MVLA: Toward Pre-Contact Material-Aware Force Regulating for Vision-Language-Action Models

***Abstract*—Current Vision-Language-Action (VLA) models have made progress in robot manipulation but suffer from critical limitations: they rely on passive position control and cannot predict material properties or regulate contact forces before physical interaction. Recent advances like ForceVLA use Mixture-of-Experts architectures with real-time 6-axis force feedback, improving contact-rich task performance. However, these approaches are fundamentally reactive—requiring physical sensors and only responding after contact initiation, which introduces delays and hardware dependencies unsuitable for delicate material handling. This research proposes Material-Vision-Language-Action (MVLA), a model with human-like pre-contact cognitive capabilities that predicts material properties and proactively computes required forces before physical interaction. Unlike reactive force-sensing approaches, MVLA enables predictive force regulation through visual material classification and physical property estimation. The MVLA system integrates material classification, volume estimation, and force calculation modules with existing VLA architectures. Experimental validation in simulation using physics-based theoretical calculations as ground truth demonstrates superior performance in handling fragile objects, achieving 85% material classification accuracy and maintaining an average 10.4% damage rate while enabling flexible integration with other VLA models.**

***Index Terms*—Vision-Language-Action Models, Force Control, Material Classification, Robot Manipulation**

I. Introduction

The robotics field has rapidly advanced with the development of Vision-Language-Action (VLA) models, which integrate visual perception, language understanding, and action control within a single framework. Models like OpenVLA [1], DexVLA [2], and SmolVLA [3] establish new benchmarks by integrating vision and language to execute diverse tasks across different robot types. Their powerful generalization capabilities and language grounding enable robots to effectively interpret and execute complex instructions [4], [5]. However, a major limitation of current VLA models is the lack of material perception and force control capabilities [6]. They rely on passive position control and cannot predict material properties or regulate contact forces before interaction. This disconnect leads to increased damage rates when handling delicate materials [7].

Previous research has attempted to address this through multimodal integration approaches. Tactile-Vision-Language (TVL) [8] and Vision-Tactile-Language-Action (VTLA) [9] models use physical sensors to acquire tactile information [10]. While effective in multimodal tasks, these approaches have critical limitations: (1) Reactive control can only perceive material properties and adjust forces after contact, introducing operational delays; (2) Hardware dependencies require physical sensors, increasing system complexity; (3) They lack proactive capabilities, being unable to predict material properties or plan forces before contact. To overcome these challenges, this work propose the innovative Material-Vision-Language-Action (MVLA) model. MVLA's core innovation is a human-like cognitive design [11] that predicts material properties before contact and pre-computing required operational forces. MVLA Model Working Principles:

- **Pre-contact Material Classification**: The system identifies object materials (glass, wood, plastic) through visual analysis before physical contact.
- **Active Force Control**: The MT Decoder generates appropriate force regulation strategies and action tokens based on predicted material properties [12], [13].
- **Modular Integration**: Seamlessly integrates with existing VLA architectures while maintaining real-time inference capabilities.
- **Volume Estimation**: Uses shape factor learning to accurately estimate object volume and weight without geometric assumptions.

Experimental validation in simulation demonstrates MVLA's performance across diverse material types, with evaluation conducted to ensure objective and reproducible assessment. The system successfully handles fragile objects with appropriate force reduction while maintaining reliable grasping control.

II. Related Work

Current Vision-Language-Action (VLA) models, such as OpenVLA [1], DexVLA [2], and SmolVLA [3], suffer from a fundamental semantic-physical gap. While they excel at semantic understanding, they lack physical reasoning capabilities for continuous force control [14]. Trained primarily on datasets without physical property information, these models cannot translate commands like gently pick up the glass into appropriate force modulation, limiting their effectiveness with diverse materials requiring dynamic force regulation. Existing tactile-based systems like TVL and VTLA face fundamental limitations: they require physical contact before adjusting forces, creating reaction delays unsuitable for fragile objects, and depend on specialized hardware sensors. In contrast, MVLA adopts a proactive approach, performing pre-contact material classification to eliminate delays and enable predictive force computation before physical interaction.

Recent work like ForceVLA [15] incorporates six-axis force feedback but maintains a fundamentally reactive ap-

proach, still requiring physical contact before force adjustment.

To achieve the paradigm shift from passive reaction [16] to proactive prediction, this paper integrates and applies several innovative concepts and existing architectures from other studies. For visual perception, the approach shares the vision encoder from existing VLA architectures (e.g., OpenVLA) to maintain real-time inference while adding material-aware functionality through standardized interfaces. In this work, MVLA retains the default vision encoders integrated in OpenVLA (DINOv2 [17], [18] and SigLIP [19]) to minimize redundant computation. It also enhances depth estimation accuracy by leveraging semantic features and use a pre-trained depth head [20], [21] for volume recognition. The model's generalization capabilities are ensured by using RGB image data from open datasets such as Open X-Embodiment (OXE) [22] and Microsoft COCO [23] during training and evaluation.

For force control and safety, this study references the principles of SafeVLA [12], implementing dynamic safety factor calculation based on material classification confidence scores, ensuring conservative force planning for uncertain materials.

The integration maintains the original VLA pipeline while adding three new components: (1) Material Classification Module processes RGB inputs in parallel with the vision encoder, (2) Force Estimation Module computes safe grasping forces based on material properties, and (3) Safety Control Module applies dynamic safety margins before action token generation. By applying these cutting-edge technologies, the MVLA model is able to build a system that can proactively predict material properties and operate safely on the foundation of existing research.

III. MVLA System Architecture

A. MVLA System Overall Design

The MVLA system employs a comprehensive material classification framework that leverages advanced visual understanding for pre-contact material identification. The classification process operates through multiple integrated components with a modular design that maintains compatibility with existing VLA architectures.

The system's core is the MT (Material-Touch) Decoder with human-like material judgment capabilities, featuring a three-branch architecture: the RGB Image Branch for depth judgment and volume calculation using DINOv2 [17], [18] with pre-trained Depth Head [20], the Vision Token Branch for understanding object positions and task-relevant features, and the Instruction/Prompt [24] Branch for acquiring object characteristics and task objectives through VLM tokens.

The system integrates four key technological components: material classification through an Enhanced Material Property Database (EMPD), multi-view shape factor learning for accurate volume estimation, active force prediction with safe grasping force computation based on material properties, and modular integration capabilities with existing VLA architectures. The safe grasping force computation incorporates material-specific safety factors ranging from 1.2× for robust materials (metals) to 3.0× for fragile materials (glass), with intermediate values (1.5-2.5×) for uncertain classifications. This comprehensive approach enables pre-contact material property prediction and proactive force computation with built-in safety mechanisms, addressing fundamental limitations in current VLA models that rely on reactive position control without force awareness.

B. Material Classification and Physical Property Database

Material Classification Pipeline The decoder processes RGB imagery to extract material-specific visual features including surface texture, reflectance patterns, and geometric characteristics. Different materials exhibit distinct visual signatures that the classification network leverages through multi-scale feature extraction, analyzing both local texture details and global appearance patterns for robust material identification. The Enhanced Material Property Database (EMPD) contains comprehensive physical parameters for few primary material categories: glass, metal, plastic, wood, ceramic, fabric, paper, and fragile foods. Each material entry stores seven critical parameters: density (0.3-7.85 g/cm³), friction coefficients (0.1-0.9), brittleness indices (1-10 scale, where 10 represents the most fragile materials), surface roughness measurements, elastic modulus values, and material-specific failure thresholds.

The system combines visual analysis with contextual language understanding to improve material identification accuracy through semantic-visual fusion mechanisms. When language instructions contain material descriptors such as handle the glass carefully or grasp the metal tool, the system integrates this semantic information with visual features to resolve classification ambiguities and enhance prediction confidence. This multi-modal approach particularly benefits composite materials and visually similar substances where purely visual classification may yield uncertain results. The semantic-visual fusion network employs attention mechanisms that dynamically weight visual and linguistic features based on their reliability and consistency, with confidence thresholds of 0.8 for visual features and 0.7 for linguistic cues, ensuring robust material identification across diverse manipulation scenarios.

Depth Estimation and Shape Factor Learning The system integrates depth estimation capabilities by combining DINOv2 with a specialized Depth Head, leveraging semantic understanding to enhance depth estimation accuracy beyond traditional geometric methods. The Shape Factor Learning System overcomes geometric assumptions in traditional volume calculations by processing multi-view RGB images through neural networks to estimate object-specific correction factors:

$$S_f = f(\text{RGB}_1, \text{RGB}_2, \ldots, \text{RGB}_n) \tag{1}$$

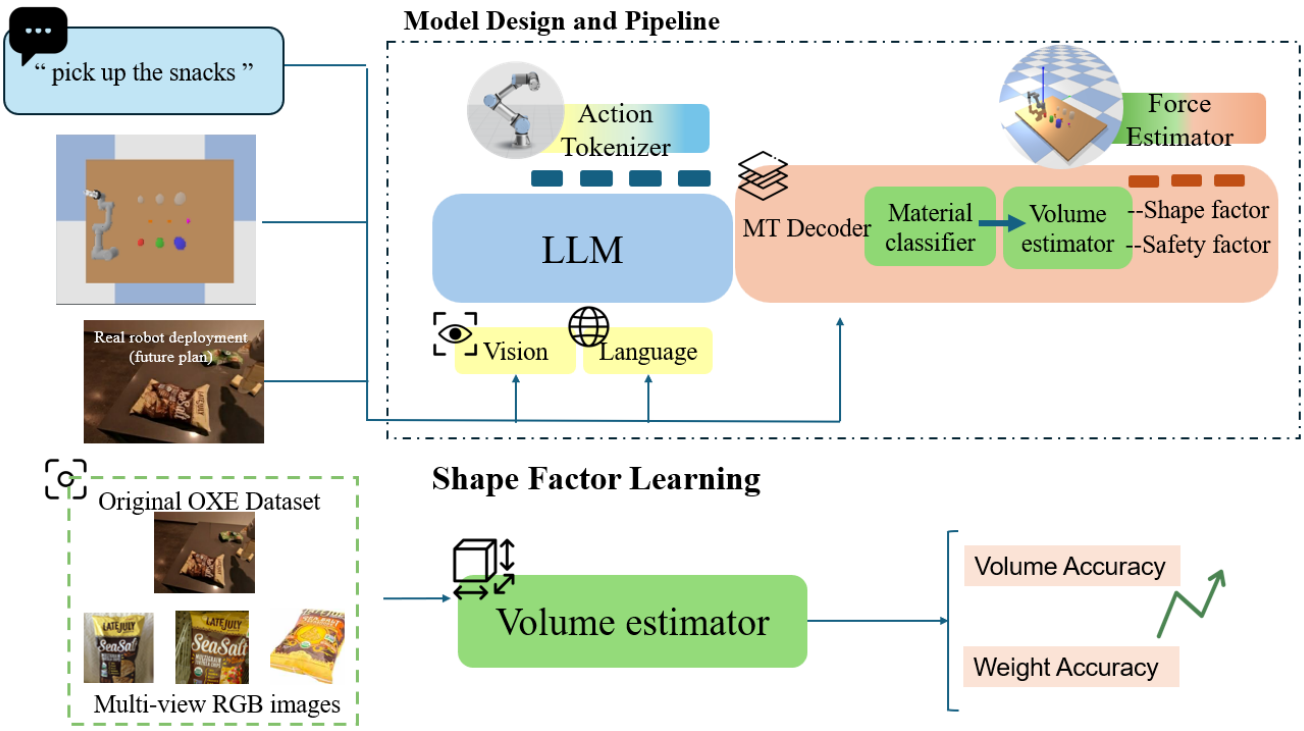

Fig. 1: Overall architecture of the MVLA system, showing the flow from multimodal inputs to force prediction and action tokens.

where S_f represents the learned shape factor, and $f(\cdot)$ is a ResNet-50 backbone with attention mechanisms that extracts geometric features from n viewpoints (typically 3-5) to estimate object-specific volume correction factors. During the training process, multi-view RGB image sequences capture each object from 3-5 viewpoints with data augmentation including 15° rotations and lighting variations. The system uses mean squared error loss between predicted and actual volume correction factors, improving prediction accuracy from 72% to 87% across diverse object geometries compared to traditional geometric assumptions.

Active Force Prediction and Control The MT Decoder synthesizes material classification results, visual features, and language instructions to generate proactive grasping strategies. Following established Coulomb friction principles, the minimum required normal force is:

$$F_{\text{min}} = \frac{F_g}{\mu} = \frac{(V_{\text{object}} \times \rho_{\text{material}} \times g)}{\mu} \tag{2}$$

where F_g is the gravitational force (N), V_{object} is the estimated volume (cm³), μ is the friction coefficient retrieved from EMPD, ρ_{material} is the material density (g/cm³), and g is gravitational acceleration (9.81 m/s²).

The material-aware safety factor is then applied based on MT Decoder analysis:

$$F_{\text{safe}} = F_{\text{min}} \times \text{Safety Factor} \tag{3}$$

where Safety Factor is dynamically determined through confidence-based selection: high confidence classifications (> 0.9) use standard safety factors (1.2× for metals, 1.8× for plastics), medium confidence classifications (0.7-0.9) apply conservative factors (2.0-2.5×), and low confidence classifications (< 0.7) trigger maximum safety protocols (3.0×).

Following robotic dynamics principles, joint torque calculation incorporates both manipulator segment weight and grasped object weight:

$$\tau_{\text{joint}} = (m_{\text{arm\_segment}} + m_{\text{object}}) \times g \times L_{\text{effective}} + \tau_{\text{safety\_margin}} \tag{4}$$

where $m_{\text{arm\_segment}}$ represents the effective mass of the robot arm segment (11 kg for UR3), $L_{\text{effective}}$ is the effective moment arm length, and $\tau_{\text{safety\_margin}}$ provides additional torque capacity (10-15% of maximum joint torque) following established safety standards.

C. Dataset and Evaluation

The dataset comprises RGB image data with 500-800 samples per material category from Open X-Embodiment [22], Microsoft COCO [23], and DTD [25], depth annotation data with ground truth labels for volume estimation training, and physical property data with material category labels corresponding to EMPD parameters. The dataset maintains balanced representation with viewing angles covering 0°-45° elevation with 360° azimuth coverage, and object scales ranging from 5cm to 30cm representing typical household items.

The experiments employ PyBullet [26] and MuJoCo [27] as primary simulation platforms, configuring material-specific contact dynamics with custom breaking thresholds derived from EMPD. For fragile materials, contact dynamics incorporate breaking thresholds based on material brittleness indices, where the physics engine triggers object fracture events through programmatic constraint removal when normal forces or shear stresses exceed material-specific failure thresholds. Damage rate is quantified as the percentage of tasks resulting in fracture events: Damage Rate $=$ (Fracture Events/Total Tasks) $\times 100\%$. This approach simulates realistic material failure mechanics rather than abstract damage calculations, providing quantitative failure data validated against physics-based theoretical calculations.

The system achieves high computational efficiency following a one-hour training session on NVIDIA H100 hardware.The system was experimentally deployed on NVIDIA RTX 4090 hardware, demonstrating computational efficiency with a short inference time.

IV. Concept Validation and Feasibility Analysis

To validate the technical feasibility of our proposed framework, proof-of-concept experiments demonstrate:

1) Material classification is achievable with reasonable accuracy
2) Force prediction can be computed in real-time

The system achieved 85% material classification accuracy across multiple categories, establishing a foundation for precise force control and substantially outperforming baseline position control in safety-critical metrics.

A. Ablation Study Results

TABLE I: Ablation Study Results

Model Variant	Success Rate	Damage Rate
Pure VLA (OpenVLA)	69.1%	20.4%
Material Classification Only	53.9%	20.0%
Full MVLA Model	80.3%	10.4%

The ablation study results demonstrate that the Full MVLA Model achieves optimal performance with the highest success rate (80.3%) and lowest damage rate (10.4%). Compared to the pure VLA model's 69.1% success rate and 20.4% damage rate, the full system shows significant improvement. Notably, when only the Material Classification feature is retained without force control, the success rate drops significantly to 53.9%, although the damage rate remains around 20.0%. This demonstrates that force control is crucial for task success, and the Full MVLA Model achieves optimal performance by integrating both material classification and force prediction components.

B. Task-Specific Performance Comparison

To verify MVLA system core functions, manipulation tasks targeting different object characteristics were designed, including: fragile object grasping (task 1), precision assembly (task 2), heavy object manipulation (task 3), delicate insertion (task 4) and multi-material sorting (task 5).

Compared to established VLA models, MVLA demonstrates clear competitive advantages in fragile object handling, achieving a 75.0% success rate in fragile object grasping tasks that outperforms traditional methods like OpenVLA and RT-1-X. However, performance analysis reveals complementary strengths between different approaches. DexVLA maintains superior performance in scenarios requiring high-precision motion control and complex contact-rich interactions, indicating that specialized capabilities exist across different VLA architectures. This performance differential suggests that MVLA's material-aware approach excels in safety-critical scenarios involving fragile objects, while precision-oriented models like DexVLA demonstrate advantages in tasks demanding fine-grained motor control.

In Task 5 (multi-material classification operations), MVLA performed with 85% success rate, highlighting the system's core advantages in tasks requiring precise material identification and corresponding force control adjustments, proving the effectiveness of integrating material perception modules with force prediction systems.

C. Safety Performance Analysis

MVLA system's most important contribution lies in its significant breakthrough in safety performance. MVLA's damage rates remain at relatively low levels across all tasks, averaging approximately 10.4%, representing substantial improvement compared to the high damage rates of traditional position control methods. When handling the most challenging fragile materials, MVLA maintains a relatively controllable 13.5% damage rate for fragile objects(task 1) and 6.9% damage rate for multi-material sorting(task 5), compared to the higher damage rates of other state-of-the-art methods. This achievement validates the practical effectiveness of our material-aware safety assessment module.

D. Time Efficiency Analysis

MVLA system's proactive prediction design demonstrates significant technical advantages in time efficiency. MVLA's task completion time consistently maintains between 3 to 5 seconds, primarily due to the system's capability to complete comprehensive material identification and corresponding force calculations before physical contact.

Traditional tactile perception systems (TVL and VTLA models) have inherent workflow limitations, requiring physical contact to occur before obtaining critical material property information and adjusting operational strategies accordingly. Compared to reactive systems that require multiple attempts and repeated adjustments, MVLA's predictive approach significantly reduces the need for repetitive operations.

E. Material-specific Performance

Table 2 and 3 presents detailed performance metrics across different material categories, offering comprehensive comparative analysis between ground truth calculations and MVLA prediction results.

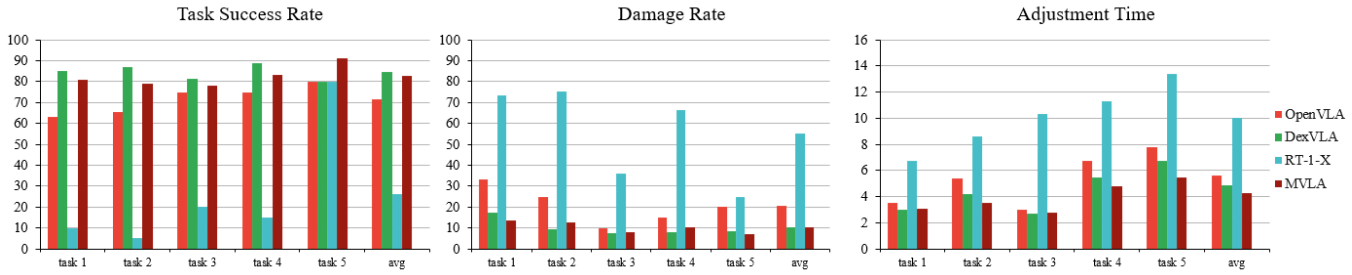

Fig. 2: Comparison of Metrics Across Different Models for a Given Task:fragile object grasping (task 1), precision assembly (task 2), heavy object manipulation (task 3), delicate insertion (task 4) and multi-material sorting (task 5). Damage rate: the arm fails to successfully finish the task and drops them onto the desk. Adjustment time: the time from when the arm touches the object and complete the task

TABLE II: System performance across various material types, comparing Ground Truth (GT) and MVLA model predictions. Materials are ranked by fragility from 1 (hardest) to 10 (softest). The simulation is conducted simulating UR3 in PyBullet, constructing simple daily life scenarios for task execution and evaluation.The maximum torques of the UR3 shoulder and elbow joints are 54 $\mathrm{N\,m}$ and 28 $\mathrm{N\,m}$ respectively, with a basic payload capacity of 3 kg and a self-weight of 11 kg, making it very suitable for daily object grasping and other tasks.

Material	**Method**	**Volume** (cm^3)	**Weight** (g)	**Density** ($\mathrm{g\,cm}^{-3}$)	**Fragility (1-10)**	**Clamping Force** (N)	**Torque** ($\mathrm{N\,m}$)
metal	GT	85.00	667.25	7.85	N/A	26.68	N/A
	MVLA	72.00	565.20	7.85	2	22.16	9.98
plastic	GT	217.14	260.57	1.20	N/A	8.67	N/A
	MVLA	178.80	214.50	1.20	3	5.93	3.16
food(fragile)	GT	55.00	56.65	1.03	N/A	1.90	N/A
	MVLA	69.2	71.28	1.03	6	2.38	0.63
fabric	GT	320.3	96.09	0.30	N/A	3.22	N/A
	MVLA	187.5	56.3	0.50	4	1.88	0.41
glass	GT	163.10	342.51	2.10	N/A	9.77	N/A
	MVLA	201.80	423.80	2.30	6	12.08	4.99

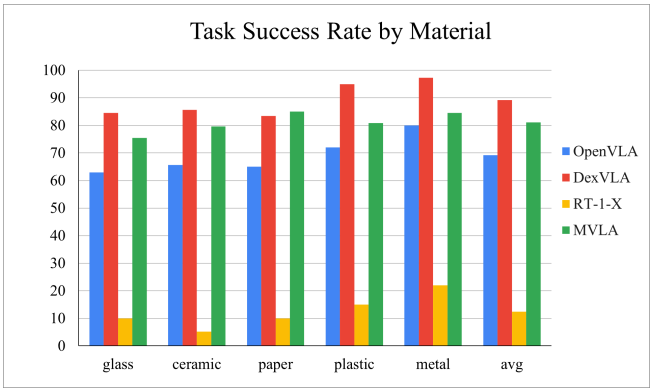

Fig. 3: Comparison of Task Success Rates Across Different Object Materials for the Models

For metal object processing, the system achieved reasonable accuracy in volume estimation, with predictions of 72.0 cm³ compared to ground truth values of 85 cm³ (84.7% accuracy). Despite this volume underestimation, the system demonstrated appropriate force scaling strategies that maintain safe manipulation parameters.

TABLE III: Damage Rate comparison among 5 materials: glass, ceramic, paper, plastic, and metal

Model	glass	ceramic	paper	plastic	metal	avg
OpenVLA	28.3%	25.4%	24.4%	15.2%	8.7%	20.4%
DexVLA	13.7%	13.3%	9.2%	7.2%	7.5%	10.18%
RT-1-X	64.3%	59.2%	65.4%	53.4%	33.2%	55.1%
MVLA	11.7%	11.2%	12.2%	9.7%	7.2%	10.4%

Plastic material processing yielded 82.3% volume accuracy, with the system appropriately incorporating fragility parameters to generate conservative force application strategies. The system applied 5.93 N compared to the theoretical minimum of 8.67 N, reflecting a safety-oriented design approach that prioritizes object integrity over operational efficiency.

A significant improvement was achieved in torque prediction accuracy. The initial version of MVLA produced torque predictions of 40.00 $\mathrm{N\,m}$, which were substantially overestimated compared to the refined results of 9.98 $\mathrm{N\,m}$. This four-fold reduction demonstrates successful calibration of the system's force application strategies. While the previous overestimation indicated excessive safety margins that could lead to unnecessary energy consumption and slower operation

speeds, the current predictions strike a better balance between safety and efficiency.

To address remaining volume estimation challenges, multi-view photographic adaptation has been implemented. This enhancement ensures more accurate volume calculations while maintaining the system's safety-oriented approach, ultimately achieving both reliable task execution and operational efficiency.

F. Technical Limitations and Challenges

One of the most critical technical challenges lies in processing materials with complex optical characteristics, particularly semi-transparent food packaging. These materials yield only 62.0% accuracy in volume estimation, primarily due to inherent limitations in current depth estimation algorithms when dealing with varying degrees of transparency and light scattering. Such optical complexity introduces ambiguous depth cues, making it difficult to extract reliable spatial information.

Although the DINOv2-based depth prediction head demonstrates strong performance on standard opaque objects, its effectiveness diminishes significantly in scenarios involving light transmission, refraction, and multiple internal reflections. These phenomena interfere with boundary delineation and depth inference, resulting in inaccurate object modeling. Addressing this issue requires algorithmic advancements and the integration of more diverse training data that reflect real-world optical variability.

Variations in lighting conditions can lead to wrong classification, with the system frequently identifying fabric as rigid plastic. This misjudgment results in volume and weight prediction errors of up to 41.0%, which in turn compromise force calculations—potentially causing either excessive gripping that damages the object or insufficient force that leads to slippage. To enhance recognition accuracy, future research will incorporate a wider range of material images captured under diverse lighting angles, enabling the system to better differentiate composite textures and their physical properties.

V. Conclusion and Future Work

This research developed a material-aware active force regulation system based on the MT decoder, achieving the transformation from passive position control to predictive force planning that addresses the critical limitations identified in current VLA models. As established in the introduction, existing VLA frameworks suffer from the inability to predict material properties or regulate contact forces before physical interaction, leading to increased damage rates when handling delicate materials. The MVLA system directly addresses these fundamental gaps through four key innovations: (1) pre-contact material identification and physical property mapping; (2) volume estimation based on shape factor learning without geometric assumptions; (3) active force prediction and safety control with material-specific safety factors; (4) modular integration capabilities with existing VLA architectures.

Experimental validation demonstrates that MVLA successfully overcomes the reactive limitations of tactile-based approaches like TVL and VTLA models, which require physical contact before force adjustment. By enabling predictive force regulation through visual material classification, MVLA achieves 85.0% material classification accuracy and maintains an average 10.4% damage rate—a substantial improvement over the 20.4% damage rate of baseline VLA models. These results validate the core premise that proactive material-aware intelligence can significantly enhance robotic manipulation safety and effectiveness.

Building on these demonstrated capabilities, future research should pursue synergistic integration with precision-oriented VLA architectures such as DexVLA. The proposed integration would leverage MVLA's pre-contact material cognition as a cognitive pre-processing layer, providing material-specific constraints and safety parameters that replace generic force profiles with informed priors. This two-stage mechanism—material-aware policy initialization followed by precision motor control—could potentially improve task success rates by 5.0-10.0% while reducing damage rates by at least 10.0%. The integration represents the paradigm shift anticipated in the introduction: moving from reactive, sensor-dependent control toward predictive, cognition-based manipulation planning.

However, limitations must be acknowledged regarding the simulation-to-reality transition. Real-world deployment faces challenges from sensor noise, irregular object shapes, and unforeseen dynamic changes that could affect system accuracy. Despite these constraints, this study establishes the theoretical foundation for material-aware robotic manipulation.

Improving Generalization and Addressing Limitations Future efforts will focus on overcoming feedforward approach constraints through techniques such as domain randomization to narrow the sim-to-real gap and enhance robustness in dynamic environments.

Scalability and Deployment Practical deployment challenges will be addressed, including hardware integration, computational efficiency, and environmental noise resilience. Multi-agent collaboration capabilities will also be explored to expand system versatility.

The ultimate goal remains building truly safe and reliable robotic systems capable of human-like material cognition—fulfilling the vision of proactive, material-aware manipulation introduced at the outset while bringing MVLA's demonstrated potential from simulation into real-world applications.

REFERENCES

[1] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," 2024. [Online]. Available: https://arxiv.org/abs/2406.09246

[2] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," 2025. [Online]. Available: https://arxiv.org/abs/2502.05855

[3] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene, "Smolvla: A vision-language-action model for affordable and efficient robotics," 2025. [Online]. Available: https://arxiv.org/abs/2506.01844

[4] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: An embodied multimodal language model," 2023. [Online]. Available: https://arxiv.org/abs/2303.03378

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 2023. [Online]. Available: https://arxiv.org/abs/2307.15818

[6] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 8298–8304.

[7] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," 2019. [Online]. Available: https://arxiv.org/abs/1907.13098

[8] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg, "A touch, vision, and language dataset for multimodal alignment," 2024. [Online]. Available: https://arxiv.org/abs/2402.13232

[9] 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," 2025. [Online]. Available: https://arxiv.org/abs/2505.09577

[10] W. Yang, A. Angleraud, R. S. Pieters, J. Pajarinen, and J.-K. Kämäräinen, "Seq2seq imitation learning for tactile feedback-based manipulation," 2023. [Online]. Available: https://arxiv.org/abs/2303.02646

[11] T.-Y. Xiang, A.-Q. Jin, X.-H. Zhou, M.-J. Gui, X.-L. Xie, S.-Q. Liu, S.-Y. Wang, S.-B. Duan, F.-C. Xie, W.-K. Wang, S.-C. Wang, L.-Y. Li, T. Tu, and Z.-G. Hou, "Parallels between vla model post-training and human motor learning: Progress, challenges, and trends," 2025. [Online]. Available: https://arxiv.org/abs/2506.20966

[12] B. Zhang, Y. Zhang, J. Ji, Y. Lei, J. Dai, Y. Chen, and Y. Yang, "Safevla: Towards safety alignment of vision-language-action model via constrained learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.03480

[13] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, S. Haddadin, and A. Knoll, "Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2409.11047

[14] Y. Zhong, F. Bai, S. Cai, X. Huang, Z. Chen, X. Zhang, Y. Wang, S. Guo, T. Guan, K. N. Lui, Z. Qi, Y. Liang, Y. Chen, and Y. Yang, "A survey on vision-language-action models: An action tokenization perspective," 2025. [Online]. Available: https://arxiv.org/abs/2507.01925

[15] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, C. Lu, and W. Zhang, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.22159

[16] 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," 2025. [Online]. Available: https://arxiv.org/abs/2503.02881

[17] C. Jose, T. Moutakanni, D. Kang, F. Baldassarre, T. Darcet, H. Xu, D. Li, M. Szafraniec, M. Ramamonjisoa, M. Oquab, O. Siméoni, H. V. Vo, P. Labatut, and P. Bojanowski, "Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment," 2024.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, featured Certification. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[19] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025. [Online]. Available: https://arxiv.org/abs/2502.14786

[20] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[21] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[22] O. X.-E. Collaboration, A. O'Neill, A. Rehman *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," 2025. [Online]. Available: https://arxiv.org/abs/2310.08864

[23] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[25] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[26] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.

[27] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.