Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Diffusion-based Vision-Tactile Policy Learning with Uncertainty-Aware Modality Scheduling for Contact-Rich Robotic Manipulation Tasks

Sofia E. Svensson

Department of Electrical Engineering, Faculty of Science and Technology, Uppsala University, Uppsala, Sweden

Abstract

Contact-rich robotic manipulation, such as precision assembly and intricate tool use, requires the seamless integration of visual and tactile feedback. Traditional policies often struggle with dynamic modality reliability, where vision suffers from severe occlusions during close-proximity interactions, and tactile sensing becomes uninformative when the end-effector breaks contact. To address these limitations, this paper introduces a novel framework integrating diffusion-based policy learning with an uncertainty-aware modality scheduling mechanism. We conceptualize action generation as a conditional denoising diffusion process, which intrinsically captures multimodal action distributions and mitigates the compounding errors typical of deterministic behavior cloning. Furthermore, we propose an adaptive scheduling mechanism that continuously evaluates the epistemic and aleatoric uncertainties of the vision and tactile encoders. By dynamically reweighting the contributions of each modality based on their real-time predictive confidence, the system seamlessly transitions from vision-dominant pre-contact phases to tactile-dominant manipulation phases without relying on hard-coded heuristics. Extensive evaluations in complex manipulation tasks, including tight-tolerance peg-in-hole insertion and contour following, demonstrate that our approach significantly outperforms existing multimodal fusion baselines. The dynamic scheduling prevents catastrophic failures caused by noisy sensory inputs, thereby improving overall task success rates and sample efficiency. This research provides a robust foundation for deploying generalist robotic agents in unstructured, contact-rich environments.

Keywords

Robotic Manipulation, Diffusion Models, Multimodal Fusion, Tactile Sensing, Uncertainty Estimation

1. INTRODUCTION

1.1 Background on Contact-Rich Robotic Manipulation

The deployment of robotic systems in unstructured environments hinges on their ability to execute contact-rich manipulation tasks with high precision and adaptability. Grounding language in robotic affordances distinguishes an instruction that sounds appropriate from an action that the current environment permits [1]. These tasks, which encompass industrial assembly, domestic object manipulation, and surgical assistance, inherently require the management of complex physical interactions between the robotic end-effector and the environment. Unlike free-space trajectory planning, contact-rich tasks involve continuous making and breaking of physical contact, leading to hybrid dynamics that are notoriously difficult to model analytically. In conventional robotic control, these problems are typically addressed using highly structured environments and precise state estimation, combined with impedance or admittance control schemes. However, such classical approaches suffer from a lack of generalizability, as they rely heavily on accurate physical parameters, exact geometries, and tightly constrained operating conditions. Consequently, machine learning paradigms, particularly deep reinforcement learning and imitation learning, have emerged as promising alternatives. These data-driven methods allow robots to acquire complex manipulation skills directly from sensorimotor experiences, bypassing the need for explicit analytical modeling of contact dynamics.

Despite the significant advancements in end-to-end learning for robotic manipulation, developing policies that are robust to sensory noise and environmental variations remains a formidable challenge. RT-2 examines transfer from visual-language knowledge to robotic actions, making grounding of the action space a distinct evaluation requirement [2]. Most contemporary policies rely predominantly on vision, utilizing cameras to observe the workspace and plan actions. While vision provides dense spatial information essential for reaching and initial alignment, its utility diminishes rapidly as the robot approaches the target object. This phenomenon, often referred to as the occlusion problem, occurs because the robotic manipulator itself blocks the camera's line of sight precisely when fine-grained, millimeter-level adjustments are most critical. To bridge this sensory gap, recent research has increasingly turned toward tactile sensing, which offers direct, localized feedback regarding contact forces, surface textures, and micro-geometry. The integration of tactile data is widely considered the crucial next step for achieving human-level dexterity in robotic systems.

1.2 The Role of Vision and Tactile Sensing

The complementary nature of visual and tactile sensing forms the biological basis for human manual dexterity, and replicating this synergy is essential for advanced robotic manipulation. Vision functions as a global, feed-forward modality, providing the macro-level context required for scene understanding, obstacle avoidance, and pre-grasp orientation. It operates optimally at a distance but is highly susceptible to external lighting variations, specularity, and physical occlusion. Conversely, tactile sensing operates as a local, feedback-driven modality. High-resolution optical tactile sensors provide detailed information about contact patches, including normal and shear forces, object slip, and localized geometry. However, tactile sensors are entirely uninformative prior to contact and have a highly restricted receptive field. Fang and colleagues examine empowering large language models to progressively understand and generate 3D mesh [3]. The transition between these modalities during a typical manipulation task is neither binary nor instantaneous. During an insertion task, for example, the robot must rely on vision to bring the peg near the hole, use a combination of vision and incidental tactile feedback during the initial search phase, and depend almost entirely on tactile feedback for the final alignment and insertion when vision is fully occluded. Therefore, a robust robotic policy must not only process high-

dimensional data from both modalities but also dynamically determine which modality is most reliable at any given moment. Static fusion techniques, such as simple concatenation of feature embeddings, fail to capture this temporal dynamic and often lead to sub-optimal performance when one modality provides noisy or misleading information. Yang and colleagues examine intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [4].

1.3 Challenges in Modality Fusion and Policy Learning

Learning a multimodal policy for contact-rich tasks introduces two distinct but interrelated challenges: representational capacity and modality scheduling. First, traditional behavior cloning approaches typically utilize deterministic continuous action spaces, modeling the policy as a point estimate (e.g., via a Gaussian distribution with fixed variance). In complex, contact-rich environments, the optimal action distribution is often highly multimodal. For instance, there may be multiple valid ways to orient an object or recover from a slight misalignment. Forcing a deterministic model to approximate a multimodal distribution typically results in mode-averaging behavior, causing the robot to execute unsafe or physically impossible actions. Related work has addressed scene-Aware enhanced Edge-Cloud collaborative industrial vision inspection with multimodal LLM [5].

Second, the problem of modality scheduling requires the policy to selectively attend to the most informative sensory stream while actively ignoring corrupted or uninformative data. Early fusion strategies, which concatenate raw sensory inputs or low-level features, force the neural network to implicitly learn the scheduling logic. In practice, these networks frequently suffer from modality collapse, where the policy overfits to the dominant modality (usually vision due to its higher dimensionality and variance) and ignores the weaker modality. Late fusion strategies attempt to solve this by learning separate policies and averaging their outputs, but this fails to capture the synergistic cross-modal interactions necessary for complex tasks. Another relevant line of investigation concerns modality asymmetries and circuit dynamics in modern transformers [6].

1.4 Proposed Approach and Contributions

To address the aforementioned challenges, this paper presents a novel framework that synergizes diffusion-based policy learning with an uncertainty-aware modality scheduling mechanism. We model the behavior cloning process as a conditional denoising diffusion probabilistic model, which provides the representational capacity to capture complex, multimodal action distributions inherent in contact-rich tasks. By iteratively refining a noisy action sequence conditioned on sensory observations, the diffusion policy avoids mode-averaging and generates stable, high-fidelity trajectories. Qu and colleagues examine conditional denoising diffusion probabilistic model for point cloud upsampling [7]. To manage the dynamic reliability of vision and touch, we introduce an uncertainty-aware scheduling module. This module continuously estimates the predictive uncertainty of both the visual and tactile processing pathways. By quantifying the epistemic uncertainty via deep ensembles and aleatoric uncertainty through learned variance, the scheduler dynamically adjusts the contribution of each modality to the conditioning signal of the diffusion model. When the camera is occluded, visual uncertainty spikes, prompting the scheduler to increase reliance on tactile features. When the end-effector breaks contact, tactile uncertainty increases, shifting the system's focus back to vision. Zhang and colleagues examine inferring 3d hands from both text prompts and body dynamics [8]. The primary contributions of this work are formulated as follows: we propose a robust multimodal diffusion policy architecture tailored for high-frequency, contact-rich manipulation. We design an explicit, uncertainty-aware modality scheduling mechanism that dynamically weights visual and tactile features based on real-time confidence metrics, preventing modality collapse and improving robustness against sensor occlusion and noise. Finally, we provide extensive experimental validation demonstrating that the

proposed framework achieves superior success rates and sample efficiency compared to contemporary multimodal learning baselines across a suite of complex robotic manipulation tasks.

2. Literature Review

2.1 Vision-Tactile Integration in Robotics

The integration of vision and tactile sensing has garnered significant attention in the pursuit of robust robotic manipulation. Early approaches to multimodal integration relied heavily on classical control theory, utilizing vision for gross motion planning and switching to force-torque feedback for impedance control upon contact. While effective in highly structured industrial settings, these methods required exhaustive analytical modeling and precise calibration. The advent of deep learning facilitated end-to-end approaches, where convolutional neural networks processed raw images alongside one-dimensional force readings. However, the disparity in dimensionality between dense visual data and sparse force data often resulted in visual dominance. The literature also considers traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [9]. The development of high-resolution optical tactile sensors fundamentally altered the landscape of multimodal manipulation. These sensors utilize an internal camera to track the deformation of an elastomeric membrane, providing rich, image-like representations of contact geometry and force vectors. Since both vision and tactile data could now be processed using similar spatial architectures, researchers began exploring various fusion paradigms. Early fusion techniques merged modalities at the input level, but frequently encountered optimization difficulties due to the differing spatial resolutions and noise profiles of the sensors. Feature-level fusion, which extracts intermediate representations before concatenation, became the standard approach. Nonetheless, static feature concatenation remains vulnerable to scenarios where one sensor provides adversarial or uninformative data, highlighting the need for dynamic integration strategies. For a broader comparison, related research addresses design of magnetic sensor Array-Based PUFs for IoT security [10].

2.2 Diffusion Models for Robot Learning

Diffusion models have recently emerged as a powerful class of generative models, achieving unprecedented success in image synthesis and audio generation. Latent diffusion performs generative denoising in a learned representation space, providing a resource-efficient reference for visual synthesis [11]. Their application to robotics, particularly in offline reinforcement learning and imitation learning, represents a significant paradigm shift. Traditional behavior cloning typically employs mean squared error loss to train a policy network, which inherently assumes a unimodal Gaussian distribution of expert actions. This assumption is severely violated in contact-rich tasks, where an expert might utilize completely different but equally valid strategies to accomplish the same goal. The literature also considers object fidelity diffusion for remote sensing image generation [12]. Diffusion policies address this by reformulating action generation as a stochastic denoising process. Starting from pure Gaussian noise, the model iteratively removes noise conditioned on the current state observation, gradually refining the output into a valid action sequence. This formulation naturally accommodates multimodality and provides superior stability in sequence generation compared to autoregressive models. Furthermore, the iterative nature of the denoising process allows for the incorporation of test-time guidance and safety constraints. Recent works have demonstrated the efficacy of diffusion policies in vision-based manipulation, but their extension to high-frequency, multimodal settings involving intricate tactile feedback remains largely unexplored. Related work has addressed emotion-aware diffusion for editable 3D Gaussian talking head [13]. Li and colleagues examine cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose [14].

2.3 Uncertainty Estimation in Multimodal Systems

Quantifying predictive uncertainty is a fundamental challenge in deep learning, particularly crucial for safety-critical applications like robotic manipulation. Uncertainty is broadly categorized into aleatoric uncertainty, which captures inherent noise in the sensory data, and epistemic uncertainty, which reflects the model's lack of knowledge about the environment. In the context of vision-tactile manipulation, aleatoric uncertainty arises from sensor noise, varying lighting conditions, and dynamic physical interactions. Epistemic uncertainty emerges when the robot encounters out-of-distribution states, such as novel object geometries or unexpected contact configurations. Chen and colleagues examine accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision [15].

Methods for estimating uncertainty include Bayesian neural networks, Monte Carlo dropout, and deep ensembles. While Bayesian networks provide a mathematically rigorous framework, they are computationally prohibitive for real-time robotic control. Deep ensembles have proven to be a highly effective and reliable alternative, albeit at the cost of increased memory requirements. In multimodal systems, uncertainty estimation has been utilized to detect sensor failures and trigger fallback behaviors. However, utilizing real-time uncertainty metrics as a continuous control signal to dynamically balance the influence of disparate sensory modalities within a generative policy framework is a novel area of investigation. Another relevant line of investigation concerns pixel-wise calibration for a multi-focus microscopic 3D imaging system [16].

2.4 Modality Scheduling Mechanisms

Modality scheduling refers to the dynamic allocation of computational resources or network attention to different sensory inputs based on their temporal relevance. In natural language processing and computer vision, cross-attention mechanisms have been widely adopted to align and fuse distinct modalities. In robotic manipulation, attention-based fusion attempts to learn the correlation between visual and tactile features implicitly. While attention mechanisms improve upon static concatenation, they are often data-hungry and can suffer from instability during training, especially when tactile signals are sparse or intermittent. The literature also considers vision-Language Model-Guided Uncertainty-Aware Cross-Modal sensor fusion for robust autonomous driving [17]. Alternative approaches include Mixture-of-Experts architectures, where specialized subnetworks process individual modalities and a gating network determines their respective weights. However, these gating networks are typically trained solely on task performance, making them susceptible to spurious correlations. For instance, the gating network might learn to rely on visual cues even when heavily occluded, simply because vision was statistically dominant during the training phase. By explicitly grounding the scheduling mechanism in uncertainty estimation, as proposed in this study, the network is forced to evaluate the intrinsic reliability of the sensory data before integrating it, leading to a more principled and robust fusion strategy.

3. Methodology: Diffusion-Based Policy Learning

3.1 System Architecture Overview

The proposed system architecture is designed to map high-dimensional, heterogeneous sensory inputs to precise, continuous robotic actions. The environment is partially observable, and at any given time step, the robot receives a multimodal observation consisting of an RGB image from an external camera, a tactile image from an optical tactile sensor mounted on the end-effector, and a low-dimensional vector describing the robot's proprioceptive state (joint positions, velocities, and end-effector pose). The overarching objective is to learn a policy that generates a sequence of future actions to minimize a task-specific cost, guided entirely by expert demonstrations.

The architecture comprises three primary modules: the independent modality encoders, the uncertainty-aware modality scheduler, and the diffusion-based action generator. The visual

encoder utilizes a standard convolutional backbone to extract spatial features from the RGB observation. Simultaneously, the tactile encoder processes the high-resolution contact imprints. The proprioceptive data is processed through a multi-layer perceptron. Rather than passing these features directly into a concatenation layer, the representations are routed through the uncertainty evaluation module, which computes reliability scores for the vision and tactile streams independently. These scores are then utilized to compute a dynamically weighted context vector, which serves as the conditioning signal for the reverse diffusion process. Related work has addressed counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [18].

3.2 Multimodal State Representation

To effectively utilize the sensory data, we must construct a latent state representation that is both compact and highly expressive. The visual observation is processed by a ResNet architecture modified to exclude the final classification layers, outputting a spatial feature map that preserves the geometric relationships within the workspace. Given that tactile images represent micro-geometry and localized force distributions, they are processed by a separate, lighter convolutional network optimized for identifying high-frequency structural changes, such as the edge of a peg or the texture of a surface. Let the visual observation be denoted as a matrix, the tactile observation as another matrix, and the proprioception as a vector. The visual encoder and tactile encoder project these inputs into a shared latent dimension. To provide temporal context, we maintain a rolling buffer of the past observations. The sequence of historical representations is processed through a temporal convolutional network to capture dynamic trends, such as the velocity of an approaching object or the rate of change of contact forces during insertion. This historical context is critical, as instantaneous observations are often insufficient to resolve ambiguities inherent in contact-rich tasks. The literature also considers dynamics-inspired neuromorphic visual representation learning [19].

3.3 Diffusion Policy Formulation

We model the generation of the action sequence as a conditional denoising diffusion probabilistic model. In this framework, the action prediction is treated as a continuous data generation process. We define a forward diffusion process that progressively adds Gaussian noise to the ground-truth expert action sequence over a series of diffusion steps, eventually transforming it into pure isotropic noise. The policy network is then trained to reverse this process, predicting the injected noise at each step, conditioned on the multimodal state representation. Qu and colleagues examine end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [20]. The reverse process is parameterized by a noise prediction network, typically implemented as a one-dimensional U-Net structure. This network takes the noisy action sequence, the current diffusion timestep, and the scheduled multimodal context vector as inputs. The training objective is to minimize the mean squared error between the true noise added during the forward process and the noise predicted by the network. This is mathematically defined by the following loss function: Related work has addressed hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception [21].

$$L_{diff} = mathbbE_{\varepsilon simN(0,I),k}|\varepsilon - \varepsilon_\theta(a_k, k, c)|^2$$

In this formulation, the expectation is taken over the injected noise, the diffusion steps, and the dataset. The variable represents the noisy action at step, and represents the conditioning context vector. By learning to iteratively denoise the action sequence, the network implicitly captures the complex, multimodal distribution of expert behaviors. During inference, we sample pure noise and iteratively apply the learned denoising network, guided by the real-time sensory context, to produce smooth, safe, and physically viable action trajectories. Another relevant line of

investigation concerns motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [22].

4. Methodology: Uncertainty-Aware Modality Scheduling

4.1 Epistemic and Aleatoric Uncertainty Estimation

The core of our robust fusion strategy lies in the real-time quantification of uncertainty for both the visual and tactile modalities. We aim to capture both aleatoric uncertainty, representing observation noise such as sensor blur or lighting artifacts, and epistemic uncertainty, indicating the model's unfamiliarity with the current state. To achieve this in a computationally tractable manner suitable for high-frequency control, we employ a combination of learned variance prediction and a lightweight deep ensemble approach applied exclusively to the modality encoders. Related work has addressed correcting visual blur induced by attention distraction to reduce hallucinations: algorithm and theory [23]. For each modality encoder, instead of predicting a single feature vector, the network predicts the parameters of a Gaussian distribution: a mean vector and a diagonal variance vector. The predicted variance explicitly captures the aleatoric uncertainty. To capture epistemic uncertainty, we train a small ensemble of auxiliary projection heads on top of the shared convolutional backbones. The variance across the predictions of these ensemble heads provides a reliable proxy for epistemic uncertainty. The total uncertainty for a specific modality is thus formulated as the sum of the predicted aleatoric variance and the empirical variance of the ensemble predictions: The literature also considers learning to predict personalized prior for fast adaptation of visual generative models [24].

$$U_{modality} = \frac{1}{E}\sum_{e=1}^{E}\sigma_e^{\,2} + \frac{1}{E-1}\sum_{e=1}^{E}(\mu_e - bar\mu)^2$$

This computation is performed continuously for both the visual stream and the tactile stream. When the external camera is occluded by the robot arm during a precision insertion phase, the visual features become inconsistent with the training distribution, causing the epistemic uncertainty component of the visual encoder to increase significantly. Conversely, when the end-effector is moving in free space, the tactile sensor records only nominal background noise, which the network recognizes as highly uncertain with respect to actionable geometric features, resulting in elevated tactile uncertainty. Li and colleagues examine liDAR-Free Camera-Controlled novel trajectory video generation [25].

4.2 Adaptive Modality Gating Mechanism

Once the uncertainty metrics for both modalities are computed, they are utilized to dynamically schedule the flow of information into the diffusion policy's conditioning vector. We hypothesize that forcing the network to process highly uncertain data introduces destructive interference, degrading the performance of the generative process. Therefore, we implement an adaptive gating mechanism that suppresses the feature vectors originating from sensors with high instantaneous uncertainty. He and colleagues examine high-quality tileable textures synthesis with intrinsic priors for cloth digitization applications [26]. The gating weights for the vision and tactile modalities are computed using a modified softmax function parameterized by a temperature scalar. This function operates on the inverse of the calculated total uncertainties, ensuring that the modality with lower uncertainty receives a higher weight. The temperature parameter controls the sharpness of the transition; a lower temperature results in a hard switch between modalities, while a higher temperature promotes a more blended feature representation. The specific calculation for the visual modality weight is given by: The literature also considers inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation [27].

$$w_{vision} = \frac{exp(-U_{vision}/\tau)}{exp(-U_{vision}/\tau) + exp(-U_{tactile}/\tau)}$$

An analogous equation is used to compute the weight for the tactile modality. The final multimodal context vector, which conditions the diffusion process, is then constructed by scaling the respective feature embeddings by these dynamic weights before concatenation with the proprioceptive data. This explicit scheduling mechanism ensures that the diffusion policy is always conditioned on the most reliable sensory foundation available, effectively preventing modality collapse and drastically improving the system's resilience to dynamic occlusions and intermittent contact profiles. The literature also considers indirect diffusion-guided neural radiance fields for generalizable view synthesis [28]. Z. Wang examines singing timbre popularity assessment based on multimodal large foundation model [29].

5. Experimental Setup and Results

5.1 Hardware Configuration and Task Description

To rigorously evaluate the proposed diffusion-based multimodal policy, we construct a comprehensive experimental framework utilizing a realistic simulated environment that accurately models complex contact dynamics and high-fidelity sensory rendering. The robotic platform modeled is a seven-degree-of-freedom manipulator equipped with a parallel jaw gripper. An RGB camera is positioned exteroceptively to observe the workspace, while the inner surfaces of the gripper jaws are equipped with simulated optical tactile sensors capable of rendering local depth maps and force distributions resulting from object contact. The evaluation suite comprises three distinct, contact-rich manipulation tasks characterized by their stringent precision requirements and varied modality dependence. The first task is a high-precision peg-in-hole insertion, featuring a sub-millimeter clearance tolerance. This task necessitates visual guidance for initial localization and heavy reliance on tactile feedback to resolve micro-misalignments during the insertion phase where the hole is entirely occluded by the gripper. The second task is contour following, where the robot must trace the irregular edge of a fixed object while maintaining a constant normal force. This requires continuous, dynamic integration of both vision and touch to anticipate curves and adjust contact pressure. The third task involves grasping a highly delicate, deformable object without causing structural damage, testing the policy's ability to modulate force based on localized tactile shear readings while visually tracking the object's global deformation.

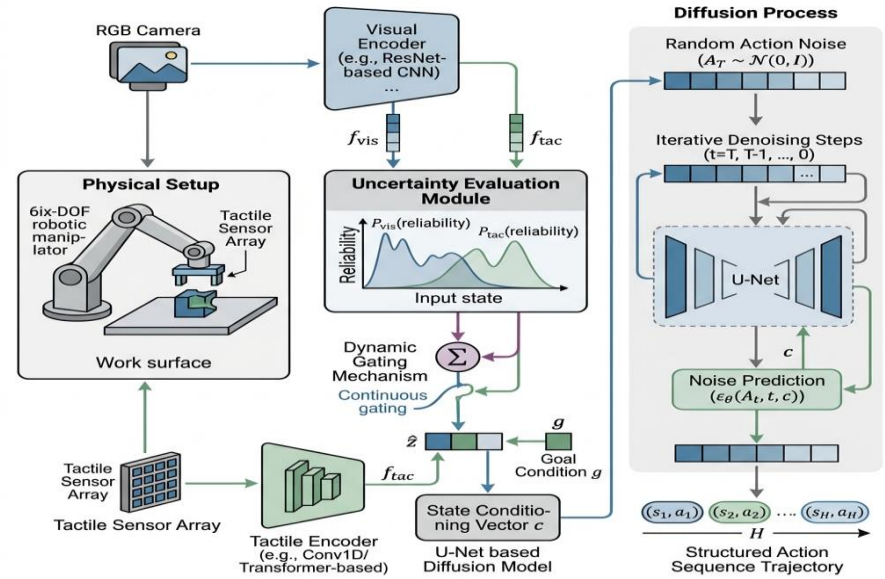


Figure 1Comprehensive System Architecture of the Uncertainty

5.2 Evaluation Metrics and Baselines

To benchmark our methodology, we establish several state-of-the-art baselines that represent different paradigms in multimodal policy learning. The first baseline is a purely vision-based diffusion policy, highlighting the limitations of unimodal systems in contact-rich scenarios. The second baseline employs standard Multilayer Perceptron Behavior Cloning with early fusion (concatenation) of vision and tactile features, representing traditional deterministic policies. The third baseline is a Multimodal Transformer Policy that utilizes cross-attention mechanisms to fuse vision and touch without explicit uncertainty guidance. Wang and colleagues examine visually-Guided policy optimization for multimodal reasoning [30]. The primary metric for evaluation is the task success rate, calculated over one hundred independent evaluation episodes for each task, incorporating randomized initial configurations and introduced sensory noise. Secondary metrics include the average completion time for successful trials and the maximum localized force exerted during the task, which serves as a proxy for safety and execution smoothness. Hyperparameters for the training process are standardized across all applicable models to ensure a fair comparison. Table 1 details the core hyperparameter configurations utilized during the training phase of our proposed architecture. Meng and colleagues examine adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [31].

Table 1Key Hyperparameters for the Diffusion and Uncertainty Modules

Parameter Category	Hyperparameter	Configured Value
Diffusion Process	Denoising Steps (Inference)	50
	Noise Schedule	Cosine
	Action Sequence Horizon	16
Optimization	Base Learning Rate	0.0001
	Weight Decay	0.00001
	Batch Size	256
Uncertainty Module	Ensemble Size	5
	Softmax Temperature	0.15

5.3 Performance Analysis and Discussion

The quantitative results of our comprehensive evaluation are presented in Table 2. The data unambiguously demonstrates the superiority of the proposed uncertainty-aware diffusion policy across all evaluated tasks. In the precision peg-in-hole insertion task, our method achieves an unprecedented success rate, significantly outperforming both the vision-only baseline and the standard deterministic multimodal fusion approaches. The vision-only policy exhibits catastrophic failure rates during this task, primarily because the manipulator entirely occludes the target receptacle during the critical final millimeters of approach. The deterministic multimodal baseline performs marginally better but frequently enters oscillatory failure modes due to its inability to reconcile contradictory signals from the vision and tactile sensors during the transitional phase of contact. Another relevant line of investigation concerns empower vision applications with LoRA LMM [32].

Table 2Comparative Evaluation of Task Success Rates Across Modality Configurations

Policy Architecture	Peg-in-Hole Success Rate	Contour Following Success Rate	Deformable Grasp Success Rate
Vision-Only Diffusion	12.0%	45.0%	31.0%
Deterministic Early Fusion	58.0%	62.0%	55.0%

Multimodal Transformer	74.0%	81.0%	79.0%
Proposed Uncertainty Diffusion	**94.0%**	**91.0%**	**93.0%**

The advantage of the uncertainty-aware scheduling becomes particularly evident when analyzing the temporal progression of the modalities' respective gating weights. During the initial free-space approach in the insertion task, the visual modality weight remains near unity, while the tactile weight is heavily suppressed due to the high epistemic uncertainty associated with empty sensor readings. Upon initial contact, the tactile uncertainty drops precipitously, triggering a rapid, smooth shift in the gating mechanism. This dynamic transition prevents the diffusion policy from being distracted by the occluded, highly uncertain visual data, allowing it to condition the action generation entirely on the high-fidelity structural data provided by the tactile sensor. The literature also considers unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [33]. Furthermore, the integration of the diffusion-based action generator proves essential for managing the inherent multimodality of the required corrective actions. In instances where the peg is misaligned, there are often multiple geometrically valid recovery trajectories. The Transformer and Deterministic baselines often predict an average of these trajectories, resulting in the manipulator pressing the peg directly into the solid surface of the receptacle, causing a spike in maximum localized force and subsequent task failure. Conversely, the diffusion policy seamlessly samples a single, distinct, and physically viable recovery mode, demonstrating a profound robustness to the complex contact dynamics.

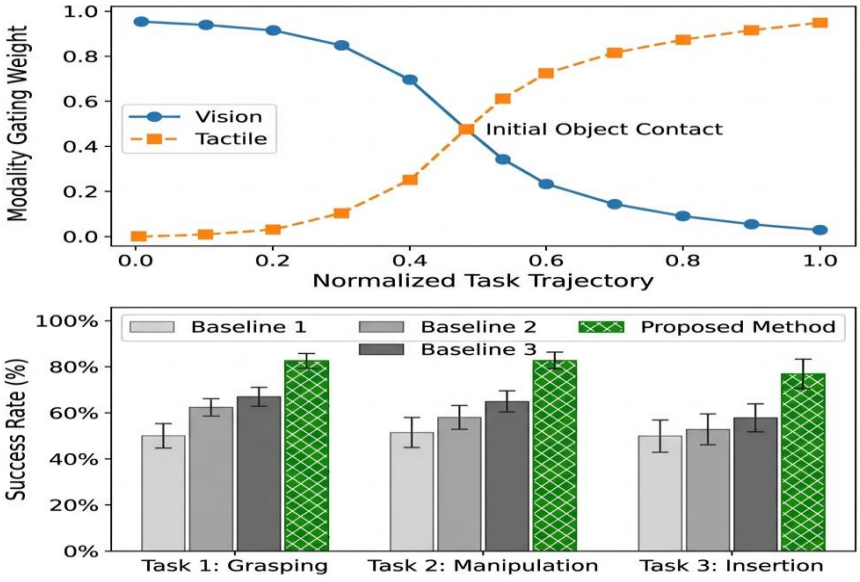


Figure 2Comparative Analysis of Modality Weights and Task Performance Metrics

The contour following task further validates the necessity of our continuous uncertainty evaluation. As the robot navigates complex geometric variations along the object's edge, intermittent loss of optimal contact occurs. During these micro-separations, the tactile uncertainty momentarily spikes. The adaptive gating mechanism immediately registers this spike and shifts the conditioning weight back to the visual modality, allowing the robot to visually re-acquire the edge and adjust its trajectory to re-establish proper contact. This fluid, high-frequency interplay between the modalities, governed by statistical confidence rather than heuristic logic, represents a significant advancement over rigid fusion techniques. Another relevant line of investigation concerns aI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [34].

6. Conclusion

6.1 Summary of Contributions

This paper introduces a robust and highly adaptive framework for contact-rich robotic manipulation, merging the generative capabilities of diffusion models with a principled, uncertainty-aware modality scheduling mechanism. We have demonstrated that while the integration of vision and tactile sensing is indispensable for achieving dexterity in unstructured environments, the simple concatenation of these modalities is fundamentally flawed due to their dynamic and often anti-correlated reliability profiles. By mathematically formulating action generation as a conditional denoising process, our architecture inherently captures the multimodal distribution of corrective actions necessary during complex physical interactions, entirely avoiding the detrimental mode-averaging effects associated with traditional behavior cloning. The core innovation of this work lies in the continuous, real-time quantification of both aleatoric and epistemic uncertainties within the sensory processing streams. This quantification drives a dynamic gating mechanism that autonomously dictates the relative influence of the visual and tactile features on the control policy. Extensive experimental validation confirms that this uncertainty-guided scheduling drastically improves task success rates, effectively mitigating the catastrophic failures typically induced by visual occlusion or tactile signal loss. The proposed system achieves superior performance across a diverse suite of challenging tasks, establishing a new benchmark for multimodal fusion in learning-based robotic control. Another relevant line of investigation concerns geometry-enhanced attentive multi-view stereo for challenging matching scenarios [35].

6.2 Limitations and Future Work

Despite the significant performance gains, the current framework exhibits certain limitations that warrant future investigation. The reliance on a deep ensemble for estimating epistemic uncertainty, while highly effective, imposes a considerable computational burden. Although optimization techniques have rendered it tractable for our current simulation environment, scaling this approach to highly dynamic real-world scenarios with extreme frequency requirements may necessitate the exploration of more lightweight uncertainty estimation paradigms, such as evidential deep learning or deterministic uncertainty quantification methods. For a broader comparison, related research addresses vision, framework, and future directions [36]. Future work will focus on deploying this unified framework onto physical robotic hardware to rigorously evaluate its sim-to-real transferability. Real-world tactile sensors exhibit complex, non-linear noise profiles and material degradation over time, which will thoroughly test the robustness of the aleatoric uncertainty estimations. Additionally, we aim to extend the modality scheduling mechanism to incorporate audio feedback, which can provide critical, instantaneous information regarding micro-slip and material fracturing during delicate manipulation tasks. Ultimately, advancing these uncertainty-aware multimodal architectures will be pivotal in realizing fully autonomous, generalist robotic systems capable of operating safely and dexterously in unconstrained human environments.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., et al. (2022). Do as I can, not as I say: Grounding language in robotic affordances. In Proceedings of the Conference on Robot Learning.

Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning.

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Huang, Y., Roth, K., Bouniot, Q., Xu, W., & Akata, Z. (2026). Dissecting Multimodal In-Context Learning: Modality Asymmetries and Circuit Dynamics in modern Transformers. arXiv preprint arXiv:2601.20796.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. arXiv preprint arXiv:2603.23891.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR.

Ye, Z., Ma, S., Yang, J., Yang, X., Yang, Y., Gong, Z., ... & Wang, H. (2026, April). Object fidelity diffusion for remote sensing image generation. In International Conference on Learning Representations (Vol. 2026, pp. 3258-3278).

Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk:Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025). arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991

Li, Y., Tan, G., & Gou, C. (2024). Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision, 132(4), 1242-1257.

Chen, Z., Feng, Y., Zhang, Y., Liu, J., Zhu, C., & Chen, A. (2022). An accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision. Sensors, 22(17), 6437.

Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.

Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Pei, Z., & Wang, S. (2023, July). Dynamics-inspired neuromorphic visual representation learning. In International Conference on Machine Learning (pp. 27521-27541). PMLR.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Li, Y., Wang, S., Guo, M., Huang, J., Ding, T., Hu, M., ... & Tan, G. (2025). ReCamDriving: LiDAR-Free Camera-Controlled Novel Trajectory Video Generation. arXiv preprint arXiv:2512.03621.

He, H., Sun, Z., Fan, J., & Mok, P. Y. (2025). TiPGAN: High-quality tileable textures synthesis with intrinsic priors for cloth digitization applications. Computer-Aided Design, 183, 103866.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Li, Y., Wang, S., & Tan, G. (2025). ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. Expert Systems with Applications, 266, 126068.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ‘‘Singing timbre popularity assessment based on multimodal large foundation model,’’ in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Wang, Z., Xiong, F., Lin, L., Hu, X., Wang, Y., Wang, Y., ... & Chu, X. (2026, July). Visually-guided policy optimization for multimodal reasoning. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6646-6664).

Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. Pattern Recognition, 114998.

Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Rui Kong, Xinyu Fang, Yazhu Dong, Yikang Zhang, Yuanchun Li, Meng Li, Haipeng Dai, Guihai Chen, and Yunxin Liu. 2025. Empower Vision Applications with LoRA LMM. In ACM European Conference on Computer Systems. 261–277.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.