Feeling the Unexpected: ResTacVLA for Contact-Rich Manipulation via Residual Tactile Representation

Pengwei Zhang[1,2], Bin Xie[3], Ce Hao[2], Xinpan Meng[1],
Xinyu Guo[2], Fang Deng[2], Long Cheng[1,*], and Tiancai Wang[3]

***Abstract*— Tactile perception is indispensable for contact-rich manipulation, yet integrating it into Vision-Language-Action (VLA) models often induces *modality collapse*, where high-bandwidth visual features overshadow sparse tactile cues. Inspired by Predictive Coding—a neural mechanism where the brain attenuates predictable inputs to prioritize surprising stimuli—we propose ResTacVLA. Rather than treating tactile data as raw input, we reformulate it as a Residual Tactile Representation capturing the discrepancy between visual priors and physical sensations. By filtering out visually predictable dynamics, this formulation transforms sparse tactile signals into dense, high-value information gain, thereby inherently resolving the bandwidth mismatch. These residuals are discretized through a Vector Quantized (VQ) bottleneck into Latent Contact Primitives that capture critical events missed by vision. Analogous to the neural surprise signal, we leverage the uncertainty of the visual prior to adaptively gate tactile integration, prioritizing residuals specifically during visually unreliable phases to explicitly prevent visual dominance. Experimental results show that ResTacVLA consistently outperforms all baselines on a diverse set of contact-rich manipulation tasks, while remaining robust to unexpected dynamic disturbances. Project website: https://awilekong.github.io/ResTacVLA/.**

I. INTRODUCTION

Empowered by large-scale embodied pretraining, robotic manipulation policies developed by Vision-Language-Action models (VLAs) have recently emerged as a dominant paradigm. However, current VLAs [1], [2] are primarily vision-centric, relying predominantly on visual perception. Although many manipulation tasks can be completed nearly perfectly using this approach, a critical bottleneck remains in contact-rich manipulation tasks—such as precision insertion [3], threaded assembly [4], and surface wiping [5]. While these tasks are relatively straightforward for humans, the visual-centric approach often struggles with severe occlusions or lacks the fidelity required to resolve fine-grained physical dynamics. The primary limitation lies in the absence of mastery over tactile perception, which is indispensable for grounding physical interactions.

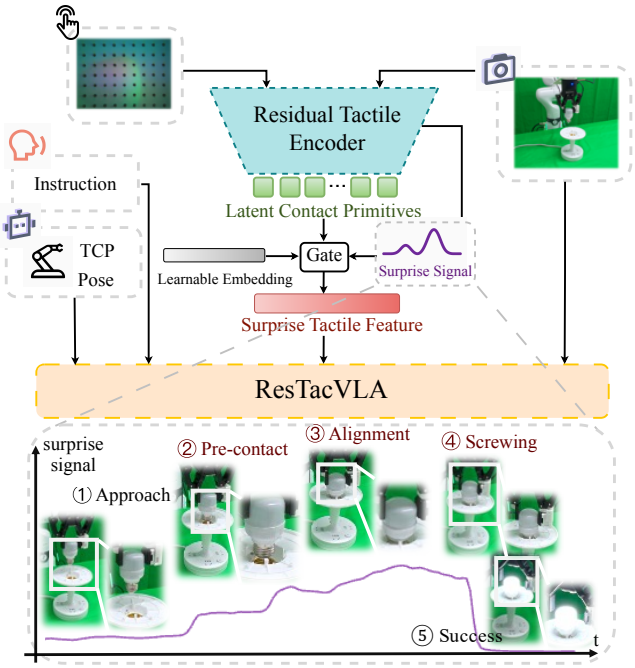


Fig. 1: Predictive Coding-Inspired Residual Tactile Fusion in ResTacVLA. The Residual Tactile Encoder extracts high-information-gain representations by modeling the discrepancy between visual priors and physical sensations. A surprise signal drives a Surprise-Aware Gate to adaptively amplify tactile cues during contact-critical phases while suppressing redundant tactile noise in free-space motion.

A naive approach to incorporating tactile perception is to simply treat it as an additional modality, directly projecting tactile inputs into a shared feature space [6], [7], [8]. However, in practice, a phenomenon known as 'Modality Collapse' [9] occurs, where the high-bandwidth, continuous visual stream naturally overshadows the event-driven, temporally sparse tactile signals [10], [11]. By overlaying the missing tactile modality during pretraining, VLAs tend to disregard the 'quiet' tactile cues in favor of the 'loud' visual features. The inefficiency of this native approach compels us to find a new way to empower VLAs with tactile perception. Recent research in cognitive neuroscience has shown that biological systems address the 'Modality Collapse' through Predictive Coding [12], [13], [14]. Specifically, instead of processing all modality inputs equally, the

This work was supported in part by the Brain Science and Brain-like Intelligence Technology – National Science and Technology Major Project under Grant 2025ZD0215600, in part by the National Natural Science Foundation of China under Grants U25A20475 and 62333023, in part by the Beijing Municipal Natural Science Foundation under Grants F2024201068 and L243014, in part by the CAS Project for Young Scientists in Basic Research under Grant YSBR-034, in part by the Zhongguancun Academy under Grant 20240307, and in part by the Fundamental Research Funds for the Central Universities.

[1]Pengwei Zhang, Xinpan Meng, and Long Cheng are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and are also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

[2]Pengwei Zhang, Ce Hao, Xinyu Guo, and Fang Deng are with Zhongguancun Academy, Beijing 100094, China.

[3]Bin Xie and Tiancai Wang are with Dexmal, Beijing 100096, China.

*Long Cheng is the corresponding author (longcheng.ia.ac.cn).

brain generates top-down predictions of expected sensory states and attenuates predictable inputs, focusing attention on the 'Unexpected'—the surprise arising from deviations from expectations. For example, humans cannot tickle themselves because the brain predicts the sensory consequences of self-generated movements and suppresses the expected signal, prioritizing only unexpected external stimuli [15]. This mechanism allows organisms to filter out redundant information and respond quickly to anomalies. Naturally, the question arises: How can we enable current VLAs to feel the unexpected?

To address this challenge, we propose a Residual Tactile Vision-Language-Action method, called ResTacVLA (Fig. 1), a novel framework that explicitly models the concept of 'feeling the unexpected' for contact-rich manipulation tasks. Rather than competing with the high-bandwidth visual modality, ResTacVLA reformulates tactile feedback as a Residual Tactile Representation—a quantitative measure of surprise relative to visual priors. To generate this representation, we introduce the Cross-Modal Predictor (CMP), which distills the discrepancy between visual expectations and physical reality, transforming sparse tactile signals into dense, high-value information. These residuals are then discretized through vector quantization into Latent Contact Primitives that capture intrinsic physical events (e.g., unexpected collisions) that vision fails to perceive. To process the Residual Tactile Representation, we employ a Surprise-Aware Gate (SAG) that adaptively regulates tactile integration, suppressing noise when vision is reliable and explicitly amplifying the tactile pathway during physically ambiguous phases.

In order to evaluate ResTacVLA, we selected five real-world contact-rich tasks, including precision insertion, screwing, and surface wiping. Experimental results demonstrate that ResTacVLA achieves state-of-the-art performance, significantly outperforming both standard VLA baselines and naive tactile fusion strategies, while exhibiting superior robustness against dynamic physical disturbances.

Our contributions are summarized as follows:

- We propose ResTacVLA, a biologically inspired framework that integrates tactile feedback through Residual Tactile Representations. By filtering out visual redundancy, we transform tactile signals into dense information gain, effectively mitigating the modality imbalance problem.
- We design a Cross-Modal Predictor (CMP) that encodes tactile signals into Latent Contact Primitives with high information gain, and a Surprise-Aware Gate (SAG) is introduced to adaptively modulate tactile injection based on task-phase characteristics.
- Through experiments on five challenging tasks, ResTacVLA achieves up to 86.7% task success and improves average performance by 34.6% over baselines, while maintaining strong robustness against dynamic disturbances.

II. RELATED WORK

A. Vision-Language-Action Models in Contact-Rich Manipulation

The emergence of Vision-Language-Action (VLA) learning has revolutionized robotic manipulation by bridging the gap between large-scale visual perception and natural language reasoning [1], [16]. Prominent frameworks, ranging from transformer-based action predictors [17], [18] to open-weight foundation models [1], [2], harness extensive datasets of aligned visual observations, linguistic instructions, and kinematic trajectories. Leveraging scaling laws in sequence modeling, these systems demonstrate strong generalization across semantically diverse manipulation scenarios. However, their capabilities remain constrained to high-level semantic understanding. While VLAs effectively determine *what* to do, they lack mechanisms to reason about *how* to perform contact-rich interactions—regulating forces, adapting to compliance, or recovering from collisions [7], [8], [19]. This 'tactile-blind' nature renders them vulnerable when occlusion obscures geometry or when physical dynamics (friction, deformation) cannot be inferred from vision alone, necessitating direct tactile feedback for robust execution.

B. Representation Learning of Vision-Based Tactile Sensing

Constructing robust representations for high-dimensional tactile data is a prerequisite for effective physical interaction. Self-supervised methods—including masked modeling [20] and compact task-agnostic encoding [21]—have established a solid foundation by extracting generalizable physical features from raw tactile streams without extensive annotation. To further bridge the semantic gap between touch and vision, cross-modal contrastive approaches such as VITaL [22] and Beyond Sight [6] align tactile embeddings with visual or linguistic latent spaces by maximizing inter-modal mutual information. Although effective at exploiting shared information, this alignment paradigm inadvertently suppresses the critical residual information unique to touch—specifically, the fine-grained contact dynamics that vision cannot predict. For contact-rich manipulation, we argue that task success hinges precisely on this residual. Consequently, rather than prioritizing cross-modal alignment, our approach explicitly models the tactile residual, capturing the discrepancy between visual expectation and physical reality to maximize information gain.

C. Tactile Integration in Generalist Policies

Strategies for integrating tactile sensing into VLA frameworks typically fall into two categories. One line of work augments policies with global force feedback [23], [24], utilizing wrist-mounted Force/Torque (F/T) sensors or joint-torque estimation. Although effective for collision detection or payload monitoring, these modalities fundamentally aggregate complex interaction dynamics into a single resultant vector, thereby forfeiting the spatial richness necessary to capture local contact geometry and distributed pressure patterns.

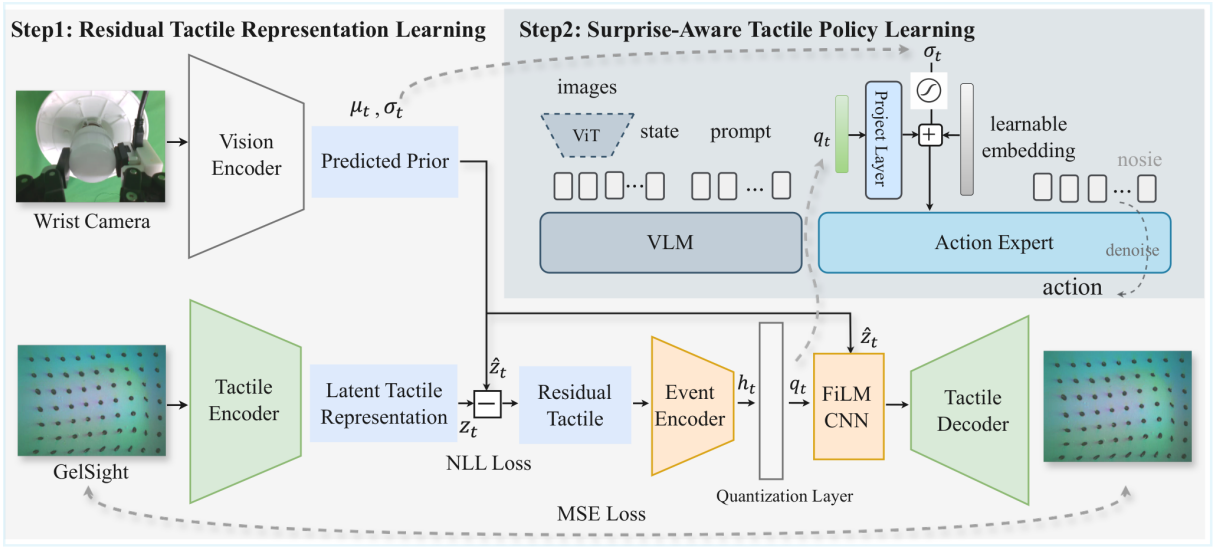


Fig. 2: Overall architecture of ResTacVLA. (a) **Step 1: Residual Tactile Representation Learning.** The Cross-Modal Predictor (CMP) is pre-trained to estimate tactile latents from wrist-camera observations, extracting residual representations that capture the discrepancy between visual priors and physical sensations. These residuals are discretized into Latent Contact Primitives via a VQ bottleneck. (b) **Step 2: Surprise-Aware Tactile Policy Learning.** With the CMP frozen, the prediction uncertainty σ_t drives a Surprise-Aware Gate (SAG) to adaptively inject contact primitives into the action expert, conditioning a flow matching policy for contact-rich manipulation.

In contrast, visuotactile-based VLA policies [7], [8], [19], [25] seek to exploit the high-resolution capabilities of vision-based tactile sensors. However, the prevailing paradigm treats tactile frames merely as 'auxiliary visual features', fusing them with scene observations through standard Transformer architectures. We argue that this naive integration induces 'Modality Collapse': the high-bandwidth, continuous stream of visual data naturally dominates sparse, event-driven tactile signals. Without explicit mechanisms to filter visual redundancy, such models tend to disregard tactile feedback in favor of dominant visual priors, thereby failing to leverage the distinctive information gain provided by touch during critical contact phases.

III. METHODOLOGY

Problem Formulation. Similar to other vision-centric VLA policies, the objective of the tactile-aware policy π is to improve the performance on contact-rich tasks by mapping inputs—including visual, tactile, and language instructions—into low-level actions. Formally, at timestep t, the policy processes both the observation O_t and the language instruction L, outputting a low-level action sequence $A_t = \{a_t, a_{t+1}, \ldots, a_{t+H-1}\}$, i.e., $\pi(A_t|O_t, L)$. The visual modality, coming from the robot's cameras, consists of base, side, and wrist visual inputs: V_t^{base}, V_t^{side}, and V_t^{wrist}; the tactile modality is represented by I_t^{tac}; and the proprioceptive state is $s_t \in \mathbb{R}^7$. These modalities are collectively denoted as $O_t = \{V_t^{base}, V_t^{side}, V_t^{wrist}, I_t^{tac}, s_t\}$. The language instructions L describe the tasks that the policy is required to perform.

A. Overview

ResTacVLA is an end-to-end multimodal robotic policy designed for contact-rich manipulation, with its overall pipeline illustrated in Fig. 2. Its core mechanism comprises two key components: a Cross-Modal Predictor (CMP) and a Surprise-Aware Gate (SAG). Rather than naively fusing raw sensory streams, the CMP reformulates tactile feedback into a Residual Tactile Representation that captures the deviation of physical reality from visual anticipation. Grounded in Predictive Coding, the CMP distills visual redundancy into discrete Latent Contact Primitives that encode intrinsic contact dynamics. These primitives are then adaptively modulated by the SAG and injected into the action denoising process, allowing the policy to explicitly refine action trajectories based on unexpected physical dynamics, while suppressing tactile noise during visually reliable phases. To acquire robust contact representations prior to policy learning, the CMP is first pre-trained independently on multi-task interaction data, and is subsequently frozen while the full tactile-augmented policy is fine-tuned end-to-end.

Building upon the $\pi_{0.5}$ framework [2], the tactile-augmented policy integrates vision, language, proprioception, and high-resolution tactile feedback to generate actions through a conditional flow-matching model [26]. Visual inputs from multiple RGB cameras and task instructions, along with proprioceptive states, are encoded by a pretrained vision-language model, e.g., PaliGemma [27] into contextual embeddings. These embeddings, combined with the modulated contact primitives, condition an iterative denoising process that predicts the action trajectory.

B. Residual Tactile Representation Learning

1) Cross-Modal Predictor and Residual Tactile Extraction: We introduce the Cross-Modal Predictor (CMP) to estimate the expected latent tactile representation exclusively from wrist camera observations. Specifically, a Vision Encoder processes V_t^{wrist} via a learnable ResNet-18 backbone followed by an MLP head, yielding a predicted mean $\mu_t \in \mathbb{R}^{3\times H'\times W'}$ and a scalar standard deviation $\sigma_t \in \mathbb{R}$. These

parameters characterize a Gaussian distribution over the predicted tactile latent $\hat{z}_t$. Concurrently, a tactile encoder, based on UniT [21], projects the actual tactile image I_t^{tac} into the same latent space, producing $z_t \in \mathbb{R}^{3\times H'\times W'}$. The residual tactile, defined as $r_t = z_t - \hat{z}_t$, isolates physical sensory components unanticipated by the visual modality. The predictor is optimized using a weighted Negative Log-Likelihood (NLL) [28] objective, which promotes accurate prediction while explicitly modeling aleatoric uncertainty:

$$\mathcal{L}_{pred} = \lambda_\sigma \log \sigma_t^2 + \frac{\|z_t - \mu_t\|^2}{\sigma_t^2}, \tag{1}$$

where hyperparameter λ_σ regulates the variance penalty to prevent uncertainty collapse. Crucially, σ_t functions as a principled measure of cross-modal surprise, which is subsequently utilized by the SAG (Sec. III-C) to adaptively modulate the contribution of tactile information.

2) Latent Contact Primitives via VQ: An event encoder f_ϕ processes the residual tactile r_t and aggregates features into a global event vector $h_t = f_\phi(r_t) \in \mathbb{R}^D$ employing convolutional residual blocks followed by global max pooling. The h_t is subsequently discretized into a Latent Contact Primitive q_t via a Vector Quantization (VQ) [29] bottleneck containing a learnable codebook $\mathcal{C} = \{c_k\}_{k=1}^K$. To mitigate codebook collapse, we implement multiple strategies during VQ training: reducing the codebook dimension [30], replacing Euclidean distance with cosine similarity [30], smoothing codebook updates via exponential moving average (EMA) [31], and periodically reinitializing inactive codebook entries [32]. The latent tactile representation is then reconstructed by modulating the predicted prior $\hat{z}_t$ with the quantized primitive q_t via FiLM [33] conditioning. The resulting $\tilde{z}_t$ is decoded back into the tactile image space, yielding the reconstructed tactile image $\hat{I}_t$. The complete CMP pipeline is jointly trained on multi-task interaction data by minimizing the following objective:

$$\mathcal{L}_{CMP} = \mathcal{L}_{rec} + \lambda_p \mathcal{L}_{pred} + \mathcal{L}_{vq}, \tag{2}$$

where $\mathcal{L}_{rec} = \|\hat{I}_t - I_t^{tac}\|_2^2$ denotes the mean squared error (MSE) for tactile reconstruction. $\mathcal{L}_{pred}$ is the cross-modal prediction loss (Eq. 1) weighted by hyperparameter λ_p, and $\mathcal{L}_{vq}$ is the standard VQ commitment loss [30].

C. Surprise-Aware Tactile Policy Learning

The prediction uncertainty σ_t, derived from the CMP, serves as a principled indicator for adaptive modality fusion. An elevated value of σ_t signifies that the visual system cannot reliably anticipate the current tactile state, implying low cross-modal mutual information and, consequently, high tactile information gain. Conversely, a low σ_t suggests that tactile feedback is largely redundant relative to visual observations. To leverage this characteristic, we compute a surprise-aware gate $g_t \in (0, 1)$ as follows:

$$g_t = \text{Sigmoid}(\text{MLP}(\sigma_t)). \tag{3}$$

This gate is then applied to modulate the tactile pathway. Specifically, the Latent Contact Primitive q_t, obtained from the frozen CMP (Sec. III-B), is projected into the token dimension of the action expert via a linear layer, yielding $p_t \in \mathbb{R}^d$. Simultaneously, a learnable embedding $e_0 \in \mathbb{R}^d$ is employed as a default 'no-contact' token. The final tactile token e_t is synthesized via gated interpolation:

$$e_t = g_t \cdot p_t + (1 - g_t) \cdot e_0. \tag{4}$$

When visual predictions exhibit high confidence (i.e., $g_t \to 0$), the output converges to e_0, effectively suppressing the tactile pathway. In contrast, under conditions of high prediction uncertainty (i.e., $g_t \to 1$), the contact primitive p_t dominates, thereby explicitly injecting physical information into the policy. To integrate tactile cues into action generation without compromising the pre-trained representations of the VLM, we directly concatenate e_t with the noise tokens as input to the action expert. With the pre-trained CMP frozen, the entire tactile-augmented policy is fine-tuned end-to-end using the standard conditional flow-matching objective [26].

IV. EXPERIMENTS

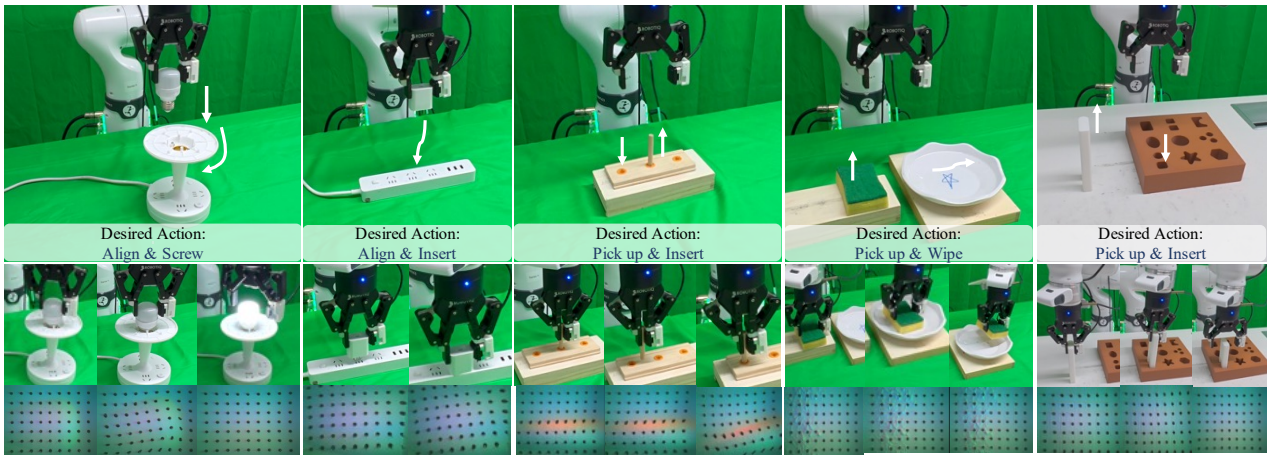
This section presents a comprehensive suite of real-world experiments to empirically validate ResTacVLA. The evaluation is structured around four core research questions:

1) In contact-rich manipulation tasks characterized by sparse tactile modalities, how does the overall effectiveness of ResTacVLA compare to standard baselines and naive tactile fusion strategies?
2) Do the proposed Residual Tactile Representation and Surprise-Aware Gating play a critical role in enhancing the capabilities of the policy?
3) Can the Latent Contact Primitives capture meaningful contact events, and the Surprise-Aware Gating adaptively regulate modality importance across different task phases?
4) Does the policy maintain robustness under unexpected physical perturbations and sensory corruptions?

A. Experimental Setup

Platform. The experimental platform comprises a Franka Research 3 robotic arm equipped with a Robotiq 2F-85 parallel gripper. A GelSight Mini tactile sensor is mounted on one fingertip of the gripper to capture high-resolution contact geometry. The robot operates within a workspace of dimensions $45 \times 60 \times 40$ cm. Three Intel RealSense D435 cameras are deployed for visual perception: one positioned frontally, one laterally, and one wrist-mounted near the gripper. All devices are connected to a workstation equipped with an Intel Core i9-10900K CPU and an NVIDIA RTX 4090 GPU, which supports both data collection and policy evaluation.

Tasks and Data Collection. To evaluate ResTacVLA, we curated five tasks that span a spectrum of contact-rich manipulation primitives, as shown in Fig. 3. These tasks represent distinct physical challenges: *Lightbulb Screwing* necessitates the detection of thread engagement and rotational resistance—dynamics imperceptible to vision alone.



(a) Lightbulb Screwing (b) Plug Insertion (c) Peg Transfer (d) Plate Wiping (e) Peg-in-Hole

Fig. 3: Overview of the five contact-rich manipulation tasks designed for evaluation. (a) **Lightbulb Screwing**: detecting thread engagement and rotational resistance—dynamics imperceptible to vision alone; (b) **Plug Insertion**: resolving sub-millimeter positional uncertainty under severe visual occlusion via tactile-guided alignment; (c) **Peg Transfer**: completing sequential fine-grained manipulation requiring high tactile sensitivity; (d) **Plate Wiping**: maintaining sustained surface contact through active force compliance to prevent end-effector hovering; (e) **Peg-in-Hole**: disambiguating contact state under occlusion with tactile precision.

Plug Insertion, *Peg-in-Hole*, and *Peg Transfer* all impose rigorous sub-millimeter tolerances under severe visual occlusion, mandating tactile-guided alignment to resolve position uncertainty. *Plate Wiping* necessitates active force feedback to guarantee sustained physical contact and effective cleaning, thereby preventing the end-effector from hovering above the surface due to inherent visual depth inaccuracies.

For each task, approximately 100 expert demonstrations are collected, during which the operator receives real-time tactile deformation imagery as tactile feedback. Evaluation is conducted over 25 trials for *Lightbulb Screwing* and *Plug Insertion*, and over 15 trials for *Peg-in-Hole*, *Peg Transfer*, and *Plate Wiping* due to their longer rollout durations, with success determined by task-specific criteria (e.g., electrical continuity for lightbulb screwing, full insertion depth for peg and plug tasks, and effective coverage for wiping). Collectively, these tasks are designed to rigorously probe the capacity of ResTacVLA to handle complex contact interactions, uncertain dynamics, and fine-grained sensorimotor coordination through the integration of vision and tactile modalities.

Evaluation Metrics and Baselines. Model performance is primarily evaluated using the task success rate across all five contact-rich tasks. In addition, for *Lightbulb Screwing*, *Peg-in-Hole*, and *Plug Insertion*, we further decompose evaluation into two phases: (1) *Alignment*—whether the object is correctly positioned at the target—and (2) *Successful Interaction*—whether the task is physically completed (i.e., electrical continuity for the lightbulb, and full insertion depth for the peg and plug). This two-phase metric provides finer-grained insight into potential failure modes.

To comprehensively evaluate ResTacVLA and the proposed CMP and SAG components, we compare it against several carefully selected baselines derived from two foundational architectures: the state-of-the-art $\pi_{0.5}$ [2] VLA model and Diffusion Policy [34]. The specific variants include: (1) DP w/o T, a standard Diffusion Policy without tactile input; (2) DP w/ T-ResTac, Diffusion Policy augmented with the proposed Residual Tactile Representation; (3) $\pi_{0.5}$ w/o T, the standard $\pi_{0.5}$ without tactile input; (4) $\pi_{0.5}$ w/ T-ResNet, $\pi_{0.5}$ with tactile images directly encoded by a ResNet-18 backbone; and (5) $\pi_{0.5}$ w/ T-UniT, $\pi_{0.5}$ with tactile images encoded by the pretrained UniT [21] model—a strong VQVAE-based tactile representation with demonstrated effectiveness in contact-rich manipulation. The inclusion of $\pi_{0.5}$ enables comparison with a strong VLA foundation, while the tactile-augmented variants on both $\pi_{0.5}$ and Diffusion Policy are crucial for demonstrating the efficacy of the proposed Residual Tactile Representation over simpler tactile integration approaches such as direct encoding or pretrained feature extraction.

B. Main Results

Overall Performance. As presented in Table I, ResTacVLA demonstrates superior performance across all five contact-rich tasks, achieving an average success rate of 62.8%. This represents a substantial improvement over vision-only baselines, surpassing both the $\pi_{0.5}$ (28.2%) and the Diffusion Policy (18.8%) by margins of +34.6% and +44.0%, respectively. A phase-wise analysis reveals that while visual policies perform adequately during the *Alignment* phase, their success rates decline sharply during the *Interaction* phase (e.g., in *Plug Insertion*, success drops from 36.0% to 20.0%). This performance gap highlights the limitations of visual perception in resolving state ambiguities under occlusion and contact constraints. In contrast, ResTacVLA maintains consistent performance throughout the interaction sequence (68.0% $\rightarrow$ 60.0%), effectively leveraging residual tactile feedback to bridge the 'Physical Gap' where vision-based policies consistently falter.

Efficacy of Residual Tactile Representation. We further evaluate the contribution of the proposed Residual Tactile

TABLE I: Quantitative Comparison on the Contact-Rich Manipulation Benchmark.

Method	Lightbulb-A	Lightbulb-I	Plug-A	Plug-I	Peg-A	Peg-I	Transfer	Wiping	*Average*
$\pi_{0.5}$ (Vision Only)	28.0	8.0	36.0	20.0	46.7	40.0	26.7	20.0	28.2
DP w/o T	20.0	0.0	28.0	16.0	40.0	26.7	13.3	6.7	18.8
$\pi_{0.5}$ w/ T-ResNet	16.0	0.0	32.0	24.0	40.0	40.0	20.0	13.3	23.2
$\pi_{0.5}$ w/ T-UniT	28.0	<u>12.0</u>	<u>40.0</u>	<u>32.0</u>	<u>73.3</u>	53.3	**66.7**	<u>33.3</u>	<u>42.3</u>
DP w/ T-ResTac	<u>32.0</u>	<u>12.0</u>	32.0	28.0	66.7	<u>60.0</u>	40.0	<u>33.3</u>	38.0
ResTacVLA (Ours)	**56.0**	**32.0**	**68.0**	**60.0**	**86.7**	**80.0**	<u>60.0</u>	**60.0**	**62.8**

We report the success rates (%) and average performance of ResTacVLA against various vision-only and naive tactile fusion baselines across five challenging tasks. Results distinguish between initial alignment (A) and successful physical interaction (I) to highlight the resolution of physical ambiguities. (Bold: best results; Underlined: second-best)

Representation by comparing it against alternative tactile integration strategies. Integrating tactile features from a standard ResNet encoder ($\pi_{0.5}$ w/ T-ResNet) yields only marginal improvements over the vision-only baseline, and even degrades performance in complex contact tasks such as *Lightbulb Screwing* (28.0% $\rightarrow$ 16.0%). While integrating a pre-trained SSL model ($\pi_{0.5}$ w/ T-UniT) provides more consistent improvements, the gains remain limited compared to ResTacVLA. This suggests that without effective regulation, high-dimensional tactile signals can act as distractors, introducing sensory interference that disrupts policy decision-making.

Conversely, incorporating the Residual Tactile Representation into the $\pi_{0.5}$ backbone elevates the success rate to 62.8%. This significant gain confirms that explicitly modeling the residual tactile—representing the deviation between expected and actual sensations—is essential for extracting orthogonal information gain that vision cannot provide. Furthermore, the application of the Residual Tactile Representation to the Diffusion Policy architecture results in a similar +19.2% improvement, demonstrating the architecture-agnostic effectiveness of our residual representation in tactile integration.

C. Interpretability Analysis

To validate the hypothesis that the CMP learns meaningful contact semantics and the SAG provides adaptive gating mechanisms, we conduct a qualitative analysis of the internal representations and gating dynamics.

Emergent Semantics of Latent Contact Primitives. As shown in Fig. 4, we investigate the structure of the learned VQ codebook by projecting per-frame latent contact primitives into a 2D space via t-SNE [35], with each frame annotated by both its interaction phase (free-space motion vs. physical contact) and task identity. Two salient structural properties emerge after joint pre-training across all five tasks. First, primitives corresponding to free-space motion—the dominant phase of every episode—converge into a single, compact cluster that is shared uniformly across all tasks, reflecting the inherently low tactile information gain during non-contact locomotion where visual priors are highly predictive. Second, during phases of physical interaction, the representation space disaggregates into a set of task-specific local clusters organized around semantically distinct contact events, including unexpected collision and alignment success.

This two-level organization—a global collapse for uninformative phases and a fine-grained task-specific partition for critical contact events—confirms that the Residual Tactile Representation successfully distills continuous, high-dimensional sensory streams into a compact, interpretable vocabulary of contact primitives, providing structured semantic abstractions that are both physically grounded and generalizable across diverse manipulation scenarios.

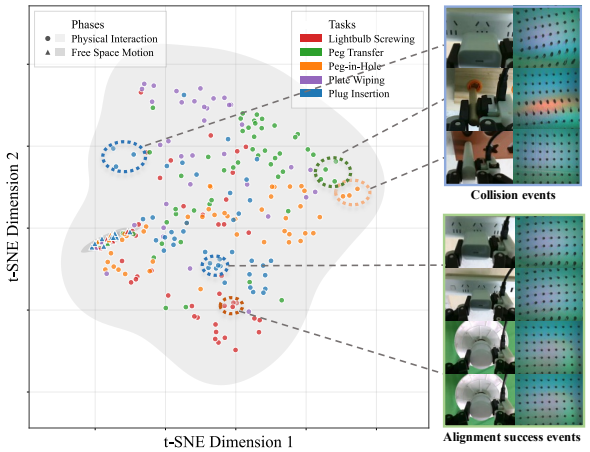


Fig. 4: t-SNE Visualization of Latent Contact Primitives. Each point represents a per-frame primitive, colored by task identity and annotated by task phase. Free-space motion frames from all five tasks converge into a single, shared compact cluster, reflecting their uniformly low tactile information gain. In contrast, physical interaction frames disaggregate into task-specific local clusters organized around semantically distinct contact events, such as unexpected collision and alignment success.

Adaptive Regulation via Surprise-Aware Gating. We further analyze the temporal evolution of the surprise-aware gate g_t during a successful *Lightbulb Screwing* trial (Fig. 5). The curve demonstrates that the gating mechanism naturally learns a phase-dependent modulation strategy consistent with human sensorimotor control. During the *Approach* phase (no contact), g_t remains attenuated (near zero), effectively suppressing tactile noise to allow the high-bandwidth visual policy to dominate trajectory planning. Conversely, upon physical contact establishment, g_t rapidly saturates, amplifying the influence of the tactile pathway. This behavior validates our predictive coding formulation: the policy actively gates tactile integration based on visual uncertainty,

prioritizing tactile feedback only when it provides critical information gain.

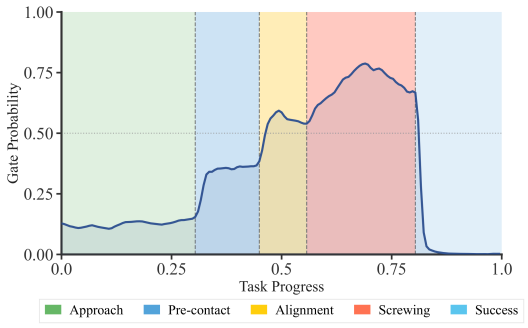


Fig. 5: Temporal Evolution of the Surprise-Aware Gate g_t During a *Lightbulb Screwing* Trial. The gate value is shown across five task phases: *Approach*, *Pre-contact*, *Alignment*, *Screwing*, and *Success*. g_t remains near zero during free-space motion and rises progressively as physical contact and thread engagement are established, validating that the gate learns to prioritize tactile feedback precisely when visual predictions become unreliable.

D. Ablation Studies

To validate the architectural design of ResTacVLA, particularly the Residual Quantization of the Cross-Modal Predictor and the Surprise-Aware Gating, we conducted comprehensive ablation studies on two representative tasks: *Plug Insertion* and *Plate Wiping*. The results are shown in Table II.

Latent Primitives vs. Continuous Residuals. We first investigate the contribution of the Vector Quantization (VQ) module by replacing the discrete Latent Contact Primitives with continuous, raw residual embeddings. As shown in Table II, this ablation results in a noticeable performance degradation (-26.7%). Qualitatively, this variant exhibits more frequent sporadic jitters and high sensitivity to grasp pose variations. This suggests that the VQ bottleneck serves as a critical information filter, distilling high-frequency tactile noise into semantic contact events that are more robust and generalizable for policy learning.

Surprise-Aware Gating vs. Fixed Fusion. We further analyze the necessity of the Surprise-Aware Gate. Replacing our adaptive mechanism with a fixed fusion strategy (constant weight of 1, i.e., always-on tactile fusion) leads to a significant drop in success rates (-13.3%). Specifically, the non-gated variant suffers from trajectory drift caused by tactile noise and initial grasp pose errors. This confirms that adaptive gating is essential for dynamically prioritizing tactile feedback during high-surprise events while preserving visual stability in free space.

E. Model Robustness

To further assess the adaptability of ResTacVLA under more challenging and varied conditions, we developed three task-specific robustness evaluations as shown in Fig. 6: (1) **Initial Grasp Perturbation** in *Plug Insertion*, adding translational (± 5 mm) and rotational ($\pm 10^\circ$) noise to simulate imperfect grasping conditions; (2) **Dynamic Perturbation** in *Peg-in-Hole*, applying random target displacements (3–5 cm) during execution to assess real-time reactivity; and (3) **Height Variation** in *Plate Wiping*, introducing surface height deviations ($+2$ cm and -2 cm) to evaluate force compliance.

TABLE II: Ablation Analysis of the ResTacVLA Architectural Components.

Configuration	Avg. Success (%)	Δ^*
ResTacVLA (Full)	**60.0**	-
w/o VQ (Continuous)[1]	33.3	-26.7
w/o Gating (Fixed)[2]	46.7	-13.3
$\pi_{0.5}$ (Vision Only)	20.0	-40.0

[*] We evaluate the performance gain Δ provided by the Residual Tactile Representation. [1]VQ-based contact discretization, and [2]Surprise-Aware Gating. Results represent the average success rate (%) across two representative tasks (*Plug Insertion* and *Plate Wiping*).

As shown in Table III, across all settings, ResTacVLA exhibited superior generalization, particularly in scenarios requiring fine physical interaction. Under *Initial Grasp Perturbation*, ResTacVLA maintained high success (52.0%), reflecting its reliance on tactile feedback to compensate for actuation misalignment beyond visual cues. In the *Dynamic Perturbation* setting, it achieved a 66.7% success rate, outperforming baselines that lacked tactile input or processed it naively. In the *Height Variation* setting, ResTacVLA achieved 53.3% (+2 cm) and 40.0% (-2 cm) success by effectively scaling its interaction forces to accommodate variable surface depths, avoiding the contact instability and unintended contact loss observed in vision-only models. These results underscore the critical role of the Residual Tactile Representation in intelligently integrating tactile cues—not just for sensing contact, but for modulating action in response to dynamic physical conditions—enabling more versatile and robust robotic manipulation.

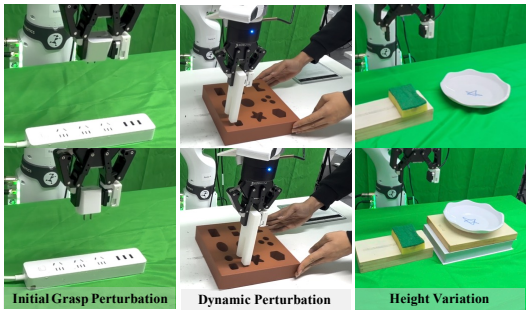


Fig. 6: Model Robustness Evaluation. Three perturbation scenarios are designed: (1) initial grasp perturbation applied to *Plug Insertion*, (2) dynamic target displacement during *Peg-in-Hole* execution, and (3) surface height variation applied to *Plate Wiping*.

V. CONCLUSIONS

In this work, we introduced ResTacVLA, a biologically inspired framework designed to efficiently integrate tactile feedback in VLA models for contact-rich manipulation. By reformulating tactile feedback as a residual representation grounded in predictive coding, our approach effectively isolates and amplifies the orthogonal information gain provided

TABLE III: Robustness Evaluation across Environmental and Actuation Perturbations.

Method	Dynamic	Height (+)	Height (-)	Grasp
$\pi_{0.5}$ (Vision Only)	26.7	33.3	0.0	8.0
$\pi_{0.5}$ w/ T-UniT	40.0	46.7	13.3	20.0
ResTacVLA (Ours)	**66.7**	**53.3**	**40.0**	**52.0**

We assess the adaptability of ResTacVLA under dynamic object displacements, significant surface height variations (± 2 cm), and initial grasp uncertainties to demonstrate its resilience in unmodeled environments. (Bold: best results)

by touch. This architecture couples latent contact primitives with surprise-driven gating, ensuring tactile cues are adaptively modulated across distinct task phases. Empirical evaluations across five challenging tasks demonstrate that ResTacVLA achieves state-of-the-art performance, exhibiting superior robustness to unexpected environmental disturbances and actuation uncertainties. These findings underscore the efficacy of residual-based sensor fusion in bridging the gap between high-level semantic planning and fine-grained physical interaction, paving the way for more adaptive and resilient generalist robot policies.

REFERENCES

[1] K. Black, N. Brown, D. Driess, A. Esmail, M. 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," *arXiv preprint arXiv:2410.24164*, 2024.

[2] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "π0.5: A vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[3] M. Heo, Y. Lee, D. Lee, and J. J. Lim, "Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1863–1891, 2025.

[4] M. Noseworthy, B. Tang, B. Wen, A. Handa, C. Kessens, N. Roy, D. Fox, F. Ramos, Y. Narang, and I. Akinola, "Forge: Force-guided exploration for robust contact-rich manipulation under uncertainty," *IEEE Robotics and Automation Letters*, 2025.

[5] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song, "Adaptive compliance policy: Learning approximate compliance for diffusion guided control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 4829–4836.

[6] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 5961–5968.

[7] J. Bi, K. Y. Ma, C. Hao, M. Z. Shou, and H. Soh, "Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback," *arXiv preprint arXiv:2507.17294*, 2025.

[8] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[9] A. Chaudhuri, A. Dutta, T. Bui, and S. Georgescu, "A closer look at multimodal representation collapse," *arXiv preprint arXiv:2505.22483*, 2025.

[10] 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," *arXiv preprint arXiv:2509.23468*, 2025.

[11] W. Chen, H. Xue, Y. Wang, F. Zhou, J. Lv, Y. Jin, S. Tang, C. Wen, and C. Lu, "Implicitrdp: An end-to-end visual-force diffusion policy with structural slow-fast learning," *arXiv preprint arXiv:2512.10946*, 2025.

[12] K. Kilteni and H. H. Ehrsson, "Predictive attenuation of touch and tactile gating are distinct perceptual phenomena," *Iscience*, vol. 25, no. 4, 2022.

[13] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[14] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[15] D. Wolpert, "Central cancellation of self-produced tickle sensation," *Nature Neuroscience*, vol. 1, pp. 635–640, 1998.

[16] H. Shi, B. Xie, Y. Liu, L. Sun, F. Liu, T. Wang, E. Zhou, H. Fan, X. Zhang, and G. Huang, "Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2508.19236*, 2025.

[17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[18] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023," *URL https://arxiv. org/abs/2307.15818*, vol. 1, p. 2, 2024.

[19] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang, "Tla: Tactile-language-action model for contact-rich manipulation," *arXiv preprint arXiv:2503.08548*, 2025.

[20] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu *et al.*, "Sparsh: Self-supervised touch representations for vision-based tactile sensing," *arXiv preprint arXiv:2410.24090*, 2024.

[21] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She, "Unit: Data efficient tactile representation with generalization to unseen objects," *IEEE Robotics and Automation Letters*, 2025.

[22] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi, "Touch begins where vision ends: Generalizable policies for contact-rich manipulation," *arXiv preprint arXiv:2506.13762*, 2025.

[23] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai *et al.*, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," *arXiv preprint arXiv:2505.22159*, 2025.

[24] Z. Zhang, H. Xu, Z. Yang, C. Yue, Z. Lin, H.-a. Gao, Z. Wang, and H. Zhao, "Ta-vla: Elucidating the design space of torque-aware vision-language-action models," *arXiv preprint arXiv:2509.07962*, 2025.

[25] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.

[26] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[27] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[28] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[30] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021.

[31] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[32] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[33] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[34] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.