Hemispheric Diffusion: A Compositional Generative Policy for Coordinated Bimanual Manipulation

Yechen Fan, *Graduate Student Member, IEEE,* Jinhua Ye, *Senior Member, IEEE,* Xianyou Ji, Chenyang Song, Haibin Wu, Gengfeng Zheng, *Senior Member, IEEE,* and Jiafu Wan, *Fellow, IEEE*

***Abstract*—Bimanual manipulation requires both adaptive multimodal perception and physically consistent inter-arm coordination. Yet existing imitation policies typically couple all sensory inputs through fixed early-fusion structures, which can amplify modality conflicts and fail under asymmetric observations or local perception degradation. We introduce Hemispheric Diffusion (HemiDiff), a compositional diffusion policy for coordinated bimanual manipulation. HemiDiff decomposes multimodal conditioning into independent modality experts and recomposes their denoising predictions through an asymmetric perception router that assigns arm-specific modality weights at inference time. A coordination energy expert further guides the denoising process toward demonstration-consistent relative motion, preserving inter-arm coordination under decoupled perception. Experiments on RLBench2 and a real bimanual robot demonstrate that HemiDiff achieves a 76% average success rate on real-world tasks, outperforming the strongest baseline by 19 percentage points while remaining robust to occlusion, sensor degradation, and viewpoint perturbations.**

***Index Terms*—Bimanual manipulation, multimodal fusion, generative policy.**

I. Introduction

HUMAN dexterous manipulation relies on coordinated bimanual actions and multimodal perception. While RGB, point clouds, semantic features, and tactile sensing provide complementary cues [1], [2], robust multimodal integration for bimanual control remains challenging.

Recent advances in bimanual manipulation, including ACT [3], diffusion policies [4], [5], [6], and VLA models [7], [8], [9], have shown strong capabilities. However, most existing frameworks still rely on early fusion [10], [11], [12] to integrate multimodal inputs, typically through feature concatenation. This design introduces two major issues: modality dominance [13], where high-dimensional global signals such as RGB suppress sparse local cues such as tactile or point-cloud observations; and sensitivity to local perception anomalies [4], [14], where partial occlusion, missing observations, or degraded modalities can corrupt the shared latent representation and cause task failure. Although recent methods such as DECO [15] attempt to alleviate these issues through decoupled conditioning, robust multimodal fusion under partial observability remains unresolved.

Furthermore, bimanual manipulation is inherently spatially asymmetric: the perceptual requirements of the left and right arms often differ across task states. For example, one arm may operate under severe occlusion while the other remains in a clear workspace. Existing methods based on global attention struggle to capture such asymmetric perceptual demands. Although routing architectures or arm-wise action modeling [1] can partially alleviate this issue, they often weaken inter-arm coordination. Without explicit kinematic constraints or state prediction [16], [17], such decoupled policies are prone to desynchronized motions, self-collisions, and antagonistic interactions, especially under partial observations.

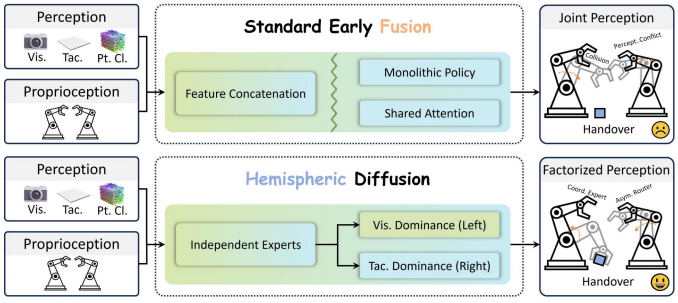

Fig. 1. Standard early fusion versus HemiDiff. HemiDiff uses independent modality experts, asymmetric routing, and coordination guidance for robust bimanual manipulation.

In this letter, we propose HemiDiff, a compositional diffusion policy for coordinated bimanual manipulation under asymmetric multimodal observations (see Fig. 1). HemiDiff decomposes multimodal conditioning into independent modality experts for RGB, point clouds, DINO features, and tactile sensing, avoiding the fragility of fixed early fusion. An asymmetric perception router recomposes their denoising predictions with arm-specific modality weights, enabling adaptive inference-time reweighting under partial observations or modality degradation. To preserve inter-arm consistency, a coordination energy expert injects demonstration-consistent relative-motion guidance during denoising. Together, these components enable robust multimodal composition while maintaining physically coordinated bimanual motion.

The main contributions of this study are summarized as follows:

- We design an asymmetric perception router that assigns arm-specific modality weights according to the current task state.
- We introduce independently trained modality experts, enabling flexible inference-time reweighting or masking of degraded modalities without retraining.
- We incorporate a coordination energy expert that injects coordination gradients derived from demonstrated relative poses during denoising, thereby preserving physically consistent bimanual motion.
- We validate HemiDiff in simulation and real-world tasks, demonstrating strong performance across diverse modality settings.

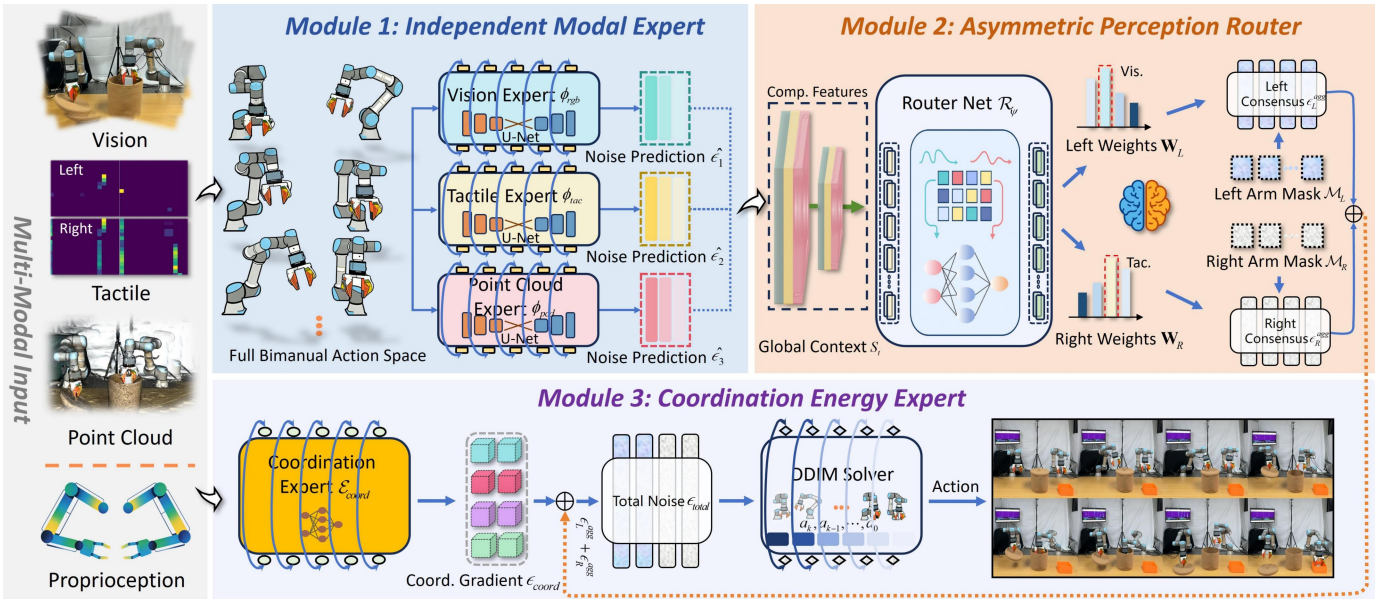

Fig. 2. HemiDiff consists of three components: independent modality experts for full-action noise prediction, an asymmetric perception router for arm-specific modality weighting, and a coordination energy expert for physical-manifold guidance during denoising. Fusion is performed at the predicted-noise/score level through arm-specific routing and masking, enabling perceptual decoupling while preserving physically consistent relative motion between the two arms.

II. Related Works

A. Learning-based Bimanual Manipulation

Data-driven imitation learning has advanced with generative policies such as Diffusion Policy (DP) [4], [14], which formulate continuous control as denoising. Large-scale VLA models such as π_0, $\pi_{0.5}$ [7], [8], and RDT-1B [9] further extend bimanual action generation. However, similar to ACT [3] and Bi3D Diffuser Actor [5], these methods typically rely on early fusion or holistic action modeling, making coordination vulnerable to local disturbances. Recent works improve bimanual coordination through object-motion modeling [17], kinematic regularization or test-time optimization [2], [16], and coordination modeling [18]. In contrast, HemiDiff introduces a coordination energy expert that injects physical-manifold gradients during denoising, preserving coordination while maintaining perceptual independence.

B. Multimodal Fusion for Robot Learning

Vision remains central to robotic perception, while 3D representations such as point clouds further improve spatial reasoning in diffusion policies [6], [19]. However, self-occlusion and mutual occlusion often degrade visual observations in dexterous manipulation, motivating the use of tactile sensing from force/torque signals to optical and piezoresistive tactile arrays [10], [12], [20]. Most multimodal methods still rely on feature concatenation [12], [21], which overlooks modality heterogeneity and may cause dominant visual features to suppress sparse tactile cues. Recent methods explore decoupled fusion, such as DECO [15] and AnyBimanual [1]. In contrast, HemiDiff uses an asymmetric perception router to dynamically reweight heterogeneous modalities and mask unreliable channels at inference time.

C. Compositional Policies and Modular Learning

Compositional generative policies provide a modular paradigm for robotic control by combining multiple components. Originating from energy-based models, where summed energy functions enable flexible constraint composition [22], this idea was later extended to diffusion models through linear combinations of score functions [4], [23]. In robotic manipulation, it has inspired modular policy learning methods such as PoCo [24], CCSP [25], and GPC [26]. Modular approaches further improve flexibility by decoupling feature representations and mitigating negative transfer. HemiDiff extends this idea to heterogeneous modalities and bimanual spatial asymmetry by composing independent modality experts with a coordination energy expert at the score-function level, enabling adaptive multimodal reweighting and explicit constraints on relative arm motion.

III. Methodology

This section presents HemiDiff (Fig. 2), including the diffusion foundation, expert construction, asymmetric routing, and coordination-guided inference.

A. Compositional Diffusion Policy Foundation

For a multimodal observation set $\mathcal{O}_t = \{m_{rgb}, m_{tac}, m_{pcd}, m_{dino}\} \times \mathcal{P}$, we model the bimanual manipulation policy as a Conditional Denoising Diffusion Probabilistic Model [4]. Given a clean action sample a_0 from expert demonstrations, the noisy action at diffusion step $k \in \{1, \dots, K\}$ is:

$$a_k = \sqrt{\bar{\alpha}_k} a_0 + \sqrt{1 - \bar{\alpha}_k}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where $\bar{\alpha}_k \in (0, 1)$ represents the coefficients accumulated from a predefined variance schedule, controlling the noise

intensity. ϵ is random noise sampled from a standard normal distribution, and $\mathbf{I} \in \mathbb{R}^{d_a \times d_a}$ is the identity matrix.

In the reverse process, we learn a denoising network ϵ_θ. For a heterogeneous multimodal observation set $\mathcal{O}_t$, the training objective minimizes the mean squared error between the predicted and actual noise:

$$\mathcal{L}_{diff}(\theta) = \mathbb{E}_{a_0,\epsilon,k,\mathcal{O}_t}\left[\|\epsilon - \epsilon_\theta(a_k, k, \mathcal{O}_t)\|^2\right] \tag{2}$$

where θ represents the parameters of the denoising network. In our framework, the total denoising network ϵ_θ is composed of multiple independently trained modality experts $\{\phi_i\}$. Based on the theory of Energy-Based Models [28], the score of the joint distribution is modeled as a weighted linear combination of modality-specific denoising scores. This implies that for an observation set $\mathcal{O}_t = \{m_1, ..., m_N\}$ consisting of N modalities, the total policy can be constructed as:

$$\epsilon_{total}(a_k, k, \mathcal{O}_t) \approx \sum_{i=1}^{N} \omega_i \cdot \epsilon_{\phi_i}(a_k, k, m_i) \tag{3}$$

where ϵ_{ϕ_i} is the independent modality expert network parameterized by ϕ_i, conditioned only on a single modality m_i and the base proprioceptive state. $\omega_i \in \mathbb{R}$ is a scalar weight coefficient regulating the contribution of each modality. However, this approach of using only a single global weight is equivalent to forcing the bimanual dependency on perception to be consistent. When the two arms are in asymmetric states, this global weight cannot simultaneously satisfy the distinct perceptual needs of both arms. This motivates the arm-specific perception routing mechanism introduced in Section III-C.

B. Expert Construction for Bimanual Asymmetry

To address the inherent spatial asymmetry in bimanual manipulation tasks, we partition the full bimanual action vector a_t in terms of left-arm and right-arm subspaces:

$$a_t = [a_{L,t}^T, a_{R,t}^T]^T \in \mathbb{R}^{d_a} \tag{4}$$

where $a_{L,t} \in \mathbb{R}^{d_L}$ and $a_{R,t} \in \mathbb{R}^{d_R}$ denote the joint-space action vectors of the left and right arms, respectively, and $d_a = d_L + d_R$. Each arm action vector includes the corresponding gripper state.

Based on this arm-wise formulation, we construct a set of modality-specific expert networks $\Phi = \{\phi_{rgb}, \phi_{tac}, \phi_{pcd}, \phi_{dino}\}$. Each expert is a complete U-Net-based denoising network, rather than a modality-specific encoder, and independently predicts noise over the full bimanual action space conditioned on a single modality and proprioception. Each expert ϕ_i is trained independently on the same demonstration trajectories, using only one sensory modality m_i and proprioception p_t as its condition, while sharing the same full bimanual action supervision. This design learns modality-specific denoising predictions without feature-level competition among heterogeneous modalities [12].

Each expert ϕ_i adopts a U-Net based diffusion architecture. For modality i the modality observation $m_{i,t}$ and proprioception p_t are first encoded into a compact condition embedding $c_{i,t} = f_{enc}^{(i)}(m_{i,t}, p_t)$, which is then fed into the corresponding expert U-Net to predict a noise estimate over the full bimanual action space:

$$\hat{\epsilon}_i = \epsilon_{\phi_i}(a_k, k, c_{i,t}) \tag{5}$$

where $f_{enc}^{(i)}$ is the feature encoder specific to modality i (e.g., ResNet-18 for vision or MLP for DINO features), and $c_{i,t}$ is the generated compact condition embedding vector. Each modality expert predicts over the full bimanual action space rather than a single arm, so as to preserve cross-arm contextual awareness during denoising. Asymmetric modality contribution is then enforced at inference time through arm-wise masking, which extracts each expert's gradient only on the corresponding subspace.

Unlike feature-level fusion, HemiDiff composes heterogeneous modalities at the predicted-noise / score level. Each modality expert independently produces a denoising prediction, and multimodal integration is performed only afterward through arm-specific routing weights and action-space masking.

C. Asymmetric Perception Routing

To achieve spatial decoupling of multimodal perception, we designed an asymmetric router network $\mathcal{R}_\psi$. This module takes the current global context state S_t (concatenated from compressed features of all modalities) as input and predicts arm-specific modality weights that determine how much each modality expert contributes to the denoising of the left-arm and right-arm action subspaces, respectively. The network first outputs raw logits $\mathbf{Z}_L, \mathbf{Z}_R$:

$$[\mathbf{Z}_L, \mathbf{Z}_R] = \mathcal{R}_\psi(S_t), \quad \mathbf{Z}_h \in \mathbb{R}^N, h \in \{L, R\} \tag{6}$$

To endow the weights with explicit probabilistic meaning, we apply a softmax normalization constraint separately to the logit vector of each arm to obtain the final modality weight vector $\mathbf{W}_h$:

$$\omega_h^{(i)} = \frac{\exp(z_h^{(i)})}{\sum_{j=1}^{N} \exp(z_h^{(j)})} \tag{7}$$

where $\mathbf{W}_L = [\omega_L^{(1)}, ..., \omega_L^{(N)}]$ and $\mathbf{W}_R$ are the modality weight vectors for the left and right arms, respectively. $z_h^{(i)}$ is the i-th component of $\mathbf{Z}_h$. As part of the overall diffusion policy, the router network $\mathcal{R}_\psi$ is optimized end-to-end via the denoising objective in Eq. (2), thereby learning an adaptive modality allocation strategy [13].

D. Inter-Arm Coordination Constraints

Fully decoupled perception routing may cause the two arms to lose coordination at the physical level. To remedy this deficiency, we introduce an energy-based coordination expert, $\mathcal{E}_{coord}$ (see Fig. 3). This expert does not rely on external vision but is based solely on proprioception p_t, aiming to learn the manifold constraints of relative motion between the two arms. We define the coordination energy function as:

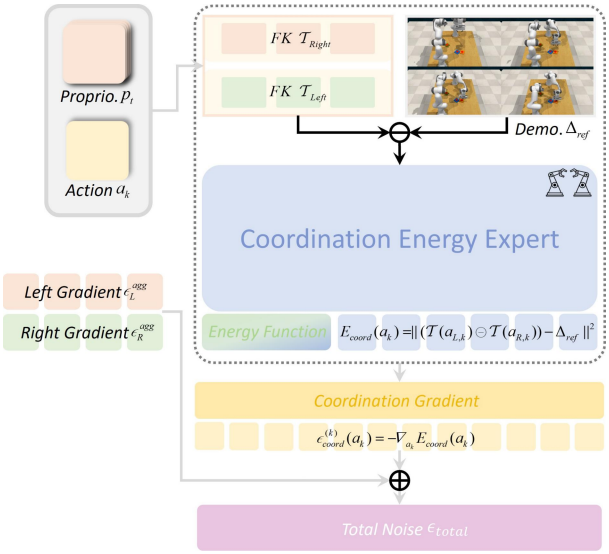

Fig. 3. The Coordination Energy Expert constructs a coordination energy function E_{coord} from the discrepancy between the generated relative pose of the two end-effectors and the demonstration-derived reference Δ_{ref}. During diffusion sampling, its gradients guide bimanual actions to maintain physically consistent coordination.

$$E_{coord}(a_k) = \|(\mathcal{T}(a_{L,k}) \ominus \mathcal{T}(a_{R,k})) - \Delta_{ref}\|^2 \tag{8}$$

where $\mathcal{T}(\cdot)$ denotes forward kinematics, $\ominus$ is the relative pose operator in SE(3) and $\|\cdot\|^2$ denotes a weighted sum of translational and rotational errors. Δ_{ref} is the expected relative transformation relationship statistically learned from demonstration data. In multi-stage tasks, we use the sliding window mean of the current time step in the demonstration trajectory as Δ_{ref} to adapt to the manipulation variations of non-rigid objects.

During the diffusion sampling process, this energy function acts as a manifold constraint in the denoising process [25], ensuring that the relative geometric relationship between the two arms does not deviate from the distribution manifold of the demonstration data even when bimanual perception is separated. The resulting coordination gradient term $\epsilon^{(k)}_{coord}$ is defined as:

$$\epsilon^{(k)}_{coord}(a_k) = -\nabla_{a_k} E_{coord}(a_k) \tag{9}$$

where the gradient of the energy function directs the action towards a valid physical manifold. In the subsequent integration step (Eq. 11), a scaling coefficient λ is applied to regulate the coordination intensity.

E. Heterogeneous Policy Consensus Inference

In the inference phase, we integrate all the aforementioned components through a spatial masking mechanism. Define action space projection masks $\mathcal{M}_L, \mathcal{M}_R \in \{0,1\}^{d_a}$, which are binary vectors with 1s for the corresponding arm's action dimensions and 0s otherwise. These masks are applied to the output gradients rather than input features, used to project the full action vector into the subspaces of the left and right arms.

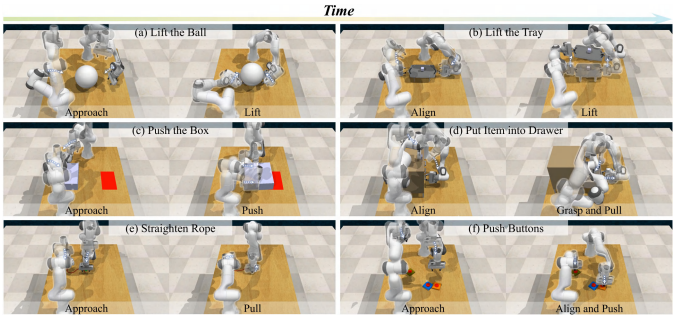

Fig. 4. Six bimanual manipulation tasks in RLBench2.

Based on Eq. (5) and Eq. (7), we compute the independent perceptual consensus gradient for a single arm according to the asymmetric weights output by the router:

$$\epsilon_h^{agg} = \mathcal{M}_h \odot \left(\sum_{i=1}^{N} \omega_h^{(i)} \cdot \hat{\epsilon}_i \right), \quad h \in \{L, R\} \tag{10}$$

where $\odot$ denotes element-wise multiplication, and $\hat{\epsilon}_i$ represents the prediction of the i-th expert over the full bimanual action space. Using the mask $\mathcal{M}_h$, we extract only the action dimensions corresponding to arm h, while setting all remaining dimensions to zero. This composition mechanism naturally enables inference-time fault tolerance: when a modality becomes unavailable or corrupted, its contribution can be masked by setting the corresponding routing weight to zero, while the remaining experts continue to provide valid denoising signals. As a result, the policy degrades gracefully without retraining.

Next, following the compositional generation paradigm [24], [26], by combining the physical constraint term from Eq. (9) and the spatial perception term from Eq. (10), we obtain the total predicted noise at step k:

$$\epsilon_{total}(a_k) = \epsilon_L^{agg} + \epsilon_R^{agg} + \lambda \cdot \epsilon^{(k)}_{coord}(a_k) \tag{11}$$

where λ is a hyperparameter controlling the intensity of physical coordination. According to Eq. (11), we use the obtained ϵ_{total} to perform a single reverse update step to obtain the denoised action a_{k-1}:

$$a_{k-1} = \frac{1}{\sqrt{\alpha_k}} \left(a_k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}} \epsilon_{total}(a_k) \right) + \sigma_k \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \tag{12}$$

where σ_k^2 denotes the posterior variance of the diffusion process, typically defined as $\sigma_k^2 = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}(1-\alpha_k)$, and $\sigma_k = \sqrt{\sigma_k^2}$, with $\bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i$. Iteratively applying Eq. 12 recovers the action a_0 while preserving perceptual decoupling and physical coordination.

IV. Experimental Study

A. Experimental Setup

We evaluate HemiDiff in both simulation and real-world environments. Each modality expert uses a DP-style U-Net denoising backbone [4], and the final policy composes their

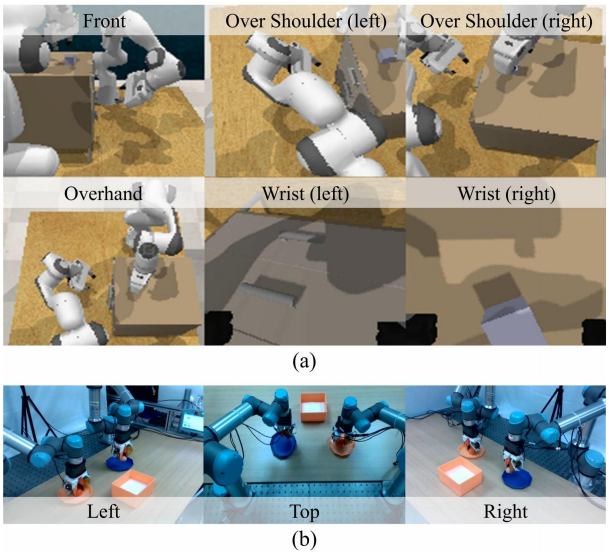

Fig. 5. Multi-view Observation (a) Six camera viewpoints used in RLBench2. (b) Three camera viewpoints used in the real-world.

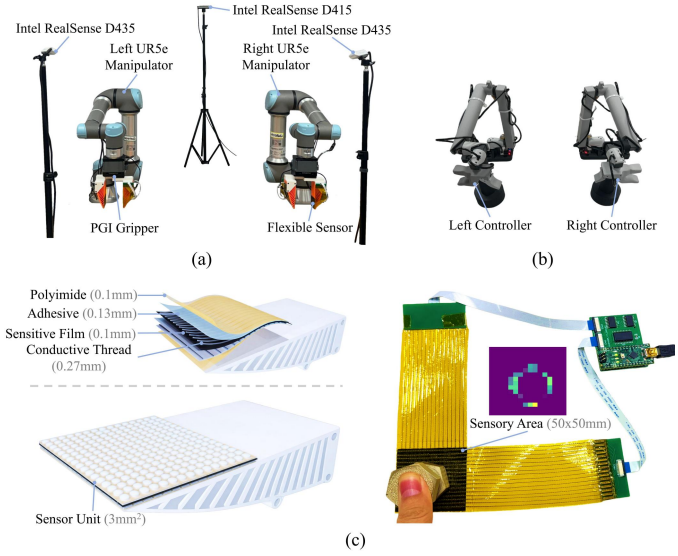

Fig. 6. Hardware Platform. (a) A bimanual robotic platform integrating visual and tactile. (b) The GELLO teleoperation controller used for data collection. (c) The multi-layer physical structure and real-world appearance of the piezoresistive tactile array.

TABLE I
TRAINING AND EXECUTION SETTINGS

Method	Optimizer	LR	Batch Size	Epochs	Pred. Horizon	Action Chunk	Sampler / Mode	Inference Steps
ACT	AdamW	1e-4	64	501	100	100	Action Chunking	–
DP	AdamW	1e-4	64	1001	16	8	DDPMScheduler	100
DP3	AdamW	1e-4	64	1001	16	8	DDPMScheduler	100
VQ-BET	AdamW	1e-4	64	501	16	5	Action Chunking	–
DP (Feature Concat)	AdamW	1e-4	64	1001	16	8	DDPMScheduler	100
HemiDiff (Ours)	AdamW	1e-4	64	1001	16	8	DDPMScheduler	100

Note: The specified hyperparameters remain constant across all modality combinations without requiring further tuning.

predictions through asymmetric routing and coordination guidance. For fair fusion-mechanism comparison, all simulation ablations are implemented within the same DP framework.

Simulation Setup: We conduct simulation experiments on RLBench2 [29], a bimanual extension of RLBench [30], covering six tasks shown in Fig. 4. The environment uses two Franka Panda arms with parallel grippers and six noise-free RGB-D cameras (128×128) for full-view coverage (Fig. 5(a)). Each task contains 100 demonstrations, and each method is evaluated over 100 randomized target configurations. Inputs include multi-view RGB images, point clouds, and pretrained DINO 3D semantic features [6]. RLBench2 is used for controlled analysis because it provides idealized RGB-D observations without tactile feedback.

Real-World Setup: The real-world platform consists of two UR5e arms with PGI grippers, four flexible piezoresistive tactile arrays [10], [11], [12] (16×16 sensing units, 2 mm resolution), and three Intel RealSense cameras (640×480) from different viewpoints (Fig. 5(b)). We evaluate six tasks in three categories: precision assembly, including Hammer into Slot and Insert Peg into Hole; occlusion handling, including Retrieve Occluded Bag and Retrieve Transparent Object; and bimanual coordination, including Handover Bread to Plate and Place Food on Plate (Fig. 7). All demonstrations are collected via GELLO teleoperation [31](Fig. 6), with 80 demonstrations for precision assembly, 70 for occlusion handling, and 60 for bimanual coordination. Each task is evaluated over 20 randomized initializations.

Baseline Methods: In simulation, we compare HemiDiff with single-modality DP policies (RGB, PCD, and

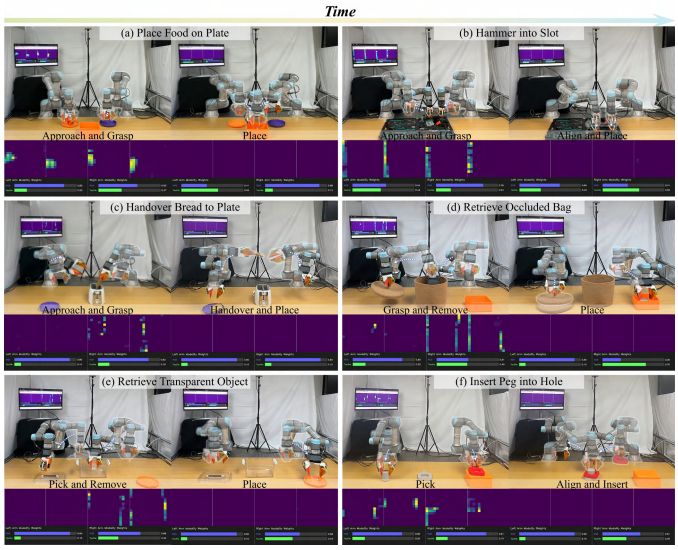

Fig. 7. Six bimanual manipulation tasks in the real world.

DINO), Feature Concat variants (RGB+PCD, RGB+DINO, and RGB+PCD+DINO), and representative bimanual baselines including PerAct2 [29], RVT-LF [32], and AnyBimanual-PerAct [1]. We also include three ablation variants: w/o Asymmetric Router, w/o Coordination Energy Expert, and w/o Modality Experts. In the real world, we compare against ACT [3], DP [4], DP3 [6], RGB+Tactile Feature Concat, and VQ-BET [33]. Target object positions are randomized during evaluation to assess robustness, as shown in Fig. 8.

Implementation Details: HemiDiff runs data collection and inference at 10 Hz. Visual observations are captured at 15 FPS, resized to 96×128, and encoded by ResNet-18. We select $W_{ref} = 5$, $\lambda = 1.0$, and $T_{obs} = 2$ via a small validation grid search on Put Item into Drawer, and keep them fixed across all tasks to avoid task-specific tuning. Detailed settings are listed

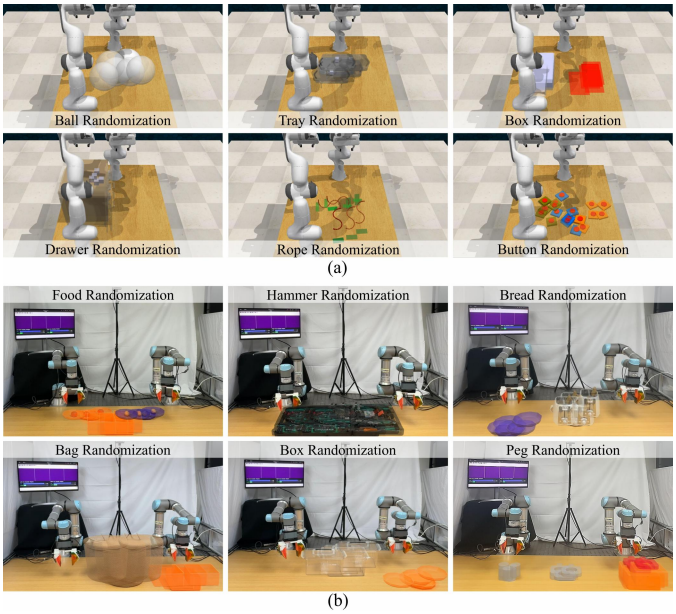

Fig. 8. Bimanual manipulation tasks are randomized in both RLBench2 and real-world settings.

TABLE II
QUANTITATIVE EVALUATION AND ABLATION RESULTS ON RLBENCH2

Method	Modality	Drawer	Tray	Box	Ball	Rope	Button	SR
Single-Modality								
DP	RGB	0.26	0.53	0.77	0.86	0.12	0.75	0.55 ±0.28
	PCD	0.04	**0.78**	0.67	0.84	0.15	0.68	0.53 ±0.30
	DINO	0.00	0.73	0.70	0.74	0.05	0.60	0.47 ±0.32
Multi-Modality								
Feature Concat	RGB+PCD	0.01	0.71	0.75	0.88	0.18	**0.88**	0.57 ±0.34
	RGB+DINO	0.15	0.68	0.79	0.85	0.14	0.76	0.56 ±0.28
	RGB+PCD+DINO	0.22	0.61	0.69	0.82	0.24	0.74	0.55 ±0.22
Ablation Variants								
HemiDiff	w/o Asymmetric Router	0.22	0.63	0.70	0.82	0.19	0.74	0.55 ±0.25
	w/o Coordination Energy Expert	0.31	0.59	0.76	0.85	0.23	0.80	0.59 ±0.26
	w/o Modality Experts	0.19	0.57	0.66	0.79	0.17	0.71	0.52 ±0.24
HemiDiff Variants								
HemiDiff (Ours)	RGB+PCD	0.35	0.76	0.80	0.91	0.25	0.81	0.65 ±0.25
	RGB+DINO	0.29	0.74	0.67	0.89	0.21	0.80	0.60 ±0.25
	RGB+PCD+DINO	**0.37**	0.76	**0.81**	**0.92**	**0.35**	0.86	**0.67** ±0.23

SR: Task success rate. Ablation variants use RGB+PCD+DINO unless otherwise specified.

in Table I. For the bimanual baselines in Table III, we follow their original implementations while aligning the RLBench2 tasks, demonstration scale, and evaluation protocol.

B. Main Performance and Robustness

As shown in Table II, RGB performs best among single-modality DP policies with a 55% average success rate, while Feature Concat brings only limited gains, indicating that early fusion does not fully exploit complementary modalities. HemiDiff consistently improves multimodal performance, with RGB+PCD+DINO achieving the best average success rate of 67%. Ablation results verify the contribution of each component: removing the asymmetric router, coordination energy expert, and modality experts reduces the average success rate to 55%, 59%, and 52%, respectively.

We further compare the best HemiDiff variant with representative bimanual baselines in Table III. HemiDiff achieves 67%, outperforming AnyBimanual-PerAct (40%), PerAct2 (35%), and RVT-LF (22%). As shown in Fig. 9, HemiDiff completes Put Item into Drawer with better perceptual allocation and more stable relative motion, whereas AnyBimanual-PerAct is more prone to coordination failure.

TABLE III
COMPARISON WITH REPRESENTATIVE BIMANUAL BASELINES ON RLBENCH2

Method	Drawer	Tray	Box	Ball	Rope	Button	SR
PerAct2	0.10	0.18	0.60	0.50	0.24	0.47	0.35 ±0.18
RVT-LF	0.10	0.08	0.52	0.17	0.06	0.39	0.22 ±0.18
AnyBimanual-PerAct	0.22	0.15	0.48	0.40	0.28	**0.88**	0.40 ±0.25
HemiDiff (Ours)	**0.37**	**0.76**	**0.81**	**0.92**	**0.35**	0.86	**0.67** ±0.23

TABLE IV
QUANTITATIVE EVALUATION RESULTS ON THE REAL-WORLD PLATFORM

Method	Precision		Occlusion		Coordination		SR
	Hammer	Peg	Bag	Transparent	Bread	Food	
ACT	0.40	0.35	0.20	0.05	0.30	0.25	0.26 ±0.11
DP	0.65	0.55	0.35	0.15	0.50	0.45	0.44 ±0.16
DP3	0.75	0.65	0.45	0.20	0.55	0.50	0.52 ±0.17
VQ-BET	0.65	0.70	0.30	0.20	0.60	0.55	0.50 ±0.18
Feature Concat	0.70	0.60	0.55	0.30	0.60	0.65	0.57 ±0.13
HemiDiff (Ours)	**0.80**	**0.80**	**0.75**	**0.65**	**0.80**	**0.75**	**0.76** ±0.05

In real-world experiments, HemiDiff achieves a 76% average success rate with only 60–80 demonstrations, outperforming all baselines (Table IV). It reaches 80% on Insert Peg into Hole, 75% on Retrieve Occluded Bag, and 65% on Retrieve Transparent Object. As shown in Fig. 9, HemiDiff completes Handover Bread to Plate by preserving interaction cues and stable coordination, while DP is more prone to object dropping due to its single shared denoising policy.

C. Generalization under Real-World Disturbances

We further evaluate HemiDiff under five real-world perturbations: human interference, visual occlusion, single-sided tactile sensor failure, viewpoint perturbation, and lighting variation (see Fig. 10). Experiments are conducted on Insert Peg into Hole using the same settings as in Table IV-A, comparing DP, Feature Concat, and HemiDiff.

As shown in Fig. 10 and Fig. 11(a), HemiDiff achieves 75%, 75%, 70%, 60%, and 65% success rates under human interference, visual occlusion, lighting variation, viewpoint perturbation, and sensor failure, respectively. The consistent gains over DP and Feature Concat indicate that HemiDiff can exploit reliable modalities under degraded perception while maintaining stable bimanual coordination under spatial disturbances.

D. Usability of Heterogeneous Data

We evaluate HemiDiff under heterogeneous visual–tactile data on Retrieve Transparent Object. The visual expert is trained with 70 demonstrations, while the tactile expert uses only 10 demonstrations. In contrast, Feature Concat requires paired multimodal data and is trained with 10 paired demonstrations. Each method is evaluated over 20 episodes with stage-wise success rates.

As shown in Fig. 11(b), HemiDiff improves Grasp success from 40% to 95% and achieves 75% and 70% on Remove and Lift, respectively. Feature Concat often fails under limited paired data, causing grasp-stage collisions (see Fig. 12). These

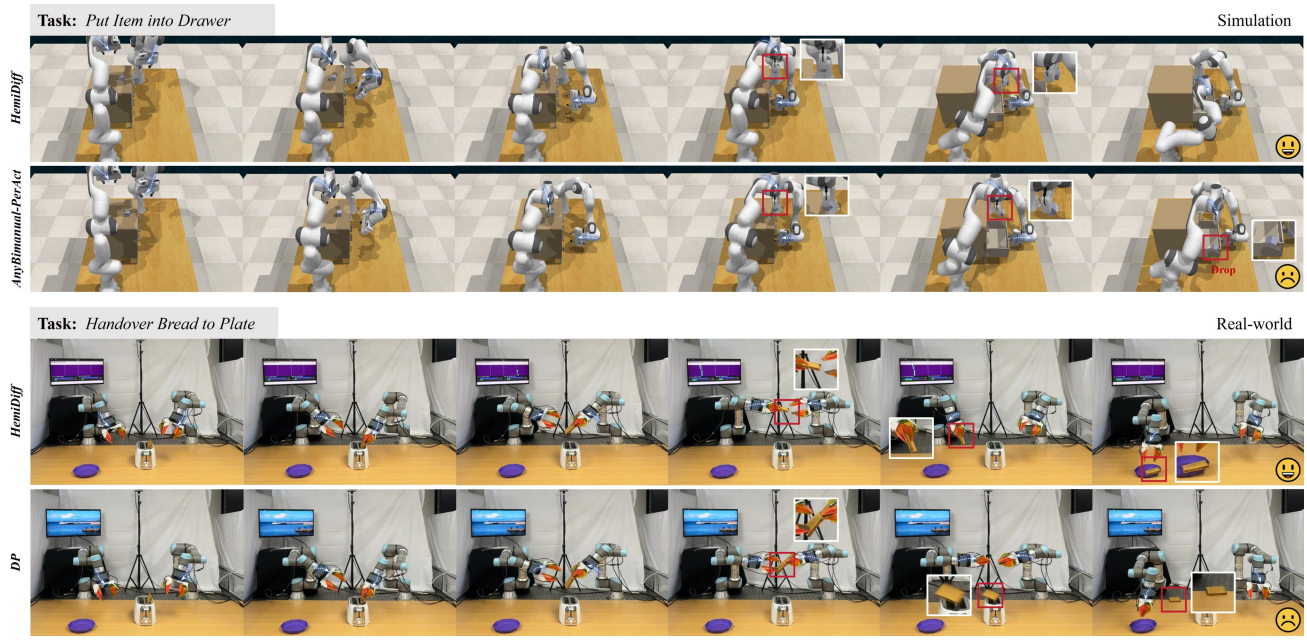

Fig. 9. Comparison of HemiDiff and representative baselines in simulation and the real world.

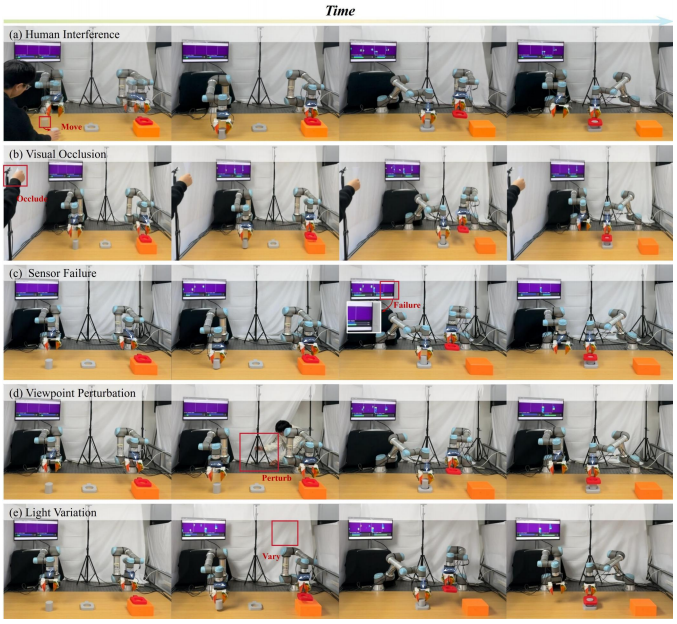

Fig. 10. HemiDiff in the Insert Peg into Hole task under five types of real-world disturbances: (a) human interference, (b) visual occlusion, (c) sensor failure, (d) viewpoint perturbation, and (e) light variation.

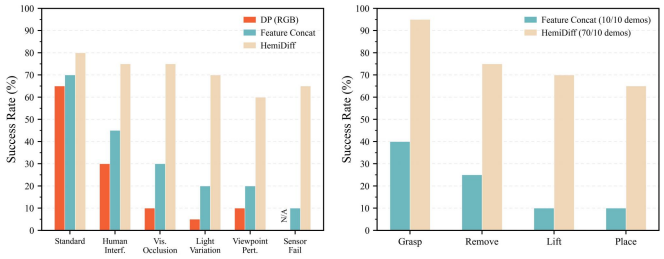

Fig. 11. Quantitative analysis. (a) Robustness under real-world disturbances. (b) Stage-wise success rates under heterogeneous visual–tactile data mixing.

results indicate that HemiDiff can better exploit unaligned and imbalanced multimodal demonstrations.

V. Conclusion

In this letter, we proposed HemiDiff, a compositional diffusion policy for coordinated bimanual manipulation under asymmetric multimodal observations. By integrating independent modality experts, an asymmetric perception router, and a coordination energy expert, HemiDiff enables arm-specific modality reweighting and coordination-aware denoising during inference. Its modular composition alleviates feature conflicts and allows flexible handling of incomplete or degraded observations without retraining. Experiments in simulation and on real robots show that HemiDiff improves multimodal fusion, preserves inter-arm coordination, and achieves robust performance under challenging perception conditions such as occlusion, transparent objects, and contact uncertainty.

References

[1] G. Lu, T. Yu, H. Deng, S. S. Chen, Y. Tang, and Z. Wang, "Anybimanual: Transferring unimanual policy for general bimanual manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 13662–13672, 2025.

[2] H. Deng, W. Guo, Q. Wang, Z. Wu, and Z. Wang, "SafeBimanual: Diffusion-based trajectory optimization for safe bimanual manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 305 of *Proc. Mach. Learn. Res.*, pp. 3218–3238, 2025.

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. Robot. Sci. Syst. (RSS)*, 2023.

[4] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, vol. 44, no. 10–11, pp. 1684–1704, 2025.

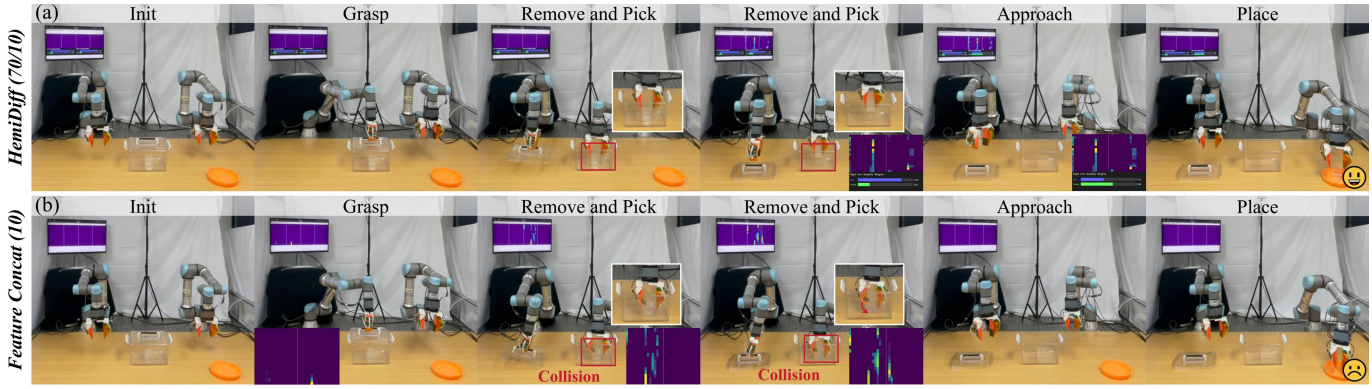

Fig. 12. Visualization of the heterogeneous data mixing experiment. (a) HemiDiff successfully completes the grasp, remove, and place stages. (b) Failure case of Feature Concat due to collisions during the grasp stage.

[5] T.-W. Ke, N. Gkanatsios, J. Xu, and K. Fragkiadaki, "Bi3D diffuser actor: 3D policy diffusion for bi-manual robot manipulation," in *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*, 2024.

[6] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations," in *Proc. Robot. Sci. Syst. (RSS)*, 2024.

[7] K. Black, N. Brown, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," in *Proc. Robot. Sci. Syst. (RSS)*, 2025.

[8] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 305 of *Proc. Mach. Learn. Res.*, pp. 17–40, 2025.

[9] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: A diffusion foundation model for bimanual manipulation," in *Int. Conf. Learn. Represent. (ICLR)*, pp. 29982–30009, 2025.

[10] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 270 of *Proc. Mach. Learn. Res.*, pp. 2557–2578, 2025.

[11] B. Huang, J. Xu, I. Akinola, W. Yang, B. Sundaralingam, R. O'Flaherty, D. Fox, X. Wang, A. Mousavian, Y.-W. Chao, and Y. Li, "VT-Refine: Learning bimanual assembly with visuo-tactile feedback via simulation fine-tuning," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 305 of *Proc. Mach. Learn. Res.*, pp. 2484–2500, 2025.

[12] X. Zhu, B. Huang, and Y. Li, "Touch in the Wild: Learning fine-grained manipulation with a portable visuo-tactile gripper," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[13] H. Chen, J. Xu, H. Chen, K. Hong, B. Huang, C. Liu, J. Mao, Y. Li, Y. Du, and K. Driggs-Campbell, "Multi-modal manipulation via multi-modal policy consensus," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2026.

[14] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proc. Robot. Sci. Syst. (RSS)*, 2024.

[15] X. Li, Y. Sun, L. Zhang, B. Huang, Y. Peng, Y. Meng, H. Jiang, S. Xie, G. Yao, A. Knoll, *et al.*, "DECO: Decoupled multimodal diffusion transformer for bimanual dexterous manipulation with a plugin tactile adapter," *arXiv preprint arXiv:2602.05513*, 2026.

[16] Q. Lv, H. Li, X. Deng, R. Shao, Y. Li, J. Hao, L. Gao, M. Y. Wang, and L. Nie, "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17394–17404, 2025.

[17] H. Chen, J. Xu, L. Sheng, T. Ji, S. Liu, Y. Li, and K. Driggs-Campbell, "Learning coordinated bimanual manipulation policies using state diffusion and inverse dynamics models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 5644–5651, 2025.

[18] T. Motoda, R. Hanai, R. Nakajo, M. Murooka, F. Erich, and Y. Domae, "Learning bimanual manipulation via action chunking and inter-arm coordination with transformers," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, pp. 1796–1801, 2025.

[19] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with 3D diffusion policies," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 2873–2880, 2025.

[20] W. Fan, H. Li, and D. Zhang, "CrystalTac: Vision-based tactile sensor family fabricated via rapid monolithic manufacturing," *Cyborg Bionic Syst.*, vol. 6, p. 0231, 2025.

[21] Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu, "Foar: Force-aware reactive policy for contact-rich robotic manipulation," *IEEE Robot. Autom. Lett.*, vol. 10, no. 6, pp. 5625–5632, 2025.

[22] Y. Du, S. Li, and I. Mordatch, "Compositional visual generation with energy based models," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6637–6647, 2020.

[23] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *Eur. Conf. Comput. Vis. (ECCV)*, pp. 423–439, 2022.

[24] L. Wang, J. Zhao, Y. Du, E. H. Adelson, and R. Tedrake, "PoCo: Policy composition from and for heterogeneous robot learning," in *Proc. Robot. Sci. Syst. (RSS)*, 2024.

[25] Z. Yang, J. Mao, Y. Du, J. Wu, J. B. Tenenbaum, T. Lozano-Pérez, and L. P. Kaelbling, "Compositional diffusion-based continuous constraint solvers," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 229 of *Proc. Mach. Learn. Res.*, pp. 3242–3265, 2023.

[26] J. Cao, Y. Huang, H. Guo, Q. Zhang, R. Zhang, W. Mai, M. Nan, J. Wang, H. Cheng, J. Sun, G. Han, W. Zhao, Y. Guo, Q. Zheng, X. Li, C. Song, P. Luo, and A. F. Luo, "Compose Your Policies! Improving diffusion-based or flow-based robot policies via test-time distribution-level composition," in *Int. Conf. Learn. Represent. (ICLR)*, 2026.

[27] M. Lauri, D. Hsu, and J. Pajarinen, "Partially observable Markov decision processes in robotics: A survey," *IEEE Trans. Robot.*, vol. 39, no. 1, pp. 21–40, 2022.

[28] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 3603–3613, 2019.

[29] M. Grotz, M. Shridhar, Y.-W. Chao, T. Asfour, and D. Fox, "PerAct2: Benchmarking and learning for robotic bimanual manipulation tasks," in *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*, 2024.

[30] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3019–3026, 2020.

[31] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 12156–12163, 2024.

[32] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "RVT: Robotic view transformer for 3D object manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 229 of *Proc. Mach. Learn. Res.*, pp. 694–710, 2023.

[33] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior generation with latent actions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 235 of *Proc. Mach. Learn. Res.*, pp. 26991–27008, 2024.