MOSS: MODE-ORIENTED SOURCE SHAPING FOR ROBOT FLOW MATCHING POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

In robotic manipulation, the same observation often admits multiple feasible actions. Generative policies such as diffusion and flow matching naturally capture such multimodal behaviors, yet which mode is executed is left to random sampling and cannot be actively controlled. We propose **M**ode-**O**riented **S**ource **S**haping (**MOSS**), which turns mode selection into an explicit choice of where generation starts. We first reveal, both empirically and theoretically, that deterministic flow transport induces mode-dependent structure in the source space. Under ideal deterministic ODE transport, distinct behavioral modes have distinct source-space preimages, while practical approximation errors can blur their boundaries due to imbalanced data, limited model capacity, and few inference steps. Motivated by this source-space structure, MOSS constructs semantically aligned mode-specific *Gaussian* sources that encode mode identity in the initial noise. A modality-strength estimator predicts the local degree of multimodality, which adaptively controls the geometry of the mode-specific sources. It contracts them toward a shared *Gaussian* where behaviors are similar and separates them near behavioral branching points. During deployment, the desired behavior is selected at inference simply by sampling from the corresponding component. Across 6 simulated and 3 real-world manipulation tasks, MOSS steers to the target mode up to 100% of the time with high task success, recovers minority modes at over 90% guidance rate under a $10:90$ imbalance dataset, and surpasses mode-conditioned flow policy by 43.3 points of guidance rate on real robots.

1 INTRODUCTION

Generative models have recently become a prominent paradigm in robot imitation learning (Pearce et al., 2023; Physical Intelligence et al., 2026). Unlike conventional policies that regress toward an optimal action, representative approaches such as Diffusion Policy (DP) (Chi et al., 2023) and Flow Matching (FM) (Lipman et al., 2023; Bai et al., 2026) generate action trajectories from random noise, allowing them to represent multiple feasible behaviors under the same observation, i.e., *multimodal* behaviors (Zhai et al., 2025). Here, a *behavior mode* denotes a semantically coherent family of action sequences that follow the same qualitative strategy for accomplishing the task (Chi et al., 2023). For instance, a robot may approach the same object from the left or the right, or interact with different targets in different orders, while still completing the task successfully. However, in existing generative policies, which mode is ultimately executed is largely determined implicitly by random sampling (Lipman et al., 2023; Zhang et al., 2026). This limits their applicability in settings that require user-specified preferences, test-time constraints, or human intervention (Zoellner et al., 2026; He et al., 2025).

Existing solutions fall into two categories. The straightforward one explicitly conditions the generative model on a mode indicator, in practice most often a language instruction, so that different conditions induce different action distributions (Yang et al., 2026; Zhang & Gienger, 2024; Liu et al., 2025). Such conditioning, however, often fails to steer the policy reliably when the training data are imbalanced (Fang et al., 2026), since high-dimensional visual scenes are strongly correlated with task execution, the model tends to learn scene-action shortcuts that bypass the weaker, low-dimensional language condition (Ren et al., 2026). The other line instead operates on the source distribution of FM, controlling multimodal behaviors by implicitly learning a set of *Gaussian* priors (Zhang et al., 2026; Machado et al., 2026). These priors, however, carry no explicit semantic

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.

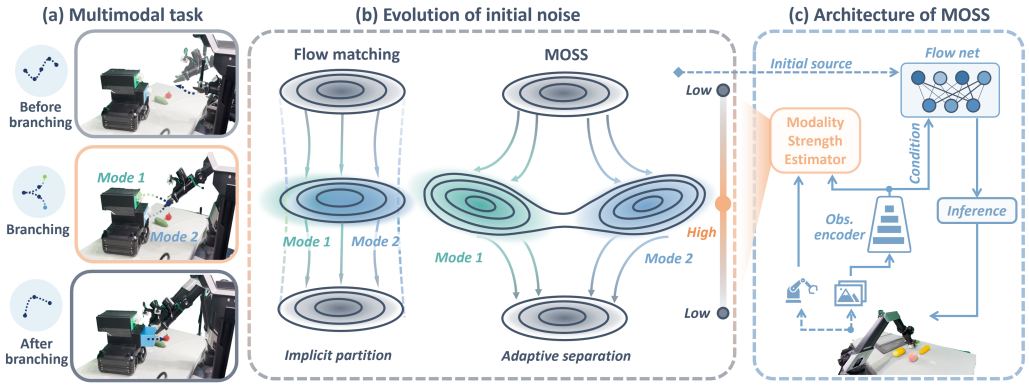

Figure 1: **Overview of MOSS. (a)** A real-world *open-drawer* task whose shared approach trajectory branches into two feasible modes. **(b)** Evolution of the initial noise. Standard FM starts from one isotropic *Gaussian* and separates modes only implicitly. MOSS shapes the source by the estimated modality strength, merging the components where the modes behave alike and separating them at branching. **(c)** Architecture of MOSS. The modality-strength estimator sets the source separation, while a single flow network is shared across modes.

correspondence to behavioral modes, so aligning them with user-defined modes still demands additional networks and laborious post-alignment (Zhang et al., 2026).

Toward a more reliable handle on mode selection, we first analyze how a well-trained FM policy behaves on multimodal tasks. We find that, even with a standard isotropic *Gaussian* source, the final behavior does not arise randomly from the entire noise space. Under deterministic Ordinary Differential Equation (ODE) transport, different behavioral modes tend to occupy distinct, albeit noticeably entangled, regions of the source space. We further show theoretically that, under ideal transport conditions, such mode-specific initial-noise regions are guaranteed to be implicitly separated. In practice, however, imbalanced mode distributions, limited model capacity, and a restricted number of inference steps leave these partitions interleaved, weakly separated, or severely imbalanced, sometimes resulting in mode preference or even mode collapse (Zhai et al., 2025; Mazza et al., 2026). This raises a natural question: *can we make the existing partitioning more explicit, and steer behavior directly through the initial distribution?*

Based on this insight, we propose **M**ode-**O**riented **S**ource **S**haping, termed **MOSS**, which enables controllable multimodal generation by dynamically shaping the source distribution of FM, as shown in Figure 1. Specifically, we first construct semantically defined mode-specific *Gaussian* sources, explicitly encoding mode identity into the initial noise and making the resulting control more robust to imbalanced mode distributions, limited model capacity, and limited inference steps. Since multiple modes typically arise only during certain critical phases rather than throughout the entire task (Jia et al., 2026a), a modality-strength estimator then predicts the local degree of multimodality from previous observation and action history, and uses it to continuously modulate the separability of the *Gaussian* components. When demonstrations from different modes behave alike, the components contract toward a shared *Gaussian* and maximize data sharing. As the robot approaches a behavioral branch, they gradually separate and give the initial noise a clear, stable mode identity.

MOSS provides several practical advantages beyond explicit mode selection. First, when different modes in the training demonstrations are highly imbalanced, the generation probabilities of standard FM tend to inherit the empirical data distribution. In contrast, our method can directly adjust the deployment probability of each mode by changing the source sampling distribution. Second, for data-scarce modes, we find that the proposed strategy can effectively leverage shared trajectory data from other modes, allowing the policy to learn more stable common motion structures while preserving clear mode boundaries, thereby improving both sample efficiency and task success for minority modes. Finally, compared with methods that apply strong guidance to shift the learned vector field at inference time, mode-specific sources still initialize transport from regions already established during training. This can potentially reduce trajectory-quality degradation caused by inference-time guidance (Yu et al., 2026), while naturally extending to currently emerging one-step denoising algorithms that do not support guidance (Ho & Salimans, 2022; Jia et al., 2026b).

We systematically evaluate our method on multiple robotic manipulation tasks with multimodal structure, including 6 simulation tasks and 3 real-world robot tasks. Compared with standard FM (Lipman et al., 2023), explicit mode-conditioned FM (Lei et al., 2026), and source-based methods (Zhang et al., 2026; Luo et al., 2026), MOSS achieves up to 100% target-mode steering while maintaining high task success rate (SR). Even under a highly imbalanced mode distribution $(10 : 90)$, our method can reliably recover the minority mode, achieving over 90% guidance rate (GR). Ablations show that adaptive source geometry improves GR over fixed *Gaussian* separation by 12-14 percentage points in average GR at the same training budget. On real-world tasks, MOSS achieves an average GR/SR of $84.2\%/85.3\%$, outperforming mode-conditioned FM by 43.3 and 12.0 percentage points in average GR and SR, respectively. These results validate the effectiveness of source-space control under visual noise, limited demonstrations, and execution errors.

2 Related work

Multimodal generative policies. Generative policies such as diffusion (Chi et al., 2023) and FM (Lipman et al., 2023) naturally model multimodal action distributions through stochastic generation. However, different behavioral modes can still become entangled or collapse during generation (Zhai et al., 2025; Mazza et al., 2026). Recent methods improve multimodal modeling through additional structure. VFP (Zhai et al., 2025) combines a variational latent prior, optimal transport, and a MoE decoder to better separate action modes, while MARS (Jia et al., 2026a) adaptively switches between stochastic and deterministic generation according to multimodality degree. None of these methods can actively control multimodality.

Source shaping for flow matching. Standard FM-based policies (Lipman et al., 2023) commonly initialize transport from a simple *Gaussian* source. Recent works instead design the source distribution to improve efficiency or structural alignment. A2A (Jia et al., 2026b) initializes flow from action-history representations to enable shorter transport and one-step inference; VITA (Gao et al., 2026) uses visual latents as the flow source; and LAFM (Machado et al., 2026) selects learned source priors according to latent actions. These methods primarily target efficient or better-aligned transport rather than explicit mode-level control. More recently, SL-FM (Zhang et al., 2026) separates behavioral branches through orthogonal source lifting while retaining a shared velocity field. MOSS similarly controls behavior through the source space, but explicitly aligns source components with semantic behavior modes and adaptively adjusts their separation according to local multimodality.

Explicit conditioning for behavioral control. Explicit conditioning provides a direct mechanism for controlling multimodal action generation by injecting the desired behavioral intent into the policy. Large VLA models such as RDT and $\pi_{0.5}$ (Liu et al., 2025; Physical Intelligence et al., 2026; 2025) use language and visual-semantic conditions to guide action generation. Lightweight generative policies can instead introduce explicit discrete mode variables; for example, PF-DAG (Lei et al., 2026) predicts a primary mode and generates continuous actions with a mode-conditioned MeanFlow policy. While effective, such explicit conditioning may introduce shortcut solutions, where the policy relies heavily on the injected mode signal rather than organizing multimodal behaviors intrinsically. In contrast, MOSS controls behavior through the source distribution without exposing the mode variable to the policy itself, and our experiments show source-based mode control yields stronger controllability than directly injecting the mode variable into the velocity-field condition.

3 Multimodal Property of Flow Matching

Standard FM learns a continuous transport from a simple source to a complex action distribution conditioned on an observation $\mathbf{o} \in \mathcal{O}$ (Lipman et al., 2023; Jia et al., 2026b). Let $d \in \mathbb{N}$ denote the dimension of the vectorized action chunk, $\mathbf{z} \in \mathbb{R}^d$ with $\mathbf{z} \sim \mu_0 = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ be the source noise, where $\mathbf{0}_d \in \mathbb{R}^d$ and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$, and let $\mathbf{y} \in \mathbb{R}^d$ with $\mathbf{y} \sim p^{\mathbf{o}}$ denote the target action distribution. Consider the linear *Gaussian* probability path

$$\mathbf{x}_t = t\mathbf{y} + \sigma_t \mathbf{z}, \qquad \sigma_t = 1 - (1 - \sigma_{\min})t, \qquad \sigma_{\min} \geq 0, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ and $t \in [0, 1]$. At $t = 1$, the probability-path endpoint is $\mathbf{x}_1 = \mathbf{y} + \sigma_{\min}\mathbf{z}$. We write $p_1^{\mathbf{o}}$ for its distribution and $\Phi_{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^d$ for the flow map induced by the ideal FM ODE from $\mathbf{x}_0 = \mathbf{z}$ to $\mathbf{x}_1$.

Empirically, we observe that FM policies trained on multimodal tasks already exhibit a structured source space: initial noises leading to different behavioral modes tend to occupy distinct regions of the *Gaussian* source, albeit noticeably entangled near their interface, as shown in Figure 2. We next show that this phenomenon arises naturally from deterministic transport theoretically.

3.1 Noise Zoning and Its Geometric Regularity

Partition the action space $\mathbb{R}^d$ into pairwise disjoint open regions $\mathcal{A}_1^{\mathbf{o}}, \dots, \mathcal{A}_K^{\mathbf{o}} \subset \mathbb{R}^d$ with semantic boundary $\mathcal{B}_A^{\mathbf{o}} \subset \mathbb{R}^d$, where $K \in \mathbb{N}$ denotes the number of behavioral modes and $p_1^{\mathbf{o}}(\mathcal{B}_A^{\mathbf{o}}) = 0$. Let $h_{\mathbf{o}} : \mathbb{R}^d \setminus \mathcal{B}_A^{\mathbf{o}} \to \{1, \dots, K\}$ denote the mode assignment of an endpoint $\mathbf{x} \in \mathcal{A}_k^{\mathbf{o}}$, i.e., $h_{\mathbf{o}}(\mathbf{x}) = k$ for $\mathbf{x} \in \mathcal{A}_k^{\mathbf{o}}$. The *noise zone* of mode k and the source-space boundary are the preimages $\mathcal{Z}_k^{\mathbf{o}} = \Phi_{\mathbf{o}}^{-1}(\mathcal{A}_k^{\mathbf{o}}) \subset \mathbb{R}^d$ and $\mathcal{B}_Z^{\mathbf{o}} = \Phi_{\mathbf{o}}^{-1}(\mathcal{B}_A^{\mathbf{o}}) \subset \mathbb{R}^d$.

Theorem 1 (Noise zoning)**.** *If $\Phi_{\mathbf{o}}$ is continuous and satisfies the exact transport condition $(\Phi_{\mathbf{o}})_{\#}\mu_0 = p_1^{\mathbf{o}}$, then $\mathcal{Z}_1^{\mathbf{o}}, \dots, \mathcal{Z}_K^{\mathbf{o}}$ are pairwise disjoint open sets with*

$$\mu_0(\mathcal{Z}_k^{\mathbf{o}}) = p_1^{\mathbf{o}}(\mathcal{A}_k^{\mathbf{o}}), \qquad \mu_0(\mathcal{B}_Z^{\mathbf{o}}) = 0. \tag{2}$$

Proof. See Appendix B.1 □

Theorem 1 shows that, up to a null boundary, the *Gaussian* source is already partitioned into semantic *noise zones*, each carrying exactly the probability mass of its behavioral mode. This partition is not imposed by design but arises from deterministic transport itself, which is the basic rationale for steering multimodal behavior through the source space: if modes already occupy distinct source regions, the generated behavior can in principle be selected by choosing where transport is initialized, without conditioning or modifying the velocity field.

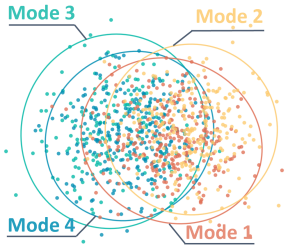

Figure 2: **Initial-noise mode separation in FM.** Experimental details are provided in Appendix A.

We now show that these zones are not arbitrarily fragmented sets but possess stable local geometry. For a concise global statement we take $\sigma_{\min} > 0$. The case $\sigma_{\min} = 0$ admits an analogous local guarantee as shown in Appendix B.4. Under standard regularity conditions on the FM velocity field $\mathbf{v}_t^{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^d$, Appendix B.2 shows that the induced ODE flow is a bi-*Lipschitz* homeomorphism. Specifically, there exists a finite constant $\Lambda_{\mathbf{o}} \in \mathbb{R}_{\geq 0}$ such that

$$e^{-\Lambda_{\mathbf{o}}}\|\mathbf{z} - \mathbf{z}'\|_2 \leq \|\Phi_{\mathbf{o}}(\mathbf{z}) - \Phi_{\mathbf{o}}(\mathbf{z}')\|_2 \leq e^{\Lambda_{\mathbf{o}}}\|\mathbf{z} - \mathbf{z}'\|_2, \qquad \forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d. \tag{3}$$

Theorem 2 (Margin-induced source separation)**.** *Suppose $\Phi_{\mathbf{o}}$ is a bi-Lipschitz homeomorphism satisfying equation 3. Then $\mathcal{Z}_k^{\mathbf{o}}$ is open and connected whenever $\mathcal{A}_k^{\mathbf{o}}$ is. Define the endpoint semantic margin $\gamma_{\mathbf{o}}(\mathbf{z}) := \operatorname{dist}(\Phi_{\mathbf{o}}(\mathbf{z}), \mathcal{B}_A^{\mathbf{o}}) \in \mathbb{R}_{\geq 0}$ and the corresponding source-space radius $R_{\mathbf{o}}(\mathbf{z}) := e^{-\Lambda_{\mathbf{o}}}\gamma_{\mathbf{o}}(\mathbf{z}) \in \mathbb{R}_{\geq 0}$, where $\operatorname{dist}(\mathbf{x}, S) := \inf_{\mathbf{s} \in S}\|\mathbf{x} - \mathbf{s}\|_2$ for $\mathbf{x} \in \mathbb{R}^d$ and $S \subset \mathbb{R}^d$. Then*

$$B(\mathbf{z}, R_{\mathbf{o}}(\mathbf{z})) \subseteq \mathcal{Z}_k^{\mathbf{o}}, \qquad \|\mathbf{z} - \mathbf{z}'\|_2 \geq R_{\mathbf{o}}(\mathbf{z}) + R_{\mathbf{o}}(\mathbf{z}'), \qquad \forall \mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}},\ \mathbf{z}' \in \mathcal{Z}_j^{\mathbf{o}},\ k \neq j, \tag{4}$$

where $B(\mathbf{z}, r) := \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u} - \mathbf{z}\|_2 < r\}$ denotes the open Euclidean ball centered at $\mathbf{z} \in \mathbb{R}^d$ with radius $r \geq 0$.

Proof. See Appendix B.2. □

Theorem 2 shows that ideal FM induces not only distinct semantic noise zones, but also locally stable ones: behaviors lying farther from the semantic boundary correspond to a larger same-mode neighborhood in the source space.

3.2 Practical FM Blurs the Noise Zoning

In practice, finite demonstrations, imperfect optimization, and finite-step ODE integration make the realized endpoint map $\widetilde{\Phi}_{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^d$ deviate from $\Phi_{\mathbf{o}}$. We measure this by the endpoint transport error $\eta_{\mathbf{o}}(\mathbf{z}) = \|\widetilde{\Phi}_{\mathbf{o}}(\mathbf{z}) - \Phi_{\mathbf{o}}(\mathbf{z})\|_2$

Proposition 3. *For any $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}}$, if $\eta_{\mathbf{o}}(\mathbf{z}) < \gamma_{\mathbf{o}}(\mathbf{z})$ then $h_{\mathbf{o}}(\widetilde{\Phi}_{\mathbf{o}}(\mathbf{z})) = h_{\mathbf{o}}(\Phi_{\mathbf{o}}(\mathbf{z}))$.*

Proof. See Appendix B.3.

Corollary 4 (Boundary-sensitive region)**.** *The guarantee vanishes once the transport error becomes comparable to the semantic margin: source points with $\gamma_{\mathbf{o}}(\mathbf{z}) = O(\eta_{\mathbf{o}}(\mathbf{z}))$ form a boundary-sensitive region in which mode confusion is likely to occur.*

Every practical factor pushes in the same direction: finite demonstrations make velocity estimation unreliable in sparsely covered regions, mode imbalance further removes supervision from minority behaviors, a network of bounded *Lipschitz* constant cannot reproduce the sharp velocity variation that a clean mode separation demands (Mazza et al., 2026), and finite-step integration adds numerical error. The resulting growth of $\eta_{\mathbf{o}}$ widens the ambiguous region and manifests during sampling as *mode intersection*, *mode preference*, or even local *mode collapse*. This exposes a fundamental limitation of leaving noise zoning to be learned implicitly from a standard *Gaussian* source: the existence of an ideal mode partition does not imply that sufficiently clear, stable, and controllable source regions emerge under finite data, limited model capacity, and finite-step inference.

4 Mode-Oriented Source Shaping

4.1 Overview

Section 3 shows that FM naturally induces mode-dependent regions in the source space, but this separation is implicit and can become weak or entangled in practice. Rather than recovering these implicit regions, MOSS deliberately imposes a simple, semantically aligned source geometry to make mode separation explicit and controllable. It enlarges the source-space separation between modes when it is actually needed, and selects the mode by choosing *where the transport starts*. How much separation is needed depends on the state: when the upcoming behavior is highly multimodal, MOSS places the mode-specific sources far apart; when the behavior is essentially determined, it lets them contract back toward a shared *Gaussian*. Figure 1 depicts its architecture.

Concretely, given an observation history $\mathbf{o}$, a target action chunk $\mathbf{y}$, and a mode label k, MOSS predicts a modality strength $q \in [0, 1]$ for the current state and draws the source sample $\mathbf{x}_0$ from the corresponding mode-specific component. In our implementation, we use $\sigma_{\min} = 0$. Training then follows standard FM (Lipman et al., 2023), with a single velocity field shared by all modes:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{y}, \qquad \mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{o}) - (\mathbf{y} - \mathbf{x}_0)\|_2^2\right]. \tag{5}$$

Mode selection is thereby moved out of the generative network and into the source space. The network learns one shared transport, while different behaviors are obtained by initializing that transport in different regions. In the following, we detail the architecture of MOSS.

4.2 Modality Strength Estimator

The degree of behavioral multimodality can vary substantially along a robot trajectory (Jia et al., 2026a). We therefore dynamically estimate whether strong mode separation is necessary at the current state.

We first construct a multimodality supervision signal r_i from the training data. For a query state i, we retrieve local neighbors with similar observation histories $\mathbf{o}_{\le t}$ and executed action histories $\mathbf{a}_{<t}$, and compare the between-mode variation of their future actions with the corresponding within-mode variation. Denoting these quantities by $\mathrm{Var}_{B,i}$ and $\mathrm{Var}_{W,i}$, respectively, we define

$$r_i = \frac{\mathrm{Var}_{B,i}}{\mathrm{Var}_{B,i} + \beta \mathrm{Var}_{W,i} + \delta}, \tag{6}$$

where $\beta > 0$ controls the relative contribution of within-mode variation, and $\delta > 0$ is a small constant for numerical stability.

When the local neighborhood contains only a single plausible future mode, $\mathrm{Var}_{B,i}$ is close to zero and the resulting r_i is small. In contrast, when multiple modes coexist and their future trajectories are clearly separated, $\mathrm{Var}_{B,i}$ increases and produces a larger modality strength. A large within-mode variation $\mathrm{Var}_{W,i}$ suppresses spurious multimodality signals caused by noisy or unstable neighbors.

We then train a lightweight MLP as the *Modality Strength Estimator*:

$$w = s_\phi(\mathbf{o}_{\le t}, \mathbf{a}_{<t}), \tag{7}$$

After training, s_ϕ is completely frozen and receives no gradient from the FM objective.

As visualized in Figure 4, the predicted strength peaks near states where multiple behavioral modes diverge and decreases once the mode has been resolved. Since the numerical scale of w can vary across tasks, we normalize it using task-wise percentile calibration computed solely from the training set, yielding a calibrated strength $q \in [0, 1]$. During rollout, we further apply temporal persistence to obtain q^{eff}, preventing the separation strength from decaying before the selected behavior is completed. For simplicity, we denote this effective strength as q hereafter. Additional estimator details are provided in Appendix C.

4.3 Adaptive Mode-Oriented Source Shaping

Given the estimated modality strength q, MOSS adaptively separates the source distributions associated with different behavioral modes. We present a unified formulation for K modes.

To define the mode-discriminative subspace, we use the multimodality supervision r_i from Section 4.2 to identify training states where future behaviors diverge. Let $\bar{\mathbf{a}}_k$ denote the empirical mean future-action chunk of mode k within this region. We then remove the component shared across modes,

$$\tilde{\mathbf{a}}_k = \bar{\mathbf{a}}_k - \frac{1}{K}\sum_{j=1}^{K} \bar{\mathbf{a}}_j, \tag{8}$$

and let $\mathbf{U} \in \mathbb{R}^{d \times d_s}$ be an orthonormal basis of the subspace spanned by $\{\tilde{\mathbf{a}}_k\}_{k=1}^{K}$, where $d_s \le K-1$ denotes the dimension of the mode-discriminative subspace.

For a requested mode k, MOSS defines its source center as

$$\mathbf{c}_k(q) = \alpha(q)\tilde{\mathbf{a}}_k, \tag{9}$$

and samples the initial condition from

$$p(\mathbf{x}_0 \mid k, q) = \mathcal{N}\Big(\mathbf{c}_k(q),\ \sigma_\parallel^2(q)\mathbf{U}\mathbf{U}^\top + \sigma_\perp^2(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\Big). \tag{10}$$

The modality strength q controls both the displacement of the source centers and their overlap, where $\sigma_\perp$ is a fixed standard deviation in the orthogonal complement of the mode-discriminative subspace. Specifically, let

$$\Delta(q) = \min_{i \ne j} \|\mathbf{c}_i(q) - \mathbf{c}_j(q)\|_2 \tag{11}$$

denote the minimum inter-mode center distance. We set

$$\sigma_\parallel(q) = \mathrm{clip}\left(\frac{\Delta(q)}{2\rho(q)}, \sigma_{\parallel,\min}, \sigma_{\parallel,\max}\right), \tag{12}$$

where both $\alpha(q)$ and the target separation ratio $\rho(q)$ increase monotonically with q. Hence, highly multimodal states induce more widely separated and less overlapping mode-specific sources, whereas small q keeps the source distributions closer to the standard Gaussian prior. For $K = 2$, the centered mode means are opposite along a one-dimensional subspace, reducing equation 10 to the symmetric dual-Gaussian construction.

Importantly, the mode label is used only to select the initial source distribution and is never provided to the velocity field $\mathbf{v}_\theta$. MOSS therefore retains a single shared FM velocity field and realizes behavioral control solely through its initial condition. At inference time, a user or high-level planner only needs to specify the desired mode k and sample $\mathbf{x}_0$ from the corresponding source distribution. Further implementation details of the source shaping procedure are provided in Appendix D.

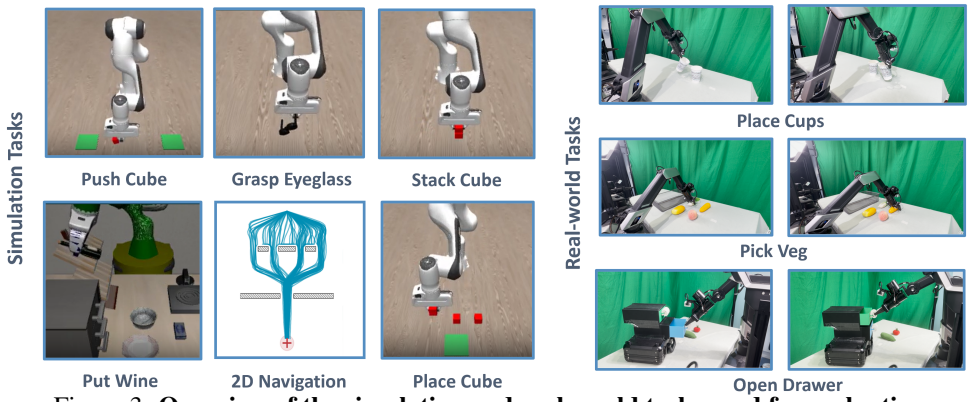

Figure 3: **Overview of the simulation and real-world tasks used for evaluation.**

Table 1: **Evaluation results on** 4 **simulation tasks.** For each requested mode, we report GR and SR (GR / SR). Best GR for each requested mode is shown in bold.

	Push Cube		Grasp Eyeglasses		Stack Cube		Put Wine	
Method	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2
Mode-Cond. FM	0.86 / 0.96	0.94 / 1.00	**1.00** / 1.00	**1.00** / 1.00	0.96 / 0.96	0.94 / 0.94	0.92 / 0.92	**0.94** / 0.96
GMM-Sourced FM	0.72 / 0.96	0.92 / 1.00	0.98 / 1.00	0.42 / 1.00	0.58 / 1.00	0.56 / 1.00	**0.94** / 0.94	0.80 / 0.94
MOSS	**0.99** / 0.99	**1.00** / 1.00	**1.00** / 1.00	**1.00** / 1.00	**0.98** / 0.98	**0.96** / 0.96	0.92 / 0.94	0.86 / 0.94

5 Evaluation

5.1 Experimental Setup

We evaluate MOSS on 6 simulated tasks and 3 real-world robotic tasks, as shown in Figure 3. The simulation benchmark consists of 4 bimodal tasks—*Push Cube*, *Grasp Eyeglasses*, *Stack Cube*, and *Put Wine*—and two multi-mode tasks, *2D Navigation* and *Place Cube* (Jia et al., 2026a; Mu et al., 2021; Liu et al., 2023). Each task admits multiple distinct yet valid strategies for successful completion. For real-world evaluation, we consider *Place Cups*, *Pick Veg*, and *Open Drawer* on a Galaxea R1 Lite robotic platform.

We compare MOSS against 8 baselines. Unconditional generative policies include FM-DiT (Lipman et al., 2023; Peebles & Xie, 2023), DDPM-DiT (Ho et al., 2020; Peebles & Xie, 2023), ACT (Zhao et al., 2023), BET (Shafiullah et al., 2022), A2A (Jia et al., 2026b), and MARS Policy (Jia et al., 2026a). We further compare against *Mode-conditioned FM*, following mode-conditioned flow-matching formulations (Lei et al., 2026), which provides the desired behavioral mode as an additional conditioning signal to the velocity field, and *GMM-Sourced FM*, following multimodal source-space formulations (Luo et al., 2026; Zhang et al., 2026), which uses data-driven *Gaussian* mixture components without explicit semantic mode supervision. Key hyperparameters are standardized across models to ensure a fair comparison. Additional experimental settings and evaluation details are provided in Appendix E.

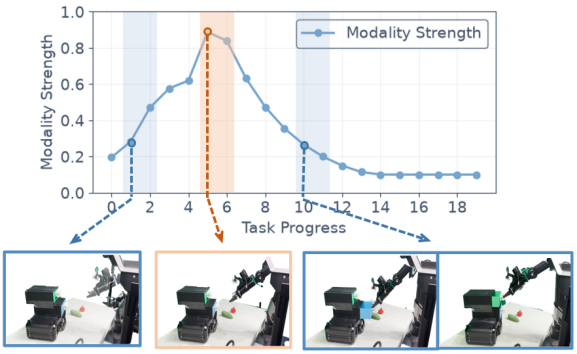

Figure 4: **Modality strength over task progress.**

Our evaluation targets three questions: (i) can MOSS steer the policy to a specified mode as expected; (ii) can it recover a minority mode under strongly imbalanced training data; and (iii) does it also bring other benefits, like improving task success and training efficiency?

5.2 Controllable Multimodal Behavior

Simulation results. Table 1 summarizes controllable behavior generation on the four bimodal simulation tasks. MOSS consistently achieves high SR and GR for both requested modes, attaining the strongest average mode controllability while maintaining high task success. Mode-conditioned

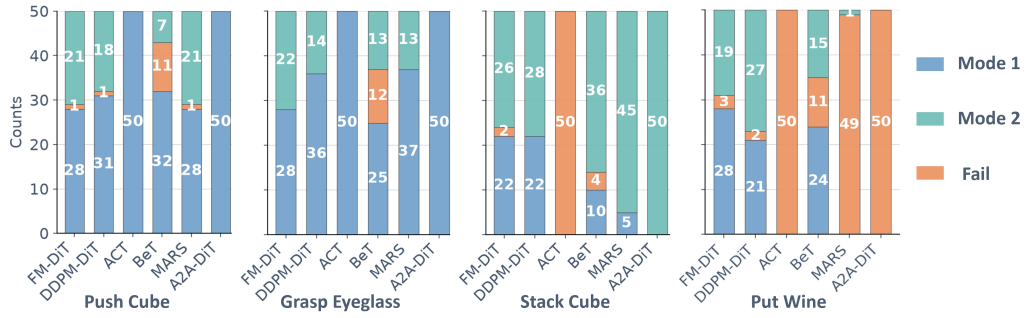

Figure 5: **Behavior distributions of unconditional policies on simulation tasks.**

FM is also competitive in simulation, whereas GMM-Sourced FM exhibits substantially more task-dependent semantic alignment. As shown in Figure 5, unconditional policies can represent multiple behavioral outcomes but provide no mechanism for selecting a desired mode, often resulting in task-dependent mode preference.

MOSS also generalizes beyond binary behaviors. As shown in Figure 6, MOSS maintains high SR and GR across multiple requested modes on *2D Navigation* and *Place Cube*. Figure 4 shows that the estimated modality strength remains low when different behaviors share similar motion and peaks near states where future trajectories diverge. Since the source separation in MOSS increases monotonically with q, these peaks trigger stronger mode-specific source separation around decision-critical stages.

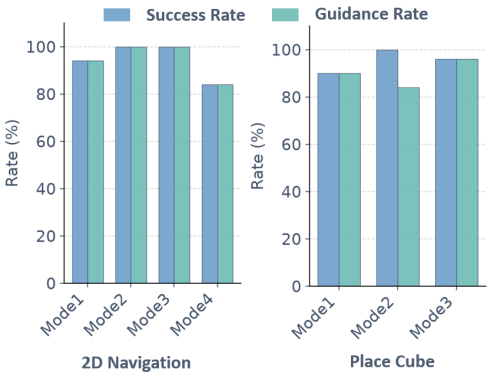

Figure 6: **Results on *2D Navigation* and *Place Cube*.**

Real-world results. As shown in Table 2, MOSS achieves strong SR and GR on all three real-world tasks and is substantially more balanced across behavioral modes than Mode-conditioned FM. For example, Mode-conditioned FM fails to control mode 1 on *Pick Veg* and mode 2 on *Open Drawer*, whereas MOSS retains controllability for both modes. These results suggest that source-space control remains effective despite the visual and execution variability encountered in real-world manipulation. Increasing classifier-free guidance (CFG) (Ho & Salimans, 2022) can partially improve Mode-conditioned FM, but does not consistently recover both SR and GR. Details are provided in Appendix F.

5.3 Robustness and Training Efficiency

To evaluate learning under biased demonstrations, we construct 10:90 mode-imbalanced datasets on *Push Cube*, *Grasp Eyeglasses*, and *Stack Cube*, and always evaluate the minority mode. MOSS and Mode-conditioned FM are trained on all 100 demonstrations, whereas FM-DiT is trained only on the 10 minority-mode demonstrations. Thus, all methods observe the same number of target-mode examples, while the former two can additionally exploit motion shared with the majority mode.

As shown in Figure 7, MOSS exhibits strong minority-mode sample efficiency, reaching high GR with relatively few target-mode sample exposures. The advantage is particularly pronounced on *Push Cube*, where MOSS approaches its final performance substantially earlier than both baselines. Compared with FM-DiT trained only on minority demonstrations, these results suggest that the

Table 2: **Real-world results.** We report average SR and mode-specific GR_1 and GR_2. Best GR for each requested mode is shown in bold.

Method	SR	GR_1	GR_2
Place Cups			
FM-DiT	0.70	–	–
Mode-conditioned FM	0.70	0.15	0.85
MOSS	0.98	**0.95**	**1.00**
Pick Veg			
FM-DiT	0.60	–	–
Mode-conditioned FM	0.65	0.00	0.60
MOSS	0.75	**0.80**	**0.65**
Open Drawer			
FM-DiT	0.75	–	–
Mode-conditioned FM	0.85	0.85	0.00
MOSS	0.83	**0.90**	**0.75**

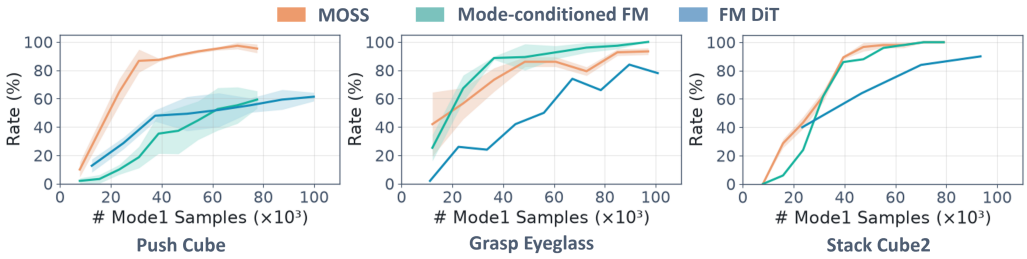

Figure 7: **Minority-mode learning efficiency under imbalanced datasets.** Further details are provided in Appendix G.

Table 3: **Ablation on 3 simulation tasks.** We report mode-specific guidance rates (GR_1 and GR_2). Best GR for each requested mode is shown in bold.

Method	Push Cube	Grasp Eyeglasses	Stack Cube
MOSS (single *Gaussian*)	0.50 / 0.68	0.76 / 0.80	0.58 / 0.70
MOSS (single-step separation)	0.64 / 0.62	0.66 / 0.72	0.54 / 0.68
MOSS (w/o adaptive)	0.86 / 0.85	0.86 / 0.90	0.78 / 0.88
MOSS (ours)	**0.99 / 1.00**	**1.00 / 1.00**	**0.98 / 0.96**

shared velocity field can reuse motion structure learned from the majority mode rather than learning the minority behavior entirely from scratch. Overall, MOSS provides robust minority-mode control while retaining high training efficiency under strongly imbalanced demonstrations. Implementation details are provided in Appendix G.

5.4 Ablation Studies

Table 3 isolates three design choices in MOSS. Specifically, the variants test whether mode-specific Gaussian sources are necessary, whether separation at only the most multimodal action position is sufficient, and whether adaptive shaping improves over fixed separation. Using a single shared Gaussian yields poor controllability, while single-step separation provides limited improvement; extending separation to the full action chunk substantially increases GR. Adaptive source shaping further improves over fixed full-chunk separation and typically reaches strong performance earlier during training. Additional analyses are provided in Appendix H.

6 Conclusion

We presented MOSS, a source-space approach for controllable multimodal FM policies. Our analysis shows that ideal deterministic FM transport naturally induces mode-dependent noise regions, while practical approximation errors can blur these implicit partitions and lead to mode preference or collapse. MOSS makes this structure explicit by associating semantic behavioral modes with mode-specific source distributions and adaptively adjusting their separation according to local multimodality, without conditioning the shared velocity field on the requested mode. Across simulated and real-world manipulation tasks, MOSS achieves strong mode controllability and task success rate, extends naturally to multiple behavioral modes, and improves minority-mode learning under highly imbalanced demonstrations. Together, these results highlight source-space geometry as an effective design dimension for controllable generative robot policies.

7 Limitations

MOSS has two main limitations. First, the current formulation assumes a fixed set of discrete behavioral modes specified before training, and therefore does not directly support open-ended mode control or continuous behavioral variation. Extending MOSS toward dynamic mode discovery and continuous mode spaces remains an important direction for future work. Second, although the source-space formulation is in principle compatible with larger generative robot policies, its performance in such large-scale settings remains to be evaluated.

AI USE STATEMENT

AI tools were used to aid the writing and polishing of this manuscript, primarily for language refinement, grammar correction, improving readability, and rephrasing existing text. Scientific ideas, methodological choices, experiments, technical claims, results, and conclusions were developed and validated by the authors. AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our results. Complete assumptions and proofs for the theoretical results are given in Appendix B. Details of the modality strength estimator are provided in Appendix C. Appendix D specifies the source-shaping parameterization, hyperparameters, and the complete training and inference procedure. Task definitions, baseline implementations, training configurations, evaluation protocols, and metric definitions are documented in Appendix E, with additional controlled analyses provided in Appendices F, G, H.

ETHICS STATEMENT

This work collects robot demonstration data through human-operated robot experiments in controlled laboratory environments. No personally identifiable information is collected, and the data are used solely for research purposes. We acknowledge that learned robotic policies may behave unpredictably under unseen conditions and should be deployed with appropriate safety validation and human oversight.

REFERENCES

Jiaqi Bai, Jindou Jia, Yuxuan Hu, Gen Li, Xiangyu Chen, Tuo An, Kuangji Zuo, and Jianfei Yang. Flash: efficient visuomotor policy via sparse sampling. *arXiv preprint arXiv:2605.15492*, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhejia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023.

Yu Fang, Yuchun Feng, Dong Jing, Jiaqi Liu, Yue Yang, Zhenyu Wei, Daniel Szafir, and Mingyu Ding. When vision overrides language: evaluating and mitigating counterfactual failures in VLAs. *arXiv preprint arXiv:2602.17659*, 2026.

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. In *Proceedings of International Conference on Learning Representations*, 2026.

Zhanpeng He, Yifeng Cao, and Matei Ciocarlie. Uncertainty comes for free: human-in-the-loop policies with diffusion models. *arXiv preprint arXiv:2503.01876*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jindou Jia, Tuo An, Yuxuan Hu, Gen Li, Jingliang Li, Bohan Hou, Xiangyu Chen, Jiaqi Bai, Bofan Lyu, and Jianfei Yang. MARS policy: multimodality only when it matters. *arXiv preprint arXiv:2605.29766*, 2026a.

Jindou Jia, Gen Li, Xiangyu Chen, Tuo An, Yuxuan Hu, Jingliang Li, Xinying Guo, and Jianfei Yang. Action-to-action flow matching. In *Proceedings of Robotics: Science and Systems*, 2026b.

Xiaohan Lei, Min Wang, Wengang Zhou, Xingyu Lu, and Houqiang Li. Primary-fine decoupling for action generation in robotic imitation. In *Proceedings of International Conference on Learning Representations*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of International Conference on Learning Representations*, 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 *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *Proceedings of International Conference on Learning Representations*, 2025.

Gaoxiang Luo, Frank Cole, Sihang Zhang, Yuxiang Wan, Yulong Lu, and Ju Sun. Flow matching for multimodal distributions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23260–23271, 2026.

Bruno Machado, Alexandre Chapin, Emmanuel Dellandrea, and Liming Chen. Flowing with purpose: latent action guided flow matching policies for robotic manipulation. *arXiv preprint arXiv:2606.23420*, 2026.

Lorenzo Mazza, Massimiliano Datres, Ariel Rodriguez, Sebastian Bodenstedt, Gitta Kutyniok, and Stefanie Speidel. Understanding multimodal failure in action-chunking behavioral cloning. *arXiv preprint arXiv:2605.22493*, 2026.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021.

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *Proceedings of International Conference on Learning Representations*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

Hanxiang Ren, Pei Zhou, Xunzhe Zhou, and Yanchao Yang. DISC: decoupling instruction from state-conditioned control via policy generation. In *Proceedings of Robotics: Science and Systems*, 2026.

Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: cloning k modes with one stone. In *Proceedings of Advances in Neural Information Processing Systems*, volume 35, pp. 22955–22968, 2022.

Shuai Yang, Hao Li, Bin Wang, Yilun Chen, Yang Tian, Tai Wang, Hanqing Wang, Feng Zhao, Yiyi Liao, and Jiangmiao Pang. Vision-language-action instruction tuning: from understanding to manipulation. In *Proceedings of International Conference on Learning Representations*, 2026.

Xuehui Yu, Fucheng Cai, Meiyi Wang, Xiaopeng Fan, and Harold Soh. Conflict-aware additive guidance for flow models under compositional rewards. In *Proceedings of International Conference on Machine Learning*, 2026.

Xuanran Zhai, Qianyou Zhao, Qiaojun Yu, and Ce Hao. VFP: variational flow-matching policy for multi-modal robot manipulation. *arXiv preprint arXiv:2508.01622*, 2025.

Fan Zhang and Michael Gienger. Affordance-based robot manipulation with flow matching. 2024.

He Zhang, Ying Sun, Pengteng Li, Ziyang Chen, Yiren Zhao, Ziyang Rao, Weiyu Guo, Yandong Guo, and Hui Xiong. Source-lifted flow matching for intervenable multimodal imitation. *arXiv preprint arXiv:2607.10206*, 2026.

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

Moritz Zoellner, Anastasios Manganaris, and Rohan Paleja. Temporal logic guidance for action-only diffusion policies with world models. *arXiv preprint arXiv:2606.22729*, 2026.

A VISUALIZATION OF INITIAL-NOISE MODE SEPARATION

We investigate whether the initial noise of a standard FM policy is predictive of the behavioral mode ultimately executed by the policy. We conduct the experiment on *2D Navigation*, where a DiT-based FM policy can reach the goal through four feasible routes, denoted as Mode 1 – Mode 4.

We perform 1,000 closed-loop rollouts using the standard Gaussian source distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. For each rollout, we record the initial noise of the first executed action chunk, which contains 8 two-dimensional actions. Since route selection is primarily reflected in vertical motion, we retain only the y-dimension, resulting in an eight-dimensional noise feature for each rollout.

A rollout is considered successful if its final position is within 0.06 of the goal at $(1.8, 0.5)$. Successful trajectories are assigned mode labels according to their mean vertical coordinate in the wall-crossing region $x \in [0.78, 0.85]$:

$$\begin{aligned} y < 0.22 &: \quad \text{Mode 1}, \\ 0.22 \le y < 0.50 &: \quad \text{Mode 2}, \\ 0.50 \le y < 0.78 &: \quad \text{Mode 3}, \\ y \ge 0.78 &: \quad \text{Mode 4}. \end{aligned}$$

After excluding failed rollouts, 939 trajectories remain, with 188, 242, 267, and 242 samples for Mode 1 – Mode 4, respectively.

For visualization, the eight-dimensional noise features are standardized and projected to two dimensions using LDA fitted with the trajectory-derived mode labels. LDA identifies directions that maximize between-mode relative to within-mode variation. Each point in Figure 2 represents the initial noise of one successful rollout, colored by its resulting behavioral mode, while the ellipses show the nominal 95% Gaussian covariance contours of each mode.

We further quantify the semantic structure of this source-space partition. For each mode, we compute the mean of its initial-noise features. The mean y-noise values for Mode 1 – Mode 4 are -0.303, -0.182, 0.135, and 0.286, respectively, increasing monotonically with the vertical location of the corresponding route. This suggests that the source-space organization learned by standard FM is aligned with the geometry of the resulting behaviors.

Together, these results show that a standard FM policy already develops a mode-relevant organization of its initial-noise space, despite being trained without explicit mode conditioning. However, the substantial overlap between neighboring modes also shows that this naturally emerging partition is insufficient for reliable mode control. This observation motivates MOSS to explicitly shape and separate mode-specific source regions for controllable multimodal generation.

To test whether this structure can directly influence generation, we construct a Gaussian source centered at each mode-specific mean while retaining the standard deviation of the original source, $\sigma = 1$, and perform 100 rollouts per source. The resulting mode distributions are reported in Table 4. Sampling around different mode-specific means induces distinct biases in the generated behaviors. In particular, the sources centered at the Mode 1 and Mode 4 means produce their corresponding modes in 62% and 56% of rollouts, respectively. The intermediate means yield more overlapping mode distributions, indicating that the naturally emerging source partition is semantically structured but not cleanly separated.

Table 4: **Results when sampling from Gaussian sources centered at the mode-specific initial-noise means of a standard FM policy.** Each row contains 100 rollouts with $\sigma = 1$.

Target mean	Mode 1	Mode 2	Mode 3	Mode 4	Stuck	Target mode
-0.303 (Mode 1)	**62**	27	8	2	1	62%
-0.182 (Mode 2)	52	**29**	9	6	4	29%
0.135 (Mode 3)	6	15	**34**	43	2	34%
0.286 (Mode 4)	3	8	29	**56**	4	56%

B PROOFS FOR SECTION 3

B.1 PROOF OF THEOREM 1

Proof. Since $\mathcal{A}_k^{\mathbf{o}} \subset \mathbb{R}^d$ is open and $\Phi_{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^d$ is continuous, its preimage

$$\mathcal{Z}_k^{\mathbf{o}} = \Phi_{\mathbf{o}}^{-1}\left(\mathcal{A}_k^{\mathbf{o}}\right) \tag{13}$$

is open.

For any $i \neq j$, the semantic regions are pairwise disjoint. Hence

$$\begin{aligned} \mathcal{Z}_i^{\mathbf{o}} \cap \mathcal{Z}_j^{\mathbf{o}} &= \Phi_{\mathbf{o}}^{-1}\left(\mathcal{A}_i^{\mathbf{o}}\right) \cap \Phi_{\mathbf{o}}^{-1}\left(\mathcal{A}_j^{\mathbf{o}}\right) \\ &= \Phi_{\mathbf{o}}^{-1}\left(\mathcal{A}_i^{\mathbf{o}} \cap \mathcal{A}_j^{\mathbf{o}}\right) = \varnothing, \end{aligned} \tag{14}$$

so the corresponding noise zones are also pairwise disjoint.

By the exact transport condition $(\Phi_{\mathbf{o}})_{\#}\mu_0 = p_1^{\mathbf{o}}$, the definition of pushforward measure gives, for every measurable $S \subseteq \mathbb{R}^d$,

$$\mu_0\left(\Phi_{\mathbf{o}}^{-1}(S)\right) = p_1^{\mathbf{o}}(S). \tag{15}$$

Taking $S = \mathcal{A}_k^{\mathbf{o}}$ yields

$$\mu_0\left(\mathcal{Z}_k^{\mathbf{o}}\right) = p_1^{\mathbf{o}}\left(\mathcal{A}_k^{\mathbf{o}}\right). \tag{16}$$

Similarly, taking $S = \mathcal{B}_A^{\mathbf{o}}$ and using $p_1^{\mathbf{o}}(\mathcal{B}_A^{\mathbf{o}}) = 0$ gives

$$\mu_0\left(\mathcal{B}_Z^{\mathbf{o}}\right) = p_1^{\mathbf{o}}\left(\mathcal{B}_A^{\mathbf{o}}\right) = 0. \tag{17}$$

This proves Theorem 1. □

B.2 FLOW REGULARITY AND PROOF OF THEOREM 2

We first justify the bi-Lipschitz property used in Theorem 2. Let $\Phi_{s,t}^{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^d$ denote the FM ODE flow from time s to t, with $\Phi_{\mathbf{o}} = \Phi_{0,1}^{\mathbf{o}}$. Define

$$L_t^{\mathbf{o}} := \sup_{\mathbf{x} \in \mathbb{R}^d} \left\|\nabla_{\mathbf{x}} \mathbf{v}_t^{\mathbf{o}}(\mathbf{x})\right\|_{\mathrm{op}}, \qquad \Lambda_{\mathbf{o}} := \int_0^1 L_t^{\mathbf{o}}\, \mathrm{d}t < \infty, \tag{18}$$

and assume that the ODE is globally well posed on $[0, 1]$.

For $\mathbf{x}_t = \Phi_{0,t}^{\mathbf{o}}(\mathbf{z})$ and $\mathbf{x}_t' = \Phi_{0,t}^{\mathbf{o}}(\mathbf{z}')$, we have

$$\left\|\mathbf{v}_t^{\mathbf{o}}(\mathbf{x}_t) - \mathbf{v}_t^{\mathbf{o}}(\mathbf{x}_t')\right\|_2 \leq L_t^{\mathbf{o}} \left\|\mathbf{x}_t - \mathbf{x}_t'\right\|_2. \tag{19}$$

Grönwall's inequality therefore gives

$$\left\|\Phi_{\mathbf{o}}(\mathbf{z}) - \Phi_{\mathbf{o}}(\mathbf{z}')\right\|_2 \leq e^{\Lambda_{\mathbf{o}}} \left\|\mathbf{z} - \mathbf{z}'\right\|_2. \tag{20}$$

Applying the same argument to the backward flow $\Phi_{1,0}^{\mathbf{o}} = \Phi_{\mathbf{o}}^{-1}$ yields

$$e^{-\Lambda_{\mathbf{o}}} \left\|\mathbf{z} - \mathbf{z}'\right\|_2 \leq \left\|\Phi_{\mathbf{o}}(\mathbf{z}) - \Phi_{\mathbf{o}}(\mathbf{z}')\right\|_2 \leq e^{\Lambda_{\mathbf{o}}} \left\|\mathbf{z} - \mathbf{z}'\right\|_2, \tag{21}$$

which establishes the bi-Lipschitz property in equation 3.

Condition $\sigma_{\min} > 0$. For the probability path in equation 1, define

$$\mathbf{m}_t^{\mathbf{o}}(\mathbf{x}) := \mathbb{E}\left[\mathbf{y} \mid \mathbf{x}_t = \mathbf{x}, \mathbf{o}\right], \tag{22}$$

$$\boldsymbol{\Sigma}_t^{\mathbf{o}}(\mathbf{x}) := \mathrm{Cov}\left[\mathbf{y} \mid \mathbf{x}_t = \mathbf{x}, \mathbf{o}\right]. \tag{23}$$

Since $\dot{\sigma}_t = \sigma_{\min} - 1$ and $\sigma_t - t\dot{\sigma}_t = 1$, the population FM velocity satisfies

$$\mathbf{v}_t^{\mathbf{o}}(\mathbf{x}) = \frac{\dot{\sigma}_t}{\sigma_t}\mathbf{x} + \frac{1}{\sigma_t}\mathbf{m}_t^{\mathbf{o}}(\mathbf{x}), \tag{24}$$

with

$$\nabla_{\mathbf{x}} \mathbf{m}_t^{\mathbf{o}}(\mathbf{x}) = \frac{t}{\sigma_t^2}\boldsymbol{\Sigma}_t^{\mathbf{o}}(\mathbf{x}), \tag{25}$$

and hence

$$\nabla_{\mathbf{x}} \mathbf{v}_t^{\mathbf{o}}(\mathbf{x}) = \frac{\dot{\sigma}_t}{\sigma_t} \mathbf{I}_d + \frac{t}{\sigma_t^3} \mathbf{\Sigma}_t^{\mathbf{o}}(\mathbf{x}). \tag{26}$$

For $\sigma_{\min} > 0$, let $\sigma_* := \min\{1, \sigma_{\min}\} > 0$. If

$$C_{\Sigma,\mathbf{o}} := \sup_{\substack{t \in [0,1] \\ \mathbf{x} \in \mathbb{R}^d}} \left\| \mathbf{\Sigma}_t^{\mathbf{o}}(\mathbf{x}) \right\|_{\mathrm{op}} < \infty, \tag{27}$$

then

$$L_t^{\mathbf{o}} \leq \frac{|\sigma_{\min} - 1|}{\sigma_*} + \frac{t C_{\Sigma,\mathbf{o}}}{\sigma_*^3}, \tag{28}$$

and therefore

$$\Lambda_{\mathbf{o}} \leq \frac{|\sigma_{\min} - 1|}{\sigma_*} + \frac{C_{\Sigma,\mathbf{o}}}{2\sigma_*^3} < \infty. \tag{29}$$

Thus, together with global well-posedness, bounded posterior covariance is sufficient for the global spatial regularity used above.

Connectedness of a noise zone. Since $\Phi_{\mathbf{o}}$ is a homeomorphism, $\Phi_{\mathbf{o}}^{-1}$ is continuous. If $\mathcal{A}_k^{\mathbf{o}}$ is connected, then

$$\mathcal{Z}_k^{\mathbf{o}} = \Phi_{\mathbf{o}}^{-1}\left(\mathcal{A}_k^{\mathbf{o}}\right) \tag{30}$$

is the continuous image of $\mathcal{A}_k^{\mathbf{o}}$ under $\Phi_{\mathbf{o}}^{-1}$ and is therefore connected. Its openness follows from continuity of $\Phi_{\mathbf{o}}$.

Same-mode neighborhood. Take $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}}$ and $\mathbf{u} \in \mathbb{R}^d$ satisfying

$$\|\mathbf{u} - \mathbf{z}\|_2 < R_{\mathbf{o}}(\mathbf{z}) = e^{-\Lambda_{\mathbf{o}}} \gamma_{\mathbf{o}}(\mathbf{z}). \tag{31}$$

By equation 3,

$$\|\Phi_{\mathbf{o}}(\mathbf{u}) - \Phi_{\mathbf{o}}(\mathbf{z})\|_2 < \gamma_{\mathbf{o}}(\mathbf{z}). \tag{32}$$

Since $\gamma_{\mathbf{o}}(\mathbf{z})$ is the distance from $\Phi_{\mathbf{o}}(\mathbf{z})$ to the semantic boundary $\mathcal{B}_A^{\mathbf{o}}$, $\Phi_{\mathbf{o}}(\mathbf{u})$ remains in $\mathcal{A}_k^{\mathbf{o}}$. Therefore,

$$B\left(\mathbf{z}, R_{\mathbf{o}}(\mathbf{z})\right) \subseteq \mathcal{Z}_k^{\mathbf{o}}. \tag{33}$$

Cross-mode separation. Recall that

$$\mathbb{R}^d = \mathcal{B}_A^{\mathbf{o}} \cup \bigcup_{k=1}^{K} \mathcal{A}_k^{\mathbf{o}}, \qquad \mathcal{B}_A^{\mathbf{o}} \cap \mathcal{A}_k^{\mathbf{o}} = \varnothing, \tag{34}$$

with the $\mathcal{A}_k^{\mathbf{o}}$ pairwise disjoint. Hence any continuous curve joining two different semantic regions must intersect $\mathcal{B}_A^{\mathbf{o}}$.

Let $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}}$ and $\mathbf{z}' \in \mathcal{Z}_j^{\mathbf{o}}$ with $k \neq j$, and define

$$\mathbf{x} = \Phi_{\mathbf{o}}(\mathbf{z}), \qquad \mathbf{x}' = \Phi_{\mathbf{o}}(\mathbf{z}'). \tag{35}$$

The line segment joining $\mathbf{x}$ and $\mathbf{x}'$ intersects $\mathcal{B}_A^{\mathbf{o}}$ at some $\mathbf{b}$. Therefore,

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}'\|_2 &= \|\mathbf{x} - \mathbf{b}\|_2 + \|\mathbf{x}' - \mathbf{b}\|_2 \\ &\geq \gamma_{\mathbf{o}}(\mathbf{z}) + \gamma_{\mathbf{o}}(\mathbf{z}'). \end{aligned} \tag{36}$$

Combining this with equation 3 gives

$$\begin{aligned} \|\mathbf{z} - \mathbf{z}'\|_2 &\geq e^{-\Lambda_{\mathbf{o}}} \left[\gamma_{\mathbf{o}}(\mathbf{z}) + \gamma_{\mathbf{o}}(\mathbf{z}')\right] \\ &= R_{\mathbf{o}}(\mathbf{z}) + R_{\mathbf{o}}(\mathbf{z}'). \end{aligned} \tag{37}$$

Together with the same-mode neighborhood result, this proves Theorem 2.

B.3 Proof of Proposition 3 and Corollary 4

Proof of Proposition 3. Let $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}}$ and define

$$\mathbf{x} := \Phi_{\mathbf{o}}(\mathbf{z}), \qquad \widetilde{\mathbf{x}} := \widetilde{\Phi}_{\mathbf{o}}(\mathbf{z}), \tag{38}$$

where $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbb{R}^d$. If

$$\eta_{\mathbf{o}}(\mathbf{z}) < \gamma_{\mathbf{o}}(\mathbf{z}), \tag{39}$$

then

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2 < \operatorname{dist}(\mathbf{x}, \mathcal{B}_A^{\mathbf{o}}). \tag{40}$$

Therefore, $\widetilde{\mathbf{x}}$ lies strictly inside the largest boundary-free ball centered at $\mathbf{x}$. Since $\mathbf{x} \in \mathcal{A}_k^{\mathbf{o}}$, it follows that $\widetilde{\mathbf{x}} \in \mathcal{A}_k^{\mathbf{o}}$. Hence

$$h_{\mathbf{o}}\left(\widetilde{\Phi}_{\mathbf{o}}(\mathbf{z})\right) = h_{\mathbf{o}}\left(\Phi_{\mathbf{o}}(\mathbf{z})\right) = k. \tag{41}$$

□

To formalize the boundary-sensitive region, let $\epsilon \in \mathbb{R}_{\geq 0}$ be a uniform endpoint-error bound and define

$$\mathcal{S}_\epsilon^{\mathbf{o}} := \left\{\mathbf{z} \in \mathbb{R}^d : \gamma_{\mathbf{o}}(\mathbf{z}) \leq \epsilon\right\}. \tag{42}$$

Suppose

$$\eta_{\mathbf{o}}(\mathbf{z}) \leq \epsilon. \tag{43}$$

For every $\mathbf{z} \notin \mathcal{S}_\epsilon^{\mathbf{o}}$, we have

$$\eta_{\mathbf{o}}(\mathbf{z}) \leq \epsilon < \gamma_{\mathbf{o}}(\mathbf{z}), \tag{44}$$

so Proposition 3 guarantees preservation of the behavioral mode.

Moreover, for any $\epsilon_1, \epsilon_2 \in \mathbb{R}_{\geq 0}$ with $\epsilon_1 \leq \epsilon_2$,

$$\mathcal{S}_{\epsilon_1}^{\mathbf{o}} \subseteq \mathcal{S}_{\epsilon_2}^{\mathbf{o}}. \tag{45}$$

Thus, as the endpoint-error bound increases, the set of source points for which mode preservation is not guaranteed expands monotonically. This establishes Corollary 4.

B.4 Local Guarantee for $\sigma_{\min} = 0$

When $\sigma_{\min} = 0$, we have $\sigma_t = 1 - t$, so the global sufficient bound in equation 28 is no longer available as $t \to 1$. We therefore do not assume a global bi-Lipschitz bound in this case. Instead, the geometric argument in Theorem 2 admits a local version on any compact source region for which the velocity Jacobian remains integrable along the corresponding trajectories.

Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact set with nonempty interior. For each $t \in [0, 1]$, define

$$\mathcal{C}_t^{\mathbf{o}}(\mathcal{K}) := \operatorname{conv}\left(\Phi_{0,t}^{\mathbf{o}}(\mathcal{K})\right) \subseteq \mathbb{R}^d, \tag{46}$$

where $\operatorname{conv}(\cdot)$ denotes the convex hull. Define the scalar

$$L_{t,\mathcal{K}}^{\mathbf{o}} := \sup_{\mathbf{x} \in \mathcal{C}_t^{\mathbf{o}}(\mathcal{K})} \|\nabla_{\mathbf{x}} \mathbf{v}_t^{\mathbf{o}}(\mathbf{x})\|_{\mathrm{op}} \in \mathbb{R}_{\geq 0}. \tag{47}$$

Assume that the ODE solution exists uniquely for every $\mathbf{z} \in \mathcal{K}$ over $t \in [0, 1]$ and that

$$\Lambda_{\mathbf{o},\mathcal{K}} := \int_0^1 L_{t,\mathcal{K}}^{\mathbf{o}} \, \mathrm{d}t < \infty. \tag{48}$$

For any $\mathbf{z}, \mathbf{z}' \in \mathcal{K}$, the line segment between $\Phi_{0,t}^{\mathbf{o}}(\mathbf{z})$ and $\Phi_{0,t}^{\mathbf{o}}(\mathbf{z}')$ is contained in $\mathcal{C}_t^{\mathbf{o}}(\mathcal{K})$. The mean-value inequality therefore gives

$$\begin{aligned} &\left\|\mathbf{v}_t^{\mathbf{o}}\left(\Phi_{0,t}^{\mathbf{o}}(\mathbf{z})\right) - \mathbf{v}_t^{\mathbf{o}}\left(\Phi_{0,t}^{\mathbf{o}}(\mathbf{z}')\right)\right\|_2 \\ &\quad \leq L_{t,\mathcal{K}}^{\mathbf{o}} \left\|\Phi_{0,t}^{\mathbf{o}}(\mathbf{z}) - \Phi_{0,t}^{\mathbf{o}}(\mathbf{z}')\right\|_2. \end{aligned} \tag{49}$$

Applying Grönwall's inequality forward and backward along these trajectories yields

$$e^{-\Lambda_{\mathbf{o},\mathcal{K}}} \|\mathbf{z} - \mathbf{z}'\|_2 \leq \|\Phi_{\mathbf{o}}(\mathbf{z}) - \Phi_{\mathbf{o}}(\mathbf{z}')\|_2 \leq e^{\Lambda_{\mathbf{o},\mathcal{K}}} \|\mathbf{z} - \mathbf{z}'\|_2. \tag{50}$$

Hence $\Phi_{\mathbf{o}}$ is bi-Lipschitz when restricted to $\mathcal{K}$.

For $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}} \cap \mathrm{int}(\mathcal{K})$, define the scalar local radius

$$R_{\mathbf{o},\mathcal{K}}(\mathbf{z}) := \min\left\{e^{-\Lambda_{\mathbf{o},\mathcal{K}}}\gamma_{\mathbf{o}}(\mathbf{z}),\ \mathrm{dist}\left(\mathbf{z}, \mathbb{R}^d \setminus \mathcal{K}\right)\right\} \in \mathbb{R}_{\geq 0}. \tag{51}$$

For every $\mathbf{u} \in B(\mathbf{z}, R_{\mathbf{o},\mathcal{K}}(\mathbf{z}))$, we have $\mathbf{u} \in \mathcal{K}$ and, by equation 50,

$$\|\Phi_{\mathbf{o}}(\mathbf{u}) - \Phi_{\mathbf{o}}(\mathbf{z})\|_2 < \gamma_{\mathbf{o}}(\mathbf{z}). \tag{52}$$

Therefore,

$$B\left(\mathbf{z}, R_{\mathbf{o},\mathcal{K}}(\mathbf{z})\right) \subseteq \mathcal{Z}_k^{\mathbf{o}}. \tag{53}$$

Similarly, for $\mathbf{z} \in \mathcal{Z}_k^{\mathbf{o}} \cap \mathcal{K}$ and $\mathbf{z}' \in \mathcal{Z}_j^{\mathbf{o}} \cap \mathcal{K}$ with $k \neq j$, the endpoint-margin argument in equation 36, together with equation 50, gives

$$\|\mathbf{z} - \mathbf{z}'\|_2 \geq e^{-\Lambda_{\mathbf{o},\mathcal{K}}}\left[\gamma_{\mathbf{o}}(\mathbf{z}) + \gamma_{\mathbf{o}}(\mathbf{z}')\right]. \tag{54}$$

Thus, although a global regularity guarantee is not assumed when $\sigma_{\min} = 0$, the same same-mode neighborhood and cross-mode separation arguments remain valid locally wherever the FM velocity field has an integrable spatial Jacobian along the relevant trajectories.

C Additional Details of the Modality Strength Estimator

C.1 Architecture and Hyperparameters

The estimator is a lightweight two-layer MLP with hidden width 256, GELU activations, and dropout 0.1, followed by a scalar sigmoid output. Its input consists of an eight-step observation history and an eight-step executed-action history. A LayerNorm operation is applied to the flattened input.

The estimator is optimized using a confidence-weighted Huber loss,

$$\mathcal{L}_{\mathrm{est}} = \frac{\sum_i c_i\, \ell_{\mathrm{Huber}}(w_i - r_i)}{\sum_i c_i + \epsilon}, \tag{55}$$

where c_i measures the reliability of the local neighborhood.

The principal hyperparameters shared across tasks are summarized in Table 5.

C.2 Neighborhood Construction and Supervision

Each observation is represented by a 512-dimensional frozen ResNet-18 feature concatenated with the task-specific normalized proprioceptive state. The flattened observation history is projected to 128 dimensions.

For each query, we first retrieve $K_{\mathrm{cand}} = 40$ candidates using the projected observation histories and then re-rank them using both observation and executed-action distances:

$$d_{ij} = \frac{d_{ij}^{\mathrm{obs}}}{\mathrm{median}(d_{i\cdot}^{\mathrm{obs}})} + \lambda_h \frac{d_{ij}^{\mathrm{act}}}{\mathrm{median}(d_{i\cdot}^{\mathrm{act}})}, \qquad \lambda_h = 1. \tag{56}$$

The $K_{\mathrm{NN}} = 20$ closest candidates are retained. The query itself and samples from the same episode within eight time steps are excluded to avoid trivial temporal neighbors. Each retained neighbor is assigned

$$\alpha_{ij} = \frac{\exp(-d_{ij}/\tau)}{\sum_{\ell \in \mathcal{N}_i} \exp(-d_{i\ell}/\tau)}, \qquad \tau = 1. \tag{57}$$

Let $\mathbf{y}_j$ denote the normalized future-action chunk of neighbor j. For each mode k, we compute its local probability and weighted future-action mean as

$$p_{ik} = \sum_{j \in \mathcal{N}_{ik}} \alpha_{ij}, \qquad \bar{\mathbf{y}}_{ik} = \frac{1}{p_{ik}} \sum_{j \in \mathcal{N}_{ik}} \alpha_{ij}\mathbf{y}_j. \tag{58}$$

Table 5: **Hyperparameters of the Modality Strength Estimator.**

Hyperparameter	Value
Observation/action history length	8 / 8
Future-action horizon	8
KNN projection dimension	128
Candidate/retained neighbors	40 / 20
MLP hidden dimensions	$256, 256$
Dropout	0.1
Optimizer and learning rate	AdamW, 3×10^{-4}
Estimator batch size	128
Calibration threshold ϵ_w	10^{-3}
Active-percentile saturation η	0.95
Persistence floor q_{floor}	0.05
Persistence factor λ_q	0.75

With $\bar{\mathbf{y}}_i = \sum_k p_{ik}\bar{\mathbf{y}}_{ik}$, the between-mode and within-mode variations are implemented as

$$\mathrm{Var}_{B,i} = \sum_k p_{ik} \left\|\bar{\mathbf{y}}_{ik} - \bar{\mathbf{y}}_i\right\|_{\mathrm{ms}}^2, \tag{59}$$

$$\mathrm{Var}_{W,i} = \sum_k \sum_{j\in\mathcal{N}_{ik}} \alpha_{ij} \left\|\mathbf{y}_j - \bar{\mathbf{y}}_{ik}\right\|_{\mathrm{ms}}^2, \tag{60}$$

The numerical stabilizer δ in the main-text supervision target is set to the 10th percentile of positive within-mode variations on the training split. The confidence weight c_i combines neighborhood quality, measured from the retrieval distance, with effective mode support, thereby suppressing unreliable targets from distant or poorly represented neighborhoods.

C.3 Calibration and Temporal Persistence

Because the numerical range of the estimator output w varies across tasks, we calibrate it using the empirical distribution of predictions on the training split. Let

$$\widehat{F}_+(w) = \frac{\sum_{i\in\mathcal{D}_{\text{train}}} \mathbb{1}[\epsilon_w < w_i \le w]}{\sum_{i\in\mathcal{D}_{\text{train}}} \mathbb{1}[w_i > \epsilon_w]} \tag{61}$$

denote the empirical CDF of active predictions. We use

$$q(w) = \begin{cases} 0, & w \le \epsilon_w, \\ \min\left(\dfrac{\widehat{F}_+(w)}{\eta}, 1\right), & w > \epsilon_w, \end{cases} \tag{62}$$

where $\epsilon_w = 10^{-3}$ and $\eta = 0.95$. Thus, inactive predictions remain zero, while the 95th percentile of active training predictions maps to $q = 1$. The calibration is computed solely from the training split and fixed thereafter.

During rollout, temporal persistence is applied as

$$q_t^{\text{eff}} = \max\left(q_t,\, q_{\text{floor}},\, \lambda_q q_{t-1}^{\text{eff}}\right), \tag{63}$$

where $q_{\text{floor}} = 0.05$ and $\lambda_q = 0.75$. This prevents a recently detected multimodal state from immediately losing source separation at subsequent policy queries. For independently sampled training windows, we use $q^{\text{eff}} = \max(q, q_{\text{floor}})$. For simplicity, we denote q^{eff} as q when there is no ambiguity.

D Additional Details of Mode-Oriented Source Shaping

D.1 Branch-Critical Region Selection.

The multimodality supervision r_i defined in Section 4.2 is also used offline to identify the training region where future behaviors are most clearly separated. Specifically, we rank training windows

by r_i and select the region with the highest multimodality. For each mode k, $\bar{\mathbf{a}}_k$ is computed as the empirical mean future-action chunk within this region. We then center these means according to equation 8 and construct $\mathbf{U}$ as an orthonormal basis of the resulting mode-difference vectors. All statistics are computed solely from the training demonstrations and are fixed before FM training and deployment.

D.2 Adaptive Separation Schedule

We first normalize the active range of the modality strength and apply a monotone gain

$$g(q) = \left[\mathrm{clip}\left(\frac{q - q_{\mathrm{floor}}}{1 - q_{\mathrm{floor}}}, 0, 1\right)\right]^{\gamma}, \qquad \gamma > 0. \tag{64}$$

The source-center scale in equation 10 is parameterized as

$$\alpha(q) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\, g(q), \tag{65}$$

and the target separation ratio as

$$\rho(q) = \rho_{\min} + (\rho_{\max} - \rho_{\min})\, g(q). \tag{66}$$

Thus, increasing q simultaneously moves the mode-specific source centers farther apart and increases their target separation relative to their spread. The corresponding $\sigma_{\parallel}(q)$ is then obtained from equation 12. The clipping bounds prevent the source distribution from becoming excessively concentrated or diffuse, while $\sigma_{\perp}$ remains fixed.

We use the same source-shaping hyperparameters across all tasks, as summarized in Table 6.

Table 6: **Source-shaping hyperparameters used in MOSS.**

$\alpha_{\min}$	$\alpha_{\max}$	$\rho_{\min}$	$\rho_{\max}$	$\sigma_{\perp}$	$\sigma_{\parallel,\min}$	$\sigma_{\parallel,\max}$	γ
0.0	1.0	1.25	2.5	1.0	0.25	1.0	0.5

D.3 Training and Inference Procedure

Algorithm 1 summarizes the complete procedure. The modality strength estimator and its calibration are fitted before FM training and remain frozen thereafter. During training, the source distribution is selected using the demonstration mode label; during deployment, it is selected using the requested mode.

E Detailed Experimental Setup

E.1 Tasks and Mode Definitions

We evaluate MOSS on 6 simulated tasks and 3 real-world tasks. The 4 bimodal simulation tasks are *Push Cube*, *Grasp Eyeglasses*, *Stack Cube*, and *Put Wine*. The 2 multi-mode tasks are *2D Navigation* and *Place Cube*. The real-world experiments comprise *Place Cups*, *Pick Veg*, and *Open Drawer*. Mode annotations identify distinct successful strategies rather than task success itself.

The mode semantics used in the three repeated simulation comparisons and the real-world experiments are summarized in Table 7.

E.2 Baselines

GMM-Sourced FM. We implement a Gaussian-mixture source baseline following the general multimodal source-space formulations explored in prior work (Luo et al., 2026; Zhang et al., 2026). The source distribution is modeled as a mixture of Gaussian components whose parameters are estimated from the training data, and samples from different components are transported using a shared flow-matching model. We adapt the source-space construction to our policy architecture and

Table 7: **Semantic definitions of the two behavioral modes used in the controlled evaluations.**

Task	Mode 1	Mode 2
Push Cube	Push around the left side	Push around the right side
Grasp Eyeglasses	Positive-y approach	Negative-y approach
Stack Cube	Cube A on Cube B	Cube B on Cube A
Place Cups	Grasp the left cup	Grasp the right cup
Pick Veg	Pick the left radish first	Pick the right corn first
Open Drawer	Open the lower blue drawer	Open the upper green drawer

Table 8: **Shared configuration of the FM-based policies.**

Hyperparameter	Value
Observation history	8 steps
Predicted action horizon	16 steps
Actions executed per query	8 steps
Dropout	0.1
FM integration steps	10
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-6}
Batch size	32
Maximum train / validation batches per epoch	250 / 250

experimental setting while keeping the remaining training configuration consistent with FM-DiT. For components without an available official implementation, we follow the formulation described in the corresponding paper and adapt it to our policy architecture.

Mode-conditioned FM. Our goal is to compare source-space mode control against direct mode conditioning under the same FM policy backbone. PF-DAG (Lei et al., 2026) contains a mode-conditioned action-generation stage, but its complete pipeline additionally predicts the primary mode, addressing the distinct problem of automatically inferring multimodal structure. Since our setting assumes that the desired semantic mode is explicitly specified at deployment, we isolate the mode-conditioned generation mechanism and directly provide the requested mode label to the FM policy. Specifically, the mode label is embedded and injected as an additional condition into the velocity network throughout transport. MOSS and Mode-conditioned FM use the same visual encoder, velocity-field architecture, action horizon, optimizer, training data, and evaluation protocol. The comparison is therefore intended to isolate where the given mode information enters the generative process—through the velocity field or through the source distribution—rather than to compare against the full PF-DAG system.

Unconditional Baselines. FM-DiT (Lipman et al., 2023; Peebles & Xie, 2023), DDPM-DiT (Ho et al., 2020; Peebles & Xie, 2023), ACT (Zhao et al., 2023), BET (Shafiullah et al., 2022), A2A (Jia et al., 2026b), and MARS Policy (Jia et al., 2026a) are reproduced following their original formulations and implementation settings.

E.3 Simulation Setup

For *Push Cube*, *Grasp Eyeglasses*, and *Stack Cube*, FM-DiT, Mode-conditioned FM, and MOSS use the same visual encoder, velocity-field architecture, prediction horizon, optimizer, and FM solver configuration. Their verified common settings are listed in Table 8. Training duration and checkpoint-selection schedules are kept identical across the compared methods for each task.

E.4 Real-World Robot Setup

The real-world experiments use a Galaxea R1 Lite robot. Each policy receives observations from a head camera and a right-wrist camera together with the seven-dimensional right-arm state, compris-

Algorithm 1 MOSS Training and Mode Intervention

Require: Labeled demonstrations $\mathcal{D} = \{(\mathbf{o}_{\le t_i}, \mathbf{a}_{<t_i}, \mathbf{y}_i, k_i)\}$, $k_i \in \{1, \dots, K\}$
Ensure: Frozen estimator s_ϕ, calibration map $\mathcal{C}$, and shared velocity field $\mathbf{v}_\theta$

Stage I: Training Modality Strength Estimator
for each query $i \in \mathcal{D}$ **do**
  Retrieve local neighbors using $(\mathbf{o}_{\le t_i}, \mathbf{a}_{<t_i})$
  Compute modality target r_i
end for
Train $s_\phi(\mathbf{o}_{\le t_i}, \mathbf{a}_{<t_i}) \approx r_i$
Fit calibration map $\mathcal{C}$ on the training split
Freeze s_ϕ and $\mathcal{C}$
Stage II: Training FM Policy
Identify the branch-critical region from $\{r_i\}$
Compute $\{\bar{\mathbf{a}}_k\}_{k=1}^K$, $\{\tilde{\mathbf{a}}_k\}_{k=1}^K$, and $\mathbf{U}$
for each minibatch $(\mathbf{o}_{\le t}, \mathbf{a}_{<t}, \mathbf{y}, k)$ **do**
  $q \leftarrow \mathcal{C}(s_\phi(\mathbf{o}_{\le t}, \mathbf{a}_{<t}))$
  $q^{\text{eff}} \leftarrow \max(q, q_{\text{floor}})$
  Sample $\mathbf{x}_0 \sim p(\mathbf{x}_0 \mid k, q^{\text{eff}})$
  Sample $\tau \sim \mathcal{U}(0,1)$ and set $\mathbf{x}_\tau = (1-\tau)\mathbf{x}_0 + \tau\mathbf{y}$
  Update only θ using the standard FM objective
end for
Stage III: Deployment
Initialize $q_{-1}^{\text{eff}} \leftarrow 0$
for each policy query t **do**
  $q_t \leftarrow \mathcal{C}(s_\phi(\mathbf{o}_{\le t}, \mathbf{a}_{<t}))$
  $q_t^{\text{eff}} \leftarrow \max\left(q_t,\ q_{\text{floor}},\ \lambda_q q_{t-1}^{\text{eff}}\right)$
  Select the desired mode k_t
  Sample $\mathbf{x}_0 \sim p(\mathbf{x}_0 \mid k_t, q_t^{\text{eff}})$
  Integrate $\mathbf{v}_\theta(\cdot, \cdot, \mathbf{o}_{\le t})$ from flow time 0 to 1 to obtain $\hat{\mathbf{y}}_t$
  Execute the current action slice and update $\mathbf{a}_{<t+1}$
end for

ing six joint positions and the gripper position. Both RGB observations are resized to 256×256. The corresponding action is also seven-dimensional and specifies the absolute positions of the six right-arm joints and the gripper.

All three real-world policies use an observation history of 8 steps and predict action chunks with a horizon of 16. Up to 8 actions from each chunk are executed before the next policy query. The robot is controlled in closed loop at 15 Hz.

For *Place Cups*, the demonstrations contain 52 episodes per mode after filtering. For *Pick Veg*, the dataset contains 55 episodes per mode. For *Open Drawer*, the converted dataset contains 51 valid episodes per mode after discarding one anomalous episode. The mode labels are assigned per episode according to the behavioral choice described in Table 7.

E.5 Evaluation Protocol and Metrics

For simulation, each explicitly requested mode is evaluated with 50 rollouts. For methods without a mode-selection interface, only task success and the empirical behavior distribution are reported.

For each real-world mode-controllable policy, we conduct 20 trials for each requested mode. The corresponding GR_k is computed over the 20 trials requesting mode k, while the reported SR is averaged over the two requested-mode conditions. FM-DiT is evaluated with 20 unconditional trials per task. Real-world results are rounded to two decimal places.

Let $S_i \in \{0, 1\}$ indicate whether rollout i completes the task, and let M_i denote its executed semantic mode. The success rate is $\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} S_i$. For a requested mode k, the guidance rate is

$\mathrm{GR}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbb{I}[S_i = 1 \wedge M_i = k]$. A rollout contributes to GR only when it both succeeds and realizes the requested semantic behavior. Consequently, $\mathrm{GR}_k \leq \mathrm{SR}_k$.

F Classifier-Free Guidance for Mode-Conditioned FM

Mode-conditioned FM uses classifier-free guidance

$$\mathbf{v}_{\mathrm{CFG}} = \mathbf{v}_{\mathrm{uncond}} + s_{\mathrm{CFG}}\left(\mathbf{v}_{\mathrm{cond}} - \mathbf{v}_{\mathrm{uncond}}\right), \tag{67}$$

where s_{CFG} is the inference-time guidance scale. The primary real-world evaluation uses $s_{\mathrm{CFG}} = 1$. We additionally increase the scale for the requested mode on which the default setting exhibits mode collapse. Table 9 reports these targeted comparisons.

Increasing CFG improves mode-1 guidance on *Place Cups* from 0.15 to 0.65. However, the effect does not generalize consistently. On *Pick Veg*, increasing the scale raises GR only to 0.20 while reducing SR from 0.60 to 0.45. Similarly, on *Open Drawer*, increasing the scale raises mode-2 GR from 0.00 to 0.25 but reduces SR from 0.85 to 0.75. These results indicate that inference-time amplification of the mode condition cannot reliably resolve the asymmetric mode collapse observed in Mode-conditioned FM.

Table 9: **CFG comparisons for modes on which Mode-conditioned FM collapses under the default guidance scale.**

Task	Requested mode	s_{CFG}	SR	GR
Place Cups	Mode 1	1	0.55	0.15
	Mode 1	3	0.70	0.65
Pick Veg	Mode 1	1	0.60	0.00
	Mode 1	3	0.45	0.20
Open Drawer	Mode 2	1	0.85	0.00
	Mode 2	3	0.75	0.25

G Imbalanced Training and Data Efficiency

For each task, mode 1 is treated as the minority mode. MOSS and Mode-conditioned FM are trained on 10 mode-1 and 90 mode-2 demonstrations, whereas unconditional FM-DiT is trained only on the same 10 mode-1 demonstrations. Evaluation always requests mode 1.

Each method is trained independently. At each checkpoint, we evaluate the requested mode and report its SR and GR. The solid curves in Figure 7 show the mean GR across the 3 seeds, while the shaded regions indicate ± 1 standard deviation computed from the seed-wise results.

The horizontal axis reports the cumulative number of sampled minority-mode training windows rather than the epoch index. This normalization allows the methods to be compared according to their exposure to target-mode data, while MOSS can additionally benefit from the majority-mode demonstrations.

H Additional Ablation Studies

H.1 Ablation Variants

The ablations in Table 3 compare 4 source-distribution variants to the source distribution:

- **Single Gaussian:** a standard Gaussian is partitioned by the sign of its projection onto the mode axis at a selected action position, without displaced Gaussian centers.
- **Single-step separation:** mode-specific Gaussian displacement is applied only at the branching point.
- **Without adaptation:** fixed mode-specific source separation is applied across the complete action chunk.
- **MOSS:** full-chunk separation is adjusted using the estimated modality strength.

The single-position variants leave most source dimensions shared, whereas the full-chunk variants maintain a consistent mode identity across the predicted action sequence.

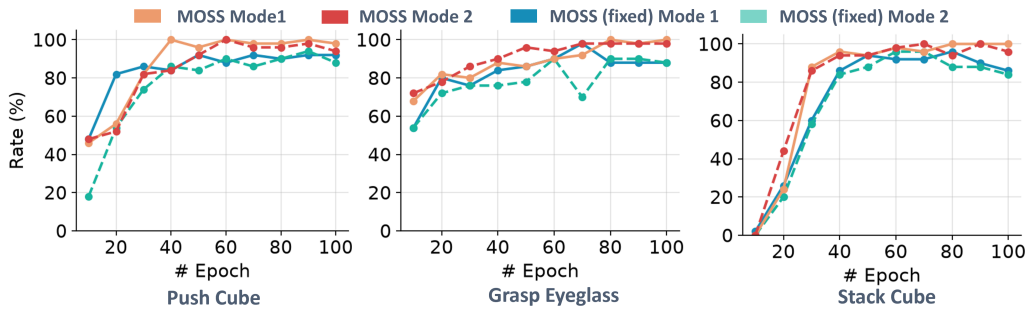

Figure 8: **Checkpoint-wise GR of fixed full-chunk source shaping and adaptive MOSS.**

Table 10: **Offline source-alignment and squared-transport statistics.** Lower values indicate better alignment or a shorter source-to-target transport.

Task	Source	D_μ	D_{tr}	$D_\parallel$
Push Cube	Fixed	2.997	174.582	2.767
	Adaptive	**2.884**	**172.823**	**0.931**
Grasp Eyeglasses	Fixed	3.032	201.079	4.649
	Adaptive	**2.945**	**200.983**	**4.565**
Stack Cube	Fixed	3.346	209.991	5.884
	Adaptive	**3.298**	**208.754**	**4.683**

H.2 Fixed versus Adaptive Training Curves

To compare convergence independently of the final checkpoints reported in Table 3, we evaluate fixed full-chunk shaping and adaptive MOSS every 10 epochs from epoch 10 to epoch 100, as shown in Figure 8. Each point contains 50 rollouts for each requested mode.

Across the three tasks, adaptive MOSS reaches consistently high guidance rates for both requested modes and remains stable at later checkpoints. At epoch 100, its mean GR across the two modes reaches 0.96, 0.99, and 0.98 on *Push Cube*, *Grasp Eyeglasses*, and *Stack Cube*, compared with 0.90, 0.88, and 0.85 for fixed shaping, respectively. The checkpoint-wise results show that the improvement of adaptive source shaping persists throughout training rather than arising from a particular final checkpoint.

H.3 Source Alignment and Transport Cost

We conduct an offline analysis in normalized action space using every available training window.

Let $\mathbf{u}$ be the normalized full-chunk mode axis, $\bar{\mathbf{a}}_k$ the empirical normalized action-chunk mean of mode k, and $\boldsymbol{\mu}_k(q_i)$ the source mean used for window i. We measure center-to-target alignment by

$$D_\mu = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\mu}_{k_i}(q_i) - \bar{\mathbf{a}}_{k_i}\|_2 . \tag{68}$$

For a normalized target action chunk $\mathbf{a}_i$ and source sample $\mathbf{z}_{i,j}$, the transport statistic is the squared displacement

$$D_{\mathrm{tr}} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|\mathbf{z}_{i,j} - \mathbf{a}_i\|_2^2 . \tag{69}$$

Its component along the mode-discriminative direction is

$$D_\parallel = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \left(\langle \mathrm{vec}(\mathbf{z}_{i,j} - \mathbf{a}_i), \mathrm{vec}(\mathbf{u})\rangle\right)^2 , \tag{70}$$

and $D_\perp = D_{\mathrm{tr}} - D_\parallel$.

Adaptive shaping reduces D_μ and total squared transport in all three tasks. The difference is most visible in $D_\parallel$, particularly on Push Cube and Stack Cube. This is expected because MOSS modifies the source primarily along the mode-discriminative direction, while the much larger orthogonal transport component is largely shared by both variants.