Touching The Future: Visuo-Tactile Reactive Robot Grasping

***Abstract*— Traditional approaches to grasping often plan for a grasp pose from visual feedback with little regard to what happens when the robot attempts to perform the actual grasp. In stark contrast, humans grasp reactively, leveraging both visual and tactile feedback to close the loop on the grasping process. In this paper, we propose Touching the Future, a multimodal action-conditioned model that learns to predict grasp success probabilities and future sensor readings using visuo-tactile feedback during physical interaction. We first train our model on a dataset collected using a hand-held gripper allowing for efficient data scaling and diversity. Next, we demonstrate the effectiveness of the model for reactive grasping via a sample-based model-predictive controller implemented on a physical robot. Our model's generalization capabilities are facilitated by using pre-trained large-scale models as backbones. Across experiments, our approach shows robust generalization to real-world grasping scenarios, outperforming point cloud based heuristic method, prior work–inspired visuo-tactile model, and ablation baselines. These results demonstrate that learning a rich latent representation from multimodal feedback improves grasping when compared to traditional methods that only learn to predict binary grasp success and/or are purely visual.**

***Index Terms* - Visuo-Tactile Perception, Multimodal Learning, Perception for Grasping and Manipulation**

I. Introduction

Grasping objects is an intuitive process for humans. Even before physically attempting a grasp, we often subconsciously predict whether we will successfully pick up and hold an object, relying on prior experience, visual information, and tactile feedback. We use vision for global and touch for local information, respectively. When an initial grasp attempt fails, humans instinctively use sensory feedback, especially touch, to reactively adjust and correct the grasp. This integration of multi-sensory signals is fundamental to achieving robust and adaptive manipulation in everyday life. Translating such capabilities to robotic systems remains a significant challenge, particularly the ability to reason about future outcomes of actions before execution.

To address this challenge, a range of strategies have been explored. Approaches based solely on vision, such as point clouds or depth maps [1], infer grasp affordances either through geometric reasoning (e.g., surface normals, curvature, or analytic metrics) or through learning-based methods [2] that directly map 3D data to grasp predictions. While effective, both often struggle in contact-rich scenarios. Other methods leverage tactile sensing to reason about stability once contact is made [3], but lack foresight before the grasp occurs. Multimodal systems combining vision and touch [4] have shown promise in bridging this gap, offering richer representations for both pre-contact and post-contact reasoning.

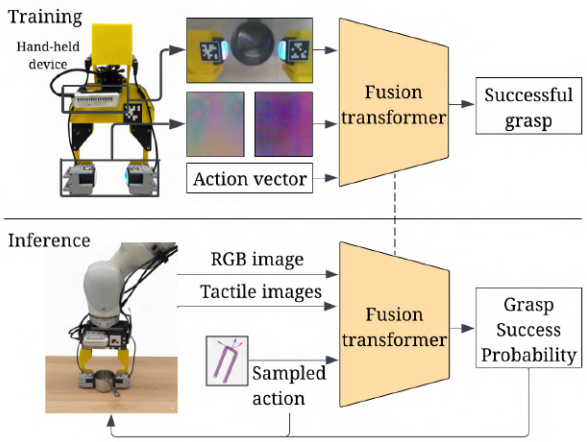

Fig. 1. Human-collected data from a hand-held device is used to train the proposed model, where visual and tactile modalities are conditioned on an action vector. At deployment, the trained model transfers to a robotic platform, which iteratively evaluates candidate grasps and selects the best action in a closed loop.

Building on prior multimodal grasping work [4], we propose an action-conditioned encoder–decoder architecture that jointly reasons over vision and tactile feedback for *reactive grasping*. Our approach differs from prior efforts along three key axes: (i) Unlike models that only classify grasp success, our approach predicts the future sensory states (both visual and tactile) given current observations and a candidate gripper action, enabling richer foresight about contact dynamics and more informed grasp outcome prediction; (ii) our approach leverages pre-trained large-scale models to facilitate generalization to unseen objects; and (iii) we leverage human demonstrations from a simple to construct UMI-like gripper [5] to scale up training data. This approach allows us to train a rich model that is then able to plan for grasp re-adjustments leveraging a sample-based model-predictive controller. Our robotic platform, shown in Fig. 1, is equipped with a wrist-mounted RGB camera and high-resolution vision-based tactile sensor (DIGIT [6]) embedded in a parallel jaw gripper. By learning to predict future sensory feedback conditioned on actions, our model develops a rich latent representation that captures the physical dynamics of grasping and predicts grasp success likelihood.

II. Related Work

A. Vision-Based Grasping

Traditional robotic grasping methods rely primarily on visual inputs to infer graspable regions. Early approaches

Manuscript 4412 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2025.

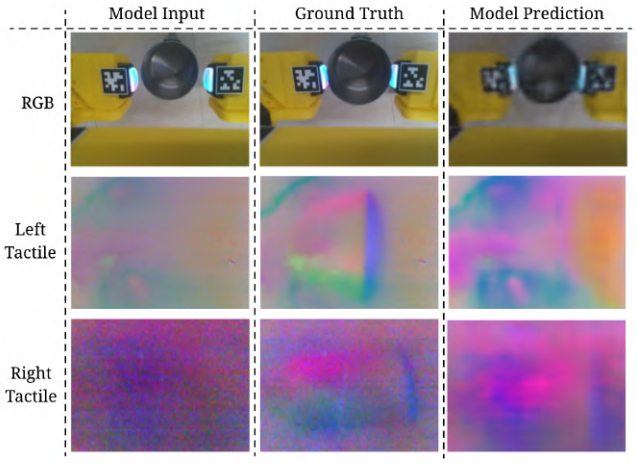

Fig. 2. Model inputs and reconstruction outputs. Each row corresponds to a sensing modality (camera, left tactile, right tactile). Columns show the input, the ground-truth signal, and the model's reconstructed output.

employed geometric or depth-based heuristics to plan grasp poses [1]. Subsequent work applied learning-based techniques directly on point clouds to predict grasp candidates [2]. More recently, methods such as Neural Grasp Distance Fields (NGDF) [7] and GIGA [8] represent valid grasp poses and object geometries in continuous latent spaces, enabling more flexible and generalizable grasp generation. CenterGrasp [9] further integrates object reconstruction and grasp prediction into a unified representation, supporting end-to-end inference. Generative models such as AutoSDF [10] and Zero-1-to-3 [11] leverage shape priors learned from sparse or single-view images to implicitly predict graspable geometries, achieving strong performance without dense 3D supervision. Despite this progress, vision-only methods often struggle in cases where appearance alone is insufficient to determine grasp stability, such as deformable, occluded, or visually ambiguous objects. Our approach extends beyond vision-only methods by incorporating tactile feedback, enabling reasoning about contact dynamics and grasp stability that cannot be inferred from appearance alone.

B. Visuo-Tactile Manipulation

Tactile sensing is central to dexterous, contact-rich manipulation. Beyond stabilizing grasps, touch enables reasoning about local geometry, compliance, and contact mode transitions that are difficult to infer from vision alone. Recent advances show how tactile signals unlock manipulation capabilities that extend far beyond appearance-based cues. For instance, Vib2Move [12] leverages fingertip micro-vibrations to reconfigure objects in hand, using controlled friction modulation and a sliding model to achieve precise in-hand repositioning. Tactile Neural De-rendering [13] introduces a generative approach to reconstruct local 3D shape directly from tactile impressions, providing pose estimates with calibrated uncertainty.

Other work has focused on representation learning and control with touch. Tactile Functasets [14] learns neural implicit representations of high-dimensional tactile datasets, yielding compact, probabilistically interpretable encodings that improve downstream tasks such as in-hand pose estimation. Complementary to this, tactile-driven non-prehensile manipulation [15] formulates extrinsic contact mode control to realize robust sliding and pivoting skills using high-resolution tactile feedback.

Together, these results highlight the value of tactile sensing for manipulation, enabling in-hand reconfiguration, local shape reconstruction, compact multimodal representations, and robust non-prehensile control. However, despite this progress, there is comparatively less emphasis on applying visuo-tactile prediction specifically to *stable grasping*. Our work builds upon these advances by shifting the focus from manipulation in general to grasp success prediction for stable grasping.

C. Visuo-Tactile Grasping

Integrating vision and touch has proven critical for achieving robust grasping in unstructured environments. By combining pre-contact visual cues with post-contact tactile feedback, visuo-tactile models can both anticipate and correct grasp failures. Prior work demonstrated that tactile signals substantially improve grasp success prediction compared to vision alone [16]. Building on this, Calandra *et al.* [4] proposed an action-conditioned visuo-tactile model *More Than a Feeling*, which learns to iteratively adjust grasps using GelSight tactile sensors. Trained on approximately 6,500 robotic grasps, this model significantly outperformed purely visual baselines, highlighting the importance of contact feedback.

Here, *action-conditioned* refers to conditioning the prediction of grasp success not only on sensory inputs, but also on the specific action vector executed (e.g., gripper translation, rotation, or finger closure). In other words, the model learns to predict whether a grasp will succeed given both the current multimodal state and the candidate action, enabling more informed and adaptive control.

In contrast to prior datasets collected in simulation or through limited autonomous exploration, our work leverages diverse *human-collected grasp demonstrations*, providing more realistic contact-rich strategies and enabling models to learn predictive representations that generalize better to real-world grasping. Building on this foundation, we introduce a multimodal encoder–decoder that predicts future visuo-tactile states, allowing the model to capture a richer latent representation of contact dynamics and improve generalization to unseen and deformable objects.

III. Methodology

A. Problem Formulation

We build upon prior visuo-tactile frameworks [4] but introduce two key innovations to further improve action-conditioned grasp success prediction. First, rather than relying solely on on-policy robotic exploration, we leverage a dataset collected through human-supervised grasps. This human-collected data introduces a richer and more diverse

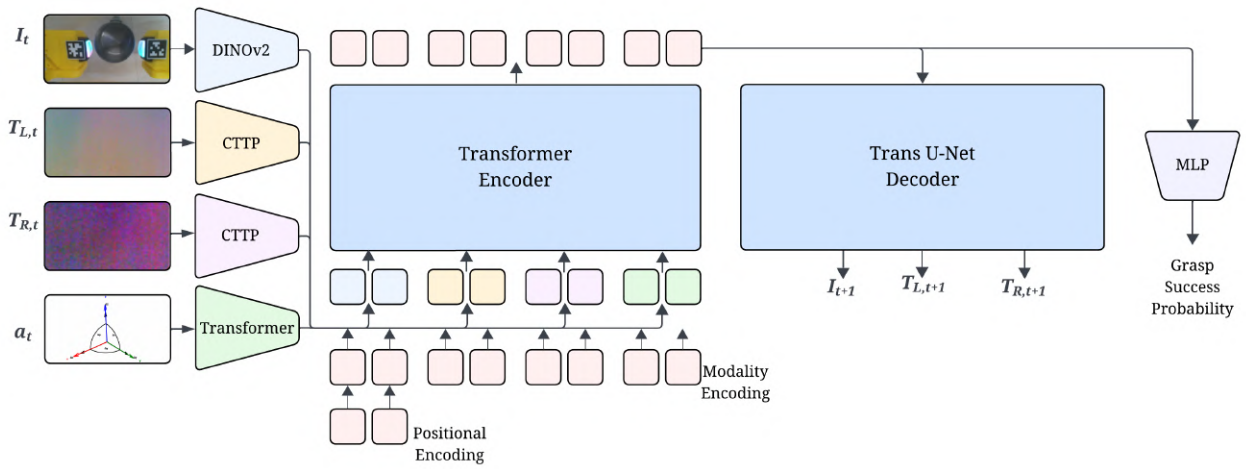

Fig. 3. Proposed model architecture. Visual and tactile modalities, along with the action vector, are encoded using pre-trained large-scale encoders and fused with a transformer. The fused latent representation is processed by two parallel heads: a TransUNet decoder for modality reconstruction and an MLP classifier for grasp success prediction.

action space that reflects realistic grasping strategies, including nuanced corrections and complex contact events that are difficult to generate autonomously. Second, instead of directly predicting grasp outcomes, we propose a multimodal encoder–decoder architecture that forecasts the future visuo-tactile sensory state conditioned on a candidate action. By explicitly learning the consequences of actions, the model forms a richer latent representation of contact dynamics and grasp stability, reasoning about physical interactions beyond what vision alone can provide.

This forward-predictive approach enables more robust action selection at runtime, allowing the robot to better anticipate grasp success. Vision provides global context before contact, while tactile sensing grounds the model in local, physical interaction cues. Together, this enables stronger generalization across both rigid and deformable objects, even when the latter were absent during training.

Formally, we cast visuo-tactile grasp prediction as a conditional future state estimation problem. At time t, let I_t be the wrist-mounted RGB image and T_t the tactile sensor image difference (contact reading minus a no-contact baseline), shown in Fig. 2. The robot's action a_t is defined as a 7-DoF gripper motion (a 6-DoF change in end-effector pose plus a change in gripper width). The proposed model f_θ predicts:

$$(\hat{I}_{t+1},\, \hat{T}_{t+1},\, \hat{P}_{\text{succ}}) = f_\theta(I_t, T_t, a_t) \tag{1}$$

where $\hat{I}_{t+1}$ and $\hat{T}_{t+1}$ are the predicted future visual and tactile observations, and $\hat{P}_{\text{succ}}$ is the probability of grasp success.

B. Model Architecture

The overall model architecture is illustrated in Fig. 3. It follows an encoder–decoder design that integrates vision, tactile, and action modalities for multimodal grasp success prediction. The network consists of three modality-specific encoders, a fusion transformer, a TransUNet-style decoder [17] for future sensory prediction, and a two-layer MLP head for grasp success estimation.

Each modality is first independently encoded:

- **Vision Encoder:** The RGB wrist image I_t is processed by a frozen DINOv2 [18] vision transformer, producing patch tokens with added positional encodings to preserve spatial structure. Freezing DINOv2 allows us to leverage its large-scale pretraining, transferring robust visual features without overfitting to our smaller dataset.
- **Tactile Encoder:** Left and right tactile difference images are encoded using a CTTP [19] encoder. Since CTTP was originally trained on Soft Bubble [20] and GelSlim [21] sensors, we fine-tune it on DIGIT data to account for the distribution shift. This ensures the encoder captures contact information specific to our hardware setup.
- **Action Encoder:** The 7-DoF delta action a_t is embedded using a lightweight transformer trained from scratch, producing action tokens tailored to our action space.

The outputs from the three encoders are concatenated, and modality encodings are added to the token sequence to enable the fusion transformer to distinguish between visual, tactile, and action inputs. These tokens are then processed by a cross-modal self-attention fusion transformer, which outputs a joint latent representation.

This fused latent space is used in two prediction heads:

- **Future Sensory Prediction:** The fused visual tokens are reshaped into a 2D feature map and decoded using a TransUNet-style decoder [17] to predict the next visual and tactile observations $(\hat{I}_{t+1}, \hat{T}_{t+1})$.
- **Grasp Success Prediction:** The latent tokens are passed through a two-layer MLP head to predict the probability of grasp success $\hat{P}_{\text{succ}}$.

Manuscript 4412 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2025.

C. Training Objectives

We jointly supervise future sensory prediction and grasp success estimation using a combination of regression and classification losses. For visual and tactile prediction, we apply mean squared error (MSE) losses between the predicted and ground-truth future observations. For grasp success, we treat it as a binary classification task and supervise the MLP head with a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{img} = \|\hat{I}_{t+1} - I_{t+1}\|_2^2 \tag{2}$$

$$\mathcal{L}_{tact\_left} = \|\hat{T}_{t+1}^{\text{left}} - T_{t+1}^{\text{left}}\|_2^2 \tag{3}$$

$$\mathcal{L}_{tact\_right} = \|\hat{T}_{t+1}^{\text{right}} - T_{t+1}^{\text{right}}\|_2^2 \tag{4}$$

$$\mathcal{L}_{tact} = \mathcal{L}_{tact\_left} + \mathcal{L}_{tact\_right}, \tag{5}$$

$$\mathcal{L}_{succ} = -\big[y_{\text{succ}} \log \hat{P}_{\text{succ}} + (1 - y_{\text{succ}}) \log(1 - \hat{P}_{\text{succ}})\big] \tag{6}$$

The total training objective is then defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{succ} + \lambda\big(\mathcal{L}_{img} + \mathcal{L}_{tact}\big) \tag{7}$$

where $\lambda = 0.1$ balances future prediction with grasp success supervision.

Fig. 4. Data collection setup. Two RGB-D cameras are used to localize the gripper (top view). The gripper then interacts with objects placed in the environment to collect grasping data (side view).

TABLE I

OFFLINE GRASP SUCCESS PREDICTION ACCURACY ON SEEN OBJECTS

Model	Accuracy (%)
Baseline	97.87%
Encoder Only	96.63%
Encoder-Decoder (Vision only)	97%
Ours (Full Architecture)	98.24%

D. Data Collection Setup

We modified a UMI [5] gripper to integrate both a wrist-mounted RealSense D435 RGB-D camera and DIGIT tactile sensors at the fingertips as shown in figures 1 and 4. The data collection setup additionally employed two external cameras observing the gripper to estimate its 6-DoF pose in the world frame. AprilTags were affixed to the gripper body and fingertips, enabling reliable tracking of both the gripper's position and orientation, as well as the grasp width of its fingers. Static cameras in the environment were used to calculate the gripper pose, while the one on the wrist was used to estimate the grasp width.

Grasp data was collected across 20 diverse objects, producing video sequences with a balanced distribution of successful and failed grasps. For each trial, the moment of initial contact was manually marked, and image frames from 5, 10, 15, and 20 time steps prior to contact were extracted and marked as pre-grasp. The difference between the end-effector poses at these two time instants was used as the action vector for training. Each grasp trial was annotated with a binary label, where 0 indicated failure and 1 indicated success. If the object could be lift up and held in air for 3s it was marked as a success. Our annotation style was similar to that of the baseline [4].

E. Experimental Setup

Our dataset consists of approximately 14,000 human-supervised grasp trials collected over 20 diverse objects. The experimental setup employed a wrist-mounted RealSense D435 RGB-D camera and DIGIT tactile sensors embedded in custom fingertip mounts for the WSG-50 gripper. For robot evaluation, we reproduced the same sensing configuration to closely emulate the data collection setup: the WSG-50 gripper was equipped with the UMI-designed fingertip mounts housing DIGIT sensors, and the RealSense D435 was mounted at the wrist.

The gripper was attached to a KUKA IIWA Med arm, and the entire control and perception pipeline was implemented in ROS1. An external RGB-D camera was additionally mounted to observe the workspace and generate point clouds, which were used for the initial point cloud–based heuristic to position the gripper near the target object prior to candidate action sampling. By maintaining hardware and sensing consistency between data collection and evaluation, the setup ensured seamless transfer of policies trained offline to real-world online robotic execution.

Manuscript 4412 submitted to 2026 IEEE International Conference on Robotics and Automation (ICRA). Received September 15, 2025.

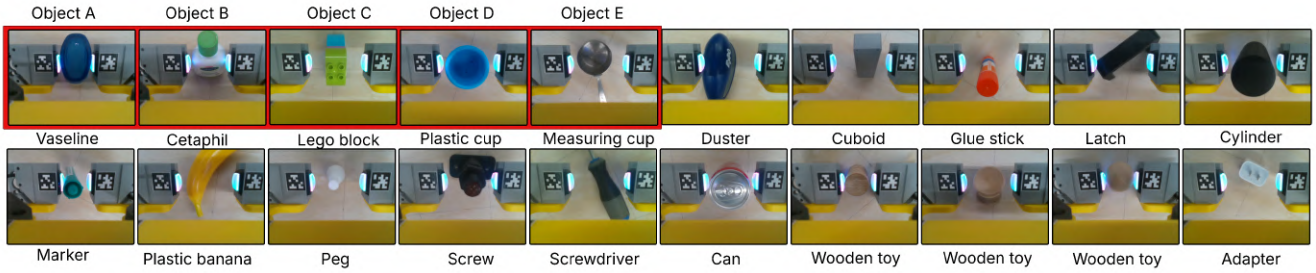

Fig. 5. Training dataset. The dataset consists of 20 objects spanning diverse shapes, colors, and textures. Shown are example images from the gripper's camera point of view before grasping. The first five objects, highlighted in red, were evaluated online on the robot as reported in Table II, and are referred to as objects A–E, respectively.

IV. Experiments and Results

A. Baselines

We split data at the object-instance level (20 objects for train/val, 5 for test). To evaluate our approach, we compare against both heuristic and learning-based baselines:

- **Heuristic**: A point cloud based non-learning baseline. The segmented point cloud of the target object is analyzed using principal component analysis (PCA). The gripper is oriented along the principal axes, and the grasp is placed at the centroid along the object's vertical axis, at the height of its centroid. This method relies purely on object geometry without tactile feedback or learning, and represents a classical vision-only heuristic approach.
- **Baseline (More Than a Feeling)**: An architecture inspired by Calandra *et al.* [4], which proposed one of the first action-conditioned visuo-tactile grasping models. Similar to their setup, this baseline encodes vision, tactile and action input using Resnets [22] and directly predicts a grasp success probability using a classification head. This baseline captures prior work but does not leverage our pre-trained large-data encoders or future-prediction based representation learning. Hence demonstrates the value in having both those elements.
- **Ours (Encoder Only)**: A multimodal encoder with direct classification. Vision, tactile, and action inputs are encoded and fused,as shown in 3, and the resulting latent representation is fed directly to an MLP classifier for success prediction. Unlike our full model, it does not attempt to predict future sensory states, and thus tests the contribution of future prediction supervision.
- **Encoder-Decoder (Vision only)**: A multimodal encoder-decoder architecture restricted to visual input. The vision stream conditioned on an action vector is encoded,and a TransUNet-style decoder architecture [17] is used to predict future visual states, while a 2-layer MLP predicts grasp success probability. This baseline tests the value of adding tactile sensing by isolating the performance of vision-only predictive modeling.
- **Ours (Full)**: The full action-conditioned encoder-decoder model. Vision, tactile, and action modalities are encoded (modality encodings are added), they are fused through a transformer, and decoded with a TransUNet to predict both future vision and tactile states. A parallel MLP predicts grasp success probability from the latent space. This model leverages multi-modal prediction to learn a rich latent representation of contact dynamics and stability while utilizing the strong priors from pre-trained models.

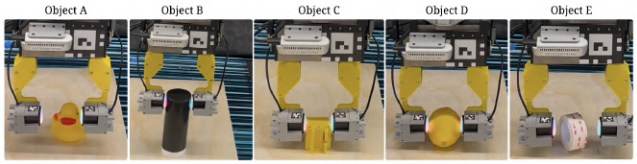

Fig. 6. Unseen test objects. Five objects are used for evaluation, including a deformable object (object A), whereas the training set only contained rigid objects.

B. Real-World Robot Evaluation

We evaluated our approach on five unseen test objects of varying shapes and materials, performing 10 grasp trials per object. A point cloud–based heuristic was first used to bring the gripper into a pre-grasp pose near the target. From this pose, candidate actions were uniformly sampled in the action space: translations of up to $\pm$5,cm along each Cartesian axis (x, y, z), rotations of up to $\pm 17^\circ$ about roll, pitch, and yaw, and gripper width variations of up to $\pm$35,mm. The model then selected and executed the highest-scoring action.

A grasp was marked as *successful* if it was *stable*, defined as the robot lifting the object off the surface and holding it securely in the air for the duration of the trial without slipping or dropping. Grasp *failures* included cases where the object slipped during closure, could not be lifted from the surface, or was dropped during the hold phase.

C. Results

As shown in Fig. I, both the baseline and our model fit the offline training data well and perform similarly in prediction accuracy. However, when deployed on the robot, the results in Tables II and III highlight a key difference: our proposed model transfers more effectively to online testing. It not only achieves higher success rates on objects seen during training,

TABLE II

REAL-WORLD ROBOT GRASP SUCCESS RATES ON SEEN OBJECTS

Model	Object A	Object B	Object C	Object D	Object E	Avg. Success (%)
Heuristic	50%	20%	40%	30%	50%	38%
Baseline	50%	60%	50%	60%	60%	56%
Encoder Only	80%	90%	70%	70%	70%	76%
Encoder-Decoder (Vision only)	80%	80%	70%	60%	80%	74%
Ours (Full Architecture)	100%	100%	90%	80%	90%	92%

TABLE III

REAL-WORLD ROBOT GRASP SUCCESS RATES ON UNSEEN OBJECTS

Model	Object A	Object B	Object C	Object D	Object E	Avg. Success (%)
Heuristic	20%	30%	90%	30%	40%	42%
Baseline	20%	70%	70%	30%	40%	46%
Encoder Only	30%	100%	80%	50%	80%	68%
Encoder-Decoder (Vision only)	30%	100%	70%	70%	80%	70%
Ours (Full Architecture)	50%	100%	100%	70%	100%	84%

but also generalizes robustly to previously unseen objects. This supports our claim that learning a rich multimodal representation space captures a general understanding of physical interactions and enables better reasoning than models trained solely for binary classification.

Compared to the baseline [4], our training setup also leveraged a richer dataset: approximately 20 objects and about 700 grasp transitions per object. Because the dataset was collected with a human-held gripper, the action space was more diverse and realistic than typical robot-collected datasets, capturing a broader range of grasp strategies and corrections. Another important difference is that the baseline modeled gripper state as binary (open or closed), whereas in our formulation the gripper width was explicitly included as part of the action vector and uniformly sampled. This allowed the model to reason over a continuous range of finger configurations, improving its ability to adapt to objects of different sizes and shapes.

We attribute the strong performance and generalization capability not only to this expanded and more varied dataset, but also to the use of expressive visual and tactile encoders, explicit action modeling, and an architecture designed to learn a rich latent representation space. Together, these elements enabled our model to achieve robust transfer and superior online performance.

Furthermore, our approach outperformed the baseline on a deformable test object as shown in Table III (Object A), even though the training set consisted only of rigid objects. This demonstrates the ability of our model to extrapolate beyond its training distribution and reason effectively about novel material properties. Qualitative observations further revealed that the vision+tactile model could reason about local contact conditions and detect slippage events, leading to more accurate grasp success predictions. In contrast, the vision-only variant often misclassified unstable grasps as successful due to the absence of tactile feedback, particularly when relative motion occurred between the object and the gripper fingers.

V. DISCUSSION

Our results demonstrate that using an encoder-decoder architecture trained to have a rich latent space outperforms at grasp success prediction and generalization compared to heuristic methods, prior work–inspired visuo-tactile models, and ablation baselines III. By leveraging a richer human-collected dataset, our approach captures a more diverse action space than typical robot-collected data, enabling the model to reason about complex contact dynamics and nuanced corrective strategies. This, combined with expressive visual and tactile encoders and a predictive encoder–decoder design, allowed our model to achieve robust transfer from offline training to real-world execution.

A key insight from our experiments is that models trained purely as binary classifiers fit the offline training data but fail to generalize effectively to online deployment. In contrast, our forward-predictive architecture learns a richer representation of grasp dynamics, enabling better reasoning across both rigid and deformable objects.

A. Limitations and Future Work

While our model demonstrates strong transfer and generalization, several limitations remain. First, the dataset, though diverse, is still relatively small in scale compared to large-scale vision-only grasping datasets. Expanding the data collection to cover a broader range of objects, including deformables, transparent items, and cluttered scenes, would further strengthen robustness. Second, our current architecture predicts one-step future sensory states. Extending this framework to multi-step prediction or full trajectory forecasting could enable planning longer-horizon manipulations. Finally, the pipeline was implemented with a single-arm setup, future work could explore multi-arm coordination, and integration with reinforcement learning for closed-loop grasp refinement.

Overall, our work shows that action-conditioned multimodal prediction is a promising direction for bridging the gap between offline learning and real-world grasp execution. We believe that combining predictive multimodal architectures with richer datasets and long-horizon reasoning will

enable more adaptive, generalizable, and human-like robotic manipulation in the future.

REFERENCES

[1] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 441–11 450.

[2] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6831–6838.

[3] M. A. Lin, E. Reyes, J. Bohg, and M. R. Cutkosky, "Whisker-inspired tactile sensing for contact localization on robot manipulators," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7817–7824.

[4] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3300–3307. [Online]. Available: https://arxiv.org/abs/1805.11085

[5] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," 2024. [Online]. Available: https://arxiv.org/abs/2402.10329

[6] M. Lambeta, P. Chou, S. Tian, B. H. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *CoRR*, vol. abs/2005.14679, 2020. [Online]. Available: https://arxiv.org/abs/2005.14679

[7] T. Weng, D. Held, F. Meier, and M. Mukadam, "Neural grasp distance fields for robot manipulation," 2022. [Online]. Available: https://arxiv.org/abs/2211.02647

[8] Z. Xu and S. Song, "Giga: Generative implicit grasp affordances," in *Proceedings of Robotics: Science and Systems (RSS)*, 2021. [Online]. Available: https://www.roboticsproceedings.org/rss17/p024.pdf

[9] E. Chisari, N. Heppert, T. Welschehold, W. Burgard, and A. Valada, "Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation," 2023. [Online]. Available: https://arxiv.org/abs/2312.08240

[10] P. Mittal, Y.-C. Cheng, M. Singh, and S. Tulsiani, "Autosdf: Shape priors for 3d completion, reconstruction and generation," 2022. [Online]. Available: https://arxiv.org/abs/2203.09516

[11] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," 2023. [Online]. Available: https://arxiv.org/abs/2303.11328

[12] X. Yi and N. Fazeli, "Vib2move: In-hand object reconfiguration via fingertip micro-vibrations," in *Robotics: Science and Systems (RSS)*, 2025. [Online]. Available: https://roboticsconference.org/program/papers/108/

[13] J. A. Eyzaguirre, M. Oller, and N. Fazeli, "Tactile neural de-rendering," 2024. [Online]. Available: https://arxiv.org/abs/2409.13923

[14] S. Li, S. Rodriguez, Y. Dou, A. Owens, and N. Fazeli, "Tactile functasets: Neural implicit representations of tactile datasets," 2024. [Online]. Available: https://arxiv.org/abs/2409.14592

[15] M. Oller, D. Berenson, and N. Fazeli, "Tactile-driven non-prehensile object manipulation via extrinsic contact mode control," 2024. [Online]. Available: https://arxiv.org/abs/2405.18214

[16] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, S. Levine, and P. Abbeel, "The feeling of success: Does touch sensing help predict grasp outcomes?" 2017. [Online]. Available: https://arxiv.org/abs/1710.05512

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[18] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Bojanowski, J. Verbeek, P. Labatut, and H. Jegou, "Dinov2: Learning robust visual features without supervision," 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[19] Z. Chen, Y. Dou, and A. Owens, "Cttp: A compact tactile transformer for perception," 2024. [Online]. Available: https://arxiv.org/abs/2410.11834

[20] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," 2019. [Online]. Available: https://arxiv.org/abs/1904.02252

[21] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," 2018. [Online]. Available: https://arxiv.org/abs/1803.00628

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385