Multi-Modal Manipulation via Multi-Modal Policy Consensus

Anonymous CVPR submission

Paper ID *****

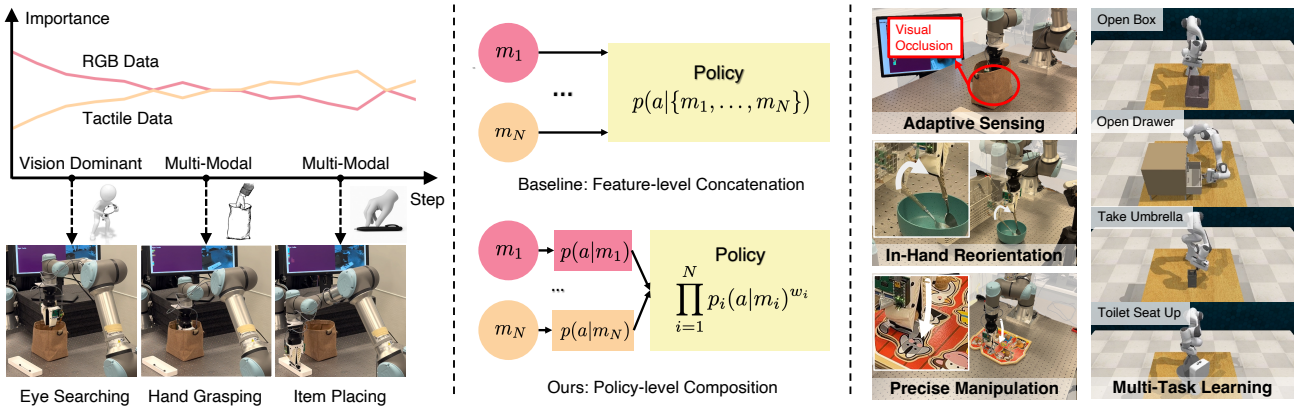

Figure 1. **Representation-Composable Policy.** (a) Perturbation-based importance analysis in the occluded marker picking task shows that vision dominates early, while tactile signals become important once occluded, demonstrating that our framework dynamically utilizes different modalities across task phases. (b) Classical feature concatenation vs. our policy-level composition, where m_i denotes a modality (e.g., RGB, point cloud, tactile, or learned visual feature). Our compositional design allows individual modality policies to be added or removed without retraining the entire network. (c) Our method unlocks key capabilities. These include *Adaptive Sensing*, retrieving an occluded marker using tactile feedback during occlusion; *In-Hand Reorientation*, reorienting a spoon within the gripper; *Precise Manipulation*, inserting a puzzle piece with fine-grained control; and *Multi-Task Learning*, consistently outperforming prior work across diverse tasks in RLBench.

Abstract

Effectively integrating diverse sensory modalities is crucial for robotic manipulation. However, the typical approach of feature concatenation is often suboptimal: dominant modalities such as vision can overwhelm sparse but critical signals like touch in contact-rich tasks, and monolithic architectures cannot flexibly incorporate new or missing modalities without retraining. Our method factorizes the policy into a set of diffusion models, each specialized for a single representation (e.g., vision or touch), and employs a router network that learns consensus weights to adaptively combine their contributions, enabling incremental of new representations. We evaluate our approach on simulated manipulation tasks in RLBench, as well as real-world tasks such as occluded object picking, in-hand spoon reorientation, and puzzle insertion, where it significantly outperforms feature-concatenation baselines on scenarios requiring multimodal reasoning. Our policy further demonstrates robustness to physical perturbations and sensor corruption. We further conduct perturbation-based importance analysis, which reveals adaptive shifts between modalities.

1. Introduction

Modern robots utilize a diverse array of modalities including RGB images, point clouds, tactile signals, and learned visual features, yet effectively integrating these data streams remains a challenge in robotics [6, 25]. These modalities can be both complementary (vision vs. touch) and overlapping (RGB-D vs. point cloud), requiring structured synergy for optimal performance. For decades, a common baseline has been the concatenation of feature-level modalities into a single high-dimensional vector, an approach that persists even in recent policies [18, 25]. Despite its popularity, this approach lacks a principled mechanism for balancing contributions across modalities and cannot easily adapt when modalities are added or missing, often resulting in suboptimal performance.

The weakness becomes apparent when different sensors are crucial at different task phases. Consider a robot trained to retrieve an object from an opaque container. For 90% of the trajectory, vision guides the robot to approach and position its gripper. Once the gripper enters the container, however, vision becomes useless while tactile signals become critical for success. Despite tactile information being essential during this crucial 10% of the task, feature concatenation struggles with this sparsity: the learning algorithm downweights the rarely-active tactile stream as noise, focusing instead on the statistically dominant visual features. Moreover, these monolithic approaches cannot adapt to new sensors or missing modalities without complete retraining. These limitations point to the need for a structured alternative that treats each modality as a distinct contributor rather than forcing premature fusion.

To address these challenges, we draw inspiration from compositional generative models [12, 13, 15, 31] and factorize robot policies into *modality-specific experts*, with a router network learning consensus among them. Each expert specializes in a single modality (e.g., vision for geometric reasoning or tactile for contact dynamics) and captures its behavioral constraints [41]. A router network learns consensus weights during training to combine these experts into a unified policy, ensuring that even rarely-active but crucial modalities (e.g., tactile) retain their influence when most needed. As illustrated in Figure 1b, our approach contrasts with classical feature-level concatenation by composing experts at the policy level, which not only mitigates sparsity issues but also allows new modalities to be added or removed without retraining the entire network. This compositional design directly resolves the limitations of prior work by preserving the contribution of each modality, adapting flexibly to sensor availability, and providing robustness to failures, while also enabling diverse manipulation skills such as those shown in Figure 1c.

Building on this compositional structure, our framework enables context-aware shifts in modality reliance: tactile experts dominate during contact-rich phases, while vision governs geometric reasoning in free space. The unified policy, formed by weighted sums of score functions from the factorized diffusion experts, yields robustness under sensor corruption and partial observability, ensuring reliable manipulation across diverse conditions.

Our contributions are as follows:

1. We introduce a framework that composes modality-specific experts through learned consensus weights, offering a principled and extensible alternative to monolithic feature concatenation. This design naturally supports incremental learning, allowing new experts to be integrated without retraining existing policies.
2. We demonstrate strong performance on the multi-task RLBench [22] simulation benchmark compared to various baselines and validate our approach on complex real-world tasks, including occluded marker picking, spoon reorientation in hand, and puzzle insertion. We also show robustness to physical perturbations, runtime disturbances, and sensor corruption.
3. We provide comprehensive analyses, including perturbation-based importance studies, that quantitatively demonstrate how our policy learns to shift reliance between modalities in response to changing task context.

2. Related Works

Multimodal Fusion in Robotics. The integration of vision and touch is critical for robust robotic manipulation, allowing robots to handle occlusions and make precise contact [3, 21, 24]. Vision provides global scene understanding while tactile sensing offers high-fidelity local information critical for tasks like grasping and insertion [11, 38]. Existing fusion approaches range from simple feature concatenation [25, 26] to sophisticated architectures including Visuo-Tactile Transformers that dynamically weigh modality features [9], stage-guided fusion [16], cross-modal attention mechanisms, and unified modality methods [20, 42]. However, all these approaches share a common limitation: they perform *feature-level* fusion to create a single conditional input for a monolithic policy. This makes them vulnerable to sparsity, where learning becomes biased toward statistically dominant but contextually irrelevant modalities (e.g., vision) while discarding essential but sparse signals (e.g., touch). Our work addresses this issue by factorizing policies into modality-specific modules and composing complete action distributions at the *policy level*.

Compositional Generation and Energy-Based Models. Our framework builds on compositional modeling, particularly the principle of combining simple models as a product of distributions to form a more expressive joint distribution [12, 30, 31, 39]. This idea is central to Energy-Based Models (EBMs), where summing energy functions corresponds to multiplying probability distributions [13, 17]. Diffusion models can be viewed as score-matching EBMs [36], making their score functions naturally composable in the same way. Beyond image generation, compositionality has been applied to trajectory modeling [1, 23], language models [14], robotic planning with constraints [32, 41], and hierarchical planning [2]. More broadly, compositional generative models extend to domains such as traffic generation [29] and human motion [35, 37], highlighting the versatility of compositionality as a modeling principle. In robotics, related work has explored combining policies across heterogeneous sources such as simulation and real-world data [40]. In contrast, our approach focuses on compositional learning across diverse modalities, enabling adaptive integration

of modalities such as vision, touch, and semantic features within a unified policy.

Diffusion Models in Robot Learning. Diffusion models have emerged as state-of-the-art policy representations in robotics, capable of modeling complex, multimodal action spaces. They have been successfully applied to imitation learning for single-arm [10, 33, 43] and bimanual manipulation [7], tool use [8], motion planning [23, 34], and reinforcement learning [1]. Extensions of diffusion planning incorporate hierarchical structures [27], and domain-specific adaptations such as autonomous driving [28]. Concurrent work by Zhang et al. [4] explores composition of RGB and point cloud policies using fixed, manually-tuned weights, but is limited to visual modalities without adaptive weighting. Our framework introduces a structured approach to compositional learning that accommodates arbitrary modalities, including tactile feedback and semantic features, through learned context-dependent routing for adaptive modality weighting.

3. Approach

Our work introduces a compositional approach to multimodal robot learning that addresses sparsity in sensory modalities through policy factorization and consensus-based composition (see Figure 2). We ground our framework in energy-based composition principles and instantiate them via policy consensus across modalities for robotic manipulation.

3.1. Problem Formulation

We consider the imitation learning setting for robot manipulation. Given a dataset of expert demonstrations $\mathcal{D} = \{(\mathbf{s}_t, \mathbf{a}_t)\}_{t=1}^{T}$, where $\mathbf{s}_t$ is the state and $\mathbf{a}_t$ the action at timestep t, the state comprises N sensory modalities:

$$\mathbf{s}_t = \{m_{1,t}, m_{2,t}, \ldots, m_{N,t}\}.$$

Our goal is to learn a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$ that leverages heterogeneous information within $\mathbf{s}_t$. Each raw modality $m_{i,t}$ is encoded into a latent embedding $\mathbf{e}_{i,t}$ through a modality-specific encoder. Here, an embedding $\mathbf{e}_{i,t}$ refers to the modality encoding together with relevant robot state information (e.g., joint angles, gripper status), providing a richer context for action prediction. We denote $\mathbf{a}_t$ as the ground-truth action in the dataset, a as a generic action candidate, and a^k as its noised version at diffusion timestep k. For each modality i, we train K_i sub-policies $p_{i,j}$, parameterized by $\theta_{i,j}$, which define energy functions $E_{\theta_{i,j}}(a, m_{i,t})$ and corresponding diffusion scores $\epsilon_{\theta_{i,j}}$. In our experiments, we set $K_i = 2$ to capture complementary behavioral modes (e.g., geometry vs. fine detail for vision, contact onset vs. sustained force for tactile), although the formulation allows general K_i. A router network R_ψ maps the embeddings $\{\mathbf{e}_{i,t}\}$ to consensus weights $\{w_i\}$, normalized by softmax so that $\sum_i w_i = 1$, indicating the relative influence of each modality.

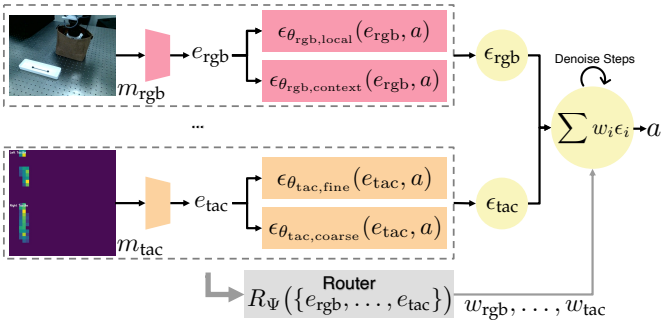

Figure 2. **Overview of Our Compositional Policy Framework.** Raw sensory modalities ($m_{\text{rgb}}, m_{\text{tac}}$) are encoded into embeddings ($\mathbf{e}_{\text{rgb}}, \mathbf{e}_{\text{tac}}$). Each modality is factorized into complementary sub-policies (e.g., $\epsilon_{\theta_{\text{rgb,context}}}(e_{\text{rgb}}, a)$, $\epsilon_{\theta_{\text{rgb,local}}}(e_{\text{rgb}}, a)$, $\epsilon_{\theta_{\text{tac,coarse}}}(e_{\text{tac}}, a)$, $\epsilon_{\theta_{\text{tac,fine}}}(e_{\text{tac}}, a)$), which produce score predictions that are averaged into a modality-specific score. A router network $R_\psi(\mathbf{e}_{\text{rgb}}, \ldots, \mathbf{e}_{\text{tac}})$ then predicts consensus weights $\{w_i\}$ to reconcile these modality-specific scores into the final composed score $\sum_i w_i \epsilon_i$, which defines the policy for action generation.

3.2. Energy-Based Policy Composition

Our approach builds on energy-based composition principles [13, 15, 31, 40], where policies are viewed as energy functions that assign low energy to preferred actions and high energy otherwise. Each base policy $p_{i,j}(a|m_{i,t})$ implicitly defines an energy function through

$$p_{i,j}(a|m_{i,t}) \propto \exp(-E_{\theta_{i,j}}(a, m_{i,t})).$$

Composing such policies corresponds to multiplying distributions, or equivalently summing their energies:

$$p(a|\mathbf{s}_t) \propto \prod_{i=1}^{N} \left[\prod_{j=1}^{K_i} \exp(-E_{\theta_{i,j}}(a, m_{i,t})) \right]^{w_i}, \tag{202}$$

$$= \exp\left(-\sum_{i=1}^{N} w_i \sum_{j=1}^{K_i} E_{\theta_{i,j}}(a, m_{i,t}) \right). \tag{204}$$

This reveals that our composition performs a weighted sum of energy functions, where the router weights $\{w_i\}$ determine each modality's influence based on the current state. Unlike feature concatenation, which forces all modalities through a shared network that tends to suppress statistically rare signals, our formulation preserves separate energy functions for each modality, ensuring that sparse but critical signals remain influential.

3.3. Compositional Policy Factorization

We factorize the policy at two levels to capture both inter- and intra-modality structure:

$$p(a|\mathbf{s}_t) \propto \prod_{i=1}^{N} p_i(a|m_{i,t})^{w_i},$$

$$p_i(a|m_{i,t}) \propto \prod_{j=1}^{K_i} p_{i,j}(a|m_{i,t}).$$

Here p_i is the composite policy for modality i, while $p_{i,j}$ are complementary sub-policies (e.g., vision experts for geometry vs. fine detail, tactile experts for contact onset vs. sustained force).

This product-of-distributions view admits a constraint satisfaction interpretation [41], where each modality-specific policy imposes behavioral constraints on the final action (e.g., geometry from vision, contact dynamics from touch). Importantly, unlike feature concatenation where sparse signals compete with dominant ones for network capacity, each p_i is trained independently on its own modality stream. This ensures that rarely active but task critical modalities, such as tactile inputs that appear only during contact, retain their influence through the learned consensus weights w_i, which provide a consistent balancing of modality influence across the policy.

3.4. Score-Based Implementation via Diffusion Models

We implement each base policy $p_{i,j}$ using Denoising Diffusion Probabilistic Models (DDPMs) [19]. Diffusion models can be interpreted as score-matching energy-based models [36], and following compositional generation principles [13, 31], sampling from a product of distributions corresponds to summing their score functions. This leads to a two-step aggregation process at each denoising step k:

Intra-Modality Composition. The composed score for modality i is obtained by averaging scores of its K_i factorized sub-policies:

$$\epsilon_{i,\text{comp}}(a^k, m_{i,t}, k) = \tfrac{1}{K_i} \sum_{j=1}^{K_i} \epsilon_{\theta_{i,j}}(a^k, m_{i,t}, k).$$

Inter-Modality Composition. The final composed score is then computed using router-weighted combination:

$$\epsilon_{\text{comp}}(a^k, \mathbf{s}_t, k) = \sum_{i=1}^{N} w_i(\mathbf{s}_t) \cdot \epsilon_{i,\text{comp}}(a^k, m_{i,t}, k).$$

The modality-specific scores $\epsilon_{i,\text{comp}}$ define gradient fields that encode each modality's behavioral constraints. The router assigns consensus weights $w_i(\mathbf{s}_t)$ to combine these fields into a unified score, balancing their contributions. This weighted composition connects energy based composition with score based diffusion, establishing a direct theoretical link between compositional energy models and diffusion-based policies.

3.5. Router Network

The router R_ψ maps modality-specific embeddings $\{\mathbf{e}_i\}$ to consensus weights $\{w_i\}$, normalized with a softmax so that they are positive and sum to one, making them interpretable as relative modality influence. Conceptually, the router reconciles the action proposals of modality-specific experts into a unified policy. The consensus weights are learned during training and fixed at execution, providing a interpretable mechanism for balancing modalities.

3.6. Advantages Over Existing Fusion

Our framework offers (1) **Robustness to sparsity**, as rarely active modalities such as tactile remain represented by separate experts rather than being suppressed. (2) **Modularity**, as experts can be trained and extended independently without retraining the entire policy. (3) **Interpretability**, with consensus weights $\{w_i\}$ directly revealing the influence of each modality. (4) **Principled consensus**, as the router provides a consistent weighting scheme grounded in energy-based composition, rather than blind feature fusion.

These properties make our framework a theoretically grounded and extensible alternative to monolithic feature concatenation for multi-modal robot learning.

4. Experiment

We evaluate our modality-composable policy framework by addressing three key research questions: (1) How does a compositional architecture compare to feature-level fusion in tasks with modality sparsity? (2) Does the compositional model capture context-dependent and modified reliance on different modalities, allowing new experts to be composed without retraining? (3) Does the policy maintain robustness under physical perturbations and sensor corruption?

4.1. Experimental Setup

Tasks. We evaluate our method on both simulation and real-world manipulation tasks. In simulation, we use RLBench [22] as a multi-task benchmark with four tasks: *open box*, *open drawer*, *take umbrella out of stand*, and *toilet seat up*. The policy is trained on a total of 200 demonstration episodes and evaluated on 200 unseen configurations.

For real-world experiments, we employ a UR5e manipulator equipped with dual cameras and tactile sensors, as shown in Figure 3(a). We evaluate three challenging manipulation tasks, with Figure 3(b–d) showing overlays of their initial testing conditions: (i) *occluded marker picking*; (ii) *spoon reorientation*; and (iii) *puzzle insertion*.

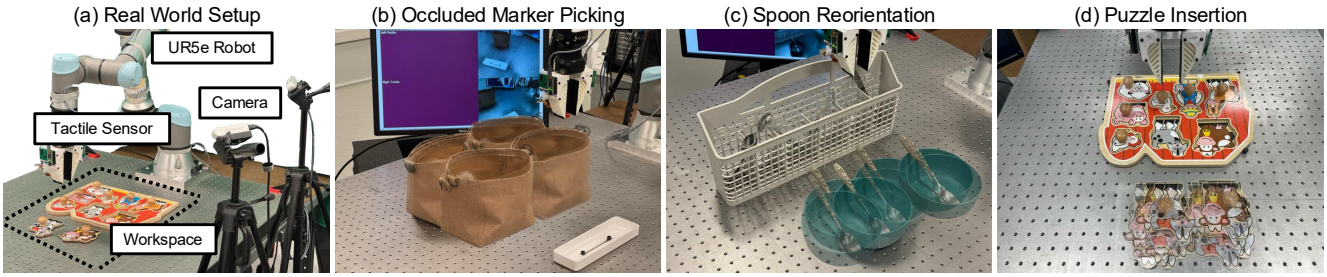

Figure 3. **Real-World Experimental Setup.** (a) UR5e manipulator equipped with dual cameras and tactile sensors. (b–d) Overlays of initial conditions for the evaluation tasks: occluded marker picking, spoon reorientation, and puzzle insertion.

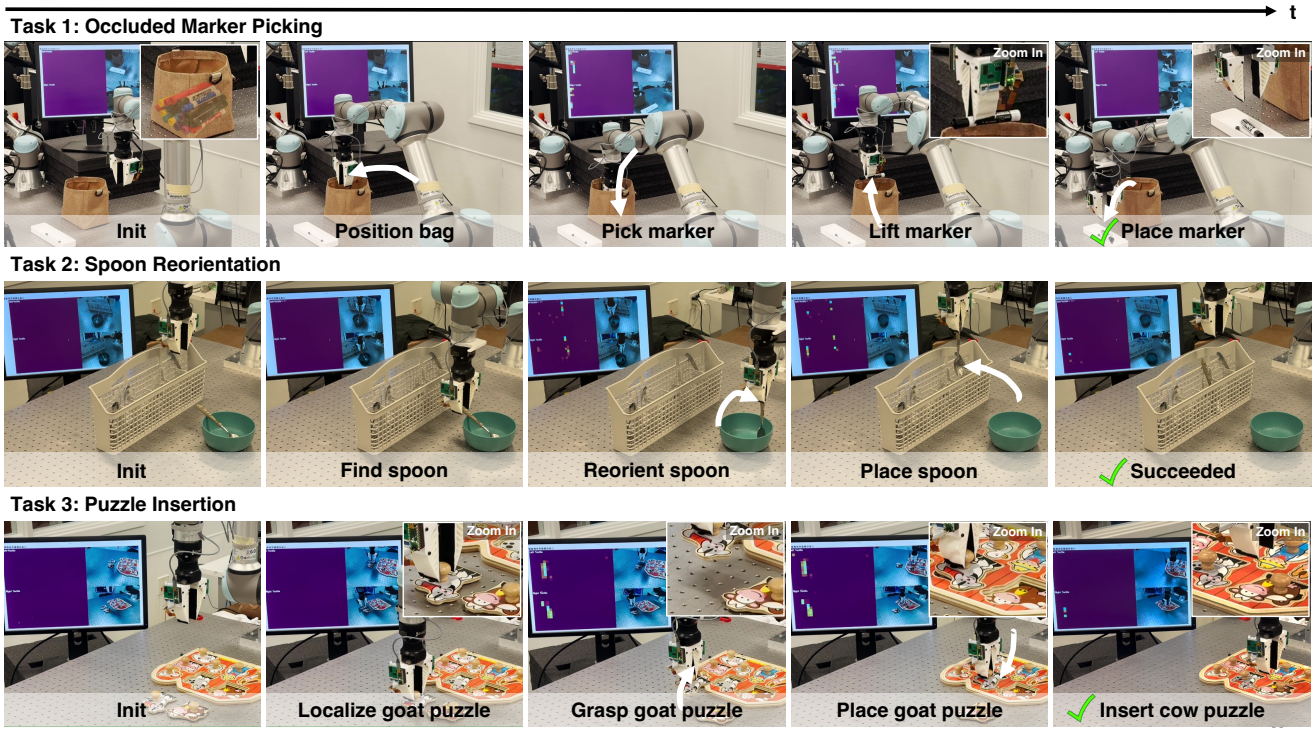

Figure 4. **Qualitative Policy Rollouts.** Representative execution traces from three tasks: Task 1 occluded marker picking, where tactile feedback guides manipulation when vision is unavailable; Task 2 spoon reorientation, demonstrating dexterous in-hand manipulation; Task 3 puzzle insertion, requiring high-precision alignment at millimeter accuracy.

We collect 80, 60, and 50 teleoperated demonstrations for these tasks, respectively.

Sensory Modalities. In simulation, we utilize RGB images from two cameras, point cloud (PCD) data from depth sensors, and 3D semantic features extracted using a pretrained DINO model [5]. For real-world experiments, we use RGB images from dual side-view Intel RealSense D415 cameras with a resolution of 96×128, along with dense tactile arrays from *FlexiTac* sensors [20]. Each finger is equipped with a tactile pad consisting of 12×32 sensing units, each with a spatial resolution of 2 mm.

Baselines. We compare against comprehensive baselines tailored to each domain:

- **Simulation:** (i) Single-modality policies trained on RGB-only, PCD-only, or DINO-only inputs; (ii) Feature concatenation combining all modality embeddings; (iii) Factorized MoE fusion using soft routing.
- **Real-world:** (i) RGB-only policy; (ii) RGB+Tactile feature concatenation baseline.

Metrics. We report success rate as the primary metric and completion time as a secondary metric for successful trials.

4.2. Main Results

Simulation Performance. Table 1 presents our simulation results. Our method achieves the highest average suc-

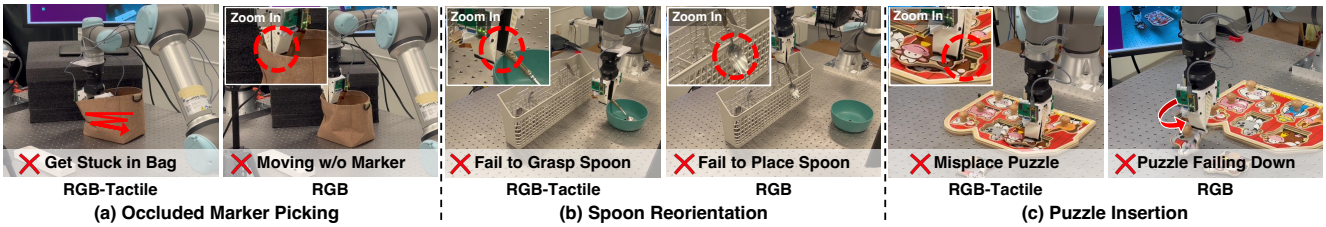

Figure 5. **Typical Failure Cases of Baseline Methods.** We show failure cases of an RGB-only policy compared with an RGB+Tactile concatenation baseline. Each task highlights the complementary roles of the two modalities: vision provides global spatial and geometric information, while tactile sensing provides contact awareness and fine-grained grasp feedback. (a) In occluded marker picking, the concatenation baseline becomes trapped without grasping, while RGB-only lacks awareness of the grasp state once occluded. (b) In spoon reorientation, the concatenation baseline fails at initial grasping, while RGB-only fails at precise placement. (c) In puzzle insertion, the concatenation baseline causes misalignment, while RGB-only suffers frequent grasp failures.

cess rate (66%), significantly outperforming both single-modality policies and the feature concatenation baseline (56%). This 18% relative improvement is achieved with minimal parameter overhead (+0.7M parameters, where M denotes millions, corresponding to only a 0.3% increase), demonstrating the efficiency of our compositional approach compared to naive fusion strategies.

Table 1. Policy performance and parameter count on RLBench tasks. Our method achieves 18% relative improvement over concatenation baseline with negligible parameter increase.

Method	Success Rate	Params (M)
Single-Modality		
RGB only	0.54	257.3
Point Cloud only	0.49	251.9
3D DINO only	0.45	251.9
Multi-Modality		
Concatenation	0.56	262.9
Ours	**0.66**	**263.6**

Real-World Performance. Tables 2–4 demonstrate consistent superiority across all real-world tasks. In contrast to baselines, our method achieves the highest success rates (65% for occluded picking, 75% for spoon reorientation, and 52% for puzzle insertion) while maintaining the lowest average completion times. Figure 4 provides qualitative evidence of these performance differences.

Notably, the RGB+Tactile concatenation baseline exhibits catastrophic failure: in *Occluded Marker Picking*, its success rate is only 5% compared to 35% for RGB-only, and in *Spoon Reorientation*, it achieves just 21% compared to 67% for RGB-only. These results confirm that naive fusion can be actively detrimental when modalities have varying informativeness. Figure 5 illustrates these systematic failure modes: the concatenation baseline gets trapped without successful grasping in occluded picking, fails at initial grasping in spoon reorientation, and causes misalignment in puzzle insertion, while RGB-only policies suffer from lack of tactile feedback during critical manipulation phases.

Table 2. Decomposed Success Rates on Real-World Puzzle Insertion Task.

Sub-Task	RGB	RGB+Tac.	Ours
Pick up goat	0.90	0.90	**1.00**
Place goat	0.90	0.90	**0.95**
Task success (goat)	0.25	0.35	**0.58**
Pick up cow	0.85	0.80	**0.95**
Place cow	0.75	0.80	**0.95**
Task success (cow)	0.30	**0.45**	**0.45**

Table 3. Performance on Occluded Marker Picking. Failed trials counted as 120s timeout.

Metric	RGB	RGB+Tac.	Ours
Success Rate	0.35	0.05	**0.65**
Avg. Time (s)	107.8	117.9	**96.5**

Table 4. Performance on Spoon Reorientation. Failed trials counted as 300s timeout.

Metric	RGB	RGB+Tac.	Ours
Success Rate	0.67	0.21	**0.75**
Avg. Time (s)	117.1	221.5	**94.8**

4.3. Analysis of Learned Modality Dependencies

To assess whether the policy leverages different modalities depending on task context, we conduct a perturbation-based importance analysis. Specifically, we measure modality dependency by injecting calibrated Gaussian noise $\mathcal{N}(0, \sigma^2)$ into each modality and computing the normalized L2 distance between perturbed and original action outputs. For stability, temporal smoothing is applied using an exponential moving average (EMA, $\alpha = 0.1$).

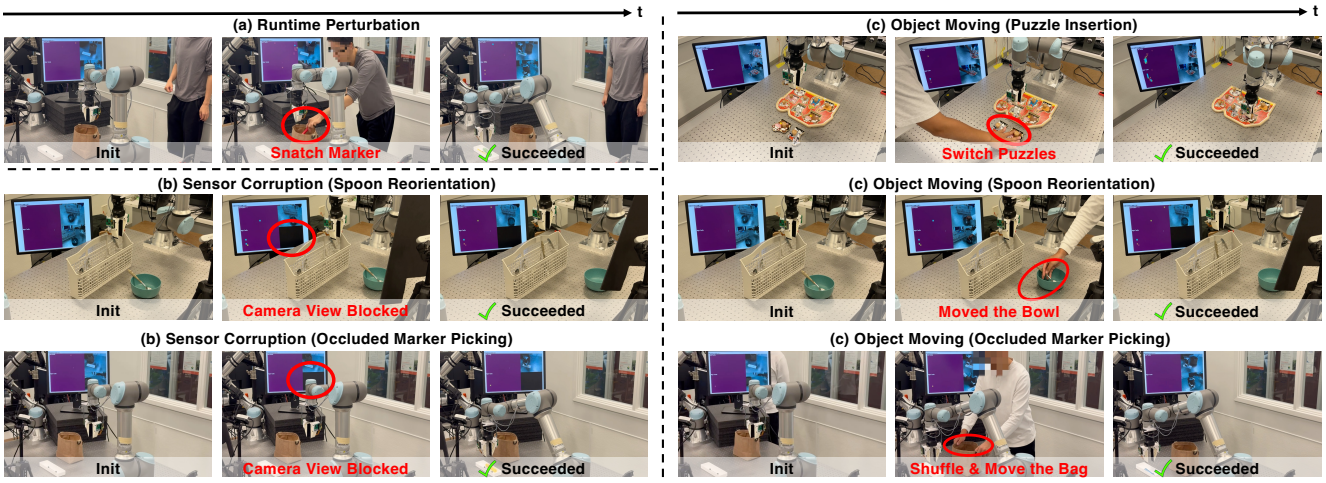

Figure 6. **Policy Robustness under Diverse Perturbations.** We evaluate three types of interventions: (a) runtime perturbation, where the marker is suddenly snatched away during execution; (b) sensor corruption, where a camera is occluded to simulate partial sensor failure; and (c) object repositioning, where task-relevant objects are reset and moved to new positions between executions. Our method maintains reliable performance across all scenarios.

As shown in Figure 1a, this analysis reveals context aware modality usage during the *Occluded Marker Picking* task, which unfolds in two distinct phases. In the first stage (t = 0 to 5 steps), vision dominates: the policy shows high sensitivity to visual perturbations, relying on visual feedback for spatial navigation and approach. In the second stage (after 5 steps), the policy shifts to multi modal coordination. As the gripper enters the occluded container, tactile sensitivity rises sharply upon contact while vision remains important, highlighting their complementary roles. This emergent two stage behavior demonstrates that our compositional architecture can transition from single modality reliance to multi modal integration based on task demands.

4.4. Robustness and Adaptation Evaluation

We evaluate the robustness and adaptation of our approach through systematic perturbation experiments across physical and sensory domains.

Physical Perturbations. We introduce two categories of physical interventions to assess policy resilience under environmental disturbances (Figure 6a, c): (a) *Runtime perturbation*: a mid-execution intervention in which the marker is suddenly snatched away during the occluded marker picking task, introducing a dynamic disturbance. (c) *Object repositioning*: objects are deliberately reset and moved to new positions between task executions (e.g., relocating the bag in the occluded-marker task, displacing the bowl and spoon in the spoon-reorientation task, or swapping the positions of the goat and cow puzzle pieces).

Sensor Corruption. We simulate partial sensor failure by occluding one camera with an opaque card, eliminating visual input (Figure 6b). This corruption is applied to both spoon-reorientation and occluded-marker tasks. Our method maintains consistent success despite the degraded sensory input, demonstrating resilience to partial sensor failure.

Incremental Learning. A key benefit of our compositional framework is the ability to combine independently trained policies without retraining. We demonstrate this by composing an RGB policy and a tactile policy using fixed consensus weights $(0.5, 0.5)$. As shown in Figure 7, while the RGB-only policy fails under occlusion, the composed policy succeeds in grasping despite never being jointly trained. This highlights that our method supports incremental integration of new sensory modalities, enabling efficient extension of existing skills.

4.5. Ablation Studies

To validate our key architectural choices, we conduct comprehensive ablation studies examining the impact of different routing and fusion strategies. Table 5 summarizes the experimental results.

Learned vs. Fixed Consensus. We first investigate the contribution of our learned consensus mechanism by replacing it with a fixed equal-weight strategy. This ablation results in a significant performance degradation of 7.6% (from 66% to 61%), demonstrating that adaptive, learned consensus weighting is important for effective policy composition.

Policy-Level vs. Feature-Level Fusion. We compare our policy-level consensus-based composition against a soft-routing MoE baseline that performs fusion at the feature level. Our method achieves a substantial improvement of 15.8% over the MoE baseline (66% vs. 57%; see Table 5). This performance gap suggests that composing

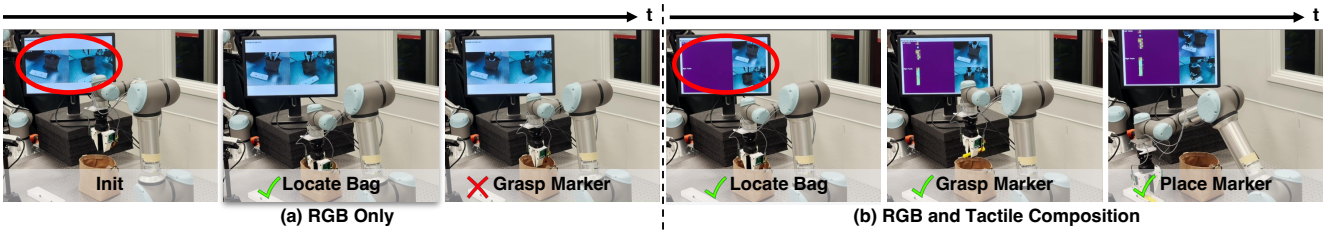

Figure 7. **Incremental Learning.** (a) An RGB-only policy fails to grasp the marker without tactile feedback. (b) By composing a pre-trained RGB policy with a tactile policy using manually set consensus weights $(0.5, 0.5)$, the combined policy successfully grasps the marker under occlusion without requiring retraining.

modalities at the action-output level better preserves modality-specific information compared to early feature selection, which may prematurely discard relevant signals needed for downstream decision-making.

Table 5. Ablation study comparing routing and fusion strategies. Results show average success rates across all evaluation tasks in simulation.

Method	Avg. Success Rate
Ours (Learned Router)	**0.66**
Ablation Variants	
Fixed Equal Weights	0.61
Factorized MoE Fusion	0.57

5. Conclusion

We presented a compositional framework for multimodal robot manipulation that factorizes policies at the modality level. The action distribution conditioned on each input modality is modeled by a separate expert policy, and their outputs are integrated through learned consensus weights from a router network. This consensus-based composition allows the policy to adaptively balance sensing modalities, preserving their expertise while enabling robust coordination. Experiments on a multi-task simulation benchmark and contact-rich real-world tasks demonstrate that our method consistently outperforms conventional feature-fusion baselines. Beyond these performance gains, our importance analysis reveals that policies dynamically shift reliance between modalities based on context, with vision handling geometric reasoning and tactile managing contact-rich phases. These compositional advantages enable incremental sensor deployment without retraining and provide natural robustness to sensor failures.

While promising, our framework opens several directions for future research. First, the current router adapts consensus weights at the policy level but does not provide fine-grained or temporally dynamic adjustments; developing more expressive consensus mechanisms could enhance adaptability. Second, our experiments focus primarily on vision and tactile inputs in controlled real-world environments, leaving extensions to additional modalities (e.g., audio, force, language) and deployment in more diverse settings as important next steps.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022. 2, 3

[2] Anurag Ajay, Song Han, Yilun Du, Shuang Li, Abhinav Gupta, Tommi Jaakkola, Joshua Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *arXiv preprint arXiv:2309.08587*, 2023. 2

[3] Antonio Bicchi and Vijay R. Kumar. Robotic grasping and contact: a review. *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, 1: 348–353 vol.1, 2000. 2

[4] Jiahang Cao, Qiang Zhang, Hanzhong Guo, Jiaxu Wang, Hao Cheng, and Renjing Xu. Modality-composable diffusion policy via inference-time distribution-level composition, 2025. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5

[6] Peixin Chang, Shuijing Liu, Haonan Chen, and Katherine Driggs-Campbell. Robot sound interpretation: Combining sight and sound in learning-based control. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 5580–5587. IEEE Press, 2020. 1

[7] Haonan Chen, Jiaming Xu, Lily Sheng, Tianchen Ji, Shuijing Liu, Yunzhu Li, and Katherine Driggs-Campbell. Learning coordinated bimanual manipulation policies using state diffusion and inverse dynamics models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 3

[8] Haonan Chen, Cheng Zhu, Shuijing Liu, Yunzhu Li, and Katherine Driggs-Campbell. Tool-as-interface: Learning

robot policies from observing human tool use. In *Conference on Robot Learning (CoRL)*, 2025. 3

[9] Yizhou Chen, Mark Van der Merwe, Andrea Sipos, and Nima Fazeli. Visuo-tactile transformers for manipulation. In *6th Annual Conference on Robot Learning*, 2022. 2

[10] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 3

[11] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. *arXiv preprint arXiv:2401.12024*, 2024. 2

[12] Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024. 2

[13] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In *Advances in Neural Information Processing Systems*, 2020. 2, 3, 4

[14] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. 2

[15] Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc, 2024. 2, 3

[16] Ruoxuan Feng, Di Hu, Wenke Ma, and Xuelong Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024. 2

[17] Nikolaos Gkanatsios, Ayush Jain, Zhou Xian, Yunchu Zhang, Christopher Atkeson, and Katerina Fragkiadaki. Energy-based Models are Zero-Shot Planners for Compositional Scene Rearrangement. In *Robotics: Science and Systems*, 2023. 2

[18] Johanna Hansen, Francois Hogan, Dmitriy Rivkin, David Meger, Michael Jenkin, and Gregory Dudek. Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8298–8304, 2022. 1

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 4

[20] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d vitac:learning fine-grained manipulation with visuo-tactile sensing. In *Proceedings of Robotics: Conference on Robot Learning(CoRL)*, 2024. 2, 5

[21] Zhe Huang, Ye-Ji Mun, Haonan Chen, Yiqing Xie, Yilong Niu, Xiang Li, Ninghan Zhong, Ha-Il You, David Livingston McPherson, and K. Driggs-Campbell. Towards safe multi-level human-robot interaction in industrial tasks. *ArXiv*, abs/2308.03222, 2023. 2

[22] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment, 2019. 2, 4

[23] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. 2, 3

[24] Roland S. Johansson and Göran Westling. Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research*, 56:550–564, 2004. 2

[25] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, page 8943–8950. IEEE Press, 2019. 1, 2

[26] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Proceedings of The 6th Conference on Robot Learning*, pages 1368–1378. PMLR, 2023. 2

[27] Wenhao Li, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. Hierarchical diffusion for offline decision making. In *International Conference on Machine Learning*, pages 20035–20064, 2023. 3

[28] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024. 3

[29] Haohong Lin, Xin Huang, Tung Phan-Minh, David S Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric M Wolff, and Hongge Chen. Causal composition diffusion model for closed-loop traffic generation. *arXiv preprint arXiv:2412.17920*, 2024. 2

[30] Nan Liu, Shuang Li, Yilun Du, Joshua B. Tenenbaum, and Antonio Torralba. Learning to compose visual relations. In *Advances in Neural Information Processing Systems*, 2021. 2

[31] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 2, 3, 4

[32] Udit Arora Mishra, Shuchen Xue, Yifan Chen, and Danfei Xu. Generative skill chaining: Long-horizon skill planning with diffusion models. In *Conference on Robot Learning*, pages 2905–2925. PMLR, 2023. 2

[33] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 3

[34] Kallol Saha, Vishal Mandadi, Jayaram Reddy, Ajit Srikanth, Aditya Agarwal, Bipasha Sen, Arun Singh, and Madhava Krishna. Edmp: Ensemble-of-costs-guided diffusion for motion planning. *arXiv*, 2023. 3

[35] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 4

[37] Shanlin Sun, Gabriel De Araujo, Jiaqi Xu, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. Coma: Compositional human motion generation with multi-modal agents. *arXiv preprint arXiv:2412.07320*, 2024. 2

[38] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, page adl0628, 2024. 2

[39] Julen Urain, Anqi Li, Puze Liu, Carlo D'Eramo, and Jan Peters. Composable energy policies for reactive motion generation and reinforcement learning. *The International Journal of Robotics Research*, 42(10):827–858, 2023. 2

[40] Lirui Wang, Jialiang Zhao, Yilun Du, Edward H Adelson, and Russ Tedrake. Poco: Policy composition from and for heterogeneous robot learning. *arXiv preprint arXiv:2402.02511*, 2024. 2, 3

[41] Zhutian Yang, Jiayuan Mao, Yilun Du, Jiajun Wu, Joshua B. Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Compositional diffusion-based continuous constraint solvers. In *Proceedings of The 7th Conference on Robot Learning*, pages 3242–3265. PMLR, 2023. 2, 4

[42] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. *arXiv preprint arXiv:2312.01853*, 2023. 2

[43] Xinyue Zhu, Binghao Huang, and Yunzhu Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *arXiv preprint arXiv:2507.15062*, 2025. 3