ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Enhancing VLA Precision in Robotic Manipulation via FiLM-based Force/Torque-Vision Integration

Gunhee Nam, *Student Member, IEEE*, and Ayoung Hong, *Member, IEEE*

***Abstract*— We propose a multimodal integration framework to enhance the precision of Vision-Language-Action (VLA) models in contact-rich robotic tasks. Although visual perception is essential for task grounding, it often lacks the force awareness required for high-precision alignment and insertion. To address this limitation, we leverage Feature-wise Linear Modulation (FiLM) to condition intermediate visual representations on 6-axis Force/Torque (F/T) data. This lightweight fusion strategy allows the model to modulate its action predictions based on real-time physical resistance without incurring significant computational overhead. Experimental results on a UR5e manipulator demonstrate that the proposed F/T-Vision integration enhances contact stability and precision in demanding manipulation tasks compared with vision-only baselines.**

I. Introduction

Vision-Language-Action (VLA) models have advanced robotic manipulation, but they struggle with contact-rich tasks such as high-precision insertion due to their heavy reliance on visual perception. Visual limitations, such as self-occlusion and the inability to perceive physical resistance, often lead to task failures. To address this, we propose a lightweight framework that integrates 6-axis Force/Torque (F/T) data into the VLA architecture using Feature-wise Linear Modulation (FiLM). By conditioning visual features on real-time force signals, our model dynamically modulates actions, significantly improving success rates and contact stability compared to vision-only baselines.

II. Proposed Method

A. Force-Conditioned VLA Architecture

We adopt $\pi_{0.5}$, a VLA model based on flow-matching, as our primary policy. While $\pi_{0.5}$ excels at semantic understanding by processing language and multi-view images via a SigLIP encoder, it lacks the physical grounding required for contact-rich tasks. In this work, we expand the input modalities of $\pi_{0.5}$ to encompass 6-axis Force/Torque (F/T) information. Integrating such signals into pre-trained VLAs faces two primary challenges. First, VLA models require substantial compute for adaptation, creating a demand for lightweight fusion strategies that remain accessible within a post-training finetuning paradigm. Second, the common baseline appends additional tokens to the VLA input, increasing the sequence length and computational cost.

This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIR)(RS-2024-00406796, HRD Program for Industrial Innovation)

Gunhee Nam, and Ayoung Hong are with the Department of Mechanical Engineering, Chonnam National University, Gwangju 61186, Republic of Korea ahong@jnu.ac.kr

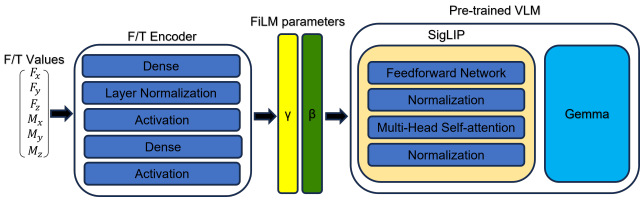

Fig. 1. Internal structure of the Force-Encoder and its integration with the SigLIP-based $\pi_{0.5}$ model.

B. ForceEncoder and Modulation Mechanism

The ForceEncoder serves as a bridge, translating raw physical signals into the modulation space of the VLA. As shown in Fig.3, the encoder consists of a multi-layer perceptron (MLP) that processes the 6-axis F/T input $f \in \mathbb{R}^6$ through the following stages:

1) Feature Extraction: The raw input is projected into a latent space $z \in \mathbb{R}^{256}$ using dense layers with Layer Normalization (LN) and SiLU activation to handle the noise and scale variance of the sensor data,

$$z = \text{SiLU}(\text{LN}(\text{FC}_{256}(f))) \tag{1}$$

2) Modulation Parameter Generation: Two separate linear heads project the latent vector z into the visual embedding dimension w (e.g., 1152):

$$\gamma = \text{FC}_{\gamma}(z), \quad \beta = \text{FC}_{\beta}(z) \tag{2}$$

We initialize the kernels of these heads to zero (Identity Initialization) so that at the start of training, $\gamma = 0$ and $\beta = 0$, ensuring a stable transition as the model learns to incorporate force awareness.

3) Global Conditioning in SigLIP: Unlike simple concatenation, we integrate the extracted force features by modulating the visual tokens within the SigLIP encoder. The generated γ and β are injected into every Transformer block of the vision tower. Within each block, the visual features F_{vis} are modulated as follows:

$$F_{mod} = (1 + \gamma) \odot F_{vis} + \beta \tag{3}$$

where $\odot$ denotes element-wise multiplication. By applying this modulation across all L layers, the model can iteratively refine its visual representations while being constrained by the real-time physical context.

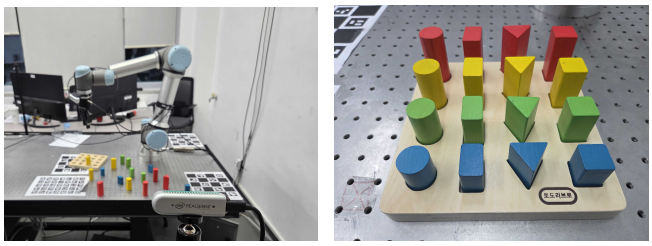

Fig. 2. Experimental Environment, (right) Task Board

III. Experimental Setup and Dataset

A. Experimental Setup view

The experimental setup for multimodal demonstration is shown in Fig.2. We use a 6-DoF UR5e robot manipulator with integrated 6-axis force/torque sensing, enabling end-effector force estimation. The visual system consists of two Intel RealSense D435 cameras, an overview camera for global scene context and a wrist-mounted camera for local alignment during high-precision tasks.

B. Multimodal Dataset Collection

To train the force-conditioned VLA policy, we constructed a high-quality dataset of 270 expert demonstrations. These demonstrations were collected using a haptic device with 1:1 end-effector mapping, ensuring precise control and high-fidelity trajectory data for complex manipulation tasks.

The task involves identifying objects of various colors and geometric shapes and inserting them into corresponding slots on a task board. This requires the model to jointly leverage semantic visual cues and real-time physical feedback. Each episode consists of synchronized data streams, including dual-view RGB images, joint states, and 6-axis F/T signals, recorded at 20 Hz. This dataset serves as the basis for learning the relationship between visual features and physical resistance in high-precision, contact-rich manipulation tasks.

IV. Experimental Results and Analysis

The proposed force-conditioned VLA was evaluated against a vision-only baseline on the UR5e platform.

A. Common Proficiency in Semantic Grounding

Both models demonstrated high success in high-level task grounding, such as identifying, approaching, and picking up target objects based on their color and geometric shape. This confirms that the underlying VLA backbone effectively processes semantic instructions and visual context for initial manipulation stages.

B. Baseline Failure (Force-agnostic Pressing)

Despite successful grasping, the vision-only baseline consistently failed during the high-tolerance insertion phase. Lacking physical feedback, it maintained constant downward pressure even when mechanical misalignment occurred. This force-agnostic execution caused the object to slip within the gripper, often leading to task failure and potential hardware strain.

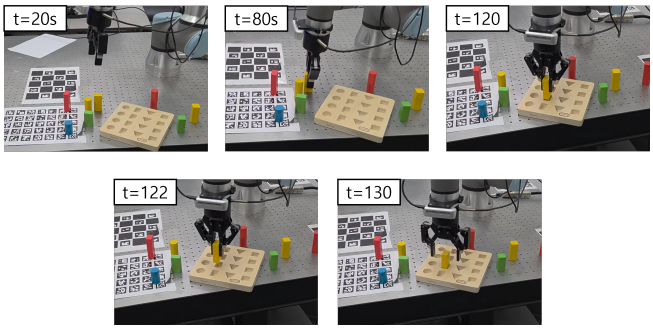

Fig. 3. Phase progression: (t=20s-80s) Approach and semantic task grounding, (t=120s) Collision detection, (t=122s) Adaptive retraction for realignment, (t=130s) Successful completion.

C. Force-aware Adaptation (Proposed)

In contrast, our model exhibited reactive behaviors upon encountering physical resistance. Instead of forced insertion, it performed subtle retractions or pauses to maintain grasp stability. This allowed the model to preserve the object's alignment, leading to successful insertions in scenarios where the baseline failed. These results validate that FiLM-based F/T integration enables real-time trajectory adaptation to physical constraints.

V. Conclusion

This paper presented a multimodal framework that enhances Vision-Language-Action (VLA) models by integrating 6-axis Force/Torque (F/T) data. By leveraging a ForceEncoder and FiLM-based modulation, we enabled the $\pi_{0.5}$ model to perceive and react to physical resistance during contact-rich tasks. Real-world experiments with a UR5e manipulator demonstrated that force awareness is essential for stable, safe interaction in high-precision scenarios. The proposed model effectively mitigated force-agnostic failures, such as gripper slippage due to excessive pressure, through adaptive, reactive behaviors. Future work will focus on scaling the multimodal dataset with diverse expert demonstrations to achieve higher success rates and robust generalization across a wider variety of manipulation tasks.

References

[1] K. Black et al., "π_0: A Vision-Language-Action Flow Model for General Robot Control," *arXiv preprint arXiv:2410.24164*, 2024.

[2] Physical Intelligence et al., "$\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization," 2025.

[3] C. Zhang et al., "VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[4] C. Morissette et al., "Tactile Modality Fusion for Vision-Language-Action Models," *arXiv preprint arXiv:2603.14604*, 2026.

[5] J. Yu et al., "ForceVLA: Enhancing VLA Models with a Force-Aware MoE for Contact-Rich Manipulation," *arXiv preprint arXiv:2505.22159*, 2025.