arXiv:2604.23272v2 [cs.RO] 11 Sep 2026

Modular Sensory Stream for Integrating Physical Feedback in Vision-Language-Action Models

Jimin Lee[1], **Huiwon Jang**[1,2], **Myungkyu Koo**[1,2], **Jungwoo Park**[1], **Jinwoo Shin**[1,2]
[1]KAIST [2]RLWRLD
{jiminl, jinwoos}@kaist.ac.kr

Abstract: Humans understand and interact with the real world by relying on diverse physical feedback beyond visual perception. Motivated by this, recent approaches attempt to incorporate physical sensory signals into Vision-Language-Action models (VLAs). However, they typically focus on a single type of physical signal, failing to capture the heterogeneous and complementary nature of real-world interactions. In this paper, we propose MoSS, a modular sensory stream framework that adapts VLAs to leverage multiple sensory signals for action prediction. Specifically, we introduce decoupled modality streams that integrate heterogeneous physical signals into the action stream via joint cross-modal self-attention. To enable stable incorporation of new modalities, we adopt a two-stage training scheme that freezes pretrained VLA parameters in the early stage. Furthermore, to better capture contact interaction dynamics, we incorporate an auxiliary task that predicts future physical signals. Through extensive real-world experiments, we demonstrate that MoSS successfully augments VLAs to leverage diverse physical signals (*i.e.*, tactile, force, and torque), integrating multiple signals to achieve synergistic performance gains. The project page is available.

Keywords: Vision-Language-Action Models, Contact-rich manipulation

1 Introduction

In the real world, humans perform dexterous manipulation by continuously integrating multiple sources of physical sensory feedback beyond visual perception. For example, during plug insertion, tactile feedback at the fingertips helps recognize grasp stability, while torque feedback from the arm provides crucial cues for detecting contact and correcting misalignment. Toward learning human-like generalizable policies, Vision-Language-Action models (VLAs) [1, 2, 3, 4, 5, 6, 7] have recently shown remarkable progress by leveraging large-scale pretrained Vision-Language Models (VLMs; [8, 9, 10]). However, despite this progress, VLAs generate actions from visual observations alone, which fundamentally limits their ability to handle contact-rich manipulation.

To mitigate this limitation, recent approaches have explored incorporating a single physical sensory modality into pretrained VLAs, either by encoding it into VLM backbones [14, 11, 15] or by directly conditioning the action decoder [12, 13]. While effective in specific settings, these approaches are not designed to grow with increasing sensory complexity; naïvely extending them to incorporate additional physical modalities does not consistently yield complementary performance gains in practice (see Figure 1). This highlights the need for a more unified adaptation framework that can effectively integrate heterogeneous physical signals for reliable dexterous manipulation.

Our approach. We present **MoSS**, a **Mo**dular **S**ensory **S**tream framework that seamlessly augments pretrained VLAs to leverage diverse physical sensory signals for action prediction. Specifically, we append decoupled modality streams to the action expert module, which processes physical inputs and exchanges information with the pretrained parameters via a joint cross-modal self-attention mechanism. To avoid corruption of pretrained parameters by the newly added streams, we adopt a two-stage training scheme. We first freeze the pretrained VLA and train only the physical signal

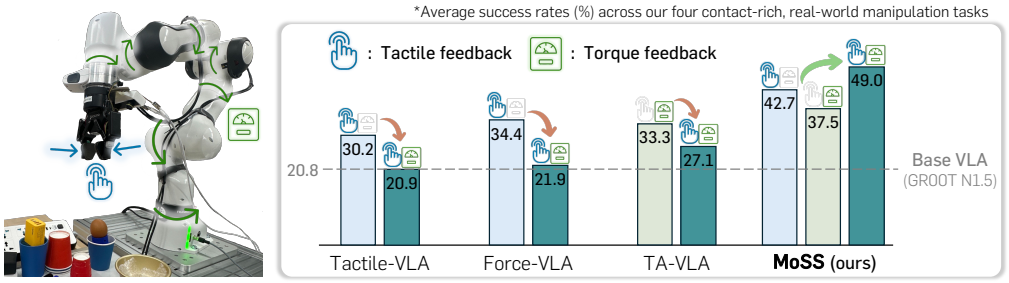


Figure 1: **Limitation of existing approaches to incorporating physical sensory signals into VLAs.** We compare the average success rates (%) of VLAs as more diverse physical sensory signals are incorporated. Existing approaches [11, 12, 13] struggle to handle multiple physical sensing modalities, leading to degraded performance when additional modalities are used together. In contrast, MoSS addresses this limitation with a unified adaptation framework that effectively integrates additional physical sensory signals into VLAs, enabling complementary use of heterogeneous physical feedback.

streams, allowing them to pre-align their representations with the existing representation space of the pretrained policy. We then jointly fine-tune the entire model end-to-end. Furthermore, we incorporate an auxiliary future physical signal prediction objective, which encourages the model to internalize physical interaction dynamics and effectively utilize physical feedback for action generation [13].

To validate the effectiveness of MoSS, we evaluate our method on contact-rich, real-world manipulation tasks that require physical interaction with surrounding environments for precise control. We show that MoSS consistently improves the performance of pretrained VLAs by effectively leveraging physical feedback from both end-effector sensing (*i.e.*, tactile) and joint-level sensing (*i.e.*, torque), compared to other baselines. Furthermore, we demonstrate that MoSS can jointly incorporate multiple physical signals within a single model, achieving cumulative performance gains by combining complementary signals. To the best of our knowledge, we are the first to present a unified VLA framework that simultaneously integrates multiple physical signals.

We summarize the main contributions of this paper below:

- We propose MoSS, a modular framework that augments pretrained VLAs to leverage physical sensory signals via decoupled modality streams and joint cross-modal attention mechanisms.
- We introduce a two-stage training scheme with auxiliary future physical signal prediction objective, which preserves pretrained VLA priors while ensuring effective use of newly added modalities.
- We design contact-rich, real-world manipulation tasks with diverse physical feedback (*i.e.*, tactile, force, and torque) across both gripper and dexterous hand embodiments, and show that MoSS achieves consistent and cumulative performance gains over the baselines.

2 Related Work

We discuss the most relevant literature here and provide further discussion in Appendix C.

Vision-Language-Action models (VLAs). The development of generalist robot policies has been fundamentally shaped by the challenge of acquiring diverse, large-scale robotic datasets. While traditional approaches [16, 17, 18] rely on task-specific demonstrations with limited generalization, VLAs address this bottleneck by leveraging pretrained Vision-Language Models (VLMs) for robotic control. Early VLAs [1, 2] generate discretized actions autoregressively as tokens, while more recent work [3, 6] has shifted toward hybrid architectures in which diffusion-based action heads are conditioned on VLM representations. Despite these advances, current VLAs predominantly rely on vision and language modalities, overlooking physical feedback (*e.g.*, tactile and torque) that is critical for contact-rich manipulation. In this work, we focus on diffusion-based VLAs and introduce an approach that integrates diverse physical sensory signals into their action generation process.

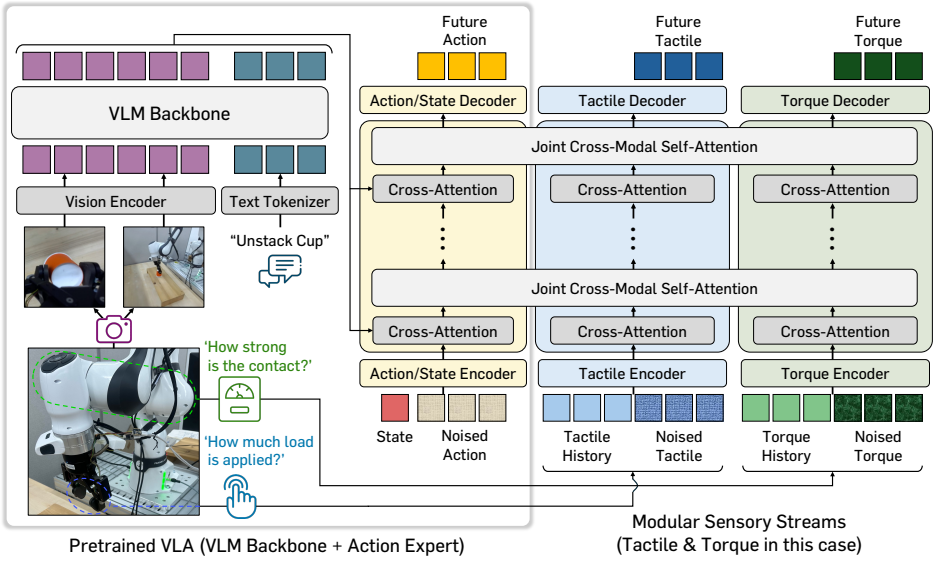


Figure 2: **Overview of Modular Sensory Stream framework (MoSS).** Building on a pretrained VLA, MoSS introduces a multimodal stream architecture that processes multiple physical sensory signals (*e.g.*, tactile and torque) in parallel. We extend the self-attention layers of the Action Expert into joint cross-modal self-attention layers, enabling each signal to be processed in a dedicated stream while allowing information exchange only through joint attention.

Vision-Language-Action models (VLAs) with physical feedback. Recent approaches have explored extending pretrained VLAs with physical feedback for dexterous manipulation using either *end-effector-level sensing*, *e.g.*, tactile or force feedback from the gripper [12, 11, 19, 15, 20, 21, 14], or *joint-level sensing*, *e.g.*, torque feedback from robot arm joints [13]. However, existing works are specialized for a single sensing modality or sensor configurations, while dexterous manipulation often benefits from complementary physical feedback. For instance, end-effector-level sensing provides fingertip-level contact perception and has been used for tasks requiring fine contact understanding, while joint-level sensing captures whole-arm interaction cues, such as object weight, which are useful for compliant and contact-aware control. In contrast, our framework provides a unified approach to incorporate multiple physical signals into pretrained VLAs, enabling generalist policies that can flexibly leverage available physical sensing modalities across tasks and hardware configurations.

3 Method

In this section, we present MoSS, an adaptation framework that augments pretrained Vision-Language-Action models (VLAs) with physical sensory modalities. Formally, let ℓ be a task instruction and $\mathbf{I}_t$ be the visual observations at timestep t. A diffusion-based VLA processes these inputs with its Vision-Language Model (VLM) backbone f_θ to obtain a unified visual-language representation $\mathbf{h}_t = f_\theta(\ell, \mathbf{I}_t)$. This feature is then passed to the action expert module V_ψ alongside the robot's proprioceptive state $\mathbf{s}_t$ to predict H future actions $\mathbf{A}_t = [\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}]$, namely action chunking [22, 17]. The model is trained with a conditional flow-matching objective:

$$\mathcal{L} = \mathbb{E}_{\tau,\epsilon}\left[\|(\mathbf{A}_t - \epsilon) - V_\psi(\mathbf{A}_t^\tau, \mathbf{s}_t \mid \mathbf{h}_t)\|^2\right], \tag{1}$$

where the noisy action chunk is given by $\mathbf{A}_t^\tau = \tau\mathbf{A}_t + (1-\tau)\epsilon$, with random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and flow-matching timestep $\tau \in [0, 1]$.

Here, our goal is to augment such VLAs to leverage a diverse set of physical feedback in addition to visual observations. To this end, we first introduce a multimodal stream architecture that processes newly added physical signals in parallel via decoupled modality streams (see Section 3.1). We then detail our two-stage training scheme to ensure stable incorporation of these modality streams (see Section 3.2), and an auxiliary future feedback prediction objective to exploit physical signals better (see Section 3.3). An overview of our framework is shown in Figure 2.

3.1 Multimodal Sensory Streams

Let $\mathbf{M} = \{\mathbf{m}_t^{(i)} \in \mathbb{R}^{d_i}\}_{i=1}^N$ be the set of N physical sensory signals (*e.g.*, tactile and torque) at timestep t. To incorporate them into a pretrained VLA, we extend its transformer-based action expert module V_ψ with modular sensory streams V_{ϕ_i}, each dedicated to processing the i-th physical signal $\mathbf{m}_t^{(i)}$. Specifically, all sensory streams take the visual-language representation $\mathbf{h}_t$, with cross-modal interaction restricted to the self-attention layers interleaved throughout the network.

$$\begin{aligned} &\text{Action stream}: \quad V_\psi(\mathbf{A}_t^\tau, \mathbf{s}_t \mid \{\mathbf{h}_t\} \cup \mathbf{M}) \\ &\text{Physical stream } i: \; V_{\phi_i}(\mathbf{m}_t^{(i)} \mid \{\mathbf{h}_t, \mathbf{A}_t^\tau, \mathbf{s}_t\} \cup \mathbf{M} \backslash \{\mathbf{m}_t^{(i)}\}) \end{aligned} \tag{2}$$

In practice, we construct each new sensory stream V_{ϕ_i} by mirroring the architecture of the original action expert module V_ψ and randomly initializing its parameters. We then replace the self-attention layers in each stream with joint cross-modal self-attention layers. Specifically, at each layer, independently computed queries, keys, and values $\{\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\}_{i=0}^N$ (with $i = 0$ denoting the action stream) are concatenated to calculate a shared scaled dot-product attention. This bidirectional design enables cross-modal reasoning (*e.g.*, modulating actions based on tactile feedback or unexpected torques), while keeping other network components decoupled to prevent gradient interference.

3.2 Two-Stage Training Strategy

We employ a two-stage training scheme designed to preserve pretrained knowledge while enabling effective learning of physical sensory representations.

Stage 1: Physical alignment. We freeze all parameters of the pretrained action expert module V_ψ and train only the randomly initialized sensory streams $\{V_{\phi_i}\}_{i=1}^N$. This allows the new streams to learn physical interaction features (*e.g.*, tactile contacts, force feedback, and joint torques) and cross-modal correspondences through joint self-attention, without disrupting the pretrained action policy.

Stage 2: Joint fine-tuning. After initializing the modular sensory streams $\{V_{\phi_i}\}_{i=1}^N$, we unfreeze the action expert module V_ψ and perform joint fine-tuning. This enables the action stream to incorporate physical feedback $\mathbf{M}$ while the sensory streams adapt to task-specific requirements, jointly learning coordinated behaviors for physically-aware manipulation.

Overall, this staged strategy is crucial for stable optimization, as directly performing end-to-end training leads to optimization difficulties due to gradient interference between the randomly initialized sensory streams $\{V_{\phi_i}\}_{i=1}^N$ and pretrained action stream V_ψ (see Section 4.3).

3.3 Training Objective

Our primary objective is the action flow-matching loss $\mathcal{L}_{\text{act}} = \mathbb{E}_{\tau,\epsilon}\big[\|(\mathbf{A}_t - \epsilon) - \hat{\mathbf{A}}_t\|^2\big]$, where $\hat{\mathbf{A}}_t = V_\psi(\mathbf{A}_t^\tau, \mathbf{s}_t \mid \{\mathbf{h}_t\} \cup \mathbf{M})$. To encourage effective use of physical sensory signals, each modality stream predicts the future signal over the action horizon from past signals $\mathbf{m}_{t-H+1:t}^{(i)}$:

$$\mathcal{L}_{\text{phy}} = \sum_{i=1}^{N} \mathbb{E}_{\tau,\epsilon_i}\big[\|(\mathbf{m}_{t+1:t+H}^{(i)} - \epsilon_i) - \hat{\mathbf{m}}_{t+1:t+H}^{(i)}\|^2\big]. \tag{3}$$

The overall objective is $\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{phy}}\mathcal{L}_{\text{phy}}$, where λ_{phy} balances the two terms. As described in Section 3.2, Stage 1 optimizes only $\mathcal{L}_{\text{phy}}$ with the action expert frozen, while Stage 2 jointly optimizes the full objective with respect to all parameters.

4 Experiments

We investigate the following questions through experiments:

- Does applying MoSS to existing Vision-Language-Action models (VLAs) effectively augment the base policy with diverse physical feedback for contact-rich manipulation, outperforming prior modality-augmented VLA baselines? (Table 1 and Figure 3)

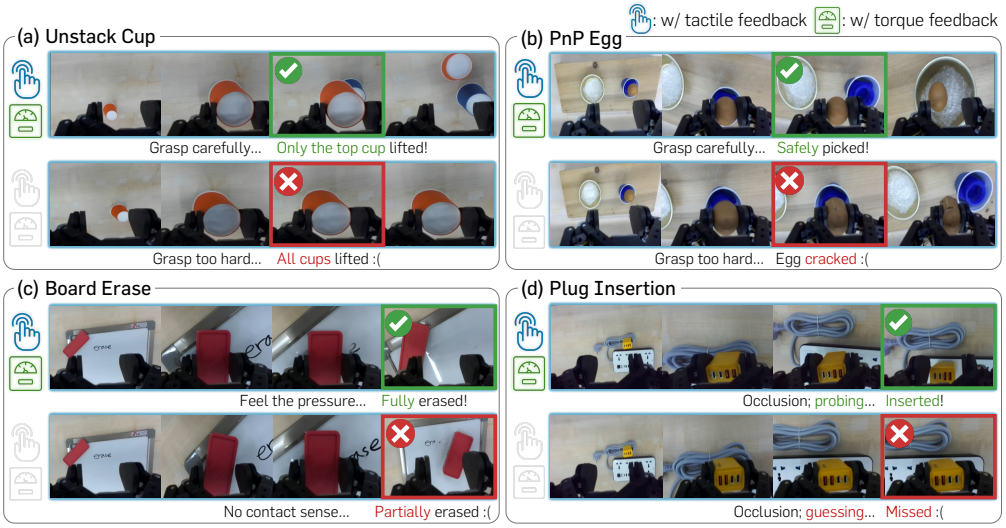


Figure 3: **Example rollouts of real-world tasks.** We provide example rollouts of the designed tasks that critically depend on physical feedback. While MoSS leverages physical feedback to successfully perform the tasks, GR00T N1.5 without physical feedback often has difficulties in (a), (b) regulating grasp force, (c) maintaining appropriate pushing force for contact, and (d) probing occluded geometry.

- Can MoSS effectively integrate multiple physical sensing signals to exploit their complementary effects? (Table 1)
- How do individual components affect performance? (Tables 2 and 3 and Figure 4)
- Does MoSS transfer to different embodiments (*e.g.*, hand) and physical feedback (*e.g.*, force signal) with distinct sensor configurations? (Table 4)

4.1 Experimental Setups

Tasks. We evaluate MoSS on real-world, contact-rich manipulation tasks where interaction with physical feedback (*e.g.*, tactile on the end-effector and torque on arm joints) is crucial for successful execution. Specifically, we design the following tasks (see Figure 3 for the visualization of the tasks, and see Appendix A.1 for more detailed descriptions):

- **Unstack Cup.** A robot should remove only the top red cup from a stack of 3 inverted cups. This task requires regulating grasping force, as gripping too strongly lifts multiple cups together.
- **PnP Egg.** A robot must pick up a fragile egg and place it in a bowl without breaking it. This task requires fine-grained feedback at the end-effector to avoid excessive force.
- **Board Erase.** A robot must erase a marked region on a whiteboard using an eraser. This task requires sensing contact to maintain consistent pressure on the board to ensure effective erasing.
- **Plug Insertion.** A robot must insert a plug into a socket, often occluded from the (wrist-mounted) camera view. This task requires torque and force cues to detect contact and guide its alignment.

Hardware setup. We construct a Franka Research 3 platform, a 7-DoF single arm robot equipped with a Robotiq 2F-85 gripper, following the DROID [24] setup. Here, we mount an AnySkin [25] tactile sensor on the gripper, while the robot arm itself supports torque sensing at each joint. We provide detailed specifications of the physical sensory signals in Appendix A.2.

Baselines. To demonstrate the generality of our approach, we apply MoSS to two state-of-the-art VLAs: π_0 [23] and GR00T N1.5 [6]. We compare MoSS with several modality-augmented VLA baselines: (i) Tactile-VLA [11] incorporates *tactile* signals on the gripper as additional input tokens to the VLM backbone; (ii) ForceVLA [12] employs an MoE-based encoder to fuse 3-axis *force* signals on the gripper with VLM features and conditions the action expert on them; (iii) TA-VLA [13] injects *torque* signals from robotic arm joints as additional tokens into a single action expert transformer

Table 1: **Results on contact-rich, real-world manipulation tasks.** We report the success rates (%) for each task using VLAs fine-tuned on a consolidated dataset that combines the training data from all tasks. For the averaged success rates (Avg.), we also report a standard deviation computed over all trials. **Bold** and underline indicate best and runner-up results, respectively.

Method	Tactile	Torque	Unstack Cup	PnP Egg	Board Erase	Plug Insertion	**Avg.**
GR00T N1.5 [6]	✗	✗	16.7	45.8	20.8	0.0	20.8 $\pm$ 4.1
+ Tactile-VLA [11]	✓	✗	29.2	45.8	33.3	12.5	30.2 $\pm$ 4.7
+ ForceVLA [12]	✓	✗	37.5	54.2	33.3	12.5	34.4 $\pm$ 4.8
+ MoSS (ours)	✓	✗	45.8	**66.7**	41.7	16.7	42.7 $\pm$ 5.0
+ TA-VLA [13]	✗	✓	29.2	45.8	37.5	20.8	33.3 $\pm$ 4.7
+ MoSS (ours)	✗	✓	33.3	50.0	41.7	**25.0**	37.5 $\pm$ 4.9
+ MoSS (ours)	✓	✓	**54.2**	**66.7**	**50.0**	**25.0**	**49.0** $\pm$ 5.1
π_0 [23]	✗	✗	12.5	50.0	29.2	16.7	27.1 $\pm$ 4.5
+ Tactile-VLA [11]	✓	✗	16.7	50.0	37.5	12.5	29.2 $\pm$ 4.6
+ ForceVLA [12]	✓	✗	25.0	62.5	50.0	16.7	38.6 $\pm$ 5.0
+ MoSS (ours)	✓	✗	**29.2**	62.5	41.7	20.8	38.6 $\pm$ 5.0
+ TA-VLA [13]	✗	✓	16.7	58.3	41.7	20.8	34.4 $\pm$ 4.8
+ MoSS (ours)	✗	✓	20.8	62.5	54.2	**29.2**	41.7 $\pm$ 5.0
+ MoSS (ours)	✓	✓	**29.2**	**66.7**	**58.3**	**29.2**	**45.9** $\pm$ 5.1

shared with action tokens, while training the policy to predict future torque signals as an auxiliary task. We adapt each method to use the physical sensory signals available on our hardware platforms.

Implementation details. We train all models by fine-tuning GR00T N1.5 and π_0 on a consolidated dataset of training data from all tasks, following their official implementation: full fine-tuning for π_0 and training only the action expert for GR00T N1.5. For GR00T N1.5 + Tactile-VLA, we additionally unfreeze the VLM backbone, since Tactile-VLA injects tactile signals into the VLM and the backbone must adapt to this unseen modality. We provide further implementation details in Appendix A.3.

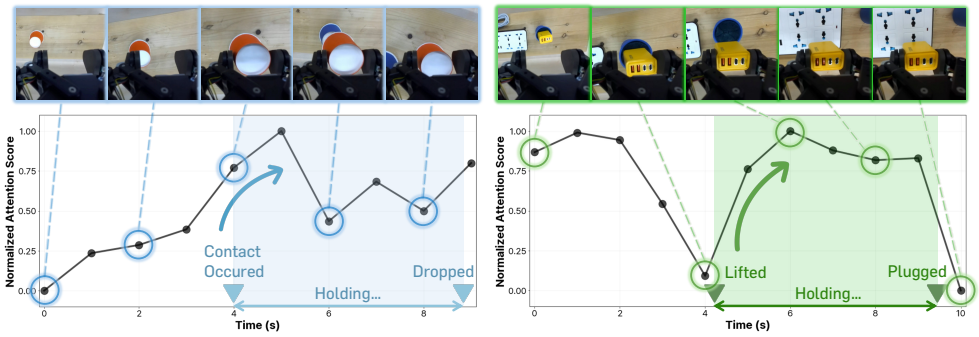
4.2 Main Results

We provide quantitative results in Table 1 and qualitative results in Figure 3. We also provide more results, including evaluation on simulated manipulation benchmarks in Appendix D.

In Table 1, we first observe that base VLAs (*i.e.*, fine-tuned VLAs without physical sensing) achieve low average success rates of 20.8% and 27.1% for GR00T N1.5 and π_0, respectively, confirming that vision and language alone are insufficient for tasks requiring fine-grained force control. In contrast, we find that MoSS achieves substantial improvements across all tasks and base models. For example, for GR00T N1.5, we achieve an average improvement of 28.2% over the base model, with particularly strong gains on Unstack Cup (+37.5%) and Board Erase (+29.2%). Similar improvements are observed on π_0, demonstrating that our approach generalizes across different VLA architectures. We also find that MoSS consistently outperforms several physical modality-augmented VLA baselines. For example, on the PnP Egg task, GR00T N1.5 + MoSS with tactile signals achieves 66.7%, but GR00T N1.5 + ForceVLA gets 54.2%. This highlights our design that maintains a separate processing stream with controlled cross-modal joint self-attention, which enables effective integration of physical sensing without corrupting action representation. Moreover, MoSS shows consistent improvements when additional physical modalities are incorporated, demonstrating that MoSS is a unified framework for adapting VLAs to multiple physical modalities.

4.3 Analysis and Ablation Study

When does the VLA attend to physical feedback? To investigate when the VLA attends to physical feedback, we visualize the attention patterns in our cross-modal attention mechanism while performing the task. In Figure 4, we observe that action tokens attend strongly to torque and tactile tokens when the generated actions require accurate, contact-intensive control (*e.g.*, when grasping



(a) Action-to-Tactile Attention over Time (b) Action-to-Torque Attention over Time

Figure 4: **Attention scores between action and physical modality stream.** We visualize the standardized attention scores between action and tactile/torque sensory streams in the joint cross-modal self-attention layers while conducting tasks. We find that action tokens attend strongly to tactile and torque tokens at moments that demand accurate, contact-intensive control, *e.g.*, (a) when tactile contact is detected, and (b) when the robot identifies the insertion point to plug in.

Table 2: **Ablation studies.** We report the success rates (%) of various variants of MoSS on contact-rich tasks using GR00T N1.5 as the pretrained VLA. Our default setting is highlighted in gray.

(a) Stream design.

Stream	Unstack Cup	PnP Egg
Single stream	33.3	50.0
Decoupled	**54.2**	**66.7**

(b) Training scheme.

Training	Unstack Cup	PnP Egg
One-stage	37.5	58.3
Two-stage	**54.2**	**66.7**

(c) Future prediction.

Future prediction	Unstack Cup	PnP Egg
w/o prediction	45.8	58.3
w/ prediction	**54.2**	**66.7**

softly), enabling the model to modulate insertion or grasping force based on resistance feedback. Notably, this demonstrates that MoSS truly leverages physical sensory signals when generating actions at moments that require accurate, contact-intensive control.

Effect of decoupled multimodal stream architecture. We compare MoSS against a naïve Diffusion Transformer (DiT) action expert that jointly predicts physical sensory signals within a single stream. In Table 2a, we observe that our decoupled multimodal architecture significantly outperforms the DiT action expert. For example, on the Unstack Cup task, it achieves 20.9% improvements over the naïve DiT baseline. These results indicate that decoupling distinct modalities to separate streams is crucial for effectively adapting VLAs to diverse physical sensory signals.

Effect of two-stage training. We compare MoSS trained with either single- or two-stage scheme, where for the single-stage training, we skip physical alignment stage and jointly fine-tunes the action and physical streams. In Table 2b, we find that a two-stage training scheme is critical for fine-tuning VLAs to adapt to new physical feedback modalities, *e.g.*, on the Unstack Cup task, it yields a 16.7% improvement over single-stage training.

Effect of future physical modality prediction. We quantify the effect of the future physical modality prediction objective in Table 2c. We compare MoSS trained with and without the future prediction objective, and find this objective provides meaningful guidance for generating more accurate actions. For instance, on the Unstack Cup task, the prediction objective improves performance by 8.4% (45.8 $\rightarrow$ 54.2%). We further visualize MoSS's prediction dynamics and analyze how anticipated physical signals guide action generation in Appendix B.2.

Inference efficiency. MoSS introduces additional processing for multiple physical sensory signals. Then, how does this extension affect inference latency? In Table 3, we analyze the inference latency of GR00T N1.5 + MoSS when incorporating multiple signals. We find that when both tactile and torque modalities are incorporated, latency increases by only 2.4 ms (1.11$\times$ slower than the base model), highlighting the efficiency of MoSS. Importantly, MoSS offers significant performance gain in real-world tasks while incurring only a marginal increase in inference latency.

Table 3: **Efficiency analysis.** We report the inference latency (ms) of GR00T N1.5 + MoSS under different physical input configurations measured per action chunk, with relative overhead shown in parentheses.

Method	Tactile	Torque	Latency (ms, ↓)
GR00T N1.5	✗	✗	21.0 (1.00×)
+ MoSS (Ours)	✓	✗	22.4 (1.06×)
	✗	✓	21.9 (1.04×)
	✓	✓	23.4 (1.11×)

Table 4: **Applicability to a dexterous hand embodiment.** We evaluate MoSS on an OpenArm + Inspire RH56F1 Hands with fingertip force (6-dim) sensors, and report success rates (%) over 24 trials per task.

Method	Pick Paper	Sort by Weight	**Avg.**
GR00T N1.5	62.5	54.2	58.4
+ MoSS (Ours)	**83.3**	**95.8**	**89.6**

4.4 Applicability to Different Embodiment

To demonstrate that MoSS is not tied to a specific embodiment or physical feedback, we further evaluate GR00T N1.5 and GR00T N1.5 + MoSS on an OpenArm robot equipped with a 6-DoF Inspire RH56F1 dexterous hand. We use a 6-dimensional fingertip force signal provided by the hand. We design two contact-rich manipulation tasks on this setup (see Figure 6 for the visualization of the tasks, Appendix A.1 for more detailed descriptions, and implementation details in Appendix A.3):

- **Pick Paper.** A robot must pick up a thin sheet of paper (or napkin) from a flat surface. This task requires fine-grained force feedback at the fingertips to grasp the paper without slipping.
- **Sort by Weight.** A robot must lift a bag and place it right if heavy or left if light. This task requires sensing fingertip force during lifting to infer the bag's weight and decide where to place it.

In Table 4, we find that MoSS consistently outperforms GR00T N1.5 across both tasks. In particular, we find a notable gain on Sort by Weight, where successful execution requires actively reasoning about the perceived force control. This indicates that MoSS effectively leverages physical feedback even on a substantially different embodiment.

5 Conclusion

In this work, we have presented MoSS, a unified framework that adapts Vision-Language-Action models (VLAs) to leverage multiple sensory signals for action prediction. By extending the action experts with new streams, MoSS can incorporate various modalities without corrupting the pretrained features. Our experiments show that MoSS improves performance on contact-rich, real-world manipulation tasks, and the performance continues to improve as more physical sensory modalities are incorporated. We hope our approach facilitates future research toward generalist robot policies that can leverage a diverse range of physical sensory modalities.

6 Limitations

Practical considerations for sensor deployment. While our experiments confirm the effectiveness of incorporating tactile, torque, and force signals into MoSS, physical sensors in general are known to exhibit wear or noise over extended use, which may need to be considered when deploying our framework in long-term, real-world scenarios. Pairing MoSS with higher-fidelity, more durable sensors and periodic recalibration would help ensure consistent performance in such settings.

Dependence on the base policy for non-contact phases. MoSS complements a pretrained VLA once meaningful physical interaction occurs, rather than enhancing VLAs' pre-contact visual perception and action generation. Accordingly, if the base policy fails to reach the object or form a stable initial grasp, MoSS provides limited corrective information. Addressing these failures will likely require complementary advances in the base policy's visual perception and motion planning. We provide a more detailed discussion of failure cases in Appendix B.3.

References

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024.

[3] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[4] C.-P. Huang, Y.-H. Wu, M.-H. Chen, Y.-C. F. Wang, and F.-E. Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025.

[5] X. Chen, H. Wei, P. Zhang, C. Zhang, K. Wang, Y. Guo, R. Yang, Y. Wang, X. Xiao, L. Zhao, J. Chen, and J. Bian. villa-x: Enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv: 2507.23682*, 2025.

[6] J. Bjorck, V. Blukis, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1.5: An improved open foundation model for generalist humanoid robots. https://research.nvidia.com/labs/gear/gr00t-n1_5/, June 2025. Accessed: 2025-09-09.

[7] J. Zheng, J. Li, Z. Wang, D. Liu, X. Kang, Y. Feng, Y. Zheng, J. Zou, Y. Chen, J. Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

[8] 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.

[9] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[10] G. Chen, Z. Li, S. Wang, J. Jiang, Y. Liu, L. Lu, D.-A. Huang, W. Byeon, M. Le, T. Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.

[11] 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. *Advances in Neural Information Processing Systems*, 2025.

[12] 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. *Advances in Neural Information Processing Systems*, 2025.

[13] 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. In *Conference on Robot Learning*, 2025.

[14] 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.

[15] 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.

[16] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in Neural Information Processing Systems*, 35:22955–22968, 2022.

[17] 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*, 2023.

[18] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto. Behavior generation with latent actions. In *International Conference on Machine Learning*, 2024.

[19] 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.

[20] A. Adeniji, Z. Chen, V. Liu, V. Pattabiraman, R. Bhirangi, S. Haldar, P. Abbeel, and L. Pinto. Feel the force: Contact-driven learning from humans. *arXiv preprint arXiv:2506.01944*, 2025.

[21] P. Lin, Y. Huang, W. Li, J. Ma, C. Xiao, and Z. Jiao. Pp-tac: Paper picking using tactile feedback in dexterous robotic hands. *arXiv preprint arXiv:2504.16649*, 2025.

[22] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

[23] 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. In *Robotics: Science and Systems*, 2025.

[24] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[25] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. In *IEEE International Conference on Robotics and Automation*, pages 16563–16570. IEEE, 2025.

[26] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[27] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.

[28] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[29] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[30] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024.

[31] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[32] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[33] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W.-t. Yih, L. Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[34] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[35] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025.

[36] E. Helmut, N. Funk, T. Schneider, C. de Farias, and J. Peters. Tactile-conditioned diffusion policy for force-aware robotic manipulation. *arXiv preprint arXiv:2510.13324*, 2025.

[37] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek. Multimodal and force-matched imitation learning with a see-through visuotactile sensor. *IEEE Transactions on Robotics*, 41:946–959, 2024.

[38] V. Pattabiraman, Y. Cao, S. Haldar, L. Pinto, and R. Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins. *arXiv preprint arXiv:2410.17246*, 2024.

[39] 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.

[40] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1105–1112. IEEE, 2025.

[41] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

A Experimental Details

A.1 Task Specifications

We collect 100 demonstrations per task for Unstack Cup, PnP Egg, Board Erase, Plug Insertion, and Pick Paper, and 60 demonstrations for Sort by Weight, with extensive variation across multiple dimensions to ensure robust generalization and avoid memorization. For the single-arm gripper setup, we use a third-person view camera and a wrist-mounted camera, both at 720×1280 resolution without depth. For the bimanual dexterous-hand setup, we use a stereo egocentric camera at the same resolution without depth. All visual observations are resized to 224×224 before being fed into the model. We use absolute joint position as the action mode. We describe the detailed scenarios as follows:

Unstack Cup: The task required lifting only the top red cup from a stack. To prevent gripper position memorization, we varied the cup sizes, using both large and small cups in different configurations. In particular, we evaluate each model over 24 trials, consisting of 12 trials with a large cup and 12 trials with a small cup. We use the instruction "*Pick up the top red cup from the stack and place it next to the blue cup.*" for this task.

PnP Egg: The egg was placed in various positions across the workspace to evaluate spatial generalization. We evaluate each model over 24 trials. We use the instruction "*Pick up the egg and place it in the gold bowl.*" for this task.

Board Erase: The whiteboard was positioned at three different heights to test height adaptability. We varied the eraser colors (red and grey) with slight differences in object size, and used markers of different colors (blue and black). Evaluations were conducted across diverse environmental conditions to assess robustness. In particular, we evaluate each model over 24 trials, consisting of one trial per combination of (height, eraser color, object size, different color). We use the instruction "*Use the (red/grey) eraser to clean the whiteboard.*" for this task.

Plug Insertion: We used chargers of three different colors (yellow, black, and white) for color and object generalization, with target insertion positions varied across trials. We evaluate each model over 24 trials, consisting of 8 trials per charger. We use the instruction "*Pick up the (yellow/white/black) charger and plug it into the socket.*" for this task.

Pick Paper: The task is performed on a height-adjustable table, allowing us to test height adaptability by varying the table height across trials. Specifically, we evaluate with three different heights: one that is in-domain with respect to the training data, and two that are out-of-domain, both lower than any height seen during training. Since the policy cannot rely on proprioception alone to localize the surface at unseen heights, this setup requires the robot to use fingertip force feedback to detect contact with the paper (or napkin). Evaluations are further conducted across diverse paper positions on the table. We evaluate each model over 24 trials. We use the instruction "*Pick up the paper.*", "*Pick up the napkin.*" for this task.

Sort by Weight: A single camping bag is used across all trials, and its weight is varied by placing 0, 1, or 2 juice pouches (each approximately 200g) inside, yielding three weight conditions. The robot must lift the bag, infer its weight from fingertip force feedback during lifting, and place it on the right if heavy or on the left if light. We evaluate each model over 24 trials across the three weight conditions. We use the instruction "*Place the bag on the right if heavy, and on the left if light.*" for this task.

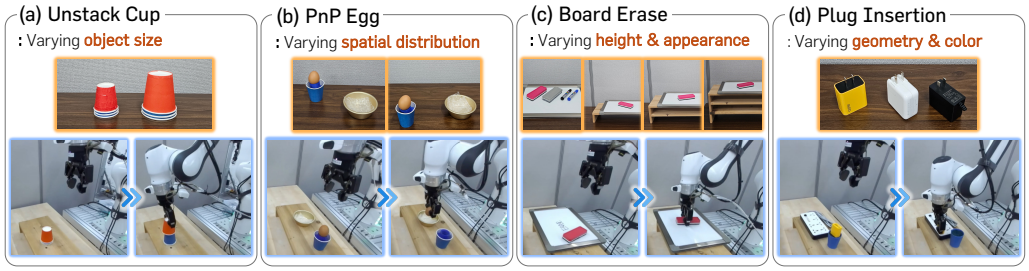


Figure 5: **Examples of the evaluation tasks.** We design contact-rich, real-world robotic manipulation tasks with systematically controlled variations in object size, spatial layout, geometry, and appearance, making visual observations alone ambiguous and physical feedback necessary for success.

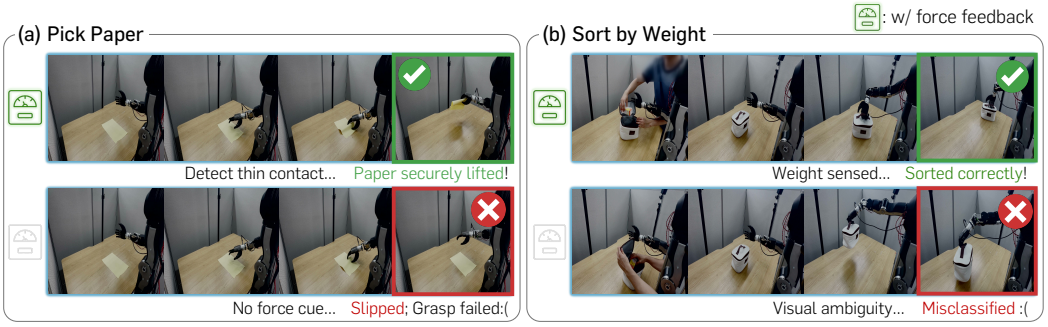
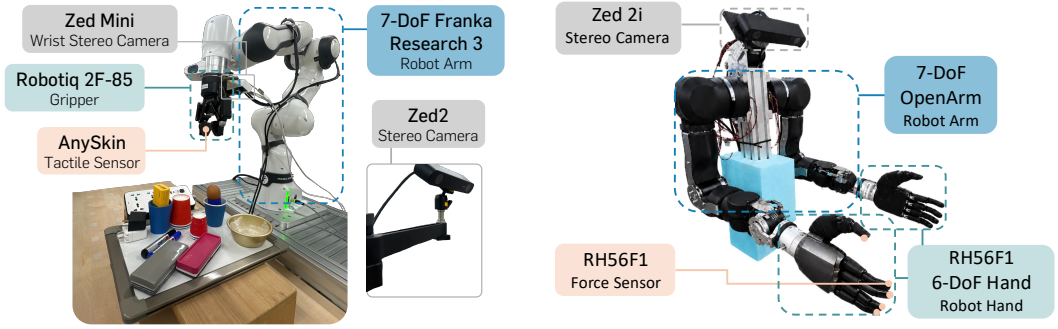


Figure 6: **Example rollouts of real-world dexterous hand tasks.** We provide example rollouts on hand tasks that critically depend on physical feedback, comparing MoSS (top, with force feedback) against GR00T N1.5 (bottom, vision only). In (a) *Pick Paper*, MoSS detects slip and regulates grasp force to lift the flat paper, while the baseline fails to maintain a stable grasp. In (b) *Sort by Weight*, MoSS infers the hidden juice box from contact force and sorts the bag to the correct side, whereas the baseline cannot disambiguate weight and sorts incorrectly.

A.2 Hardware Setup: Sensor Specifications

We visualize the hardware setup of both platforms, including sensor specifications, in Figure 7. For the main experiments, we construct a Franka Research 3 platform, a 7-DoF single arm robot equipped with a Robotiq 2F-85 gripper. Here, we mount an AnySkin [25] tactile sensor on the gripper, while the robot arm itself supports torque sensing at each joint. Specifically, for tactile signals, each AnySkin sensor provides a 15-dimensional feature vector, obtained from five sensing units that each measure a 3-dimensional force vector (*i.e.*, x, y, z). Since we attach one AnySkin sensor to each gripper finger, our tactile sensing system provides a 30-dimensional observation vector at each timestep. For the torque signals, the robot provides a 7-dimensional vector (one scalar torque value per joint) measured in Newton-meters. The dual-sensor configuration enables the system to simultaneously monitor forces applied by both gripper fingers, providing rich information about grasp stability and contact interactions with manipulated objects. For the applicability to different embodiments, we consider the OpenArm robot equipped with an Inspire RH56F1 Hands, a 6-DoF dexterous hand with fingertip force sensors. The force sensor provides a 6-dimensional force signal (one per finger, with two on the thumb).



(a) **Single-arm gripper platform.** A 7-DoF Franka Research 3 robot arm equipped with a Robotiq 2F-85 gripper and AnySkin tactile sensors mounted on the gripper fingers, together with a wrist-mounted Zed Mini stereo camera for visual observation.

(b) **Bimanual dexterous-hand platform.** Dual 7-DoF OpenArm robot arms, each equipped with a 6-DoF RH56F1 dexterous hand and an RH56F1 force sensor at each fingertips, together with a head-mounted Zed 2i stereo camera for visual observation.

Figure 7: **Overview of our hardware platforms.** We use two complementary robot setups that provide tactile, force, and torque sensing for contact-rich manipulation.

A.3 Implementation Details

We report the performance of MoSS by fine-tuning two pretrained Vision-Language-Action models, GR00T N1.5 and π_0, following their official implementations. For GR00T N1.5, we freeze the VLM backbone and only train the action expert; for π_0, we fine-tune all model parameters. The only exception is GR00T N1.5 + Tactile-VLA, where we additionally fine-tune the VLM backbone, since Tactile-VLA injects tactile signals directly into the VLM and the backbone must adapt to this previously unseen modality.

All models are trained jointly on a consolidated dataset combining the training data from all tasks, with a batch size of 32. For GR00T N1.5, we train for 60k iterations, and for π_0, we train for 30k iterations. We set the coefficient of the physical prediction loss to $\lambda_{\text{phy}} = 0.1$. The remaining training hyperparameters are summarized in Table 5.

Table 5: Hyperparameter details for training MoSS.

	GR00T N1.5 + MoSS	π_0 + MoSS
Optimizer	AdamW	AdamW
Optimizer momentum	$\beta_1, \beta_2 = 0.95, 0.999$	$\beta_1, \beta_2 = 0.9, 0.95$
Optimizer weight decay	1e-5	1e-10
Learning rate	1e-4	2.5e-5
Learning rate scheduler	Cosine decay	Cosine decay
Warmup iterations	3,000	1,000
Batch size	32	32
Training iterations (stage 1)	20,000	10,000
Training iterations (stage 2)	40,000	20,000

B Additional Analysis

B.1 Hyperparameter Sensitivity Analysis

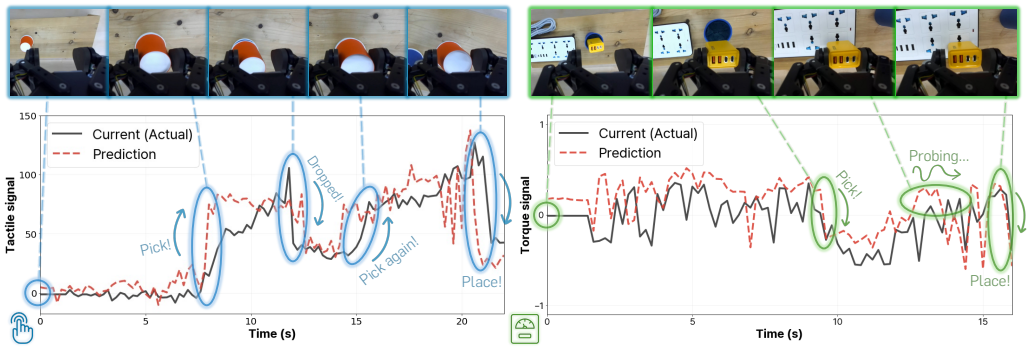
In Table 6, we analyze the hyperparameter sensitivity of training loss coefficients for the modular sensory streams λ_{phy} on Unstack Cup and PnP Egg tasks. We observe that performance is relatively robust across the range $\lambda_{\text{phy}} \in \{0.1, 0.5\}$, with peak performance at $\lambda_{\text{phy}} = 0.1$. We also find that performance degrades at $\lambda_{\text{phy}} = 1.0$, where the future prediction loss increasingly dominates the action learning signals. Our default setting is highlighted in gray.

Table 6: Results of GR00T N1.5 + MoSS for different λ_{phy} values.

λ_{phy}	Unstack Cup	PnP Egg
0.1	**54.2**	**66.7**
0.5	50.0	58.3
1.0	29.2	41.7

B.2 Prediction Dynamics Analysis

We visualize MoSS's prediction dynamics in Figure 8. We find that MoSS anticipates contact by predicting rising tactile and torque signals, prompting increased gripping force. This demonstrates how the future prediction objective provides grounded physical guidance that leads to the performance gains reported in Table 2c.



(a) Future **tactile** feedback prediction (b) Future **torque** feedback prediction

Figure 8: **Physical feedback and predicted values.** We visualize the physical sensory signals (*i.e.*, tactile and torque) predicted by MoSS while conducting tasks. We find that MoSS accurately predicts the changes in physical signals, particularly at moments when such signals are critical for successfully performing the task.

B.3 Failure Case Analysis

While MoSS demonstrates meaningful improvement on contact-rich tasks, we find several failure cases, which are visualized in Figure 9. First, in tasks such as Unstack Cup, PnP Egg, and Board Erase, we observe that when the base policy fails to reach the target object or establish a stable grasp, MoSS cannot compensate. This is because the physical feedback never becomes informative at these stages, since the gripper has not yet made meaningful contact. Moreover, in the Plug Insertion task, we observe that when the initial grasp is imperfect, it results in a misaligned plug angle that prevents successful socket insertion. As with the first case, this is also related to the grasping quality issue that occurs before physical feedback becomes effective.

We observe similar failures in the dexterous hand tasks. In the Pick Paper task, especially at lower table heights, the hand occasionally fails to apply sufficient normal force against the paper before picking up. This causes the paper to slip from the fingertips. In the Sort by Weight task, the policy must first insert the hand into the bag handle and lift the bag to obtain informative fingertip force feedback for weight estimation. However, when the hand fails to grasp the handle properly, it cannot lift the bag and thus never receives the physical feedback required for the downstream decision. These examples highlight that MoSS is most effective after meaningful contact is established, but remains dependent on the base policy for pre-contact grasping and alignment.



Figure 9: **Failure case examples.**

C More Discussion on Related Work

Multimodal architectures. Many recent efforts have attempted to design an architecture to incorporate a diverse range of modalities. In particular, the success of VLMs has been largely driven by advances in multimodal fusion architectures: early approaches [26, 27] employed cross-attention mechanisms to align visual and textual representations, while more recent work has explored richer fusion strategies, including perceiver-based architectures [28], modality-specific adapters [29, 30], and unified tokenization schemes [31, 32, 33]. These methods demonstrate that carefully designed fusion mechanisms can effectively model heterogeneous modalities with different dimensionalities and semantic structures. In the robotics domain, recent VLAs [23, 34] have successfully adapted such vision-language fusion techniques to incorporate action prediction. However, extending these architectures to physical sensing modalities poses additional challenges, as such signals differ fundamentally from visual and linguistic inputs in temporal structure, scale, and semantics. Our work draws inspiration from multimodal fusion in VLMs, but adapts it specifically to the action prediction setting, introducing a modular architecture that enables pretrained VLAs to integrate physical sensory signals while preserving their action priors.

Behavior cloning (BC) policy models with physical feedback. Prior work augments behavior cloning and diffusion-based policies [17] with physical feedback by either expanding the observation space with tactile or force inputs [35] or adding modality-specific encoders to the policy architecture [36], training the resulting multimodal policy from scratch. These approaches typically rely on a single sensing modality at a time, such as a vision-based tactile sensor (*e.g.*, GelSight, DIGIT), or a skin-based tactile array [37, 38, 39, 35], or a wrist- or fingertip-mounted force/torque sensor for contact-rich tasks such as peeling and insertion [40]. While these works demonstrate the value of physical feedback for contact-rich manipulation, they are developed as task-specific policies tied to a fixed sensor configuration, without leveraging the broad semantic and behavioral priors of pretrained VLAs. In contrast, our framework builds on pretrained VLAs and unifies heterogeneous physical feedback signals within a single architecture, enabling generalist policies that flexibly leverage available sensing modalities across various tasks and hardware configurations.

D Additional Quantitative Results

D.1 Results on Simulated Robotic Manipulation Benchmarks

In Table 7, we report the performance of VLAs on the RLBench benchmark [41], across 9 tasks, using the force signals natively supported by the simulator. Among these, 4 tasks involve contact-sensitive interactions: open window (objects occluded during manipulation), move hanger (requiring delicate contact), and lamp on / push button (requiring precise pressing). We observe that MoSS significantly improves performance on the 4 contact-sensitive tasks, by improving GR00T N1.5 by 10.9%. We also observe that MoSS maintains comparable performance on non-contact sensitive tasks. For instance, GR00T N1.5 achieves 46.8% on the non-contact-sensitive tasks, while GR00T N1.5 + MoSS achieves 47.1%. This demonstrates that MoSS effectively leverages physical signals for contact-sensitive tasks without degrading general manipulation capabilities.

Table 7: **Results on RLBench benchmark.** We report the success rates (%) for each task using VLAs finetuned on a consolidated dataset that combines the training data from all tasks. We use the force signals natively supported by the simulator. Avg. denotes averaged success rates over entire tasks. **Bold** indicates best results.

	Contact-sensitive Tasks					Non-contact-sensitive Tasks						
Method	Open Window	Lamp On	Move Hanger	Push Button	**Avg.**	Basketball in Loop	Phone on Base	Pick Up Cup	Slide Block to Target	Take Lid Off Saucepan	**Avg.**	**Overall Avg.**
GR00T N1.5 [6]	8.0	36.0	29.3	36.0	27.3	**72.0**	20.7	50.0	18.7	**72.7**	46.8	38.1
+ MoSS (Ours)	**9.3**	**47.3**	**45.3**	**50.7**	**38.2**	70.0	**22.7**	**52.7**	**21.3**	68.7	**47.1**	**43.1**

D.2 Scalability of VLAs to Multiple Physical Sensory Signals

In Table 8, we report the performance of VLAs fine-tuned to incorporate multiple physical sensory signals. We find that, while existing approaches often degrade when additional modalities are introduced, MoSS consistently improves as each new modality is added.

Table 8: **Results on using both tactile and torque inputs.** We report the success rates (%) for each single-arm, gripper-based task when *both* tactile and torque signals are provided as inputs to all methods. Values in parentheses indicate the absolute performance change compared to the corresponding single-modality setting in Table 1. **Bold** and underline indicate best and runner-up results, respectively.

Method	Tactile	Torque	Unstack Cup	PnP Egg	Board Erase	Plug Insertion	Avg.
GR00T N1.5 [6]	✗	✗	16.7	45.8	20.8	0.0	20.8
+ Tactile-VLA [11]	✓	✗	29.2	45.8	33.3	12.5	30.2
	✓	✓	12.5 (−16.7)	37.5 (−8.3)	16.7 (−16.6)	16.7 (+4.2)	20.9 (−9.3)
+ ForceVLA [12]	✓	✗	37.5	54.2	33.3	12.5	34.4
	✓	✓	25.0 (−12.5)	33.3 (−20.9)	20.8 (−12.5)	4.2 (−8.3)	21.9 (−12.5)
+ TA-VLA [13]	✗	✓	29.2	45.8	37.5	20.8	33.3
	✓	✓	20.8 (−8.4)	41.7 (−4.1)	29.2 (−8.3)	16.7 (−4.1)	27.1 (−6.2)
+ MoSS (ours)	✓	✗	<u>45.8</u>	<u>66.7</u>	<u>41.7</u>	16.7	<u>42.7</u>
	✗	✓	33.3	50.0	<u>41.7</u>	<u>25.0</u>	37.5
	✓	✓	**54.2** (+8.4)	**66.7** (−0.0)	**50.0** (+8.3)	**25.0** (−0.0)	**49.0** (+6.3)

D.3 Full results

In Table 9, we report the full performance of all VLAs, including the results of each fine-tuned model on several variations.

Table 9: **Full results on contact-rich, real-world manipulation tasks.** We report the success rates (%) for each task using VLAs fine-tuned on a consolidated dataset that combines the training data from all tasks. For the tactile modality, we mount an AnySkin [25] tactile sensor on the gripper, and for the torque modality, we use the joint torque measurements provided by the robot. Avg. denotes averaged success rates over entire tasks. **Bold** indicates best results.

			Unstack Cup			**Board Erase**			**Plug Insertion**			
Method	Tactile	Torque	Small	Big	**PnP Egg**	Low	Middle	High	Yellow	White	Black	**Avg.**
GR00T N1.5 [6]	✗	✗	0.0	33.3	45.8	0.0	12.5	**50.0**	0.0	0.0	0.0	20.8
+ Tactile-VLA [11]	✓	✗	8.3	50.0	45.8	12.5	37.5	**50.0**	25.0	12.5	0.0	30.2
+ ForceVLA [12]	✓	✗	16.7	58.3	54.2	12.5	37.5	**50.0**	12.5	25.0	0.0	34.4
+ MoSS (ours)	✓	✗	**25.0**	66.7	**66.7**	**37.5**	50.0	37.5	25.0	12.5	**12.5**	42.7
+ TA-VLA [13]	✗	✓	8.3	50.0	45.8	**37.5**	50.0	25.0	25.0	25.0	**12.5**	33.3
+ MoSS (ours)	✗	✓	8.3	58.3	50.0	**37.5**	**62.5**	25.0	25.0	**37.5**	**12.5**	37.5
+ MoSS (ours)	✓	✓	**25.0**	**83.3**	**66.7**	**37.5**	**62.5**	**50.0**	**37.5**	25.0	**12.5**	**49.0**
π_0 [23]	✗	✗	0.0	25.0	50.0	12.5	37.5	37.5	25.0	12.5	12.5	27.1
+ Tactile-VLA [11]	✓	✗	0.0	33.3	50.0	25.0	37.5	50.0	25.0	12.5	0.0	29.2
+ ForceVLA [12]	✓	✗	**16.7**	33.3	62.5	50.0	**50.0**	50.0	25.0	12.5	12.5	38.6
+ MoSS (ours)	✓	✗	8.3	**50.0**	62.5	25.0	**50.0**	50.0	25.0	25.0	12.5	38.6
+ TA-VLA [13]	✗	✓	8.3	25.0	58.3	12.5	37.5	**75.0**	37.5	12.5	12.5	34.4
+ MoSS (ours)	✗	✓	0.0	41.7	62.5	50.0	**50.0**	62.5	**50.0**	**25.0**	12.5	41.7
+ MoSS (ours)	✓	✓	8.3	**50.0**	**66.7**	**62.5**	**50.0**	62.5	37.5	**25.0**	**25.0**	**45.9**