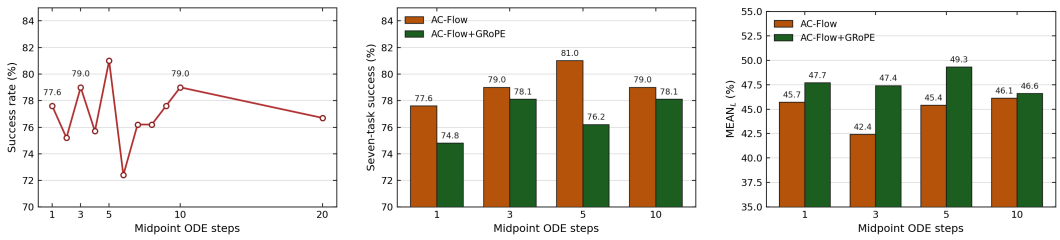
Table 2: Structure ablations on seven LIBERO tasks (30 trials per evaluation). The suite uses `open_middle_drawer` in place of the bowl task. Joint-only (S), time-only (T), and union denote joint, time, and joint+time, respectively; hist denotes the history-shaped noise source.

Variant	History	ST	Steps	Avg.
FM-DiT	no	–	10	24.3
FM + hist	yes	–	10	78.6
AC-HFlow-S (joint-only)	no	S	10	78.1
AC-HFlow-T (time-only)	no	T	10	77.1
AC-HFlow-union	no	union	10	78.6
AC-HFlow+GRoPE	yes	union	10	78.1
AC-HFlow-union + hist	yes	union	10	**79.5**

4.2 Does Explicit Joint–Time Structure Help?

Table 1 summarizes the operating-point comparison. AC-HFlow attains the highest average success in this evaluation, and even a single integration step remains competitive with the strongest long-schedule baselines. AC-HFlow+GRoPE stays within one point of AC-HFlow and remains competitive across its one-to-ten step sweep (Figure 3b; Appendix A.5, Table 7), with only small per-task differences on individual tasks. The diffusion-policy baselines and VITA reach lower averages despite 10–90 integration steps, and the timestep-only FM-DiT is markedly weaker. The contrast with FM-DiT shows the overall effect of the complete AC-HFlow framework under the same action-space flow-matching and DiT policy family, while the component-level evidence is provided by Table 2. The comparisons across policy families are not compute-matched and should not be read as controlled architectural ablations.

4.3 Component Ablations



(a) AC-HFlow as a function of the number of ODE steps (Table 5).

(b) Effect of adding GRoPE on LIBERO (Appendix A.5, Table 7).

(c) Effect of adding GRoPE on LIBERO-Plus (Appendix A.4).

Figure 3: Integration-budget ablation and the effect of GRoPE. (a) AC-HFlow seven-task Avg. as a function of the number of midpoint ODE steps, ablating the integration budget (detailed per-task numbers in Appendix A.3, Table 5). (b) AC-HFlow vs. AC-HFlow+GRoPE on LIBERO, showing that GRoPE stays on par in-distribution (Appendix A.5, Table 7). (c) AC-HFlow vs. AC-HFlow+GRoPE on LIBERO-Plus, showing that GRoPE generalizes more robustly under visual and scene shifts (best step counts in Appendix A.4, Table 6).

Table 2 separates the effects of relational factorization, positional encoding, and history-shaped source construction under the shared training recipe. The plain FM-DiT baseline is substantially weaker, confirming that a timestep-only velocity field struggles to capture the joint–time structure. Adding either the temporal (T) or the joint (S) axis independently lifts performance into the high-70s, showing that each axis is a complementary route absent from the flattened baseline; their union matches the history-only FM baseline at ten steps. GRoPE attains a mean on par with the default encoding while replacing the learned joint-slot embedding with rotation-based grid coordinates. Adding the history-shaped source yields the best Avg. of this ablation, and the source alone already rescues plain FM-DiT to a comparable level, confirming that source construction is a major independent component rather than a minor auxiliary prior. The gaps among the structured variants are

small, so these ablations indicate relative contributions rather than statistical superiority. Per-task values appear in Appendix A.6 (Table 8).

4.4 Few-Step Behavior

A dedicated step sweep for the complete AC-HFlow policy on the bowl suite ablates the number of midpoint integration steps: a single step already attains competitive success, and performance stays robust across the entire sweep, showing that the policy tolerates aggressive budget reductions (Figure 3a; per-task values in Appendix A.3, Table 5). Across the same sweep, AC-HFlow+GRoPE remains on par with the default encoding at every step count, so adding GRoPE preserves in-distribution performance (Figure 3b; per-task values in Appendix A.5, Table 7). The benefit of GRoPE becomes visible under distribution shift: on LIBERO-Plus it keeps a consistent edge over the default encoding at every step count, indicating better generalization robustness (Figure 3c; Appendix A.4). We therefore adopt **3 steps** as the primary reporting point for the complete attention-and-history policy.

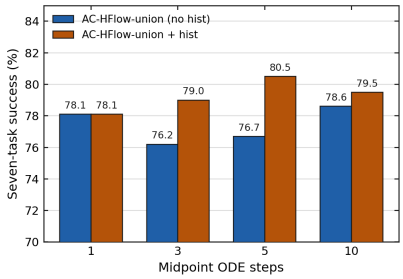


Figure 4: Union-masked attention with and without the history-shaped source across ODE steps.

Figure 4 compares the union attention with and without the history-shaped source across midpoint ODE steps. Compared with the no-history variant, the history-shaped source reaches a higher best result with fewer integration steps. Recent action statistics thus reduce the required integration budget, complementing the relational structure of the velocity field.

Action-chunk diagnostics further confirm that the generated chunks reflect the intended within-chunk joint coordination and replan-boundary continuity, and the wall-clock cost of few-step inference is several times lower than that of long-schedule diffusion baselines (Appendix A.7, Tables 10 and 11).

Table 3: LIBERO-Plus microwave success (%) across shared L1–L5 variants (seed 42; see Appendix A.2 for per-level scene counts). Avg. is the unweighted average over levels. AC-HFlow and AC-HFlow+GRoPE are shown at their best-performing step count.

Policy	Steps	L1	L2	L3	L4	L5	Avg.
AC-HFlow	10	53.6	60.0	51.4	**40.0**	**25.7**	46.1
AC-HFlow+GRoPE	5	60.7	**62.9**	**57.1**	40.0	25.7	**49.3**
VITA	10	50.0	40.0	42.9	37.1	11.4	36.3
A2A	10	**64.3**	37.1	37.1	20.0	5.7	32.8
A2A-Noise	10	46.4	45.7	42.9	22.9	17.1	35.0
FM-DiT	10	10.7	5.7	11.4	2.9	0.0	6.1
DDPM-DiT	90	53.6	42.9	37.1	37.1	22.9	38.7
DDIM-UNet	90	64.3	60.0	48.6	37.1	20.0	46.0
DDPM-UNet	90	60.7	48.6	34.3	31.4	25.7	40.1

4.5 LIBERO-Plus robustness and cross-policy analysis

We evaluate shared-ID LIBERO-Plus variants (Fei et al., 2025) for `put_mug_in_microwave` across difficulty levels L1–L5; all policies face identical scene IDs (seed 42; per-level counts in Appendix A.2), providing a controlled robustness test under visual and scene shifts.

Table 3 summarizes the generalization results under shared scene IDs. AC-HFlow+GRoPE attains the highest Avg. at only five midpoint steps, exceeding both the long-schedule diffusion policies and the other flow-based baselines, while AC-HFlow is also competitive and uses an order of magnitude fewer integration steps than the diffusion baselines. The advantage is not uniform across levels: at the easiest level (L1) several methods perform comparably, but our policies maintain consistent leads over the other flow-based baselines across L2–L5 as scene complexity increases (Figure 6). A

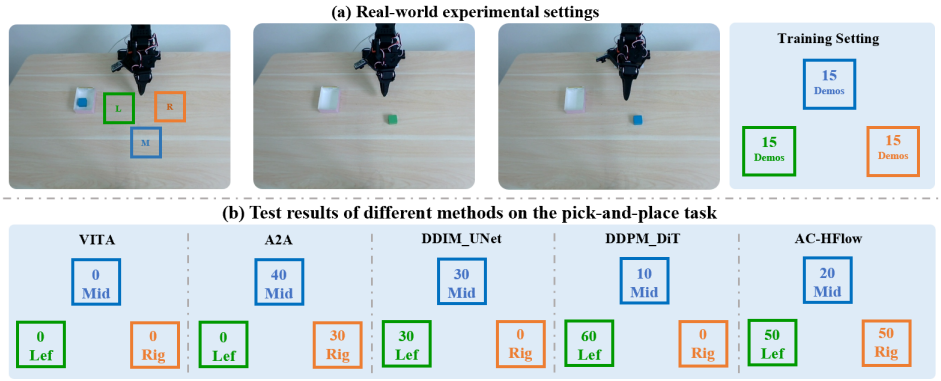


Figure 5: Real-robot success rate (%) at the left, center, and right positions for each policy. Each position is evaluated with ten trials (five blue and five green objects). AC-HFlow is the only policy that succeeds at all three positions.

step ablation on the same shared IDs confirms that this edge is consistent across integration budgets (Figure 3c; Appendix A.4): the structured velocity field captures sufficient information early in integration, and the GRoPE encoding transfers more robustly under distribution shift.

4.6 Real-Robot Validation

To assess performance beyond simulation, we deploy the trained AC-HFlow policy on a real robot arm (SO-101 / SO-ARM101 follower arm (Team, 2024)) in a tabletop manipulation task. Unlike the LIBERO OSC setup, demonstrations and closed-loop control on the physical arm use *joint-position* actions (changes in joint angles). Demonstrations were collected from three spatial positions (left, center, and right) covering blue and green objects; each method is evaluated at each position with ten trials (five per object color), testing whether the joint–time relational structure and the history-shaped source transfer to joint-space control under positional and appearance variation.

Figure 5 reports the per-position success rates on the physical system. AC-HFlow is the only policy that succeeds at all three positions, whereas each of the strongest baselines fails entirely at one position. That AC-HFlow is the only policy with non-zero success everywhere confirms that the union-masked relational attention and the history-informed flow source transfer robustly to the physical setup, complementing the simulated evaluations.

Beyond aggregate success, AC-HFlow exhibited a distinctive closed-loop behavior during fine-grained grasping: near the target block it repeatedly issued small corrective motions before committing to the grasp, although the teleoperated demonstrations are smooth and no baseline displayed comparable behavior. We attribute this emergent pre-grasp refinement to the combination of the history-shaped flow source and few-step inference: by re-conditioning on recent executed actions at every replan, the policy continuously aligns its upcoming chunk with the actual end-effector state, effectively performing closed-loop pose correction without an explicit controller. This is consistent with the closed-loop diagnostic in Appendix A.7 (Table 11), where AC-HFlow shows the smallest replan-boundary arm discontinuity among all policies. While anecdotal, this indicates that the proposed mechanism supports reactive fine-tuning that the baselines did not exhibit.

5 Conclusion

We introduced AC-HFlow, which structures action chunks as joint–time grids through Action-Consistent Attention and initializes generation from a history-shaped source rather than isotropic noise. Extensive experiments demonstrate that ACA improves temporal evolution and same-step joint coordination, while the history-shaped source enhances cross-replan continuity and enables competitive performance with very few, or even a single, integration step. Our current validation focuses on mid-sized flow policies; extending AC-HFlow to large-scale VLA and WAM systems remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods (code). We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, prove claims, propose or refine hypotheses, design experimental methodology, interpret results, or assist with translation, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for recommended tasks such as drafting and editing text for readability and formatting, including grammar checking and checking consistency of descriptions across the paper. AI-assisted code was reviewed, tested, and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

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*, 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Galliker, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 305, pp. 17–40, 2025.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

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. *arXiv preprint arXiv:2510.13626*, 2025.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning (CoRL)*, 2022.

Dechen Gao, Boqi Zhao, Andrew Lee, Ian Chuang, Hanchu Zhou, Hang Wang, Zhe Zhao, Junshan Zhang, and Iman Soltani. VITA: Vision-to-action flow matching policy. *arXiv preprint arXiv:2507.13231*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jindou Jia, Gen Li, Xiangyu Chen, Tuo An, Yuxuan Hu, Jingliang Li, Xinying Guo, and Jianfei Yang. Action-to-action flow matching. In *Robotics: Science and Systems (RSS)*, 2026. arXiv:2602.07322.

Sunshine Jiang, Xiaolin Fang, Nicholas Roy, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Siddharth Ancha. Streaming flow policy: Simplifying diffusion/flow-matching policies by treating action trajectories as flow trajectories. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 305, pp. 238–257, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 270, pp. 2679–2713, 2025.

Andrew Choong-Won Lee, Ian Chuang, Ling-Yuan Chen, and Iman Soltani. InterACT: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 270, pp. 1730–1743, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Jiahua Ma, Yiran Qin, Yixiong Li, Xuanqi Liao, Yulan Guo, and Ruimao Zhang. CDP: Towards robust autoregressive visuomotor policy learning via causal diffusion. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 305, pp. 4190–4205, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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 *Robotics: Science and Systems (RSS)*, 2025.

Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Yuan. Consistency policy: Accelerated visuomotor policies via consistency distillation. In *Robotics: Science and Systems (RSS)*, 2024.

Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems (RSS)*, 2024.

Moritz Reuss, Hongyi Zhou, Marcel Rühle, Ömer Erdinç Yağmurlu, Fabian Otto, and Rudolf Lioutikov. FLOWER: Democratizing generalist robot policies with efficient vision-language-flow models. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 305, pp. 3736–3761, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. In *Neurocomputing*, volume 568, 2024. arXiv:2104.09864.

Hugging Face LeRobot Team. LeRobot: State-of-the-art machine learning for real-world robotics. `https://github.com/huggingface/lerobot`, 2024.

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. In *Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zhendong Wang, Max Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, Ming-Yu Liu, and Yu Zeng. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. In *International Conference on Machine Learning (ICML)*, pp. 63399–63416, 2025.

Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. DiffusionVLA: Scaling robot foundation models via unified diffusion and autoregression. In *International Conference on Machine Learning (ICML)*, pp. 66558–66574, 2025.

Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14754–14762, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

A Appendix

A.1 Implementation and Reproducibility Details

Policy instantiation. The full policy corresponds to `AC-HFlow` in the accompanying implementation. The observation encoder is a ResNet-18 (He et al., 2016) without pretrained weights, with GroupNorm, ImageNet normalization, and random cropping during training. The flow transformer uses four blocks, hidden width 512, eight heads, MLP ratio 4, dropout 0.1, and flow-time embedding width 256. Each block performs horizon self-attention, ACA, and a GELU MLP in this order; all three residual branches have independently predicted AdaLN shift, scale, and zero-initialized gates. The ACA lift uses d_j=64 features per learned joint factor with seven arm joints and two gripper factors (J=9); on LIBERO these slots factorize OSC end-effector commands, while on the real robot they factorize joint-position vectors (Figures 1 and 5). AC-HFlow+GRoPE uses the same policy configuration: identical backbone, history source, and union mask, but applies equation 9 on the lifted (B, T, J, C) grid before flattening (no learned joint-slot embedding). The rotary base defaults to $b = 10\,000$.

Training hyperparameters. AC-HFlow uses AdamW (lr=5×10^{-4}, β=$(0.95, 0.999)$, ε=10^{-8}, weight decay 10^{-6}), cosine learning-rate decay with 500 warmup steps, EMA (power 0.75), batch size 64, and 90 epochs. Baselines follow the same epochs, warmup, and batch size where their family allows. Demos are loaded as OSC zarr trajectories with ImageNet-normalized 128×128 RGB and the shared action normalizer.

Optimization target and solver. Training samples flow time uniformly and minimizes an elementwise MSE velocity loss over unconditioned action entries. The structure ablations in Table 2 additionally evaluate the no-history union and the GRoPE variant at ten midpoint steps for a common reporting point, while the primary AC-HFlow result evaluates the history-enabled field with three midpoint steps. The policy predicts a horizon of 16 normalized actions and then replans.

History buffer. At evaluation, the buffer stores the previously predicted eight-action execution chunk and is cleared at episode reset. At training time, aligned dataset history is normalized with the same action normalizer as the target chunk. Invalid anchors fall back to isotropic Gaussian noise. For valid anchors, the configuration uses one eight-action history chunk to compute the per-control statistics in equation 10, a standard-deviation floor of 10^{-3}, and full-horizon source shaping. During training, $p = 0.85 + 0.10\xi$ with $\xi \sim \mathrm{Beta}(5, 5)$; evaluation fixes p=0.9. The affine standardization in equation 11 is enabled in the reported configuration.

Union-mask construction. Tokens are flattened with control index outermost, $r(j, t) = jT + t$. For every query (j_q, t_q), the implementation allows keys satisfying Eq. equation 6 and sets all remaining logits to $-\infty$ before scaled dot-product attention. Default AC-HFlow uses equation 8 and

Table 4: LIBERO-Plus microwave per-level scene counts (seed 42; total n=168).

Level	L1	L2	L3	L4	L5	Total
Scenes	28	35	35	35	35	168

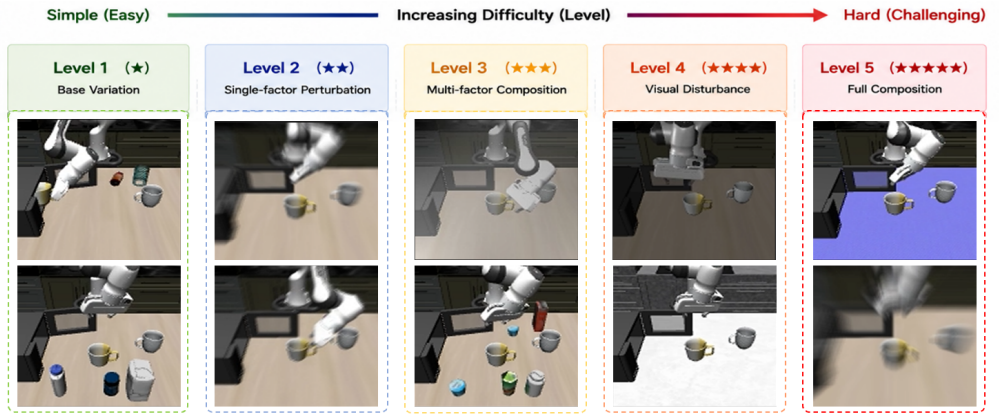


Figure 6: LIBERO-Plus generalization across difficulty levels. Higher levels introduce stronger scene perturbations (layout, lighting, camera, background, and object variations), and success degrades accordingly. AC-HFlow and AC-HFlow+GRoPE retain the strongest success at every level and degrade more gracefully than the flow-based baselines under the shared scene IDs.

$\mathrm{RoPE}(\cdot; r(j,t), b)$ on the flattened sequence; AC-HFlow+GRoPE applies equation 9 on the lifted grid before flattening. An internal residual connection and LayerNorm are applied in joint-feature space before projection back to the DiT width.

A.2 LIBERO-Plus Scene Counts

Table 4 reports the per-level scene counts used in the LIBERO-Plus microwave evaluation. All variants are drawn with seed 42 and capped at $\min(35, |\mathrm{pool}|)$ per level; L1 is pool-limited while L2–L5 are capped at 35. To keep the generalization evaluation objective, we directly evaluate the LIBERO-trained weights on the corresponding LIBERO-Plus scenes without any Plus-specific fine-tuning.

Figure 6 visualizes the per-level success of all evaluated policies on the shared LIBERO-Plus microwave variants: performance degrades as the perturbation level increases, and AC-HFlow as well as AC-HFlow+GRoPE retain the strongest success at every level.

A.3 Inference-Step Sensitivity

Table 5 reports every evaluated midpoint step count for the history-enabled union model. This sweep keeps the bowl task (unlike the drawer-based structure suite) and varies only the midpoint step count at evaluation. All cells use 30 episodes, the official OSC controller, 128×128 RGB, and five warmup no-ops. Performance is non-monotonic in solver steps: three steps match the best ten-step Avg. (79.0%) while one step already reaches 77.6%.

A.4 LIBERO-Plus Microwave Step Sensitivity

Table 6 reports AC-HFlow and AC-HFlow+GRoPE at their best-performing step counts on the shared LIBERO-Plus microwave variants, using the same shared IDs (seed 42; per-level counts in Table 4).

Table 5: Complete inference-step sweep for AC-HFlow with history on seven LIBERO tasks (bowl suite; 30 trials). Success is in percent.

Steps	both moka	bowl	microwave	pudding	soup	stove	stove+moka	Avg.
1	56.7	96.7	90.0	56.7	60.0	100.0	83.3	77.6
2	53.3	93.3	76.7	53.3	66.7	100.0	83.3	75.2
3	53.3	100.0	90.0	60.0	73.3	100.0	76.7	79.0
4	50.0	100.0	93.3	43.3	60.0	96.7	86.7	75.7
5	56.7	96.7	90.0	56.7	80.0	100.0	86.7	**81.0**
6	46.7	100.0	93.3	40.0	56.7	96.7	73.3	72.4
7	53.3	100.0	90.0	43.3	60.0	100.0	86.7	76.2
8	56.7	100.0	86.7	56.7	60.0	100.0	73.3	76.2
9	60.0	100.0	86.7	43.3	70.0	100.0	83.3	77.6
10	53.3	96.7	93.3	60.0	66.7	100.0	83.3	79.0
20	60.0	90.0	86.7	43.3	70.0	100.0	86.7	76.7

Table 6: AC-HFlow and AC-HFlow+GRoPE at their best-performing midpoint step count on the shared LIBERO-Plus microwave variants (n=168; seed-42 shared IDs; per-level counts in Table 4). Success is in percent; Avg. is the unweighted average over L1–L5.

Policy	Steps	L1	L2	L3	L4	L5	Avg.
AC-HFlow	10	53.6	60.0	51.4	40.0	25.7	46.1
AC-HFlow+GRoPE	5	60.7	62.9	57.1	40.0	25.7	**49.3**

A.5 AC-HFlow+GRoPE Evaluation Details

Evaluation protocol. AC-HFlow+GRoPE follows the same LIBERO OSC setup as AC-HFlow (Appendix A.8): 128×128 RGB, five warmup no-ops, and **30 trials per run**. Policies are trained for 90 epochs; each table cell reports success for the given task and midpoint step count. The eight evaluated training tasks are `open_middle_drawer`, `turn_on_stove`, `put_bowl_on_plate`, `put_mug_in_microwave`, `put_mug_pudding_plate`, `put_soup_cheese_in_basket`, `put_both_moka_pots_on_stove`, and `turn_on_stove_moka_pot`. Main-text seven-task tables use `open_middle_drawer` in the drawer column together with stove, microwave, pudding, soup, moka, and stove+moka (bowl is omitted from that aggregate).

LIBERO-Plus microwave protocol. Generalization rows reuse shared variant IDs with seed 42 (n=168; per-level counts in Table 4). Unlike LIBERO, each Plus variant is evaluated with **one trial per scene** (168 rollouts total). Each policy is evaluated on the same scene IDs. AC-HFlow+GRoPE Plus sweeps at steps $1/3/10$ were collected on identical shared IDs; the additional five-step sweep and the AC-HFlow sweeps at steps $1/5$ were evaluated following the same per-level rollout protocol. Avg. is the unweighted average of per-level success rates L1–L5.

Table 7 lists per-task AC-HFlow+GRoPE success for every reported integration step.

A.6 Structural Ablations

Table 8 exposes the per-task values underlying the means in Table 2, with `open_middle_drawer` replacing bowl. All variants share the Setup training recipe (AdamW, EMA, batch 64, 90 epochs) and DiT backbone; only the ST visibility mask, history-source flag, and positional encoding differ. All variants are evaluated at ten midpoint ODE steps. Per task, the timestep-only FM-DiT fails on most goals (e.g., 0.0 on both-moka, pudding, soup, and 3.3 on stove+moka), whereas every structured variant recovers high per-task success on drawer, microwave, stove, and stove+moka; the history-shaped source alone brings FM-DiT to 78.6% Avg. by rescuing the goal-oriented tasks. AC-HFlow-union + history reaches the best Avg. (79.5%), with per-task gains most visible on microwave and pudding. Table 9 gives the per-task step-efficiency comparison underlying Figure 4 in the main text. Each reported entry uses 30 trials per evaluation.

Table 7: AC-HFlow+GRoPE per-task success (%) on seven LIBERO tasks (30 trials per run). Columns match Table 1; Avg. is the unweighted average over the seven listed tasks.

Steps	drawer	stove	micro.	pudd.	soup	moka	s+moka	Avg.
1	100.0	100.0	93.3	46.7	60.0	40.0	83.3	74.8
3	100.0	100.0	90.0	53.3	63.3	53.3	86.7	78.1
5	100.0	100.0	90.0	46.7	60.0	50.0	86.7	76.2
10	100.0	100.0	93.3	50.0	66.7	50.0	86.7	78.1

Table 8: Per-task structure ablation with `open_middle_drawer` replacing bowl (30 trials per evaluation). S and T denote joint-only and temporal-only visibility. "Hist" denotes history-shaped flow initialization. All variants use ten midpoint ODE steps.

Variant	Hist	Steps	both moka	drawer	microwave	pudding	soup	stove	stove+moka	Avg.
FM-DiT	no	10	0.0	80.0	20.0	0.0	0.0	66.7	3.3	24.3
FM-DiT + history	yes	10	73.3	96.7	83.3	56.7	53.3	100.0	86.7	78.6
AC-HFlow-S	no	10	63.3	100.0	83.3	56.7	60.0	100.0	83.3	78.1
AC-HFlow-T	no	10	60.0	96.7	86.7	50.0	63.3	100.0	83.3	77.1
AC-HFlow-union	no	10	50.0	100.0	83.3	56.7	73.3	100.0	86.7	78.6
AC-HFlow+GRoPE	yes	10	50.0	100.0	93.3	50.0	66.7	100.0	86.7	78.1
AC-HFlow-union	yes	10	53.3	100.0	93.3	60.0	66.7	100.0	83.3	**79.5**

A.7 Efficiency and Action-Chunk Diagnostics

Closed-loop task completion time. Table 10 reports wall-clock time and environment steps to finish a successful episode on `turn_on_stove` and `put_mug_in_microwave`. Each policy uses its operating-point checkpoint and step count; metrics are averages over successful trials only among 10 closed-loop rollouts (failures excluded). Environment step counts are similar across methods on stove, whereas wall-clock time is dominated by per-chunk inference: long-schedule diffusion is several times slower than few-step flow policies despite comparable control horizons.

Action-chunk diagnostics. Table 11 summarizes action-level diagnostics aligned with the two relation types modeled by Action-Consistent Attention (ACA) in equation 6: the spatial (inter-joint) coupling within each timestep and the temporal (inter-step) evolution along each joint factor within a chunk. The last column additionally reports the jump between consecutive chunks at replan boundaries, which is related to the history-shaped source rather than to the intra-chunk edges of ACA.

Interpretation (Table 11). AC-HFlow lifts the velocity field onto a joint–time grid and applies union-masked ACA (equation 6) after horizon self-attention in every DiT block. The mask admits two edge types on $\mathcal{G}_T \times \mathcal{G}_J$: causal *inter-step* links along each joint factor and directed *inter-joint* links within each timestep. The diagnostics in Table 11 probe whether these inductive biases are reflected in the *output* action chunks rather than only in training loss.

Same-step control coupling (inter-joint edges). The first four columns target the within-step clique in equation 6: at a fixed t, attention is directed along joint indices, coupling the ordered joint factors of that step. c_{adj} measures Pearson correlation between adjacent arm controls inside a predicted chunk; $\mathrm{spat_st_tv}$, $|x-y|$, and $|y-z|$ measure the same coordination in magnitude form (mean adjacent-joint gaps). AC-HFlow attains the best value in every spatial column, including the largest open-loop c_{adj}=0.090, whereas the baselines cluster near zero or negative (VITA -0.191, A2A -0.013, A2A-Noise -0.035). This pattern is consistent with the design goal of equation 6: without an explicit within-step factorization, horizon attention treats each timestep as a single token and does not require neighboring controls to move coherently, so same-step dimensions can drift or anti-correlate even when the chunk is temporally smooth. Union-masked ACA instead routes messages among ordered joint factors at the same step, and the diagnostics show that the resulting chunks exhibit stronger same-step coupling both statistically (c_{adj}) and geometrically (smaller $|x-y|$ and $|y-z|$ gaps).

Table 9: Per-task step-efficiency ablation (30 trials per evaluation; drawer suite). Avg. is the unweighted average of the seven per-task success rates. The history-shaped source does not decrease the seven-task Avg. at any midpoint ODE step.

Variant	Steps	both moka	drawer	microwave	pudding	soup	stove	stove+moka	Avg.
AC-HFlow-union	1	60.0	96.7	86.7	56.7	63.3	100.0	83.3	78.1
AC-HFlow-union	3	50.0	96.7	86.7	60.0	56.7	100.0	83.3	76.2
AC-HFlow-union	5	46.7	100.0	86.7	50.0	66.7	100.0	86.7	76.7
AC-HFlow-union	10	50.0	100.0	83.3	56.7	73.3	100.0	86.7	78.6
AC-HFlow-union + hist	1	56.7	100.0	90.0	56.7	60.0	100.0	83.3	78.1
AC-HFlow-union + hist	3	53.3	100.0	90.0	60.0	73.3	100.0	76.7	79.0
AC-HFlow-union + hist	5	56.7	93.3	90.0	56.7	80.0	100.0	86.7	**80.5**
AC-HFlow-union + hist	10	53.3	100.0	93.3	60.0	66.7	100.0	83.3	79.5

Table 10: Success-conditioned closed-loop completion cost (10 trials; failures excluded). Env. steps are control steps to success; time is wall-clock seconds including policy inference and simulation.

		stove		microwave	
Policy	Steps	Env. steps	Time (s)	Env. steps	Time (s)
AC-HFlow	3	78.2	1.13	252.0	3.06
VITA	10	73.2	0.85	296.2	2.85
A2A	10	79.7	0.92	255.4	2.56
A2A-Noise	10	77.3	0.89	252.6	2.49
DDPM-DiT	90	77.0	3.04	254.3	9.48
DDIM-UNet	90	74.2	3.78	268.8	13.08
DDPM-UNet	90	74.4	3.83	246.0	13.16

Table 11: Action-chunk diagnostics. The first four columns probe the two relations encoded by ACA: spatial same-step joint coupling (c_{adj}, $\mathrm{spat_st_tv}$, $|x-y|$, $|y-z|$); the last column reports the jump between consecutive chunks at replan boundaries. c_{adj} is measured open-loop (OL) on `put_both_moka_pots_on_stove`; the other columns are measured closed-loop (CL) on `turn_on_stove` (10 trials). Gripper excluded; best per column bolded.

Policy	$c_{\mathrm{adj}}\uparrow$	$\mathrm{spat_st_tv}\downarrow$	$\lvert x-y\rvert\downarrow$	$\lvert y-z\rvert\downarrow$	$\mathrm{replan_disc_arm}\downarrow$
AC-HFlow	**0.090**	**0.301**	**0.424**	**0.564**	**0.095**
VITA	-0.191	0.315	0.462	0.598	0.096
A2A	-0.013	0.302	0.448	0.567	0.125
A2A-Noise	-0.035	0.310	0.451	0.582	0.109
DDPM-DiT	0.014	0.307	0.441	0.579	0.107
DDIM-UNet	0.008	0.327	0.474	0.621	0.105
DDPM-UNet	-0.062	0.318	0.457	0.600	0.108

Replan-boundary continuity (cross-chunk). $\mathrm{replan_disc_arm}$ measures $\|C_{k+1}[0]-C_k[-1]\|$ on arm dimensions—the jump between the last action of one predicted chunk and the first action of the next after replanning. This probes temporal coherence *across* chunk boundaries; it is complementary to the causal inter-step edges inside each chunk and reflects the effect of the history-shaped source, which aligns each new chunk's start with the previously executed motion. AC-HFlow again ranks first (0.095), with a clearer margin over A2A (0.125) and A2A-Noise (0.109) than over VITA (0.096). We attribute the improvement to the combination modeled in equation 6 and equation 11: causal temporal message passing within each lifted chunk encourages self-consistent intra-chunk evolution, while History Flow Matching shapes the flow source from recent executed actions so consecutive replans start from a history-aligned endpoint rather than independent noise. Flat-sequence policies that lack the joint–time lift (VITA, diffusion UNet baselines) do not expose causal inter-step structure explicitly on joint factors, and their replan discontinuities remain in a similar 0.10–0.13 band despite strong closed-loop success on stove.

What the table does not claim. These metrics do not replace task success and are not exhaustive smoothness scores: AC-HFlow is not uniformly best on lag-one autocorrelation or jerk-based smoothness in our broader benchmark logs, where history-conditioned flat flow policies can be sharper. Table 11 instead isolates the two structures AC-HFlow was designed to make explicit—directed within-step coordination and causal inter-step evolution within a chunk—together with the cross-chunk replan continuity shaped by the history source, and shows that the union-masked relational velocity field leaves a measurable signature in generated actions, supporting the action-consistency motivation stated in the introduction and the structure ablations in Table 2. Because open-loop and closed-loop columns come from different tasks and the closed-loop averages use only ten rollouts, the comparisons should be read as qualitative audit evidence rather than as high-power statistical tests.

A.8 Evaluation Protocol

All reported LIBERO evaluations use the official OSC action interface, 128×128 RGB observations, a maximum of 600 environment steps, five no-op warmup steps, and 30 trials per run unless the LIBERO-Plus table states otherwise. Training follows the shared AdamW/EMA recipe in Appendix A.1; evaluation loads EMA weights when available and applies each policy's configured midpoint (or diffusion) step count unless a sweep overrides it. AC-HFlow+GRoPE LIBERO cells follow the same protocol as AC-HFlow (Appendix A.5). The structure ablations in Tables 2 and 8 average over both-moka, `open_middle_drawer`, microwave, pudding, soup, stove, and stove-moka (drawer replaces the bowl task used in earlier logs). The main comparison in Table 1 uses drawer together with stove, microwave, pudding, soup, moka, and stove-moka; AC-HFlow+GRoPE step sweeps are eval-only (Appendix A.5). LIBERO-Plus comparisons use shared variant IDs generated with seed 42; per-level scene counts are provided in Table 4 (n=168 total), so all policies encounter the same perturbations. Plus microwave RoPE evaluations use one trial per shared variant (168 episodes) rather than 30 LIBERO trials (Appendix A.5). Closed-loop task-completion times in Table 10 average successful trials only (10 rollouts per policy) on stove and microwave. Action-chunk diagnostics in Table 11 follow the task/protocol stated in the table caption. Since all compared policies are trained and evaluated under the same protocol, differences of a few percentage points should be interpreted cautiously.