2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Symmetry-Aware Fusion of Vision and Tactile Sensing via Bilateral Force Priors for Robotic Manipulation

Wonju Lee[1,†], Matteo Grimaldi[1], and Tao Yu[2]

***Abstract*— Insertion tasks in robotic manipulation demand precise, contact-rich interactions that vision alone cannot resolve. While tactile feedback is intuitively valuable, existing studies have shown that naïve visuo-tactile fusion often fails to deliver consistent improvements. In this work, we propose a Cross-Modal Transformer (CMT) for visuo-tactile fusion that integrates wrist-camera observations with tactile signals through structured self- and cross-attention. To stabilize tactile embeddings, we further introduce a physics-informed regularization that encourages bilateral force balance, reflecting principles of human motor control. Experiments on the TacSL benchmark show that CMT with symmetry regularization achieves a 96.59% insertion success rate, surpassing naïve and gated fusion baselines and closely matching the privileged "wrist + contact force" configuration (96.09%). These results highlight two central insights: (i) tactile sensing is indispensable for precise alignment, and (ii) principled multimodal fusion, further strengthened by physics-informed regularization, unlocks complementary strengths of vision and touch, approaching privileged performance under realistic sensing.**

I. Introduction

Robotic insertion is a long-standing benchmark in contact-rich manipulation, requiring accurate perception and fine-grained control under uncertainty. Vision-based policies, enabled by deep architectures such as CNNs [1] and Vision Transformers [2], excel at global scene understanding and object localization. However, in insertion tasks they often fail to capture subtle physical interactions—such as micro-slippage, compliance, or misalignment during contact—and remain sensitive to occlusion, lighting, and incomplete geometry [3]–[5].

Tactile sensing complements vision by directly measuring local contact states. GelSight [6], DIGIT [7], and TacTip [8] encode surface deformations that can be transformed into geometric and force-distribution descriptors. Empirical studies confirm their value in grasp stability [9], slip detection [10], and dexterous manipulation [11]. For insertion, tactile feedback is particularly crucial: contact forces signal misalignments and socket interactions that vision alone cannot infer.

Fig. 1 illustrates the complementary roles of different observation modalities. Vision (left) provides coarse global alignment but misses fine corrections. Tactile sensing (center) captures local force patterns essential for precise adjustments. Visuo-tactile fusion (right) combines these strengths, yielding robust insertion behavior. Quantitatively, augmenting either low-dimensional state inputs or wrist-camera observations with tactile signals improves insertion success by $+2.2\%$ and $+2.8\%$, respectively (Table I), underscoring the indispensability of contact feedback. We further observe in extended experiments on the screw task (Appendix VII) that tactile sensing alone achieves perfect success, reinforcing its role as a primary modality in contact-rich manipulation.

Yet, effectively leveraging tactile signals is challenging. TacSL [12] reported negligible or even negative gains with direct feature concatenation, highlighting two issues: (i) the difficulty of synchronizing heterogeneous representations, and (ii) the risk of diluting modality-specific cues. These limitations motivate the need for structured fusion strategies that respect the role of each modality and exploit their complementarity in a task-aware manner.

Recent works have investigated attention mechanisms for this purpose: self-attention for intra-modal correspondences [13], cross-modal alignment [14]–[16], timing-aware fusion [17], and force-guided weighting [18], [19]. Together, these advances suggest that principled attention-based fusion, rather than naïve concatenation, is key to exploiting visuo-tactile synergy.

In this paper, we propose a *symmetry-aware, physics-informed visuo-tactile fusion framework* tailored for robotic insertion. Our design is motivated by human motor control, where bilateral force balance ensures stability [20], [21]. Inspired by these principles, we introduce a **bilateral force regularization** that explicitly enforces vertical symmetry between left and right finger forces. This physics-informed inductive bias stabilizes grasps and reduces lateral misalignments during insertion.

The symmetry-regularized tactile embeddings are fused with visual features via a Cross-Modal Transformer (CMT), which applies hierarchical self- and cross-attention to integrate global visual context with local, physically consistent tactile feedback. By combining structured attention with physics-informed regularization, our framework yields smoother, more robust insertion trajectories compared to vision-only, naïve fusion, and TacSL baselines.

Our contributions are threefold:

- **Methodological novelty**: We propose a Cross-Modal Transformer (CMT) that integrates vision and tactile cues via hierarchical self- and cross-attention, addressing synchronization challenges that hinder naïve fusion.
- **Physics-informed regularization**: We introduce a bilateral force-symmetry constraint inspired by human motor control [20], [21], which encodes a physics-

† denotes the corresponding author
[1]DexAI, Emergent Business Unit, Analog Devices Inc., Limerick, Ireland (wonju.lee@analog.com, matteo.grimaldi@analog.com)
[2]DexAI, Emergent Business Unit, Analog Devices Inc., Boston, MA, USA (tao.yu@analog.com)

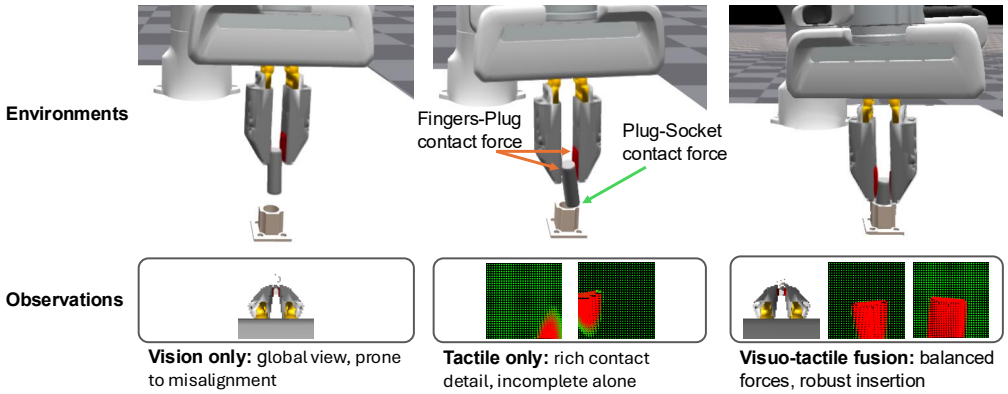

Fig. 1. Comparison of observation modalities for robotic insertion policies. **Left**: Vision-only input provides global alignment cues but lacks local precision. **Center**: Tactile-only input encodes fine-grained force signals critical for corrective actions. **Right**: Visuo-tactile fusion integrates coarse visual guidance with detailed tactile feedback, achieving robust insertion by exploiting complementary strengths.

based inductive bias for balanced grasping and stable insertion.

- **Experimental validation**: We demonstrate superior insertion performance over naïve and gated fusion, achieving 96.59%, nearly matching the privileged wrist+force configuration (96.09%).

II. Related Work

A. Vision-based Manipulation

Vision has long been the primary sensing modality for robotic manipulation due to its scalability and the maturity of visual deep learning. CNN- and Transformer-based policies [1], [2] achieve strong results in recognition and pick-and-place. However, vision-only pipelines often struggle in insertion tasks due to depth ambiguity, occlusion, and the absence of physical cues such as contact forces [3]–[5]. Large-scale visuomotor benchmarks such as RoboNet [22] and RLBench [23] reinforce this limitation, motivating the integration of tactile feedback in contact-rich domains.

B. Tactile Sensing for Contact-rich Tasks

Tactile sensing directly captures local interaction states, including contact geometry, slip, and force distributions. Vision-based tactile sensors such as GelSight [6], DIGIT [7], and TacTip [8] provide high-resolution deformation maps that CNN backbones can transform into compact descriptors. These signals have been shown to improve grasp stability [9], [24], slip detection [10], and dexterous in-hand manipulation [11]. Beyond optical sensors, force–torque sensors [25] and capacitive arrays [26] have been applied to insertion and assembly, though they offer lower spatial resolution. High-frequency tactile datasets further highlight the importance of transient slip and compliance cues [27]. For insertion in particular, augmenting visual or low-dimensional policies with tactile input consistently improves success, underscoring the indispensability of tactile sensing for alignment and frictional dynamics (Table I).

C. Visuo-Tactile Fusion Strategies

The central challenge in visuo-tactile fusion lies in reconciling heterogeneous feature spaces while preserving modality-specific information. Naïve concatenation, as reported in TacSL [12], often yields negligible or negative gains. To address this, structured fusion approaches have emerged. VTFSA [13] applies self-attention across visuo-tactile features, while ConViTac [14] and ViTacFormer [15] employ cross-modal Transformers for alignment. Adaptive methods such as AdapTac [18] and its extensions [19] use predictive force cues to reweight modalities dynamically. More recently, VTLA [28] and OmniVTLA [29] integrate language to guide fusion and improve task generalization. Despite these advances, most existing methods treat tactile signals as features to be adaptively weighted, rather than embedding explicit physical principles into the learning process.

D. Symmetry and Physical Priors

Human motor control studies highlight the role of bilateral symmetry and modular control in achieving stable and coordinated movements. Morasso [20] showed that reaching movements are spatially organized, while Bizzi et al. [21] demonstrated modular synergies in spinal control. In robotics, Ilonen et al. [30] exploited symmetry priors for tactile object reconstruction, and Su et al. [31] applied symmetry constraints for dual-arm manipulation. More broadly, physics-informed machine learning [32] has emerged as a paradigm for embedding physical constraints directly into neural architectures. Our work contributes to this line by introducing bilateral force symmetry as a physics-informed regularizer for visuo-tactile policy learning, rather than for perception or state estimation. This transforms tactile signals into structured, physically meaningful representations that stabilize grasps and reduce jamming during insertion.

E. Positioning of Our Work

Prior work has (i) highlighted the limitations of vision-only policies, (ii) shown the utility of tactile sensing, and (iii) explored structured visuo-tactile fusion strategies. However, existing approaches either assume symmetry for state estimation and 3D reconstruction [30] or employ force-guided attention to reweight modalities adaptively [18], [19]. These methods focus on object modeling or predictive force tasks rather than direct policy learning. In contrast,

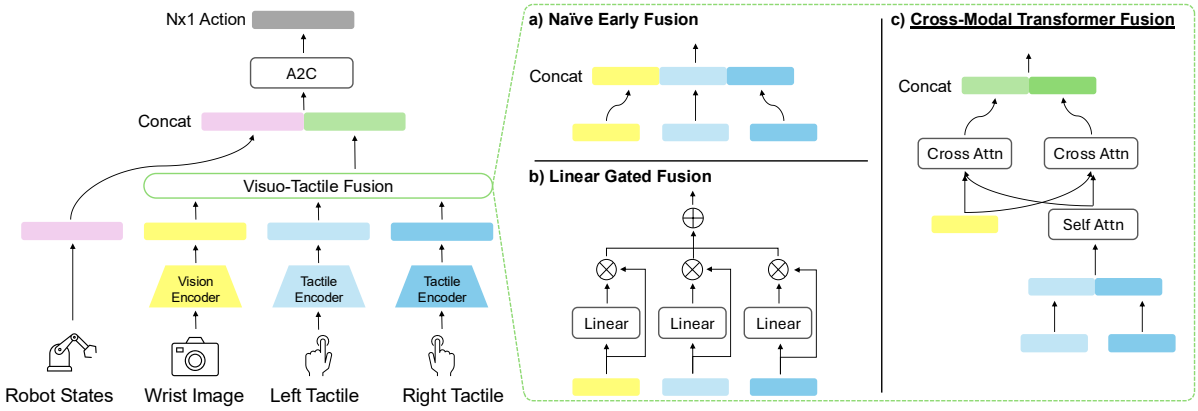

Fig. 2. Overview of visuo-tactile fusion architectures. (a) Naïve concatenation of embeddings, which risks diluting modality-specific signals. (b) Gated fusion with linear layers that adaptively weight neuronal contributions. (c) The proposed Cross-Modal Transformer (CMT), which embeds symmetry-aware tactile encoding and integrates vision and touch via cross-attention.

our framework leverages bilateral symmetry as a physics-informed regularization that stabilizes tactile embeddings and integrates them with vision through hierarchical attention. This combination yields a principled, policy-driven approach that advances beyond naïve concatenation and prior adaptive fusion mechanisms, enabling robust performance in contact-rich insertion.

III. Proposed Method: Symmetry-Aware Visuo-Tactile Fusion

We present a novel visuo-tactile fusion framework for robotic insertion that combines global alignment cues from vision with local corrective feedback from tactile sensing. Our method introduces a **physics-informed symmetry prior** to regularize tactile embeddings, which encourages balanced bilateral forces and mitigates jamming during insertion. This prior is integrated into a **Cross-Modal Transformer (CMT)** architecture that employs hierarchical self- and cross-attention, ensuring structured fusion between modalities. The overall framework is summarized in Fig. 2.

A. Problem Formulation

We formulate insertion as a partially observable Markov decision process (POMDP) defined by (S, A, O, T, R), where S is the latent state, A continuous robot actions, O multimodal observations, T transition dynamics, and R the sparse reward function. The observation O comprises (i) RGB wrist camera inputs for global alignment and (ii) tactile force fields from the gripper fingers for local contact sensing. The policy $\pi_\theta(a|o)$ aims to maximize the expected discounted return as

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{t_{\max}} \gamma^t r_t\right], \tag{1}$$

with $\gamma \in (0, 1)$ and reward $r_t = 1$ only upon successful insertion.

B. Residual Tactile Encoding with Symmetry Priors

Let the raw tactile forces at time t be $\bar{f}_t^L, \bar{f}_t^R \in \mathbb{R}^d$. To account for possible object asymmetry, we define residual forces relative to calibrated reference signals by

$$f_t^L = \bar{f}_t^L - f_{ref}^L, \quad f_t^R = \bar{f}_t^R - f_{ref}^R, \tag{2}$$

where f_{ref}^L, f_{ref}^R are prior forces obtained either through a short calibration contact or set to zero for symmetric objects. This formulation generalizes symmetry-aware balancing: symmetric objects correspond to $f_{ref}^L = f_{ref}^R = 0$, while asymmetric objects can be handled by anchoring to calibrated reference distributions.

The calibration is performed in a pre-insertion step by gently grasping the object to record the stable, pre-contact force distribution. This measured force profile serves as a physics-informed baseline, allowing the policy to learn corrective actions based on deviations from this nominal state rather than on the absolute force values themselves. While our current evaluation focuses on symmetric objects, this residual formulation is designed to be robust to object asymmetry and variations in grasp pose, generalizing the core principle of force balancing to a wider class of manipulation tasks.

The residuals are encoded via backbones into embeddings h_t^L, h_t^R. To capture bilateral consistency, we perform self-attention over concatenated tactile features as

$$z_t^T = \text{Attn}(W_q[h_t^L; h_t^R],\, W_k[h_t^L; h_t^R],\, W_v[h_t^L; h_t^R]), \tag{3}$$

yielding a residual-aware tactile representation that normalizes for asymmetry before fusion. This step enforces coherent intra-modal structure before cross-modal fusion.

C. Visuo-Tactile Cross-Attention

Camera observations v_t are embedded as $z_t^V = \phi^V(v_t) \in \mathbb{R}^k$. We then apply cross-attention with vision as *query* (guiding alignment) and tactile as *key/value* (providing corrective feedback) as

$$z_t^{VT} = \text{Attn}(W_q^V z_t^V,\, W_k^T z_t^T,\, W_v^T z_t^T). \tag{4}$$

This asymmetry reflects our design choice: vision provides global context, while tactile sensing supplies fine-grained local corrections. By stacking intra-tactile self-attention and visuo-tactile cross-attention, the CMT achieves **hierarchical fusion**, structurally modeling the roles of each modality.

TABLE I

SUCCESS RATES UNDER PRIVILEGED AND REDUCED SETTINGS ACROSS DIFFERENT SENSORY MODALITIES AND FUSION STRATEGIES. IMPROVEMENTS OVER THE REDUCED BASELINE ARE SHOWN IN PARENTHESES.

Method	Privileged	Reduced	Contact forces	Wrist	Tactile	Success rate (%)
Privileged	✓					96.74 ± 1.63
+ Contact forces	✓		✓			98.96 ± 0.83 (+2.22)
Tactile		✓			✓	91.41 ± 5.51
Wrist		✓		✓		93.23 ± 2.00
Wrist + Contact forces		✓	✓	✓		96.09 ± 1.41 (+2.86)
Fusion - Naïve [12]		✓		✓	✓	92.97 ± 1.41
Fusion - Gated ($\lambda_{\text{sym}} = 0$)		✓		✓	✓	94.53 ± 2.73 (+1.56)
Fusion - CMT ($\lambda_{\text{sym}} = 0$)		✓		✓	✓	96.22 ± 0.98 (**+3.25**)
Fusion - Gated + Symmetry regularization ($\lambda_{\text{sym}} = 1$)		✓		✓	✓	95.05 ±1.76 (+2.08)
Fusion - CMT + Symmetry regularization ($\lambda_{\text{sym}} = 1$)		✓		✓	✓	96.59 ±2.11 (**+3.62**)

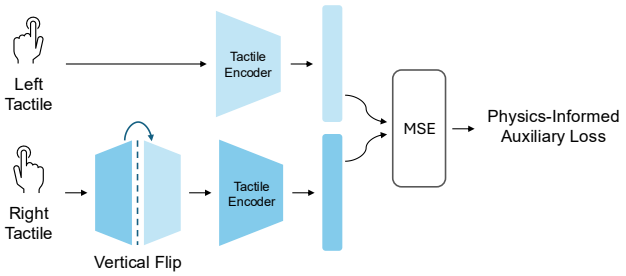

Fig. 3. Physics-informed symmetry regularization. The right tactile map is vertically flipped and encoded as $\tilde{h}_t^R$, then compared with h_t^L. The mean squared error loss penalizes deviations, encouraging bilateral consistency. This auxiliary objective stabilizes grasp forces before insertion and reduces lateral misalignment during insertion.

D. Physics-Informed Symmetry Regularization

Inspired by biological motor control principles [20], [21], we introduce a physics-informed auxiliary loss that enforces bilateral force balance (Fig. 3). Specifically, the right tactile map is vertically flipped and encoded as $\tilde{h}_t^R$, which is compared against h_t^L as

$$\mathcal{L}_{\text{sym}} = \mathbb{E}_{t\sim\mathcal{D}}\big[\|h_t^L - \tilde{h}_t^R\|_2^2\big]. \tag{5}$$

This regularization serves two functions: (i) *pre-insertion*, it suppresses asymmetric grasp forces and stabilizes initial contact; (ii) *during insertion*, it reduces lateral misalignments that otherwise cause jamming. We therefore interpret $\mathcal{L}_{\text{sym}}$ as a **physics-informed inductive bias**, injecting physical consistency into the learned policy. This stabilizes grasp forces and mitigates lateral misalignments, while generalizing from symmetric to asymmetric manipulation tasks.

E. Policy Optimization with PPO

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} + \lambda_{\text{sym}}\mathcal{L}_{\text{sym}}, \tag{6}$$

where $\mathcal{L}_{\text{PPO}}$ is the clipped surrogate objective of PPO, and λ_{sym} balances task reward with regularization. The stochastic policy outputs Gaussian actions with bounded variance

$$\sigma_\theta(o_t) \leftarrow \text{clamp}(\sigma_\theta(o_t), \sigma_{\min}, \sigma_{\max}), \tag{7}$$

ensuring training stability across random seeds.

Our framework unifies three innovations: (i) intra-modal tactile self-attention for symmetry-aware encoding, (ii) vision-guided cross-modal attention for structured fusion, and (iii) physics-informed symmetry regularization for stable and precise insertion. Together, these components advance visuo-tactile fusion beyond naïve concatenation or gated fusion, achieving near-privileged performance while retaining robustness in realistic, contact-rich scenarios.

IV. EXPERIMENTS

We evaluate the proposed symmetry-aware visuo-tactile fusion framework on the robotic insertion task. We first describe the experimental setup, including environment and sensors, followed by the baselines considered for comparison. We then present quantitative results on insertion success rates and qualitative analyses highlighting the role of tactile sensing and symmetry priors. All experiments are conducted with three different random seeds, and reported values correspond to averages across these runs with standard deviations.

Following the evaluation framework established in TacSL [12], we adopt the same insertion task setup and success criteria (cf. their Table V, Table VI, and Table VII). This ensures consistency in benchmarking and enables a direct comparison of our method against prior visuo-tactile fusion approaches. A detailed description of the experimental setup, sensor configurations, and additional results is provided in the Appendix VI. To foster reproducibility, all training code and task configurations will be released publicly in the official IsaacGymEnvs repository[1].

A. Quantitative Analysis

Table I reports insertion success rates across different observation modalities and fusion strategies. A first observation is that incorporating contact force sensing consistently improves performance. In the privileged setting with compact state features, adding contact forces increases success from 96.74% to 98.96%, a $+2.22\%$ absolute gain. A similar trend is observed in the vision-based policy: augmenting the wrist camera with contact force raises success from 93.23% to 96.09% ($+2.86\%$). These results indicate that contact feedback conveys information—such as micro-slippage and compliance—that vision alone cannot provide.

[1]https://github.com/isaac-sim/IsaacGymEnvs

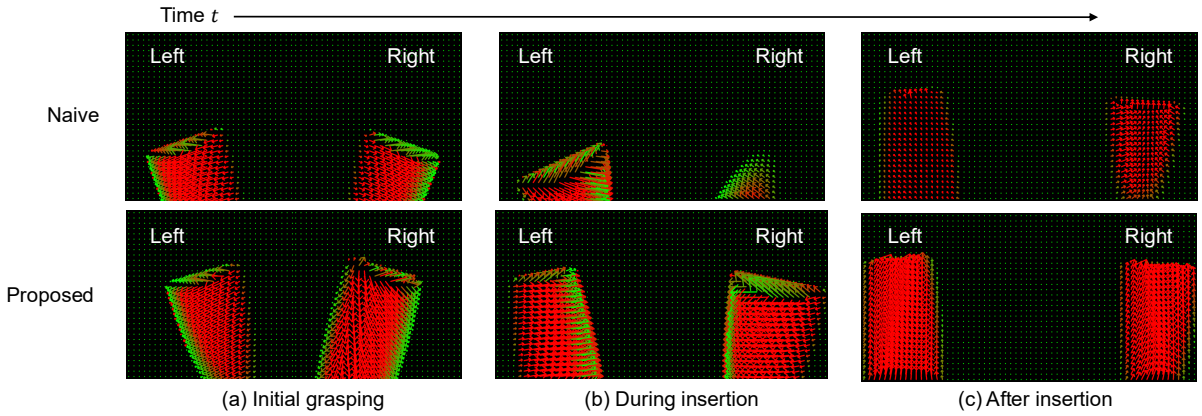

Fig. 4. Evolution of bilateral force fields during insertion under two fusion strategies. **Top**: Naïve fusion does not enforce symmetry; contact with the socket induces pronounced left–right imbalance, triggering unstable corrections and occasional re-grasping. **Bottom**: The proposed symmetry-aware CMT maintains balanced force distributions throughout the episode, reducing unnecessary lateral contact and yielding a straighter, smoother insertion trajectory aligned with the table normal. This illustrates how explicitly modeling bilateral symmetry stabilizes contact-rich manipulation under visuo-tactile fusion.

Tactile sensing also demonstrates strong standalone performance: the tactile-only policy reaches 91.41%, showing that rich geometric and force-distribution information is directly embedded in tactile signals, even without visual context. This robustness suggests that tactile feedback can sustain task execution under degraded or occluded visual conditions.

The most significant improvement arises in visuo-tactile fusion. While naïve concatenation and gated fusion yield only moderate gains over unimodal policies, the proposed CMT achieves 96.22%, substantially surpassing both baselines. Notably, this performance nearly matches the privileged "wrist + contact force" configuration (96.09%), underscoring that structured cross-modal attention can extract benefits previously attainable only with privileged supervision. Furthermore, adding the proposed symmetry regularization boosts performance in both gated and CMT architectures. For gated fusion, the success rate increases from 94.53% to 95.05%, while for CMT it rises from 96.22% to 96.59%. These gains, though modest, indicate that symmetry-aware balancing acts as a stabilizing inductive bias, reducing variability across seeds and further narrowing the gap to privileged sensing.

In summary, three insights emerge: (i) contact forces are indispensable for precise insertion across all policy types; (ii) tactile sensing is independently informative and offers resilience when vision is limited; and (iii) principled fusion via CMT unlocks the full potential of multimodal perception, narrowing the gap to privileged force feedback.

B. Qualitative Analysis

Fig. 4 contrasts the evolution of left/right tactile force fields for naïve versus symmetry-aware visuo-tactile fusion. In the naïve case (top row), contact events with the socket wall disturb the balance between the two fingers; the resulting lateral forces and torques manifest as asymmetric flow patterns over the tactile arrays, often followed by corrective re-grasping and attitude oscillations. These behaviors are typical when fusion does not exploit structure, echoing prior observations that simple concatenation frequently fails to capitalize on multimodal complementarity in contact-rich tasks [12].

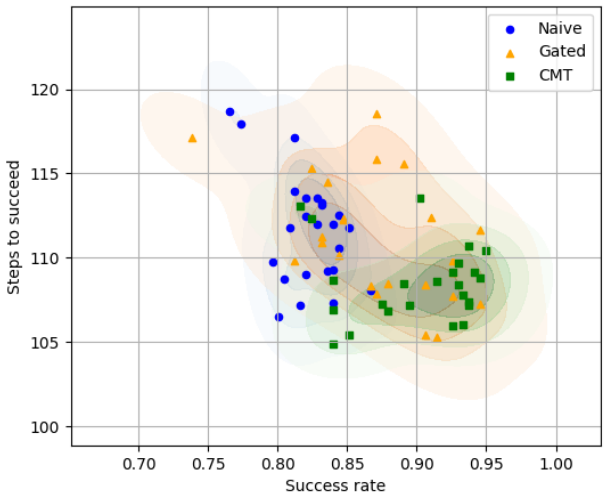

Fig. 5. Distributions of insertion performance for Naïve (blue), Gated (orange), and CMT (green). Scatter points denote individual trials, with kernel density contours indicating outcome distributions in terms of success rate (x-axis) and steps to succeed (y-axis).

In contrast, the proposed CMT-based fusion (bottom row) couples a symmetry-aware tactile encoder with cross-modal attention. By vertically mirroring the right tactile map and aligning it to the left in feature space, the model learns bilateral consistency even under transient contacts; cross-attention then gates these tactile cues with visual context so that pose corrections are made before sustained contact occurs. The resulting force fields remain balanced as the gripper transitions from pre-insertion to insertion, yielding trajectories that are close to straight-down motion, with near-zero net lateral torque—akin to human practice of pre-aligning the plug and then inserting cleanly to minimize incidental contact.

To quantify these qualitative observations, Fig. 5 reports the distribution of steps required to complete insertion. The naïve fusion policy requires 111.63 steps on average, whereas CMT reduces this to 108.48 steps (-2.83%). Impor-

TABLE II

INFERENCE LATENCY, MEMORY USAGE, AND THROUGHPUT OF DIFFERENT METHODS. LATENCY IS MEASURED OVER 1,000 FORWARD PASSES AFTER WARM-UP, AND THROUGHPUT IS COMPUTED AS $1000/\text{LATENCY}$.

Method	Latency (ms)	Memory (MB)	Throughput (fps)
Naïve	5.42	19.24	184.50
Gated	5.51	17.43	181.49
CMT	6.52	21.45	153.37

tantly, this reduction directly follows from the balanced force distributions observed in Fig. 4 by maintaining symmetry throughout insertion, CMT reduces corrective oscillations, which shortens the trajectory and yields more efficient execution. This provides a clear causal link between qualitative stability and quantitative improvements.

C. Computation Analysis

Finally, we assess the computational efficiency of the fusion models. Table II reports latency, throughput, and memory usage, averaged over 1,000 forward passes. Throughput is computed as $1000/\text{latency}$ (ms).

The naïve model achieves the lowest latency (5.42 ms, 184.5 fps). The gated model shows comparable runtime (5.51 ms, 181.5 fps) while achieving the smallest memory footprint (17.43 MB). The proposed CMT, which incorporates cross-modal Transformer layers, incurs slightly higher cost (6.52 ms, 153 fps, 21.45 MB), yet remains well within real-time control requirements of 60 Hz.

Although CMT incurs $\sim$20% higher latency than the naïve baseline, this trade-off is justified by its substantial $+3.62\%$ improvement in success rate (Table I). Importantly, all models exceed 150 fps, indicating that performance differences are not constrained by computation but rather by the ability to exploit multimodal structure effectively.

Overall, naïve and gated fusion offer minimal overhead, but the proposed CMT provides the best trade-off between computational efficiency and performance, making it the most practical choice for real-world deployment.

V. CONCLUSION

We proposed a visuo-tactile fusion framework for robotic insertion that combines cross-modal attention with a physics-informed balancing loss. Experiments demonstrate three consistent insights: (i) tactile sensing is indispensable for precise alignment, (ii) tactile-only policies remain robust under degraded vision, and (iii) structured fusion via a Cross-Modal Transformer achieves near-privileged performance in real time. These findings establish tactile sensing as a **core modality for contact-rich manipulation**.

While validated here on symmetric plug insertion, the formulation generalizes naturally to asymmetric objects through reference calibration and data-driven invariances. Future work will extend evaluation to such tasks. By uniting the indispensability of tactile feedback with the stability of physics-informed regularization, this work lays the foundation for **general visuo-tactile policies** that adapt across diverse geometries and real-world assembly scenarios.

TABLE III

ENVIRONMENT RANDOMIZATION BOUNDS (ADAPTED FROM TACSL APPENDIX C, TABLE V). EACH PARAMETER IS UNIFORMLY SAMPLED WITHIN THE SPECIFIED RANGE.

Parameter	Range
End-effector X (m)	[0.4, 0.6]
End-effector Y (m)	[-0.1, 0.1]
End-effector Z (m)	[0.1, 0.2]
End-effector Euler-X (rad)	[3.04, 3.24]
End-effector Euler-Y (rad)	[-0.1, 0.1]
End-effector Euler-Z (rad)	[-1.0, 1.0]
Socket X (m)	[0.4, 0.6]
Socket Y (m)	[-0.1, 0.1]
Socket Z (m)	[0.0, 0.02]
Peg-in-gripper Z-pos (m)	[-0.0125, 0.0125]
Peg-in-gripper X-rot (rad)	[-0.628, 0.628]
Socket XYZ noise (m)	[-0.005, 0.005]
Compliance stiffness noise (N/m)	[150, 350]
Compliance damping noise (N/(m/s))	[0.0, 1.0]
Joint damping noise (N/(m/s))	[-1.5, 1.5]

TABLE IV

POLICY ARCHITECTURE. VISION INPUT IS $64 \times 64 \times 3$, TACTILE INPUT IS $32 \times 32 \times 3$ (3 CHANNELS FOR f_x, f_y, f_z). ALL MODALITIES SHARE THE SAME CNN ENCODER STRUCTURE; OUTPUT IS PROJECTED TO A 128-D EMBEDDING. FUSION MODULE DIFFERS BY METHOD.

Layer / Module	Configuration	Output Dim.
Encoder (per modality)		
Input (Vision)	$64 \times 64 \times 3$	$(B, 3, 64, 64)$
Conv1	8×8, stride 2, 32 channels	$(B, 32, 29, 29)$
Conv2	4×4, stride 1, 64 channels	$(B, 64, 26, 26)$
Conv3	3×3, stride 1, 64 channels	$(B, 64, 24, 24)$
Spatial SoftArgMax	64 channels $\times$ 2D coords	$(B, 128)$
Input (Tactile)	$32 \times 32 \times 3$	$(B, 3, 32, 32)$
Conv1	8×8, stride 2, 32 channels	$(B, 32, 13, 13)$
Conv2	4×4, stride 1, 64 channels	$(B, 64, 10, 10)$
Conv3	3×3, stride 1, 64 channels	$(B, 64, 8, 8)$
Spatial SoftArgMax	64 channels $\times$ 2D coords	$(B, 128)$
Fusion (varies by method)		
Naïve Fusion	Concatenate [vision, left, right]	$(B, 384)$
Gated Fusion	Weighted sum [vision, left, right]	$(B, 128)$
CMT Fusion	Self-attn + cross-attn	$(B, 256)$
Policy Head		
RNN	2-layer LSTM, 256 hidden units	$(B, 256)$
MLP	[256, 128, 64] + ELU	$(B, 64)$
Policy Mean	Linear $\rightarrow \mathbb{R}^6$	$(B, 6)$
Policy Log-Std	Linear $\rightarrow \mathbb{R}^6$	$(B, 6)$
Value Head	Linear $\rightarrow \mathbb{R}^1$	$(B, 1)$

To foster reproducibility, all code and configuration files will be released, enabling direct replication of our results.

APPENDIX

VI. TRAINING DETAILS AND EXPERIMENTAL SETUP

We follow the training setup and experimental design of TacSL [12], and reproduce the key details here for completeness. This ensures that our results are directly comparable to their benchmarks, and that the proposed visuo-tactile fusion framework can be reproduced independently of their paper.

A. Environment Randomization

Table III summarizes the task randomization levels applied when generating initial states for the insertion task. The end-effector is randomized around a nominal home pose, the peg is initialized with random offsets inside the gripper,

TABLE V

A2C HYPERPARAMETERS.

Parameter	Value
Optimizer	Adam
Learning rate	1.0e-4
Rollout / Horizon length	512 steps
Mini-batch size	512
PPO / Mini epochs	4
Discount factor γ	0.99
GAE parameter λ	0.95
Clip ratio ϵ	0.2
Max gradient norm	1.0
Entropy coefficient	0.0
Bounds loss coefficient	0.0001
Value loss / Critic coefficient	2

and the socket is placed with randomized position in front of the robot. Additionally, contact parameters and damping coefficients are randomized, and observation noise is applied to socket localization to simulate imperfect perception.

B. Policy Model Architecture

Following TacSL [12], we use a lightweight 3-layer CNN encoder with a Spatial SoftArgMax layer. The CNN extracts local features from visual or tactile inputs, while the SoftArgMax produces differentiable keypoint-like embeddings, preserving geometric information critical for insertion. On top of this shared encoder, we implement three fusion strategies (naïve, gated, and CMT), combined with a recurrent and feedforward policy head. The complete architecture is summarized in Table IV.

C. Training Hyperparameters

We train all policies with PPO using the hyperparameters listed in Table V. These are directly taken from TacSL [12] Appendix C (Table VII).

D. Code and Configuration Availability

For completeness, we follow the official TacSL task configuration provided in their public repository. The full environment specification for the insertion task (including randomization, observation spaces, and reward shaping) is available in the TacSL branch of the IsaacGymEnvs repository at TacSLTaskInsertion.yaml[2].

VII. TACTILE SENSING FOR SCREW TASKS

To further assess the role of tactile sensing, we extend our evaluation to the screw task, which involves grasping a nut and performing screwing motions onto a bolt. We follow the experimental setup and environment specifications described in Factory [33], ensuring consistency with prior benchmarks.

Our results, summarized in Fig. 6, reveal that tactile sensing alone is sufficient to solve the screw task with perfect reliability. Specifically, the tactile-only policy achieves a 100% success rate across all evaluated trials, outperforming both the privileged contact-force baseline and the wrist-camera policy. This advantage likely stems from the tactile arrays providing a $3 \times 32 \times 32$ measurement of surface deformations, capturing rich geometric and force details. In contrast, the privileged baseline only observes a compact 3×1 contact-force vector.

[2]https://github.com/isaac-sim/IsaacGymEnvs/blob/tacsl/isaacgymenvs/cfg/task/TacSLTaskInsertion.yaml

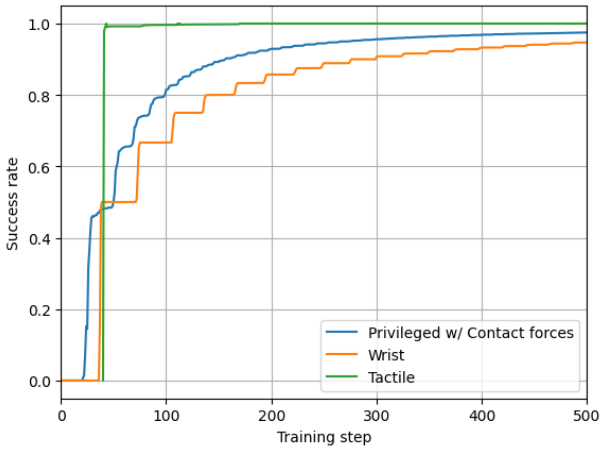

Fig. 6. Insertion performance for the screw task under three policies: Privileged (blue), Wrist-camera (orange), and Tactile (green). The high-resolution tactile policy achieves perfect success, outperforming even the privileged contact-force baseline, demonstrating the benefit of detailed tactile feedback for precise contact-rich manipulation.

These findings suggest that for manipulation tasks dominated by fine-grained contact interactions, high-resolution tactile feedback can serve as a complete substitute for vision. While vision may still be beneficial for broader alignment or multi-step tasks, Fig. 6 demonstrates that tactile sensing alone is sufficient for precise screw insertion, highlighting its potential as a primary modality for contact-intensive operations.

Expanding the task scope to include both *insert* (lifting and aligning the nut onto the bolt) and *screw* (rotational insertion of the nut) stages will likely restore the necessity of multimodal fusion. In such settings, vision can provide global pose and alignment context, while tactile sensing ensures precise contact tracking during screwing. We expect that principled visuo-tactile fusion will thus be critical for scaling toward real-world robot assembly, where robustness, generalization, and efficiency are essential for industrial deployment.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learning Representations (ICLR)*, 2021.

[3] F. Gu, Y. Zhou, Z. Wang, S. Jiang, and B. He, "A survey on robotic manipulation of deformable objects: Recent advances, open challenges and new frontiers," *arXiv preprint arXiv:2312.10419*, 2023.

[4] L. Bergamini, M. Sposato, M. Pellicciari, M. Peruzzini, S. Calderara, and J. Schmidt, "Deep learning-based method for vision-guided robotic grasping of unknown objects," *Adv. Engineering Informatics*, 2020.

[5] W. Qi, H. Fan, H. R. Karimi, and H. Su, "An adaptive reinforcement learning-based multimodal data fusion framework for human–robot confrontation gaming," *Neural Networks*, 2023.

[6] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, 2020.

[8] N. F. Lepora, "Soft biomimetic optical tactile sensing with the TacTip: A review," *IEEE Sensors*, 2021.

[9] J. M. Yau, S. S. Kim, P. H. Thakur, and S. J. Bensmaia, "Feeling form: the neural basis of haptic shape perception," *J. Neurophysiology*, 2015.

[10] X. Hu, A. Venkatesh, Y. Wan, G. Zheng, N. Jawale, N. Kaur, X. Chen, and P. Birkmeyer, "Learning to detect slip through tactile estimation of the contact force field and its entropy properties," *Mechatronics*, 2024.

[11] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, "Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing," *Robotics and Autonomous Systems*, 2025.

[12] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "TacSL: A library for visuotactile sensor simulation and learning," in *Robotics: Science and Systems (RSS)*, 2025.

[13] S. Cui, R. Wang, J. Wei, J. Hu, and S. Wang, "Self-attention based visual-tactile fusion learning for predicting grasp outcomes," *IEEE Robotics and Automation*, 2020.

[14] Z. Wu, Y. Zhao, and S. Luo, "ConViTac: Aligning visual-tactile fusion with contrastive representations," *arXiv preprint arXiv:2506.20757*, 2025.

[15] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik, "ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation," *arXiv preprint arXiv:2506.15953*, 2025.

[16] J. Li, T. Wu, J. Zhang, Z. Chen, H. Jin, M. Wu, Y. Shen, Y. Yang, and H. Dong, "Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation," *arXiv preprint arXiv:2505.13982*, 2025.

[17] J. Lenz, T. Gruner, D. Palenicek, T. Schneider, and J. Peters, "Analysing the interplay of vision and touch for dexterous insertion tasks," *arXiv preprint arXiv:2410.23860*, 2024.

[18] J. Li, T. Wu, J. Zhang, Z. Chen, H. Jin, M. Wu, Y. Shen, Y. Yang, and H. Dong, "Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation," *arXiv preprint arXiv:2505.13982*, 2025.

[19] P. Jin, B. Huang, W. W. Lee, T. Li, and W. Yang, "Visual-force-tactile fusion for gentle intricate insertion tasks," *IEEE Robotics and Automation Letters*, 2024.

[20] P. Morasso, "Spatial control of arm movements," *Experimental Brain Research*, vol. 42, no. 2, pp. 223–227, 1981.

[21] E. Bizzi, S. F. Giszter, E. Loeb, F. A. Mussa-Ivaldi, P. Saltiel, "Modular organization of motor behavior in the frog's spinal cord," *Trends in Neurosciences*, 1995.

[22] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, C. Finn, "RoboNet: Large-scale multi-robot learning," in *Conference on Robot Learning (CoRL)*, 2019.

[23] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark," *IEEE Robotics and Automation Letters*, 2020.

[24] Z. Xu and Y. She, "LeTac-MPC: Learning model predictive control for tactile-reactive grasping," *IEEE Trans. Robotics*, 2024.

[25] A. Al-Yacoub, Y. C. Zhao, W. Eaton, Y. M. Goh, and N. Lohse, "Improving human robot collaboration through Force/Torque based learning for object manipulation," *Robotics and Computer-Integrated Manufacturing*, 2021.

[26] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using finger vision," in *IEEE-RAS Int. Conf. on Humanoid Robots*, 2017.

[27] D. Sliwowski, S. Jadav, S. Stanovcic, J. Orbik, J. Heidersberger, and D. Lee, "Demonstrating REASSEMBLE: A multimodal dataset for contact-rich robotic assembly and disassembly," in *Robotics: Science and Systems (RSS)*, 2025.

[28] 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.

[29] Z. Cheng, Y. Zhang, W. Zhang, H. Li, K. Wang, L. Song, and H. Zhang, "OmniVTLA: Vision-tactile-language-action model with semantic-aligned tactile sensing," *arXiv preprint arXiv:2508.08706*, 2025.

[30] J. Ilonen, J. Bohg, and V. Kyrki, "Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing," *Int. J. Robotics Research*, 2014.

[31] Z. Su, X. Huang, D. Ordoñez-Apraez, Y. Li, Z. Li, and Q. Liao, "Leveraging symmetry in RL-based legged locomotion control," in *IEEE/RSJ Int. Confer. Intelligent Robots, Systems (IROS)*, 2024.

[32] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, 2021.

[33] C. Chi, A. Y. Wang, T. Xiao, S. Liu, Y. Zhu, A. Garg, and S. Song, "Factory: Fast contact for robotic assembly," in *Robotics: Science and Systems (RSS)*, 2022.