

DexTact: A Visuo-Tactile Foundation Model for Dexterous Hand Grasping

Xinmin Fang¹, Lingfeng Tao², Zhengxiong Li¹

Abstract—Dexterous hand grasping has advanced rapidly through large-scale datasets and learned policies, yet these approaches predominantly rely on visual or geometric modalities, with tactile sensing remaining largely unexplored for dexterous manipulation, limiting their robustness to visual perturbations. We present **DexTact**, the first visuo-tactile foundation model for dexterous hand grasping. We construct a large-scale dataset of 174,500 grasping demonstrations across nearly 200 objects, 5 environments, and 10 lighting conditions, collected in simulation. Built on this dataset, we train a visuo-tactile-language-action (VTLA) model that jointly processes RGB images, 3D contact forces from all finger joints, and natural language instructions to generate coordinated grasping actions. Experiments demonstrate that **DexTact** achieves 84.2% grasping success on seen objects and 68.6% on unseen objects, while generalizing effectively to novel environments (81.6%) and lighting conditions (75.1%). Ablation studies reveal that vision and tactile modalities serve complementary roles: vision dominates object-level generalization, while tactile sensing provides robustness to visual perturbations, reducing the lighting generalization gap from 16.6% (w/o Tactile) to 4.2% (full model). We further validate favorable scaling behavior following power-law trends ($R^2 \geq 0.999$), with seen-object success improving from 32.3% to 84.2% as data scales from 100 to 174,500 demonstrations; the fitted curves project that each $3\times$ data increase yields approximately 5 percentage points of additional improvement, establishing the continued scalability of visuo-tactile foundation models for dexterous manipulation.

I. INTRODUCTION

Dexterous hand manipulation is a critical capability for achieving general-purpose robotic systems. Compared with parallel-jaw grippers, multi-fingered hands offer significantly greater dexterity, enabling fine-grained grasping and in-hand manipulation of diverse objects [1], [2]. However, the high-dimensional action space and complex contact dynamics of dexterous hands pose substantial challenges for learning robust and generalizable grasping policies.

Recent efforts have made substantial progress in scaling dexterous grasping through large datasets and learned policies. DexGraspNet [2] generates over one million grasp poses across thousands of objects, while UniDexGrasp [1] proposes a universal grasping pipeline that generalizes to novel objects through goal-conditioned reinforcement learning. More recently, DexGraspVLA [3] demonstrates that vision-language-action models can enable language-conditioned dexterous

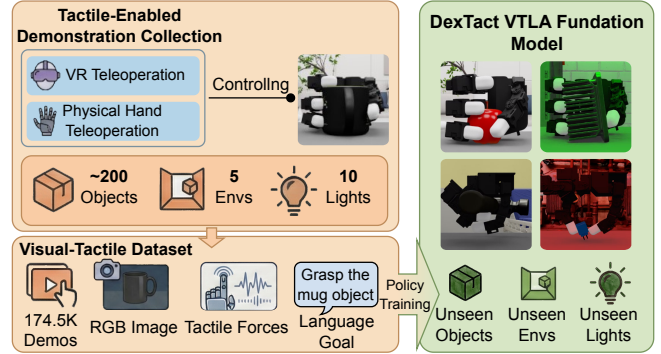


Fig. 1: DexTact integrates tactile sensing with vision and language to enable robust and generalizable dexterous hand grasping across unseen objects and novel environments.

grasping. Despite this progress, existing large-scale approaches rely exclusively on visual or geometric modalities, which introduces inherent limitations for dexterous grasping. Besides, these datasets primarily consist of simple or geometrically regular objects, limiting the generalization to complex, irregular shapes encountered in practice. Vision-only policies are also vulnerable to domain shifts in lighting and environment, as they lack a complementary modality to compensate for visual perturbations. In contrast, human grasping fundamentally depends on the synergistic integration of vision and touch to achieve stable, adaptive manipulation [4].

Although several works have explored visuo-tactile learning for manipulation [5], [6], [7], [8], these efforts have mostly focused on simple parallel-jaw grippers. Introducing tactile sensing to high-degree-of-freedom (DoF) dexterous hands poses new and distinct challenges. First, the high-dimensional action space makes the mapping from tactile signals to coordinated finger actions considerably more complex. Second, multi-finger coordination requires integrating tactile information from multiple fingertips simultaneously, demanding architectures that can effectively process high-dimensional, multi-finger tactile inputs. This raises a question: *Can we effectively incorporate tactile sensing into a foundation model for dexterous hand grasping and achieve practical generalization at scale, spanning diverse objects, environments, and lighting conditions?*

In this paper, we propose **DexTact**, a visuo-tactile foundation model for dexterous hand grasping. As shown in Fig. 1, **DexTact** addresses this gap by incorporating tactile sensing into a foundation model that jointly leverages vision, tactile, and language for dexterous grasping. We first construct a

*This work is partially supported by NSF Awards #2426469, #2426470, #2602731, #2436428, Colorado OEDIT AI Grant Award, and the NVIDIA Academic Grant.

¹X. Fang and Z. Li are with the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217 USA (email: xinmin.fang@ucdenver.edu, zhengxiong.li@ucdenver.edu).

²L. Tao is with the Department of Robotics and Mechatronics Engineering, Kennesaw State University, InDex Lab. (email: ltiao2@kennesaw.edu).

large-scale visuo-tactile dataset by collecting 1,745 unique grasping trajectories across nearly 200 custom objects. These trajectories are expanded across 5 environments, 10 lighting conditions, and 2 camera viewpoints to produce 174,500 demonstrations with diverse visual conditions. Built on this dataset, we train a visuo-tactile-language-action (VTLA) model that jointly processes RGB images, tactile forces, and language instructions to generate dexterous hand grasping actions. Through extensive experiments, we demonstrate that incorporating tactile sensing consistently improves grasping performance over vision-only baselines, particularly under visual domain shifts, achieving 84.2% and 68.6% success rates on seen and unseen objects, respectively. We further discover that VTLA model performance scales predictably with both data volume and visual domain diversity, exhibiting power-law and logarithmic growth patterns analogous to neural scaling laws.

Our contributions are summarized as follows:

- We propose **DexTact**, the first visuo-tactile foundation model for dexterous hand grasping, and systematically demonstrate that the visuo-tactile model outperforms its vision-only counterpart, particularly under visual domain shifts in lighting and environment.
- We discover that VTLA model performance scales predictably with both data volume and visual domain diversity, following power-law and logarithmic growth patterns analogous to scaling laws. The fitted curves project that each $3\times$ data increase yields approximately 5 percentage points of additional improvement at the current scale.
- We construct the first large-scale visuo-tactile dexterous hand grasping dataset, containing 174,500 demonstrations across nearly 200 objects spanning 8 categories, including non-convex and geometrically irregular objects such as tools, electronic components, and mechanical parts, with diverse environments and lighting conditions.
- We design and implement a complete tactile-enabled dexterous hand data collection framework supporting multiple teleoperation methods, enabling efficient collection of multi-modal demonstration data in simulation.
- We conduct extensive experiments demonstrating that **DexTact** exhibits favorable scaling behavior, with seen-object success rising from 32.3% at 100 demos to 84.2% at 174.5K demos, while generalizing to unseen objects (68.6%) and environments (81.6%). Tactile sensing serves as a grounding anchor that is directly coupled to the physical contact state: when visual inputs are degraded by environmental perturbations, tactile signals prevent the policy from drifting, reducing the lighting generalization gap from 16.6% (vision-only) to 4.2%.

II. RELATED WORK

A. Dexterous Hand Manipulation

Dexterous hand manipulation has been studied through both reinforcement learning (RL) and data-driven approaches.

RL-based methods [1], [9], [10] combine grasp proposal generation with goal-conditioned learning and curriculum strategies to achieve generalization across diverse objects. On the data-driven side, large-scale grasp datasets [2], [11], [12], [13] and unified frameworks [14], [3] have enabled training generalizable policies across objects and hand morphologies. However, these approaches rely solely on visual and proprioceptive feedback without tactile sensing, making their grasping policies sensitive to visual domain shifts across environments and lighting conditions. Our work addresses this gap by integrating tactile signals into both the data collection pipeline and the learned policy.

B. Visuo-Tactile Learning for Manipulation

Visuo-tactile learning has been explored across both gripper-based and dexterous hand settings. For grippers, 3D-ViTac [5] combines 3D visual representations with tactile sensing for fine-grained manipulation, MimicTouch [6] leverages human-inspired vision-touch sensing for contact-rich tasks, and UniVTAC [15] provides a unified simulation platform for visuo-tactile learning and benchmarking. For dexterous hands, T-Dex [7] proposes touch-based representations for in-hand manipulation, Some [8], [16] integrates tactile feedback into hand-arm teleoperation, PP-Tac [17] explores tactile policy learning on dexterous platforms, and Ye et al. [18] combine visual-tactile pretraining with online multitask learning. While these works validate the utility of tactile feedback, they typically involve few objects and limited demonstrations, leaving the scalability of visuo-tactile approaches largely unexplored. **DexTact** addresses this gap by constructing a large-scale visuo-tactile dataset spanning nearly 200 objects and training a generalizable foundation model for dexterous grasping.

C. Foundation Models for Robot Manipulation

Foundation models have emerged as a promising paradigm for robot manipulation by leveraging large-scale data and pretrained representations. Vision-language-action (VLA) models [19], [20], [21] demonstrate that training on large-scale, multi-task robot data enables generalist control, with DexGraspVLA [3] extending this paradigm to dexterous grasping. However, existing foundation models for robot manipulation rely exclusively on visual and language modalities, without incorporating tactile sensing. This renders their policies sensitive to visual domain shifts in lighting and environment, and leaves them without direct contact feedback to assess grasp stability during manipulation. **DexTact** addresses this gap by combining large-scale tactile data collection, multi-modal fusion, and generative action modeling in a unified framework for dexterous hand grasping.

III. SYSTEM DESIGN

This section presents the design of **DexTact**, a visuo-tactile foundation model for dexterous hand grasping. We first describe our tactile-enabled demonstration collection framework (Section III-A), then introduce the large-scale visuo-tactile dataset (Section III-B), and finally detail the model

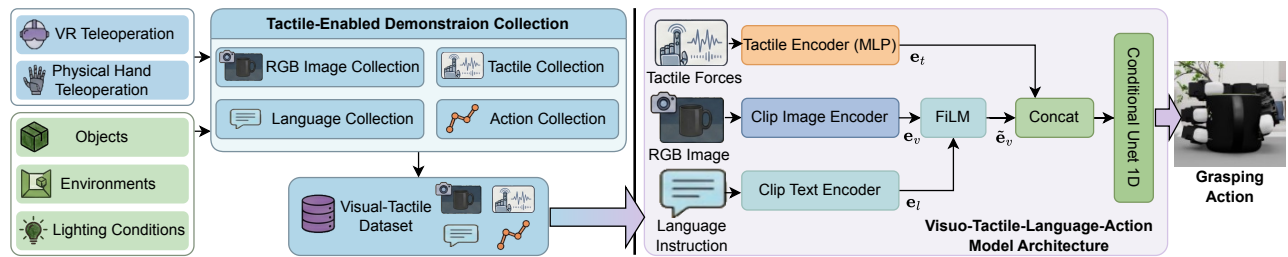


Fig. 2: Overview of DexTact. Left: visuo-tactile data collection via VR and physical hand teleoperation. Right: the VTLA model fuses visual, tactile, and language inputs to generate grasping actions.

architecture (Section III-C). Fig. 2 illustrates the overall pipeline from data collection to model inference.

A. Tactile-Enabled Demonstration Collection

We design a demonstration collection framework that captures synchronized visual, tactile, and proprioceptive data during grasping with the LEAP Hand [22], a 16-DoF anthropomorphic hand, in NVIDIA Isaac Sim [23].

Teleoperation Methods. To collect diverse and high-quality grasping demonstrations, we implement two complementary teleoperation approaches:

VR-based teleoperation. The first method uses Meta Quest 3 headsets to track the operator’s hand pose in real time. The tracked hand skeleton is mapped to the simulated LEAP Hand through a joint-angle-based retargeting algorithm that converts changes in the operator’s finger joint angles to corresponding joint commands on the robotic hand. Flexion and abduction angles of the thumb, index, middle, and ring fingers are mapped to the corresponding LEAP joints with per-joint scaling and limit clamping; the human little finger is unused. This approach enables rapid demonstration collection for simple objects, as operators can intuitively grasp objects using natural hand motions.

Physical hand teleoperation. The second method uses a physical LEAP Hand as the input device, directly mapping its joint angles to the simulated hand. This approach provides higher fidelity for demonstrations that require precise multi-finger coordination, such as grasping thin or irregularly shaped objects.

Tactile Data Representation. During each demonstration, we record 3D contact force vectors at each of the 16 body parts of the LEAP Hand. Each force vector consists of three components (F_x, F_y, F_z), yielding a 48-dimensional tactile observation at every timestep. These contact forces are computed by the physics engine and reflect the interaction between each finger link and the grasped object. The forces are visualized to the operator in real time during teleoperation; validating these simulated force profiles against real tactile sensors remains future work. The tactile signal provides complementary information to vision by encoding contact location, force magnitude, and force direction across the hand.

Data Format. Each demonstration timestep comprises: (1) an RGB image captured from a fixed third-person camera, (2) a 48-dimensional tactile vector, (3) a natural language

instruction describing the target object (e.g., “grasp the mug object”), and (4) a 16-dimensional action vector representing the target joint angles. The action space directly corresponds to the position commands for all 16 joints.

B. Visuo-Tactile Grasping Dataset

Using the collection framework described above, we construct a large-scale visuo-tactile dexterous hand grasping dataset. To our knowledge, this is the first dataset that combines tactile sensing with dexterous hand manipulation at this scale. As shown in Fig. 2, the LEAP Hand is mounted on a platform with the target object placed in front. Depending on the object geometry, the hand is oriented either sideways for upright objects or palm-down for flat-lying objects. All scenes are rendered across diverse background environments and lighting conditions to promote generalization to novel visual settings.

Scale. The dataset is built from 1,745 unique grasping trajectories collected across 188 unique objects using both teleoperation methods described above. Each trajectory is then expanded across all 5 environments, 10 lighting conditions, and 2 camera viewpoints, yielding 174,500 demonstrations. While the visual conditions vary across demonstrations, the underlying grasp trajectories (joint actions and tactile forces) remain identical within each expansion group; only the RGB observations and environment context change. Each demonstration captures a complete grasp sequence with synchronized visual, tactile, and action data at every timestep, where each timestep corresponds to a 0.05 s simulation interval. In total, the dataset contains over 16.8 million timesteps, corresponding to approximately 234 hours of interaction data. The average demonstration length is 97 timesteps (4.8 s), while individual demonstrations range from 37 to 669 timesteps depending on object complexity and grasp strategy.

Object Diversity. The 188 rigid objects, generated using Hunyuan3D 2.0 [24], span 8 categories: hand tools (26), power tools (16), electronics (28), mechanical (26), medical (24), containers (23), office (17), and accessories (28). These categories encompass objects with diverse geometric and physical properties, from small components such as capacitors and compression springs to large items such as power supplies and VR headsets. Fig. 3 shows representative objects from each category, and Fig. 4 summarizes the demonstration count and total demonstration length per category.



Fig. 3: Representative objects from each of the eight categories in the DexTact dataset. The objects span a wide range of geometric and physical properties.

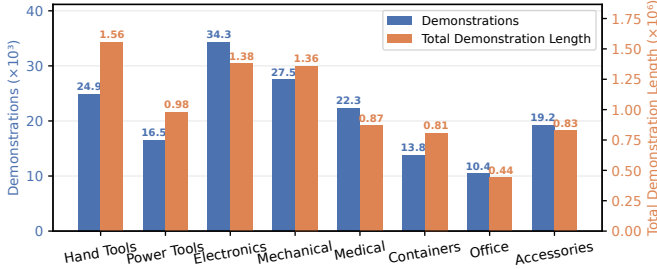


Fig. 4: Distribution of demonstrations and total demonstration lengths across object categories.

Environmental Diversity. To promote visual generalization, demonstrations are collected across 5 distinct simulated environments (factory, hospital, living room, office, and warehouse), each featuring different background scenes and spatial layouts. We further augment visual diversity through 10 lighting conditions: five levels of white illumination ranging from bright to dim, and five colored lights (red, green, blue, yellow, and cyan). Each of the demonstration trajectories is expanded across all 5 environments and 10 lighting conditions, yielding 50 visual variants per demonstration. Combined with two camera viewpoints, this expansion produces the full set of 174,500 training demonstrations. This systematic environmental diversity encourages the policy to generalize across visual domain shifts rather than overfitting to specific scene configuration.

C. Visuo-Tactile-Language-Action Model Architecture

We design a multimodal architecture that fuses visual, tactile, and language inputs to generate dexterous grasping actions. The model comprises four components: modality-specific encoders, a conditioning mechanism, an action generation backbone, and a flow matching training objective. The model totals 379M parameters, comprising 151M frozen CLIP ViT-B/32 [25] weights and 228M trainable parameters in the tactile encoder (0.05M), FiLM layers (1.6M), and ConditionalUNet1D backbone (226M). Fig. 2(B) illustrates the model design.

Tactile Encoder. Encoding tactile signals from dexterous hands presents distinct challenges compared with gripper-

based settings. The sensing is distributed across 16 body parts with heterogeneous roles, where fingertips provide primary contact feedback while proximal joints convey force transmission patterns. Moreover, contact signals are inherently sparse during the approach phase, with substantial forces arising only upon object contact. We process the 48-dimensional tactile observation through a two-layer MLP (256 and 128 units with ReLU activations), producing a compact 128-dimensional tactile embedding $\mathbf{e}_t$. The lower dimensionality relative to the 512-dimensional CLIP visual embeddings reflects the complementary nature of tactile feedback: rather than providing object-level recognition, tactile sensing contributes contact-level information that enhances grasp stability, particularly under visual domain shifts such as lighting variations.

Visual and Language Encoders. We use frozen pretrained CLIP ViT-B/32 [25] encoders for both visual and language modalities. The CLIP image encoder processes the RGB observation into a visual embedding $\mathbf{e}_v$, while the CLIP text encoder converts the natural language instruction into a language embedding $\mathbf{e}_l$. Using a pretrained vision-language model provides strong semantic representations and enables the policy to associate visual object appearances with language descriptions without task-specific pretraining.

Multimodal Fusion via FiLM. We employ Feature-wise Linear Modulation (FiLM) [26] for cross-modal conditioning. The language embedding $\mathbf{e}_l$ is projected through an MLP to produce scale (γ) and shift (β) parameters that modulate the visual features via $\tilde{\mathbf{e}}_v = \gamma \odot \mathbf{e}_v + \beta$. The modulated visual features $\tilde{\mathbf{e}}_v$ are then concatenated with the tactile embedding $\mathbf{e}_t$ to form the global conditioning vector $\mathbf{c}$ for action generation.

Action Generation with Flow Matching. The action generation backbone is a ConditionalUNet1D architecture adapted from Diffusion Policy [27] with channel dimensions (384, 768, 1536). We train this backbone using flow matching [28], which learns a continuous velocity field to transport samples from a Gaussian prior to the data distribution. Compared with denoising diffusion, flow matching offers straighter transport paths and requires fewer inference steps. At each decision step, the model conditions on an observation window of the 6 most recent timesteps, each comprising an RGB image and a tactile vector, to capture the temporal evolution of both visual and contact states during the grasp. The model predicts action chunks of 16 future timesteps; during inference, we integrate the velocity field over 16 steps using a first-order Euler ordinary differential equation (ODE) solver and execute the first 8 predicted actions before replanning with updated observations. Although the observation window spans only 0.3 s, replanning every 8 steps allows the policy to reassess the contact state continuously over the full grasp.

IV. EXPERIMENTS

We evaluate DexTact on simulated dexterous grasping tasks, focusing on generalization across objects and environments, scalability, and the contribution of tactile sensing.

A. Experimental Setup

All experiments are conducted in NVIDIA Isaac Sim [23] using the LEAP Hand [22] with 16 DoF. The hand operates in joint position control mode. As during data collection, the global hand pose and wrist orientation are pre-set according to object geometry, and the policy controls the 16 finger joints. We train all models using the AdamW optimizer with a learning rate of 1×10^{-4} , weight decay of 0.01, batch size of 128, and cosine annealing schedule for 300 epochs.

B. Objects

We evaluate DexTact on 70 objects divided into 50 seen objects and 20 unseen objects. The 50 seen objects are drawn from the training dataset, while the 20 unseen objects are newly collected and excluded from the 174,500 training demonstrations, enabling evaluation of zero-shot generalization to entirely novel objects. Both sets span six categories to ensure consistent evaluation across object types: (1) **Tools**, including drills, pliers, hammers, and wrenches; (2) **Electronics**, such as circuit boards, motors, transformers, and controllers; (3) **Containers**, including cans, mugs, beakers, and spray bottles; (4) **Mechanical**, such as pulleys, gears, and valve wheels; (5) **Household**, including everyday objects like fruits, flashlights, and staplers; and (6) **Instruments**, such as multimeters, cameras, and medical syringes. For both seen and unseen objects, the language instruction follows the same template and names the target object.

C. Environments and Lighting

We evaluate generalization across diverse visual conditions by selecting a subset of environments and lighting conditions from the full dataset and partitioning them into seen and unseen sets.

Environments. The seen environment set includes four distinct scenes: **warehouse**, **factory**, **hospital**, and **living room**, each featuring unique table surfaces, background textures, and spatial configurations. The unseen set comprises **bedroom** and **corridor**, which feature distinct background textures and spatial layouts not encountered during training, testing the policy’s robustness to novel visual scenes.

Lighting Conditions. The seen lighting set consists of five conditions: **bright white**, **dim white**, **pure red**, **pure blue**, and **pure green**. The unseen lighting set includes three novel color combinations not present during training: **magenta**, **dark magenta**, and **yellow-green**. This partition ensures that the unseen conditions require interpolation or extrapolation beyond the training distribution.

D. Evaluation Metrics

We evaluate performance using the following metrics: (1) **Success Rate.** After the policy executes the grasp, the supporting platform is lowered and a 2-second hold period begins. A grasp is considered successful if the object remains in the hand without falling throughout this period. (2) **Generalization Gap.** We define the generalization gap as the difference in success rate between seen and unseen conditions (objects or environments). (3) **Contact Stability.**

This metric measures the consistency of contact forces during the grasp phase. We compute the standard deviation of the total contact force magnitude over the grasp duration; lower values indicate more stable grasps. (4) **Completion Steps.** The number of action steps required from the initial approach to a successful stable grasp. Fewer steps indicate more efficient grasping behavior.

V. RESULTS

A. Overall Performance

We evaluate DexTact trained on 174,500 demonstrations across multiple generalization dimensions: seen versus unseen objects, environments, and lighting conditions.

Object Generalization. We assess whether the learned policy generalizes across object categories, including objects unseen during training. Table I(a) presents the per-category success rates for both seen and unseen objects. On seen objects, DexTact achieves a strong overall success rate of 84.2%. For context, UniDexGrasp++ [9] reports 78.2% on novel objects under fixed visual conditions; since the protocols and object sets differ, these numbers are not directly comparable. Instruments and Containers achieve the highest success rates (93.8% and 93.4%, respectively), as these categories predominantly contain objects with regular cylindrical or compact geometries that facilitate stable multi-finger wrapping. Household and Mechanical follow with strong performance (87.5% and 85.7%), while Electronics exhibits the lowest success rate (79.8%), as this category contains thin circuit boards and small components that provide limited contact surfaces for multi-finger grasping.

For unseen objects, the model achieves a 68.6% overall success rate without any fine-tuning, resulting in a generalization gap of only 15.6%. The generalization gap varies across categories: Tools and Mechanical exhibit the smallest gaps (9.9% and 13.7%), suggesting that the distinct geometric features of tools and mechanical components transfer effectively to novel instances within these categories. In contrast, Household and Electronics show larger gaps (22.2% and 18.9%), indicating that the greater intra-category diversity and irregular geometries within these groups present challenges for zero-shot generalization.

Beyond functional categories, we further analyze generalization along geometric properties by partitioning the evaluation objects into convex and non-convex groups based on their surface geometry. As reported in Table I(a), convex objects achieve higher success rates on both seen (84.8%) and unseen (73.7%) splits compared with non-convex objects (82.0% and 64.9%, respectively). The generalization gap for convex objects (11.1%) is notably smaller than that for non-convex objects (17.1%), suggesting that convex geometries afford more predictable contact surfaces that enable the policy to transfer grasping strategies more readily to novel instances. Non-convex objects, with their concavities and irregular contact regions, demand more precise finger placement, which presents greater challenges for zero-shot generalization.

TABLE I: Overall generalization performance. Success rates (%) with 50 trials per (object, environment, lighting) combination. (a) By object category with seen environments and lighting. (b) By environment with seen objects and lighting. (c) By lighting condition with seen objects and environments.

(a) Object Category			(b) Environment			(c) Lighting Condition		
Category	Seen	Unseen	Split	Environment	Succ. Rate	Split	Lighting	Succ. Rate
Tools	82.9	73.0	<i>Seen</i>	Warehouse	83.8	<i>Seen</i>	Bright White	82.4
Electronics	79.8	60.9		Factory	82.2		Dim White	76.3
Containers	93.4	74.7		Hospital	82.4		Pure Red	76.4
Mechanical	85.7	72.0		Living Room	77.0		Pure Green	77.2
Household	87.5	65.3		Average	80.3		Pure Blue	77.5
Instruments	93.8	75.0	<i>Unseen</i>	Average	80.3	<i>Unseen</i>	Average	79.3
Convex	84.8	73.7		Bedroom	82.2		Magenta	76.8
Non-Convex	82.0	64.9		Corridor	80.9		Dark Magenta	74.9
Overall	84.2	68.6		Average	81.6		Yellow-Green	73.6
							Average	75.1

Environment Generalization. We evaluate the policy’s robustness to different scene backgrounds and spatial configurations. Table I(b) reports per-environment success rates across four seen environments and two unseen environments. Among the seen environments, warehouse achieves the highest success rate (83.8%), followed by hospital (82.4%), factory (82.2%), and living room (77.0%). The variance across seen environments remains modest, indicating that the policy does not overfit to any single scene configuration.

For the unseen environments, bedroom and corridor achieve 82.2% and 80.9%, respectively. These environments feature different background textures compared with the training set, yet the model maintains an average unseen success rate of 81.6%, exceeding the seen average of 80.3%. This result suggests that training across diverse environmental configurations enables the model to learn background-invariant representations.

Lighting Generalization. We assess the model’s robustness to varying illumination conditions, including novel color combinations absent from training. Table I(c) reports success rates across individual lighting conditions. Performance on seen lighting conditions ranges from 76.3% to 82.4%, with bright white achieving the highest success rate. Pure color lighting (red, green, blue) presents greater challenges due to the reduced visual information compared to white light.

On unseen lighting conditions, magenta achieves the highest unseen success rate (76.8%), as it represents a combination of training colors (red and blue). Yellow-green shows the lowest unseen performance (73.6%), likely because it introduces a color component distant from the pure training colors. Dark magenta achieves 74.9%, where the combination of novel color mixing and reduced intensity compounds the challenge. The overall generalization gap between seen and unseen lighting (4.2%) remains modest, indicating that the model learns lighting-invariant representations rather than overfitting to specific illumination patterns.

B. Scaling Analysis

We investigate how DexTact performance scales along two axes: the number of demonstrations and the diversity of

visual domains. Table II (top section) reports success rates across all generalization dimensions.

Data Scaling. We train models on 100, 10,000, and 174,500 demonstrations while fixing all 100 visual domain configurations. Performance across all six metrics follows a Chinchilla-style scaling law of the form $y = A - B \cdot N^{-\alpha}$, with $R^2 \geq 0.999$ in all cases, indicating highly predictable scaling behavior analogous to that observed in large language models. Seen-object success rises from 32.3% to 67.6% to 84.2%, while unseen-object performance scales most steeply (22.3% to 43.0% to 68.6%), narrowing the generalization gap from 24.6% at 10K to 15.6% at 174.5K. The fitted asymptotic bounds ($A \in [78.8, 156.9]$) confirm that performance has not saturated, indicating significant room for further improvement with additional data. Extrapolating the fitted curves, a $3\times$ increase in data from the current scale is projected to yield approximately 5 percentage points of additional seen-object improvement, while a $10\times$ expansion would yield roughly 11 percentage points. Since unseen-object performance exhibits the steepest scaling exponent, the generalization gap is expected to narrow even faster, suggesting that the current dataset size remains far from the regime of diminishing returns.

Domain Scaling. We vary the number of visual domain configurations (environment–lighting combinations) from 5 to 50 to 100 while fixing the total demonstrations at 10,000. Performance follows a logarithmic growth pattern ($R^2 > 0.97$), with the strongest effect on lighting generalization: unseen lighting performance increases from 54.7% to 67.8% (13.1% gain), while unseen environment performance improves from 62.3% to 70.7% (8.4% gain). The marginal returns diminish notably from 50 to 100 domains compared with 5 to 50, suggesting that moderate domain diversity captures the majority of the generalization benefit. Extrapolating the logarithmic fits, each subsequent doubling of domain configurations is projected to yield a constant absolute improvement of approximately 2–3% in unseen lighting performance, indicating that domain diversity, in contrast to data volume, approaches diminishing returns more rapidly.

TABLE II: Scaling analysis and modality ablation results. 50 trials per (object, env., lighting) combination. Success rates (%) across seen (S) and unseen (U) splits for objects, environments, and lighting conditions.

Setting	Config.	Object		Env.		Lighting	
		S	U	S	U	S	U
Data Scale	100 demos	32.3	22.3	33.8	44.4	31.5	15.2
	10K demos	67.6	43.0	68.2	66.7	66.3	67.8
	174.5K demos	84.2	68.6	80.3	81.6	79.3	75.1
Domain Scale	5 domains	63.4	43.4	67.0	62.3	61.2	54.7
	50 domains	66.9	46.1	66.7	67.6	64.4	59.2
	100 domains	67.6	47.7	68.2	70.7	66.3	67.8
Modality (10K)	Full VT	67.6	43.0	68.2	66.7	66.3	67.8
	w/o Tactile	66.6	39.8	61.3	56.4	62.2	42.1
	w/o RGB	61.8	25.7	58.1	58.2	57.4	50.5
Modality (70K)	Full VT	74.9	65.3	77.2	76.7	74.0	72.4
	w/o Tactile	73.5	63.6	72.9	69.8	71.5	54.9
	w/o RGB	64.9	38.0	66.0	66.6	64.3	62.1

C. Ablation Study

To understand the contribution of each sensory modality, we conduct ablation experiments comparing the full visuo-tactile model against two variants: without tactile input (w/o Tactile) and without RGB input (w/o RGB), at both 10,000 and 70,000 demonstrations.

Table II (bottom section) presents the modality ablation results at two data scales. The full visuo-tactile model consistently achieves the highest performance across all metrics and data scales. At 70K, the full model attains 74.9% on seen objects, exceeding w/o Tactile and w/o RGB by 1.4% and 10.0%, respectively. The advantage is most pronounced under unseen lighting, where the full model (72.4% at 70K, 67.8% at 10K) substantially surpasses both single-modality variants, indicating that multimodal fusion achieves robustness that neither modality provides alone.

Examining the single-modality variants reveals a clear division of roles between the two modalities. For object generalization, vision serves as the dominant modality: at 70K, w/o Tactile outperforms w/o RGB on both seen objects (73.5% vs. 64.9%) and unseen objects (63.6% vs. 38.0%), confirming that visual input is essential for object recognition and grasp planning. For lighting robustness, tactile sensing becomes the critical modality: w/o RGB achieves 62.1% under unseen lighting compared with 54.9% for w/o Tactile, as tactile feedback is inherently invariant to visual changes. This invariance is further evidenced by the seen-to-unseen lighting drop: w/o Tactile suffers a 16.6% decrease (71.5% to 54.9%), while w/o RGB shows only a 2.2% decrease (64.3% to 62.1%). The same trends hold at 10K demonstrations, where w/o Tactile (66.6%) again surpasses w/o RGB (61.8%) on seen objects, while w/o RGB (50.5%) exceeds w/o Tactile (42.1%) under unseen lighting. These consistent patterns across data scales confirm the complementary nature of the two modalities: vision provides spatial awareness and object recognition, while tactile feedback contributes grasp stability

and robustness to visual domain shifts.

Beyond success rate, the full visuo-tactile model also achieves the lowest contact force variance (1.16 at 70K, 1.30 at 10K) and the fewest completion steps (304.81 at 70K, 324.91 at 10K) compared with both single-modality variants, where w/o RGB consistently exhibits the highest instability (1.45, 1.75) and slowest execution (310.27, 330.21). These results indicate that multimodal fusion not only improves grasp success but also yields more stable contact dynamics and more efficient motion planning, further supporting the complementary benefits of combining vision and tactile.

VI. DISCUSSION AND KEY TAKEAWAYS

Takeaway 1: Tactile sensing serves as a lightweight yet effective grounding signal. The tactile encoder adds only 0.05M parameters to a 379M-parameter model, yet it reduces the lighting generalization gap from 16.6% (w/o Tactile) to 4.2% (full model). This disproportionate impact arises because tactile feedback is inherently invariant to visual perturbations: contact forces reflect the physical grasp state regardless of scene appearance. Consequently, when lighting shifts degrade visual features, the tactile channel prevents the policy from drifting. This result suggests that even a minimal tactile encoder can substantially improve robustness in visuo-linguistic manipulation policies, without requiring large-scale cross-modal pretraining or billion-parameter architectures.

Takeaway 2: VTLa scaling follows predictable power-law trends with a memorization-to-generalization phase transition. Data scaling exhibits Chinchilla-style power-law behavior ($R^2 \geq 0.999$) across all six evaluation metrics. Notably, the generalization gap follows a non-monotonic trajectory: it starts small at 100 demonstrations (10.0%, uniformly low performance), peaks at 10K (24.6%, object-specific memorization), and then narrows at 174.5K (15.6%, transferable primitives). This trajectory suggests a phase transition beyond a critical data threshold, where the policy shifts from memorizing object-specific strategies to learning generalizable manipulation primitives. The fitted asymptotic bounds confirm that performance has not saturated at the current scale, projecting approximately 5 percentage points of additional improvement per $3\times$ data increase.

Takeaway 3: Tactile benefits are orthogonal to model scale. DexTact achieves 84.2% on seen objects and 68.6% on unseen objects with a 379M-parameter architecture, which is an order of magnitude smaller than prevalent VLA models such as π_0 [29] (3.3B) and DexGraspVLA [3] (10B). While these larger models benefit from greater representational capacity, the tactile modality addresses a fundamentally different limitation: sensitivity to visual domain shifts. Our ablation confirms that tactile integration reduces the lighting generalization gap regardless of data scale (at both 10K and 70K demonstrations), suggesting that tactile grounding and model scaling improve performance along complementary axes. This orthogonality implies that incorporating tactile sensing into larger VLA architectures could yield compounding benefits.

Takeaway 4: Visual complexity matters more than training familiarity. The seen living room environment (77.0%) underperforms the unseen bedroom (82.2%) and corridor (80.9%), indicating that the policy is more sensitive to background texture complexity than to whether a scene was encountered during training. Visually cluttered backgrounds appear to hinder object localization regardless of training exposure. This finding carries a practical implication for data collection: prioritizing texture and layout diversity during domain randomization may yield greater generalization benefits than simply increasing the number of training environments.

VII. CONCLUSION

We presented **DexTact**, a visuo-tactile foundation model for dexterous hand grasping that integrates visual, tactile, and language modalities at scale. We constructed the first large-scale visuo-tactile dexterous grasping dataset comprising 174,500 demonstrations across nearly 200 objects, 5 environments, and 10 lighting conditions, collected through a flexible teleoperation framework supporting both VR-based and physical hand interfaces. Our model exhibits favorable scaling behavior, with seen-object success rising from 32.3% at 100 demos to 84.2% at 174.5K demos, while generalizing to unseen objects (68.6%) and environments (81.6%). Notably, tactile sensing reduces the lighting generalization gap from 16.6% (w/o Tactile) to 4.2%, confirming the complementary role in bridging visual domain shifts. Future work will focus on sim-to-real transfer to validate **DexTact** on a physical LEAP Hand platform.

REFERENCES

- [1] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, T. Liu, L. Yi, and H. Wang, “UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [2] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, “DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 359–11 366.
- [3] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, Y. Yang, and Y. Chen, “DexGraspVLA: A vision-language-action framework towards general dexterous grasping,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.
- [4] R. S. Johansson and J. R. Flanagan, “Coding and use of tactile signals from the fingertips in object manipulation tasks,” *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.
- [5] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3D-ViTact: Learning fine-grained manipulation with visuo-tactile sensing,” in *Conference on Robot Learning (CoRL)*, 2024, pp. 5799–5816.
- [6] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, “MimicTouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation,” in *Conference on Robot Learning (CoRL)*, 2024, pp. 3153–3172.
- [7] I. Guzey, B. Evans, S. Chintala, and L. Pinto, “Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play,” *arXiv preprint arXiv:2303.12076*, 2023.
- [8] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, “Learning visuotactile skills with two multifingered hands,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [9] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, and H. Wang, “UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [10] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, A. Gupta, and D. Fox, “Learning robust real-world dexterous grasping policies via implicit shape augmentation,” in *Conference on Robot Learning (CoRL)*, 2023.
- [11] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang, “DexGraspNet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes,” in *Conference on Robot Learning (CoRL)*, 2024, pp. 5768–5783.
- [12] J. Ye, K. Wang, C. Yuan, R. Yang, Y. Li, J. Zhu, Y. Qin, X. Zou, and X. Wang, “Dex1B: Learning with 1b demonstrations for dexterous manipulation,” in *Robotics: Science and Systems (RSS)*, 2025.
- [13] Y. Zhong, Q. Jiang, J. Yu, and Y. Ma, “DexGrasp Anything: Towards universal robotic dexterous grasping with physics awareness,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 584–22 594.
- [14] H.-S. Fang, H. Yan, Z. Tang, H. Fang, C. Wang, and C. Lu, “Any-DexGrasp: General dexterous grasping for different hands with human-level learning efficiency,” *arXiv preprint arXiv:2502.16420*, 2025.
- [15] B. Chen, W. Wan, T. Chen, X. Guo, C. Xu, Y. Qi, H. Zhang, L. Wu, T. Xu, Z. Li, Y. Wu, R. Li, X. Yang, P. Luo, W. Sui, and Y. Mu, “UniVTAC: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking,” *arXiv preprint arXiv:2602.10093*, 2026.
- [16] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang, et al., “World action models are zero-shot policies,” *arXiv preprint arXiv:2602.15922*, 2026.
- [17] P. Lin, Y. Huang, J. Ma, C. Xiao, W. Li, and Z. Jiao, “PP-Tac: Paper picking using omnidirectional tactile feedback in dexterous robotic hands,” in *Robotics: Science and Systems (RSS)*, 2025.
- [18] Q. Ye, Q. Liu, S. Wang, J. Chen, Y. Cui, K. Jin, H. Chen, X. Cai, G. Li, and J. Chen, “Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity,” *Science Robotics*, vol. 11, no. 110, p. eady2869, 2026.
- [19] A. Brohan, N. Brown, J. Carbajal, et al., “RT-1: Robotics transformer for real-world control at scale,” in *Robotics: Science and Systems (RSS)*, 2023.
- [20] —, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conference on Robot Learning (CoRL)*, 2023, pp. 2165–2183.
- [21] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. 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,” in *Conference on Robot Learning (CoRL)*, 2024, pp. 2679–2713.
- [22] K. Shaw, A. Agarwal, and D. Pathak, “LEAP Hand: Low-cost, efficient, and anthropomorphic hand for robot learning,” in *Robotics: Science and Systems (RSS)*, 2023.
- [23] NVIDIA, “NVIDIA Isaac Sim,” <https://developer.nvidia.com/isaac/sim>, 2024.
- [24] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, et al., “Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation,” *arXiv preprint arXiv:2501.12202*, 2025.
- [25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021.
- [26] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [27] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” in *Robotics: Science and Systems (RSS)*, 2023.
- [28] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [29] K. Black, N. Brown, D. Driess, et al., “ π_0 : A vision-language-action flow model for general robot control,” in *Robotics: Science and Systems (RSS)*, 2025.