Phy-Tac: Toward Human-Like Grasping via *Phy*sics-Conditioned *Tac*tile Goals

***Abstract*—Humans naturally grasp objects with minimal level required force for stability, whereas robots often rely on rigid, over-squeezing control. To narrow this gap, we propose a human-inspired physics-conditioned tactile method (Phy-Tac) for force-optimal stable grasping (FOSG) that unifies pose selection, tactile prediction, and force regulation. A physics-based pose selector first identifies feasible contact regions with optimal force distribution based on surface geometry. Then, a physics-conditioned latent diffusion model (Phy-LDM) predicts the tactile imprint under FOSG target. Last, a latent-space LQR controller drives the gripper toward this tactile imprint with minimal actuation, preventing unnecessary compression. Trained on a physics-conditioned tactile dataset covering diverse objects and contact conditions, the proposed Phy-LDM achieves superior tactile prediction accuracy, while the Phy-Tac outperforms fixed-force and GraspNet-based baselines in grasp stability and force efficiency. Experiments on classical robotic platforms demonstrate force-efficient and adaptive manipulation that bridges the gap between robotic and human grasping. The dataset and code will be open-source.**

***Index Terms*—Grasp stability, tactile sensing, diffusion model, LQR control.**

I. Introduction

Humans naturally achieve stable grasping through active tactile regulation that applies only the necessary amount of force to maintain stability while avoiding object damage [1]. In contrast, robotic grasping still relies heavily on rigid control or over-squeezing strategies, which compromise safety, energy efficiency, and adaptability, especially when interacting with fragile or compliant materials [2]. Therefore, achieving force-optimal stable grasping (FOSG), where the grasp remains stable with the minimal necessary contact force, remains a fundamental challenge for robotic manipulation.

Most existing approaches treat grasping pose planner and force regulation as separate stages. Pose planners like [3] usually optimize geometric force-closure or heuristic confidence, without reasoning about whether the chosen contacts can sustain an optimal force distribution. On the other hand, force regulations often intervene only after slip occurs or just provide a fixed contact force rather than proactively regulating contact force toward an optimal regime [4]. Consequently, robots lack a unified mechanism that jointly reasons over contact planning and contact force optimization within a single loop, thereby preventing globally consistent force optimization and perception-control coordination.

However, humans achieve force-optimal grasping through an anticipatory and feedback-rich process as shown in Fig. 1-a. They first select geometrically favorable contact regions (*S1*), estimate the minimal required force from prior experience (*S2*), and then refine it using fingertip tactile feedback (*S3*). Normally, humans form *S2* and *S2* as closed loops to directionally regulate force in time for contacting state changes to avoid

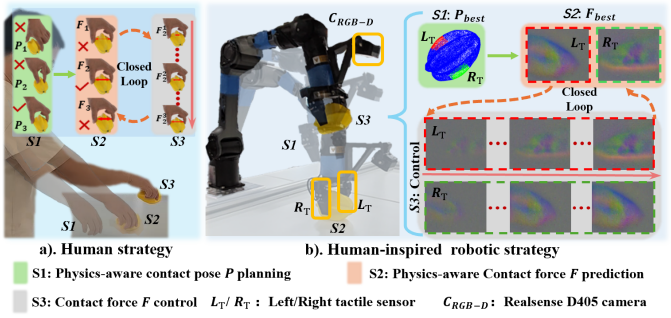

Fig. 1. Comparison of human grasping strategy (a) and our Phy-Tac framework (b) for FOSG. Humans first select an optimal contact pose (*S1*), estimate the required grasping force (*S2*), and refine it to a just-enough level to maintain stability (*S3*). Likewise, Phy-Tac unifies pose planning (*S1*), tactile state prediction (*S2*), and force regulation (*S3*) to achieve the same principle, where *S2* and *S3* form a closed loop for optimal force regulation.

sliding occurs. Inspired by this principle, our work shown in Fig. 1-b aims to endow robots with a similar capability to select grasp pose, estimate optimal force, and regulate contact force proactively rather than reactively, thereby bridging the gap between human dexterity and robotic manipulation.

To this end, we propose a human-inspired Phy-Tac for FOSG that integrates grasping pose planning, tactile prediction, and grasping force regulation within a unified tactile–control pipeline as shown in Fig. 2. The key idea is to predict the tactile imprint corresponding to a FOSG and to drive the gripper toward this tactile imprint with minimal actuation. Specifically, a physics-conditioned contact selector identifies grasp poses with optimal force distribution, a physics-conditioned latent diffusion model (Phy-LDM) predicts the optimal tactile state, and a latent-space LQR controller efficiently converges to that state. Our main contributions are summarized as follows:

- We propose a human-like Phy-Tac, which unifies grasp planning, tactile estimation, and force regulation into a single closed loop centered on a force-optimal tactile goal, enabling proactive and minimal-force grasping rather than reactive stabilization.
- We develop a physics-conditioned Phy-LDM that predicts the tactile imprint corresponding to optimal stable force, and align it through a LQR servo to achieve FOSG.
- We provide a physics-conditioned tactile dataset that covers objects' physical attributes and tactile features, providing a foundation for physically consistent tactile study.

II. Related Work

Grasp pose planning. Recent end-to-end methods [3, 5, 6] leverage large-scale datasets to train grasping policy satisfying

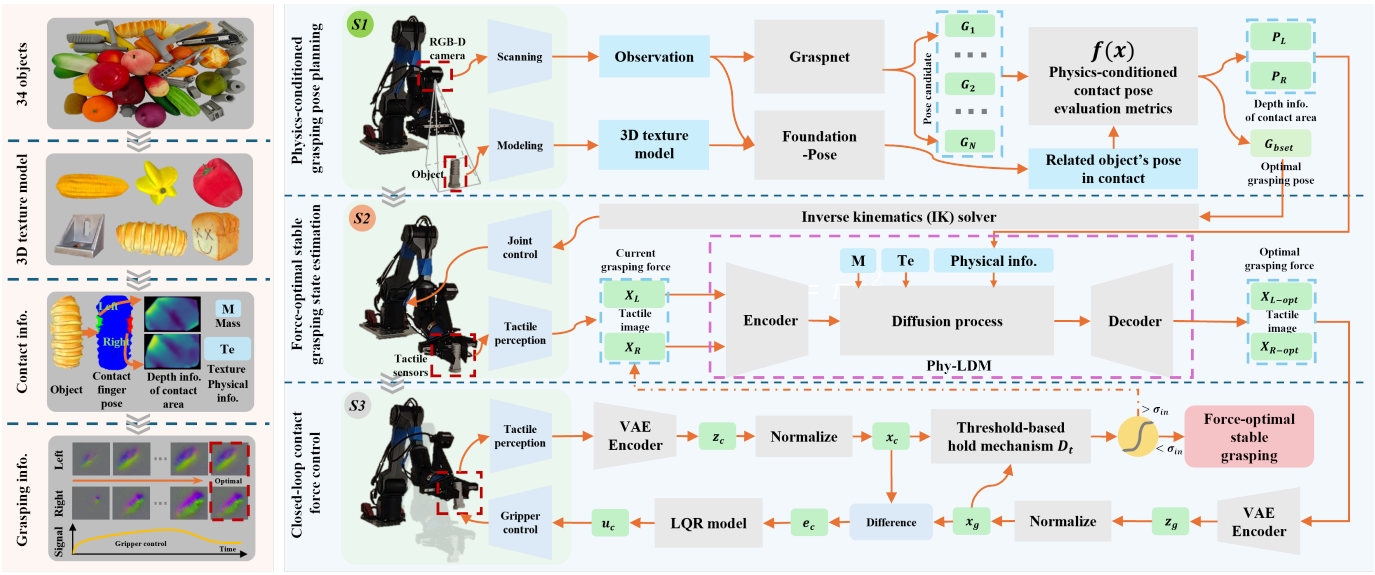

Fig. 2. The description of human-inspired stable grasping method Phy-Tac. a). Our physics-conditioned SG dataset consists of objects' 3D texture models, contact region information, control signals, and tactile states. b). Our force-optimal stable grasping strategy Phy-Tac contains 3 steps, i.e., grasping pose planning, grasping state estimation, and contact force regulation.

geometric force closure or maximizes pose confidence. Unfortunately, they do not optimize contact forces with respect to the physical properties of the contact region. Therefore, [7] proposes a physics-informed, large multimodal model that incorporates object physical information into grasp pose generation, which achieves over a 10% improvement. However, it requires the creation of extensive, task-specific datasets for training, which is both labor-intensive and costly. Therefore, we explore a more efficient strategy to incorporate objects' geometric features into grasp pose planning.

Tactile sensor in grasping. Tactile sensors play a critical role in force regulation during robotic grasping by providing rich contact information, e.g., force distribution. The vision-based sensors (e.g., GelSight [8]) are particularly notable in tactile sensors [9] for their unique capability to convert contact-induced deformations into high-resolution images as shown in Fig. 3. This property is widely used in grasping tasks, such as contact detection [10–12], and force estimation [13–15], and grasping control [16–18]. Motivated by these advantages, we employs a vision-based tactile sensor for FOSG task. Although previous studies have explored context-conditioned contact estimation [19–21], these coarse-grained text conditions fail to estimate fine-grained trends in contact property, e.g., optimal force required for stable grasping. Therefore, we leverages fine-grained physical information in contact region together with the initial contact state to guide fine-grained contact force prediction.

Diffusion model. Diffusion models synthesize of high-fidelity and diverse samples by progressively adding noise to data and learning the reverse denoising process [22]. Compared with GAN-based approaches [23], diffusion models offer superior training stability, detail preservation, and controllability, which has led to their increasing adoption in robotic perception tasks. For tactile domain, diffusion models have been applied to tactile image generation for robotic manipulation [24–26]. These works demonstrate the strong potential of diffusion models in modeling fine-grained tactile generation task. Therefore, we use physics-conditioned latent diffusion model to generate the tactile imprint of the optimal stable state, providing a clear and high-fidelity tactile prior for achieving FOSG.

Tactile dataset. There are many tactile datasets [27–29] available for different usage, such as tactile-driven image stylization. Although these well-collected datasets have been used for grasp stability [30], they still face several challenges. First, they lack the detailed information of grasping objects, e.g., material and mass, which is important effector for tactile state as shown in Fig. 3. Second, the object 3D information providing by third-view cameras cannot accurately describe the contact region, which reduces the prediction accuracy for optimal contact state. Moreover, only initial and stable grasping states cannot describe the complex grasping process [31]. Therefore, we provide a physical-conditioned tactile dataset to solve these problem for FOSG study.

III. Methodology

We propose a human-inspired unified framework Phy-Tac for FOSG, which embeds force optimality as a guiding principle throughout the entire grasping process (Fig. 2). Instead of treating tactile feedback as a passive signal, our method formulates an explicit tactile goal representing the force-optimal stable state and actively drives the system toward it. The pipeline consists of four components: Sec. III-A builds a physics-conditioned tactile dataset; Sec. III-B plans grasp poses with contact-patch physics; Sec. III-C predicts the optimal tactile imprint as the control target; and Sec. III-D realizes that target via a latent-space servo that achieves stability with just-enough force.

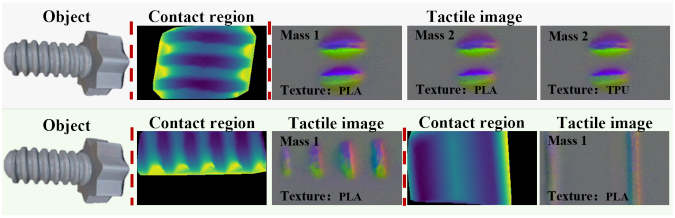

Fig. 3. A example to show the influence of physical condition for contact tactile state. Specifically, the physical factors are mass M, texture T, and contact region P_t.

A. Physics-conditioned tactile dataset

To enable modeling of tactile mechanisms that respect real-world contact physics, we construct a physics-conditioned tactile dataset that explicitly couples tactile signals with object physical attributes and stable grasp states. This dataset contains 34 objects with per-object material label T, mass M, and an accurate 3D model O scanned by an EinScan-HX device as shown in Fig. 2-a. For each grasp, we extract the contact-region point cloud P_c and align it to the fingertip frame P_t by a rigid transform T_T^C. We keep points within a rectangular fingertip window, as Eq. 1:

$$P_t = \{p_t(x,y,z) | p_t = T_T^C p_c, |x| \leq \frac{w}{2}, |y| \leq \frac{h}{2}\}. \quad (1)$$

Where, h and w are the hight and wight of tactile sensor contact region. This operation removes out-of-gripper points and standardizes the contact geometry, which mainly contributes to the the tactile feature. During contact, we synchronously record gripper commands and gripper feedback state $\{u_c, u_f\}$ together with the tactile image x_c of the gripper fingertip. All tactile signals are collected using an Xsense G1-WS device, capturing 400×700 RGB image at 40Hz. Specifically, we explicitly provide the tactile image x_s in FOSG condition. The FOSG condition is obtained by manual detection during data collection while gradually reducing the contact force until the target slips. Generally speaking, x_c is a dynamic tactile image reflecting the contact information during the grasping process (shown in Fig. 1-b *S3*), which is not necessarily the optimized x_s (shown in Fig. 1-b *S2*).

Totally, our physics-conditioned tactile dataset provides about 8.6K paired entries $\{O, M, T, P_t, u_c, u_r, x_c, x_s\}$ across 0.3K stable grasps.

B. Contact-optimized pose planning

Human grasping behavior often involves subtle finger adjustments before applying significant gripping force to achieve uniform contact pressure, thereby reducing local stress concentration and potential slip. Inspired by this observation, we further refine the top N grasp candidates with score S generated by GraspNet through introducing a human-inspired pose planning strategy thereby promoting an optimal contact force distribution. This enables the gripper to select poses that inherently favor stable and distribution-optimal interactive force before tactile optimization begins. Prior studies on contact modeling [32] indicate that the geometry of the contact region critically influences the distribution of grasping forces. Specifically, sharp or highly curved regions cause localized stress concentrations and potential slippage, whereas gradual surfaces promote uniform force distribution and stable contact. To capture these physically meaningful geometric properties, we define a physics-inspired geometric consistency metric $\mathcal{F}(\mathcal{P}_t)$ to evaluate each candidate pose P_t as Eq. 2.

$$\begin{cases} \mathcal{F}(\mathcal{P}_t) = \alpha \cdot S_{rough} + \beta \cdot (1 - C_N) + \gamma \cdot U_C, \\ W_p = \delta \cdot (1 - S) + (1 - \delta) \cdot \mathcal{F}(\mathcal{P}_t). \end{cases} \quad (2)$$

where S is the grasp candidate's score estimated by the GraspNet; S_{rough} quantifies the local surface roughness, representing the frictional stability of the contact region; C_N measures the consistency of surface normals across contact points, indicating the alignment of potential contact forces; and U_C describes the curvature uniformity, which reflects the evenness of stress distribution. Together, these three descriptors approximate the mechanical plausibility of contact stability without requiring explicit stress computation.

To balance the physically motivated geometric consistency with the data-driven confidence from GraspNet, we introduce a combined heuristic W_p. where δ controls the trade-off between learning-based confidence and physics-informed geometry. A smaller W_p indicates a pose that simultaneously exhibits high network confidence and strong force-distribution consistency.

C. Optimal tactile imprint estimation

We propose a physics-conditioned latent diffusion model (Phy-LDM) for generating tactile image for FOSG by following procedures.

Problem statement. Given the current tactile observation x_c, the depth image of the contact patch P_t, and the object's physical attributes (i.e., mass M and texture Te), we aim to synthesize the tactile imprint x_s of the force-optimal stable state. Specifically, x_s is the tactile image recorded in FOSG state (Set. III-A). We formulate state estimation as a tactile state synthesis problem by a learning generator f_θ as Eq. 3.

$$\hat{x}_g = f_\theta(x_c, P_t, M, Te) \approx x_s. \quad (3)$$

Therefore, the controller in Sec. III-D can servo the gripper toward the corresponding goal latent with minimal actuation.

Variational latent space. The VAE serves as a fundamental component for learning compressed latent space of tactile images, enabling efficient downstream diffusion-based generation. The pair (E, D) denotes the encoder–decoder of a stable-diffusion-style AutoencoderKL [33]. For a tactile image x, we obtain the latent space z through $z = E(x) \in \mathcal{R}^{C_z \times H_z \times W_z}$ and the estimated tactile image by $\hat{x} = D(z)$. The latent channel C_z and spatial size are set by configuration and the down-sampling factor. The training protocol implements a optimization strategy to balance pixel-wise reconstruction fidelity and latent regularization. Specifically, the pixel-wise reconstruction fidelity is ensured by the L1 loss, while the latent regularization is guaranteed by the loss $\mathcal{L}_{KL}$ as shown in Eq. 4. The composite loss function combines two components through linearly warmed-up KL weight ω_{KL} shown in Eq. 5.

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q_\phi(z \mid x) \,\|\, \mathcal{N}(0, I)). \quad (4)$$

$$\omega_{\mathrm{KL}}(epoch) = \min\left(1, \frac{epoch}{E_{\mathrm{warm}}}\right) \lambda_{\mathrm{KL}}. \tag{5}$$

Where E_{warm} is the warm-up length, and λ_{KL} is the target KL weight.

Physics-conditioned latent diffusion model. To explicitly couple tactile generation with grasping physics, we design a physics-conditioned Latent Diffusion Model (Phy-LDM). Unlike conventional conditional diffusion that relies purely on visual or text cues, Phy-LDM injects physics-informed conditions, i.e., contact geometry and object attributes, into denoising step of the latent diffusion process. This design allows the model to generate tactile goal states that obey force–balance and contacting constraints, producing stable yet minimally loaded contact patterns. Specifically, the current tactile and depth images are encoded as $z_t = E(x_t)$ and $z_{cdp} = E(P_t)$. The object mass M and texture Te are projected into embeddings e_M and e_T, then fused as the physical conditioning vector $C = [e_M, e_T]$. The C is injected into the U-Net via cross-attention to modulate feature propagation.

The diffusion process learns to generate optimal tactile imprint through iterative denoising, conditioned on physical constraints. The forward diffusion process in Eq. 6 progressively corrupts the target latent representation z^0 through additive noise following a variance-preserving schedule.

$$z^t = \sqrt{\bar{\alpha}^t} z^0 + \sqrt{1-\bar{\alpha}^t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I). \tag{6}$$

The reverse process employs a DDIM sampler [34] to generate samples through deterministic reverse diffusion, as formalized in Eq. 7.

$$z^{t-1} = \sqrt{\frac{\bar{\alpha}^{t-1}}{\bar{\alpha}^t}}\left(z^t - \sqrt{1-\bar{\alpha}^t}\varepsilon_\theta\right) + \sqrt{1-\bar{\alpha}^{t-1}}\varepsilon_\theta. \tag{7}$$

The noise prediction ϵ_θ in Eq. 8 is performed by a conditioned U-Net architecture, which incorporates multi-modal conditioning through cross-attention mechanisms. This architecture integrates both the current grasp state z_a and physical prompt information.

$$\begin{cases} \epsilon_\theta = \mathrm{UNet}([z_t, z_{in}, z_{cdp}], t, C), \\ \mathcal{L}_{\mathrm{obj}} = \mathbb{E}_{z_0, z_{in}, z_{cdp}, C, t, \varepsilon}\left[\|\varepsilon - \varepsilon_\theta\|^2\right]. \end{cases} \tag{8}$$

The model parameters are optimized by minimizing the objective function $\mathcal{L}_{\mathrm{obj}}$, which ensures accurate noise prediction conditioned on the physical constraints.

D. LQR Grasp Servo

To drive the robotic gripper towards a stable grasp, we formulate a closed-loop controller in the latent tactile space. The control objective is to minimize the discrepancy between the current tactile observation and the synthesized target tactile imprint, while avoiding excessive gripper motions.

State representation. The current tactile image x_c is encoded into a low-dimensional latent vector $z_c = E_p(x_c)$ by pre-trained control-aware VAE encoder. The target tactile state z_g is obtained by encoding the synthesized target tactile image x_s from the Phy-LDM. The current latent error is defined as e_c. Since the latent space is explicitly designed to capture controllable directions, the normalized control state is directly expressed as Eq. 9. Where S_{scale} denotes a scale vector obtained from the empirical distribution of latent values for normalization.

$$e_c = diag(1/S_{scale})(z_c - z_g) \in R^m. \tag{9}$$

LQR modeling. Around the goal state, the dynamics of e_c under incremental gripper displacement $\Delta u_c = u_c - u_{c-1}$ are approximated by a linear model in Eq. 10

$$\begin{cases} e_{c+1} = Ae_c + B\Delta u_c + d + w_c, \\ J = \sum_{t=0}^{\infty}(e_c^T Q e_c + \Delta u_c^T R \Delta u_c). \end{cases} \tag{10}$$

Where A and B are identified from data using recursive least squares, d denotes a small constant bias, and w_t is process noise. The identified dynamics include a small constant bias d and process noise w_c. In practice, we omit d and w_c during controller synthesis, since their effects are consistently small relative to the latent error magnitude. Residual bias and noise are tolerated by the robustness of the LQR feedback and by the threshold-based stopping rule introduced below. The control objective J minimizes both the latent error and the control error with the $Q \geq 0$ and $R > 0$. Solving the discrete algebraic Riccati equation, we get the control signal $\Delta u = -Ke_c$.

Threshold-based hold mechanism. As exact tactile matching is unattainable, a threshold-based stopping rule is introduced. The normalized error distance $D_c = \|e_c\|_2$ is monitored within a sliding window of duration τ. A grasp is considered achieved if $D_c \leq \delta_{in}$ for all frames in the window and safety conditions are satisfied, and the controller switches to hold mode.

IV. Experimental study

To verify the efficiency of our Phy-Tac method, we evaluate each procedure in our strategy by comparing with state-of–the-art methods in both qualitative and quantitative experiments.

A. Physics-conditioned grasping pose planning

As the method described in III-B, we select 4 candidates with top scores generated by GraspNet as shown in Fig. 4. Subsequently, the corresponding depth images under contact region for each gripper finger are captured. Visually, the contact image of first candidate is sharper and exhibits a smaller contact region when compared with the third candidate. Therefore, this grasping attempt is more likely to slip or rotate, resulting failure of grasping based on the studies for considering contact geometry in stable grasp, e.g. [32]. To avoid this problem, a physics-conditioned function is used for selecting force distribution optimal pose for stable grasping instead of intuitive analysis. The quantitative results in Fig. 5-a shows the detailed parameter values and the weighted cost W_p for each candidate with the constant values $\alpha = 0.2$, $\beta = 0.6$,$\gamma = 0.2$, and $\delta = 0.5$. Specifically, physics-conditioned parameters S_{rough}, C_N, and U_C are normalized across the candidates. Following our principle for grasping pose selection, the third candidate is the best one ($W_p \leq 0.2$) which is optimal for contact force distribution.

Furthermore, a significant mismatch rate ($\geq 20\%$) exists between the candidate with the optimal contact force distribution (best W_p) and the top-ranked candidate from the

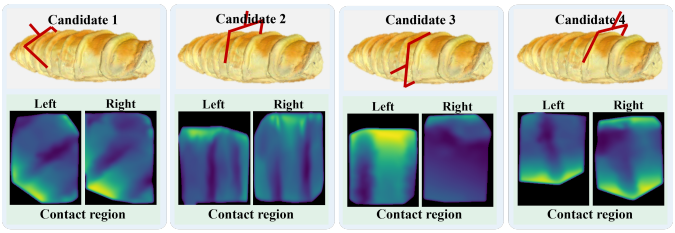

Fig. 4. The contact description of the selected candidates with top score. The first row in each sub-figure is the grasping pose generated by GraspNet for bread. The second row is the depth information of contact region in left/right finger for generated grasping pose.

grasp pose generator (best S) during our experiment for four object groups as shown in Fig. 5-b. This significantly high rate ($\geq$ 20%) underscores the necessity of a physics-based evaluation to identify grasps with optimal force distribution, rather than directly adopting the initial output from GraspNet. Furthermore, the trend observed in the mismatch rate as the number of candidates increases substantiates the rationale for selecting the top four candidates from GraspNet in our experiments. Specifically, beyond four candidates, the rate of mismatch plateaus ($\leq$ 10%), indicating diminished returns with additional candidates.

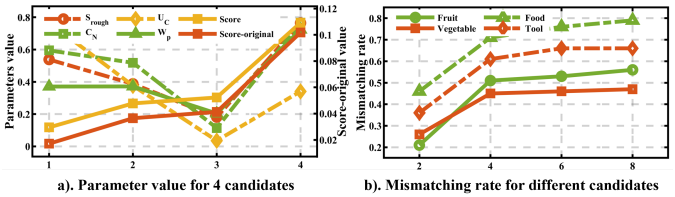

Fig. 5. The parameter value of the selected grasping candidates generated by GraspNet. a). The score and evaluating metrics of each candidates for grasping pose planning. b). The value of mismatching rate is used to select suitable candidate numbers in contact-optimal pose planning process.

B. Physics-conditioned contact state generation

To evaluate the effectiveness of our physics-conditioned contact state (tactile image) generation method for stable grasping with optimal contact force, we compare it with several state-of-the-art generative models. Specifically, five quantitative metrics are employed for evaluation, including pixel-level fidelity, i.e., mean absolute error (MAE) and root mean square error (RMSE), structural similarity, i.e., structural similarity index (SSIM), distribution consistency, i.e., learned perceptual image patch similarity (LPIPS), reconstruction quality, i.e., peak signal-to-noise ratio (PSNR).

Comparative study. Compared with the four classical image generation baselines, our Phy-LDM achieves significant and consistent improvements across all evaluation metrics, as shown in Table I. Specifically, Phy-LDM reduces both MAE and RMSE by over 30–40% compared with the best diffusion-based baseline (LDM [33]), and simultaneously achieves higher PSNR ($\geq$ 42) and SSIM ($\geq$ 0.98), indicating superior pixel-level accuracy and structural preservation. The LPIPS score also drops notably by over 17%, demonstrating enhanced perceptual quality and realism. This performance gain stems from the explicit integration of contact geometry and object properties into the generative process. By embedding these physical priors into the latent space, Phy-LDM can model the intrinsic relationship between contact mechanics and tactile appearance, rather than relying solely on data-driven visual correlations. As a result, it generates tactile images that more faithfully reproduce fine-grained pressure distributions, deformation patterns, and texture variations across different contact scenarios as shown in Fig. 6.

TABLE I
THE COMPARATIVE RESULTS WITH SEVERAL CLASSICAL IMAGE GENERATION METHODS.

Method	MAE↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
CGAN [23]	0.083	0.093	32.137	0.915	0.143
CVAE [35]	0.063	0.079	34.082	0.935	0.119
DDIM [34]	0.011	0.021	38.089	0.965	0.085
LDM [33]	0.015	0.016	39.776	0.968	0.081
Phy-LDM	**0.009**	**0.011**	**42.087**	**0.988**	**0.067**

Here, the LDM uses a lightweight stable-diffusion–style architecture due to limited training data.

To further validate the generalization capability of Phy-LDM across different object categories, we evaluate its performance on four representative groups: fruit, food, vegetable, and tool, as shown in Fig. 7. The model consistently maintains low MAE ($\leq$ 0.01), RMSE ($\leq$ 0.03), and LPIPS ($\leq$ 0.1), while achieving high PSNR ($\geq$ 36) and SSIM ($\geq$ 0.93), demonstrating stable tactile image generation under diverse contact conditions. Notably, Phy-LDM achieves its best performance on tool-type objects, while showing relatively lower accuracy on food objects. This discrepancy primarily arises from the highly deformable and soft nature of food items (e.g., bread), where the contact geometry dynamically changes during interaction. Although our model introduces geometric priors to constrain the generation process, the non-rigid deformation of soft materials leads to slight mismatches between predicted and real contact states. In contrast, Tool-type objects exhibit rigid and stable contact geometries, allowing Phy-LDM to leverage the encoded physical priors more effectively, resulting in sharper and more faithful tactile reconstructions. These results highlight that the incorporation of contact geometry and material properties enables Phy-LDM to generalize well across categories, while the residual performance gap reveals potential directions for future improvements in modeling deformable object interactions.

For comparing with state-of-the-art tactile state estimation method, several experimental study is implemented. As shown in Table II, our physics-conditioned method substantially outperforms existing approaches, generating more accurate tactile images for FOSG. Specifically, Phy-LDM reduces both MAE and RMSE over 60%, LPIPS over 40%, improves the PSNR over 11%, SSIM over 2%. Therefore, this superiority is reflected not only in pixel-level, but also in structural similarity and distribution consistency in the generated tactile image. To visually show the each method's performance, a significant example is provided in Fig. 8. Although the other method can generate tactile image closely aligning with the provided text descriptions, they are difficult to produce one as the real tractile image in given contact situation as Phy-LDM.

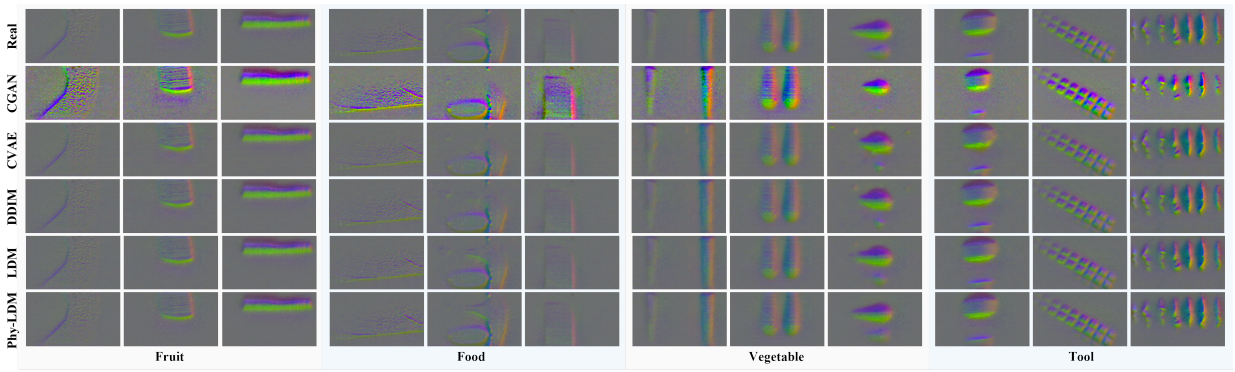

Fig. 6. The predicted tactile results of different state-of-the-art methods for FOSG. This comparative results contains four type of objects, i.e., fruit, food, vegetable, and tool.

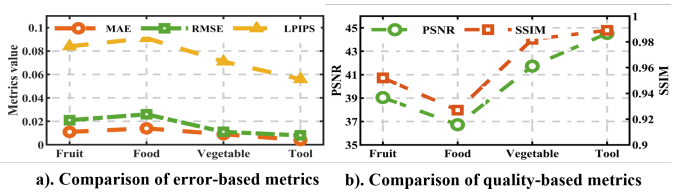

Fig. 7. The evaluation value of different metrics for four types of grasp objects under our Phy-LDM method.

The reason is that the pure text description cannot reflect the contact information in detail, especially the geometry in contact region which is one of the main factors to influence the contact pattern in tactile image. Instead, our method introduces the fine-grained information in generation process to overcome this limitation. Therefore, the tactile image generated by Pyh-LDM can accurately reflect the contact state.

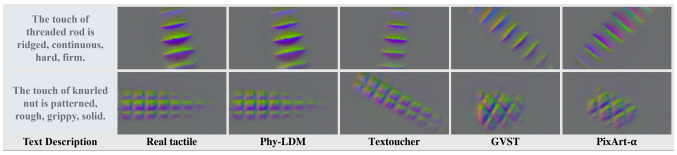

Fig. 8. The comparative result between Phy-LDM and other representative methods. Phy-LDM can generate more accurate tactile image based on the physical contact information.

TABLE II
THE COMPARATIVE RESULTS WITH STATE-OF-THE-ART TACTILE IMAGE ESTIMATION METHODS.

Method	MAE ↓	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GVST [26]	0.042	0.057	30.094	0.943	0.235
PixArt-α [36]	0.037	0.042	34.438	0.924	0.197
TextToucher	0.031	0.035	36.814	0.952	0.142
Phy-LDM	**0.010**	**0.014**	**41.087**	**0.974**	**0.083**

Ablation study. To analyze the contribution of each conditional component to generation performance, we conducted an ablation study as summarized in Table III. The results clearly show that removing the contact geometry descriptor (z_{cdp}) and contact state representation (z_{in}) leads to the most significant performance degradation across all metrics, i,e., MAE, RMSE, and LPIPS increased by 35%, 31%, and 28% respectively; PSNR and SSIM decreased by 14%, and 4% respectively. This observation highlights that accurate tactile generation relies heavily on spatial and structural cues describing the contact configuration, which determine the overall texture layout and fine-grained deformation patterns in tactile imprint. In contrast, excluding the object's physical properties, such as mass (e_M) and texture embedding (e_T), results in a smaller performance drop, i,e., MAE, RMSE, and LPIPS increased by 18%, 15%, and 12% respectively; PSNR and SSIM decreased by 2%, and 1% respectively. This suggests that, within the limited variability of our dataset, object-level physical attributes play a secondary but complementary role. They refine the realism and local consistency of generated images, but the tactile texture formation is primarily driven by geometric contact priors. Overall, this experiment confirms that incorporating contact geometry and instantaneous interaction states is crucial for achieving physically consistent tactile image synthesis, while the integration of material and mass information further enhances perceptual fidelity and generalization.

TABLE III
THE ABLATION RESULTS OF PHY-LDM METHOD.

Condition	MAE↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Lacking z_{cdp}	0.018	0.019	34.742	0.942	0.108
Lacking z_{in}	0.014	0.016	36.371	0.955	0.094
Lacking e_T	0.012	0.013	39.932	0.969	0.079
Lacking e_M	0.011	0.013	40.882	0.976	0.077
Phy-LDM	**0.009**	**0.011**	**42.087**	**0.988**	**0.067**

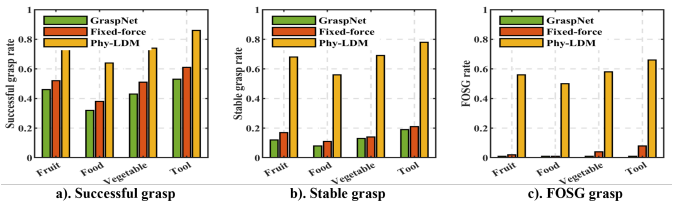

Fig. 9. The experimental results for three grasp methods in four types of objects. Three evaluating metrics are calculated, i.e., successful grasp rate, stable grasp rate, and FOSG rate.

C. Grasp experiment

To demonstrate the effectiveness of our Phy-Tac, we implement our strategy on two robotic arms and compared it with other classical methods, i.e., GraspNet and fixed-force method. Specifically, the contact pose of our and fixed-force method are generated by the GraspNet model. Furthermore, we consider three grasping states, i.e., successful grasp (SuG) which requites to catch the object, stable grasp (StG) which means successful grasp without slip and rotation, and FOSG which ensures optimal contact force in stable grasp.

As shown in Fig. 9, the successful rates of the three states over 50 grasp attempts indicate that the widely used fixed-force strategy can significantly improve the SuG rate ($\geq 7\%$). However, its effectiveness in enhancing states StG and FOSG is limited ($\leq 1\%$). The reason to cause this results is that the given constant force can ensure that contact between gripper and object is real, which is not promised by original GraspNet. Because the applied force is constant, this strategy cannot ensure the pre-setting is suitable for stable grasp Fortunately, our method Phy-Tac can solve this shortcoming by it unique design, i.e., taking the force state (tactile image) into consideration. A real-world experiment show in Fig. 10 clearly demonstrates this: under GraspNet, the gripper makes contact but slips due to insufficient force; the fixed-force method achieves a grasp but with excessive deviation from the optimal force. In contrast, our Phy-Tac drives the gripper to apply the optimal contact force, successfully achieving FOSG.

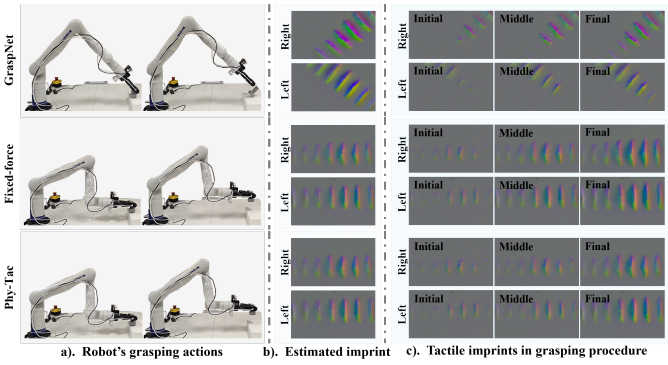

Fig. 10. A robotic experiment for three grasping strategies. a). the initial contact and lifting states. b). the estimated imprint by Phy-LDM in initial contact condition. c). the tactile imprints during the grasping process.

Considering the challenges of tactile modeling for object with different material properties, especially the deformable items, our Phy-Tac method demonstrates consistent and robust performance. Among the four types of grasped objects, the three evaluation metrics of Phy-Tac show only slight variations as shown in Fig. 9. For example, the differences between categories such as food (soft bread) and tools (hard screws) are not significant. This is mainly because our method actively adjusts the gripper configuration to achieve optimized contact forces, thereby reducing the likelihood of deformation during stable grasping. Consequently, Phy-LDM attains high success rates across objects with varying material properties.

To evaluate the LQR grasp servo, we visualize the error distance D_t of one successful grasp for each type of object.

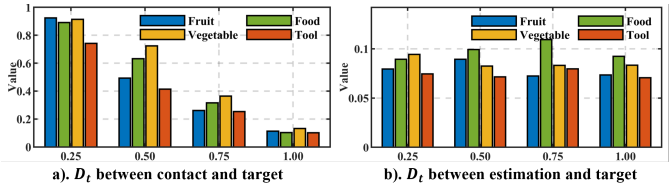

Fig. 11. The error distance D_t during the grasp process. The x-axis represents the percentage of the grasping process, and the y-axis represents the error distance D_t.

Specifically, the error distance D_t between current tactile state during the grasp process and optimal tactile state under given grasp condition is shown in Fig. 11-a. In practice, we set the threshold δ_{in} as 0.15. The results indicate that our grasp servo can efficiently drive the robotic gripper to suitable position for achieving optimal tactile state, which requires $D_t \leq \delta_{in}$. Therefore, our method is robust for all four types of objects. In Fig. 11-b, we indicate the D_t for FOSG between the estimated optimal tactile in grasp process and the optimal tactile state under given grasp condition. The small fluctuations of D_t during the grasp process indicates that our Phy-LDM can estimated the optimal tactile for FOSG robustly, which benefits from incorporating object and contact-region physical information. In summary, by leveraging physically informed tactile representations and an LQR-based servo controller, our method consistently converges to the optimal tactile state with minimal error, highlighting the critical role of physical priors in achieving robust and smooth grasp control.

V. Conclusion

This paper presented a human-like framework, Phy-Tac, for FOSG that unifies pose planning, tactile prediction, and latent-space force control. A physics-conditioned latent diffusion model predicts the optimal tactile imprint, while an LQR-based controller drives the gripper toward force-optimal stability. Experiments on diverse rigid and compliant objects demonstrate that the proposed method significantly improves grasp stability and force efficiency compared with baseline strategies. Overall, this work establishes a pathway from stability-driven to force-optimal manipulation, contributing to safer and more adaptive tactile intelligence in robotic hands.

References

[1] R. Newbury, M. Gu, L. Chumbley *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023.

[2] X. Mao, Y. Xu, R. Wen *et al.*, "Efficient tactile sensing-based learning from limited real-world demonstrations for dual-arm fine pinch-grasp skills," in *2024 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5112–5119.

[3] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proc. of the IEEE/CVF ICCV*, 2019, pp. 2901–2910.

[4] J. Yang, M. Chen, W. Chen *et al.*, "Vt-vt: a slip detection model for transformer-based visual-tactile fusion," *Advanced Robotics*, vol. 38, no. 17, pp. 1177–1187, 2024.

[5] R. Wang, J. Zhang, J. Chen *et al.*, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," *arXiv preprint arXiv:2210.02697*, 2022.

[6] Z. Weng, H. Lu, D. Kragic *et al.*, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, 2024.

[7] D. Guo, Y. Xiang, S. Zhao *et al.*, "Phygrasp: generalizing robotic grasping with physics-informed large multimodal models," *arXiv preprint arXiv:2402.16836*, 2024.

[8] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[9] T. Li, Y. Yan, C. Yu *et al.*, "A comprehensive review of robot intelligent grasping based on tactile perception," *Robotics and Computer-Integrated Manufacturing*, vol. 90, p. 102792, 2024.

[10] Y. Liu, X. Xu, W. Chen *et al.*, "Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1106–1113, 2023.

[11] N. Zhang, R. Sui, L. Zhang *et al.*, "A robust incipient slip detection method with vision-based tactile sensor based on local deformation degree," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 17 200–17 213, 2023.

[12] S. Zhang, Y. Sun, J. Shan *et al.*, "Tirgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6307–6314, 2023.

[13] D. Ma, E. Donlon, S. Dong *et al.*, "Dense tactile force estimation using gelslim and inverse fem," in *2019 Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5418–5424.

[14] B. Sundaralingam, A. S. Lambert, A. Handa *et al.*, "Robust learning of tactile force estimation through robot interaction," in *2019 Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9035–9042.

[15] C. Lin, H. Zhang, J. Xu *et al.*, "9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 923–930, 2023.

[16] M. Matak and T. Hermans, "Planning visual-tactile precision grasps via complementary use of vision and touch," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 768–775, 2022.

[17] M. Costanzo, "Control of robotic object pivoting based on tactile sensing," *Mechatronics*, vol. 76, p. 102545, 2021.

[18] J. Lloyd and N. F. Lepora, "Goal-driven robotic pushing using tactile and proprioceptive feedback," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 1201–1212, 2021.

[19] J. Tu, H. Fu, F. Yang *et al.*, "Texttoucher: Fine-grained text-to-touch generation," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7455–7463.

[20] F. Yang, C. Feng, Z. Chen *et al.*, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 26 340–26 353.

[21] C. Gungor, D. Eppinger, and A. Kovashka, "Towards generalization of tactile image generation: Reference-free evaluation in a leakage-free setting," *arXiv preprint arXiv:2503.06860*, 2025.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[24] X. Lin, W. Xu, Y. Mao *et al.*, "Vision-based tactile image generation via contact condition-guided diffusion model," *arXiv preprint arXiv:2412.01639*, 2024.

[25] S. Rodriguez, Y. Dou, M. Oller *et al.*, "Touch2touch: Cross-modal tactile generation for object manipulation," *arXiv preprint arXiv:2409.08269*, 2024.

[26] F. Yang, J. Zhang, and A. Owens, "Generating visual scenes from touch," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, 2023, pp. 22 070–22 080.

[27] F. Yang, C. Ma, J. Zhang *et al.*, "Touch and go: Learning from human-collected vision and touch," *arXiv preprint arXiv:2211.12498*, 2022.

[28] Y. Li, J.-Y. Zhu, R. Tedrake *et al.*, "Connecting touch and vision via cross-modal prediction," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[29] L. Fu, G. Datta, H. Huang *et al.*, "A touch, vision, and language dataset for multimodal alignment," *arXiv preprint arXiv:2402.13232*, 2024.

[30] S. Kanitkar, H. Jiang, and W. Yuan, "Poseit: A visual-tactile dataset of holding poses for grasp stability analysis," in *2022 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 71–78.

[31] K. Nakahara and R. Calandra, "Learning gentle grasping using vision, sound, and touch," *arXiv preprint arXiv:2503.07926*, 2025.

[32] M. A. Roa and R. Suárez, "Grasp quality measures: review and performance," *Autonomous robots*, vol. 38, pp. 65–88, 2015.

[33] R. Rombach, A. Blattmann, D. Lorenz *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[34] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[35] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[36] J. Chen, J. Yu, C. Ge *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *ICLR*, 2024.