Ensuring Force Safety in Vision-Guided Robotic Manipulation via Implicit Tactile Calibration

Lai Wei*,[1], Jiahua Ma*,[1], Yibo Hu[2], and Ruimao Zhang[1]

Abstract— **In dynamic environments, robots often encounter constrained movement trajectories when manipulating objects with specific properties, such as doors. Therefore, applying the appropriate force is crucial to prevent damage to both the robots and the objects. However, current vision-guided robot state generation methods often falter in this regard, as they lack the integration of tactile perception. To tackle this issue, this paper introduces a novel state diffusion framework termed SafeDiff. It generates a prospective state sequence from the current robot state and visual context observation while incorporating real-time tactile feedback to refine the sequence. As far as we know, this is the first study specifically focused on ensuring force safety in robotic manipulation. It significantly enhances the rationality of state planning, and the safe action trajectory is derived from inverse dynamics based on this refined planning. In practice, unlike previous approaches that concatenate visual and tactile data to generate future robot state sequences, our method employs tactile data as a calibration signal to adjust the robot's state within the state space implicitly. Additionally, we've developed a large-scale simulation dataset called SafeDoorManip50k, offering extensive multimodal data to train and evaluate the proposed method. Extensive experiments show that our visual-tactile model substantially mitigates the risk of harmful forces in the door opening, across both simulated and real-world settings.**

I. Introduction

Vision-guided robotic manipulation [2], [3], [7], [13], [14], [22], which involves generating a feasible motion trajectory solely from visual information to carry out specific manipulation tasks, has become increasingly prevalent in the field of robotics and embodied AI. By leveraging visual inputs to perceive and interpret the environment, this approach allows robots to plan and execute intricate movements. However, current research predominantly emphasizes the success rate of these manipulation tasks, often neglecting the critical factor of how much force the robot should apply during the process. In practice, the inappropriate force exerted not only risks damaging the manipulated object but also puts unnecessary strain on the robot's joints. Hence, addressing this issue is essential, and we refer to this consideration as ensuring **force safety**.

From the perspective of state planning, inappropriate force mainly arises when the generated state fails to meet the specific physical properties of the manipulated object. Taking the door-opening task shown in Fig. 1 as an example, the door can only move along the arc-shaped trajectory determined by its physical properties, such as size, opening angle, and its position relative to the robot. This indicates that all states of the robot's end-effector, generated by the state planning model, must strictly adhere to this arc-shaped trajectory to ensure force safety. Otherwise, the robot controller might exert **harmful forces** in an attempt to reach states outside this trajectory, which could result in damage to both the door and the robot. To be brief, we define the state that lies on this arc-shaped trajectory as **safe state**, while those outside are deemed unsafe. In this paper, our primary focus is on planning safe states to ensure force safety throughout the robotic manipulation process.

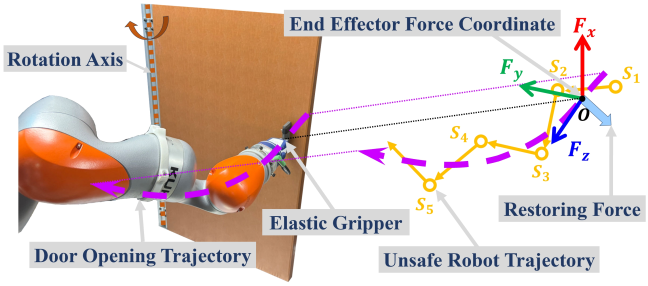

Fig. 1: The restoring force exerted by the robot's end-effector can be decomposed into three components: F_x, F_y, and F_z. The component F_z is tangent with the door's opening trajectory and is termed the effective force. The forces lying in the xOy plane are orthogonal to the trajectory. These forces might cause damage to both the robot and the door and are therefore referred to as **harmful forces**.

An intuitive solution for the above safe door-opening state planning can be found in bionics: when opening the door, humans estimate future door-opening states based on the door's physical properties—such as size, opening angle, and so forth—using visual perception, and then modify them in real-time based on the forces sensed through tactile feedback during the actual door-opening process. Inspired by this, we aim to dynamically integrate real-time tactile feedback to refine the vision-guided generated states. However, this solution remains challenging due to the intricate, nonlinear dynamics between the current force feedback and the refinement of future states. These dynamics are influenced by factors like the robot's manipulability, positions relative to the door, and other physical considerations in the real world.

To address such an issue, we develop a diffusion-based

*indicates equal contribution

[1] Lai Wei, Jiahua Ma, and Ruimao Zhang (Corresponding Author) is with School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, 518172, China, laiwei1@link.cuhk.edu.cn, jiahua_ma_sjtu@163.com, ruimao.zhang@ieee.org

[2] Yibo Hu is with the State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China, boyihu@zju.edu.cn

model named **SafeDiff** to plan safe states, leveraging the effectiveness of diffusion models in approximating complex distributions. In this work, we utilize offline demonstrations collected from the simulator to learn the aforementioned dynamics of the door opening and embed this knowledge into the state representation. This allows us to perform implicit calibration on vision-guided states online, utilizing real-time tactile feedback obtained during inference. Such a process enables the generated states to progressively satisfy the constraints imposed by the door's properties, thereby ensuring force safety during the entire door-opening process.

Based on this, our SafeDiff demonstrates strong generalization capabilities and brings several benefits as well. (1) **Robustness to External Disturbances.** SafeDiff excels in handling environmental disturbances online. It can continuously correct the state changes induced by these disturbances during the manipulation. It ensures force safety, effectively adapting to the changing environmental conditions, while other methods often fail to open the door, let alone maintain force safety. This highlights SafeDiff's ability to calibrate its planned state dynamically based on real-time tactile feedback. (2) **Few-shot Sim-to-Real Transfer.** SafeDiff demonstrates exceptional capability in transferring knowledge from simulation to the real world, even with limited real-world data. This significantly reduces the need for extensive real-world training, streamlining the transition from simulation to practical applications.

The main **contributions** of this work are three-fold. 1) We introduce a novel benchmark centered on force safety during robotic manipulation. This benchmark includes three physically sound and computationally efficient metrics that can effectively assess the safety of state planning models. For the door opening, we create a large-scale safety-related simulation dataset, **SafeDoorManip50k**, to validate the effectiveness of the involved models. 2) We propose a diffusion-based model, **SafeDiff**, which dynamically incorporates real-time tactile feedback to calibrate vision-guided robot states implicitly. 3) Extensive experiments cross both simulated and real-world settings show that our method outperforms existing approaches in safe state planning, greatly reducing the risk of object damage during robotic manipulation.

II. Related Works

A. Vision-based Robotic Manipulation

Numerous studies on vision-based robotic manipulation have addressed tasks such as object grasping [7], [14], articulated object manipulation [13], [22], and object reorientation [3]. These works emphasize improving the robot's environmental perception through various visual modalities to enhance task success rates. For instance, [2], [7], [14] proposed using RGB-only images for robust robotic manipulation, while SAGCI [20], RLAfford [13], and Flowbot3D [9] rely solely on point clouds for observations. Additionally, [29], [32] integrated both RGB images and point clouds to promote the performance on specific manipulation tasks. However, the objects manipulated by robots are often fragile, especially articulated ones. In view of this, vision-based manipulation is challenging to apply in real-world applications because it cannot accurately reflect the force safety status of the manipulated objects. Therefore, it is of great significance for robots to incorporate tactile feedback such that it can dynamically adjust the planned states and handle objects in a safer and more controllable manner.

B. Multimodal Tactile Feedback for Enhanced Manipulation

Various learning-based approaches have employed tactile feedback to enhance robotic manipulation. For instance, [23] introduced a tactile perception-driven method that enables robots to learn how to grasp objects without relying on visual input. Numerous studies focus on grasp stability [4], [6], [8], as well as regrasping [5], [25]. A few methods [15], [26]–[28] combine reinforcement learning with tactile feedback to formulate manipulation strategies. And very few approaches leverage the combined benefits of both vision and touch. For example, [10] integrated prior knowledge with dynamic model adaptation to locally compensate for changing dynamics, while [16] developed a self-supervised learning framework that fuses visual and tactile inputs for peg insertion, improving learning efficiency. However, the majority of these works used tactile feedback to improve manipulation effectiveness rather than to guide safe planning.

C. Datasets for Door Opening

In recent years, a primary approach for door manipulation tasks has been to build simulation environments that emulate real-world conditions. Studies such as [9], [11], [12], [21], [30], [32], [33] have introduced a variety of simulated door-opening mechanisms, including pushing, pulling, and even those involving latching mechanisms. Moreover, datasets like PartNet-Mobility [31] and AKB-48 [19] offer diverse collections of articulated objects, including doors, but their focus on visual data collection overlooks crucial modalities such as tactile information, limiting their effectiveness for safe door-opening states planning. To address these shortcomings, we developed a comprehensive door manipulation environment with multi-modal inputs and provided a large-scale door-opening dataset to support safe manipulation planning.

III. Methodology

A. Preliminary

We begin by briefly reviewing the diffusion models, a class of generative models that synthesize data by reversing a Markovian process where Gaussian noise is progressively added to data samples. These models consist of two primary phases: the forward process and the reverse process. In the forward process, the original data is systematically corrupted, transitioning from a structured state to pure Gaussian noise over a predefined number of steps, described by the equation $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon$, where ϵ is Gaussian noise and α_t are variance-preserving coefficients. The reverse process entails learning to undo the noise addition to recover the original data from its noisy state. This involves training a

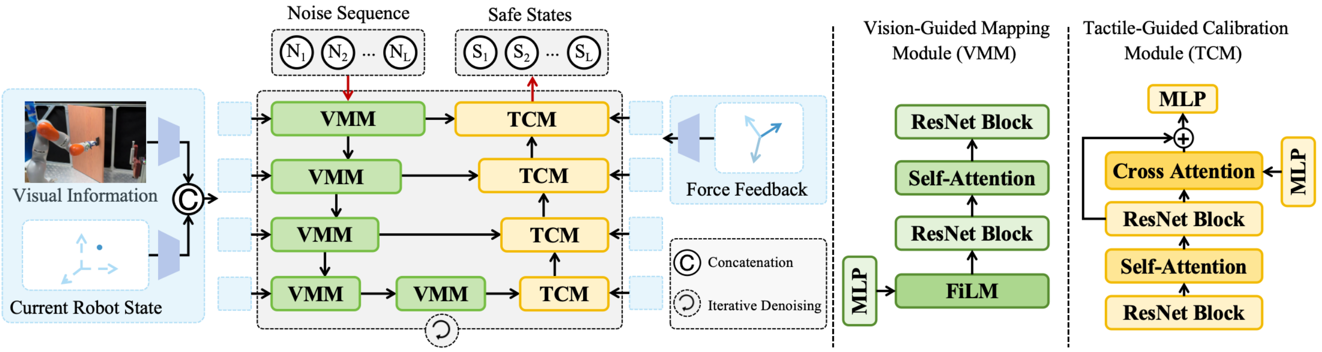

Fig. 2: Our framework takes a noise sequence as input, visual information, current robot state, and its corresponding force feedback as conditions and outputs the final safe states through T denoising iterations. The architecture consists of an encoder and a decoder. The encoder is composed of a series of multi-scale Vision-Guided Mapping Modules (VMMs) that integrate visual data using FiLM [24] and generate state representations initially. The decoder comprises a stack of Tactile-Guided Calibration Modules (TCMs) which can refine the state representations based on tactile feedback.

neural network to estimate the reverse conditional distribution $p(x_{t-1}|x_t)$, utilizing advanced deep learning techniques. A typical application of the diffusion model in robotic manipulation is Decision Diffuser [1], which makes decisions using a return-conditional diffusion model, allowing policies to generate behaviors satisfying individual constraints.

B. The Overall Framework

Motivated by the Decision Diffuser [1], the proposed **SafeDiff** aims to generate a consistent robot state sequence $\mathbf{S} = \{S_k\}_{k=1}^{L}$ that ensures force safety conditioned on visual-tactile information experienced during manipulation, thereby preventing any potential damage to the door. As shown in Fig. 2, we employ an encoder-decoder architecture for our diffusion model. Given the visual representation $\mathbf{O}$ of the current scene context, typically obtained from the image $\mathbf{I}$, and the current robot state $\mathbf{R}$, the initial input to the model is a set of Gaussian noise $\mathbf{N} = \{N_k\}_{k=1}^{L}$ with length L. After T iterations, the model produces a sequence of L consecutive robot states $\mathbf{S}$. Notably, we have opted to replace the action sequence generated in [1] with a sequence of robot states. This option stems from the fact that, while the door opening trajectory is predictable, conventional control actions do not inherently guarantee force safety. Instead, each robot state is closely correlated with the current state's potential harmful force magnitude. Consequently, using robot states facilitates a more robust and efficient model training when integrating tactile feedback.

To harness visual and tactile information effectively to generate safe and reasonable robot states, we first introduce the Vision-Guided Mapping Module (VMM) to construct the encoder for our state diffusion model. This module translates the robot's current state, denoted as $\hat{\mathbf{S}}$, and the visual scene context $\mathbf{O}$, including the door size and relative position to the robot, into a comprehensive state space representation. Although the diffusion model can initially estimate robot state trajectories based on these visual cues, it falls short of guaranteeing force safety during the manipulation process. To tackle this, we further introduce a Tactile-Guided Calibration Module (TCM) to act as the decoder of our model. Drawing inspiration from human adaptability in responding to tactile feedback and adjusting actions accordingly, this module is designed to capture the intricate, nonlinear dynamics between the current force feedback, represented by $\mathbf{F}$, and the projected residuals of future states. For more details about the module design, please refer to Sec. III-C.

C. Network Architecture Design

Visual-Guided Mapping Module (VMM) As shown in Fig. 2, we stack a series of VMMs with different temporal scales to construct the encoder for our state diffusion model. In this module, we initially generate the robot state representation by using visual information and the current robot state. Firstly, we use a Multi-Layer Perceptron (`MLP`), followed by a Resnet block (`Res`), to extract current scene context from the input image $\mathbf{I}$ and current state $\hat{\mathbf{S}}$. And following FiLM [24], we regard such extracted current scene context as affine coefficients and map Gaussian noise inputs $\mathbf{N}$ into the initial state representation. Then, a self-attention (`Sttn`) and a Resnet block (`Res`) are used to enhance the temporal coherence of these state representations:

$$[\alpha,\ \beta] = \texttt{MLP}(\mathbf{I},\ \hat{\mathbf{S}}) \tag{1}$$

$$\mathbf{S}^* = \texttt{Res}(\alpha \cdot \mathbf{N} + \beta) \tag{2}$$

$$\mathbf{S}^* = \texttt{Res}(\texttt{Sttn}(\mathbf{S}^*)) \tag{3}$$

where α and β denote the affine coefficients, and $\mathbf{S}^*$ denotes the state representation with the specific temporal scale in the corresponding VMM.

Tactile-Guided Calibration Module (TCM) Similar to the previous module, we utilize a series of TCMs with different temporal scales to form the decoder for our state diffusion model. In this module, we calibrate the robot state representation $\mathbf{S}^*$ to a safer one by introducing tactile information. Before calibration, we use a combination of two `Res` and one `Sttn` to further enhance the temporal

TABLE I: Quantitative evaluation of our method and existing models on the **simulation** scenarios from our **SafeDoorManip50k**, highlighting the effectiveness of our method in safe state planning. Ours (V) denotes our method utilizing only visual data as input, while Ours (V+T) incorporates both visual data and tactile calibration. The symbols ✓and ✗ indicate whether the door manipulated is seen or unseen.

	Seen (?)	SuR (%) ↑	AHF (N) ↓	Threshold - 5 N		Threshold - 10 N		Threshold - 15 N	
				SaR-95 (%) ↑	SaR-80 (%) ↑	SaR-95 (%) ↑	SaR-80 (%) ↑	SaR-95 (%) ↑	SaR-80 (%) ↑
Li et al. [17]	✓	96.40	7.68	0.10	0.66	6.09	43.12	56.85	**94.19**
Ours (V)	✓	**99.98**	6.27	0.83	7.87	22.41	57.65	66.86	88.26
Ours (V+T)	✓	**99.98**	**5.00**	**6.10**	**25.28**	**49.25**	**78.73**	**78.79**	93.00
Li et al. [17]	✗	95.27	7.51	0.00	0.66	8.82	47.55	58.15	**92.76**
Ours (V)	✗	**99.78**	13.08	0.87	2.57	8.15	21.35	33.44	51.75
Ours (V+T)	✗	99.33	**5.08**	**5.13**	**24.90**	**55.54**	**79.33**	**79.87**	91.54

coherence of $\mathbf{S}^*$. And then, we extract safety context from the input force feedback $\mathbf{F}$ using a MLP. Essentially, harmful forces and states errors can be regarded as 2 physical forms of force insecurity in different spaces (i.e. the former is in force space and the latter is in state space). Based on this, we use a cross attention block (Cttn) to map such extracted safety context into implicit state residual which can be used to calibrate the initial state trajectory generated by the encoder.

$$\mathbf{S}^* = \texttt{Res}(\texttt{Sttn}(\texttt{Res}(\mathbf{S}^*))) \tag{4}$$

$$\mathbf{S}^* = \mathbf{S}^* + \texttt{Cttn}(\mathbf{S}^*, \texttt{MLP}(\mathbf{F})) \tag{5}$$

IV. Dataset

To overcome the gap in datasets, we establish the first dataset for ensuring force safety in door opening manipulation planning, named **SafeDoorManip50k**. Drawing on the open-source assets detailed in [18], we constructed a diverse collection of 57 doors, each featuring unique structural designs and distinct color textures. Notably, due to functional limitations of the *Isaac Gym*, x-axis harmful forces are inaccurate with the original door handle. Consequently, we made modifications to the collision mesh of the door handle model, enabling accurate readings of the harmful forces in the x-axis. These doors were subsequently divided into a set of 45 seen doors and a set of 12 unseen doors.

In the *Isaac Gym* simulation environment, we established an assembly of doors and robots, where the type, size, and position of the doors, mechanical properties of hinges, stiffness of robots, as well as the lighting conditions, were randomized via random strategies in each scene. The label for the sampled demonstration is derived as follows: the door handle's pose in the world coordinate system is accessed via the simulation engine interface, and upon acquiring this pose, the ground truth for the current door opening angle is established by applying the predefined offset between the robot end-effector's and the door handle's coordinate systems.

We sampled a total of $47,727$ training demonstrations on the seen-door set and labeled them accordingly. For testing, employing random strategies akin to those used during training, we sampled $4,580$ scenarios on the seen-door set and $4,438$ on the unseen-door set.

V. Benchmark

A. Evaluation Metrics

We propose a set of novel evaluation metrics specifically designed to comprehensively assess the model's performance in safe state planning. These metrics address the shortcomings of existing methods for evaluating safe manipulation, offering a more precise and multifaceted assessment of the model's capabilities.

Success Rate (SuR) Following [18], we leverage SuR to quantify the effectiveness of the state planning model in robotic manipulation tasks, by calculating the proportion of test scenarios in which the model can successfully complete the task out of the total number of test scenarios.

Safety Rate (SaR-95) and Sub-Safety Rate (SaR-80) Safety Rates are utilized to evaluate the scenario-wise force safety of the state planning model in robotic manipulation tasks. A test scenario is considered safely manipulated only if the harmful forces magnitude $\|\mathbf{F}_{\texttt{harmful}}\|$ within the whole continuous manipulation trajectory remain below a specific force magnitude threshold $\mathbf{f}$ at all times. But it is worth mentioning that the force safety of a state planning model only depends on the state generated by itself, rather than other states in that manipulation trajectory. Thus, we discretize the above defined as,

$$\|\mathbf{F}_{\texttt{harmful}}\|^k < \mathbf{f}, \quad \forall k \in [1, L] \tag{6}$$

where L denotes the length of states planned by the model. To eliminate the impact of noise and make the safety evaluation more robust, we relaxed the above definition to the following two metrics: relaxed safety and sub-safety. Concretely, a test scenario can be considered manipulated by the model in a relaxed safe manner when 95% of its generated states have the harmful force below $\mathbf{f}$. Therefore, the safety of the state planning model can be computed as,

$$\texttt{SaR-95} = \frac{\mathbf{Num}_{\texttt{safe}}}{\mathbf{Num}_{\texttt{success}}} \tag{7}$$

where $\mathbf{Num}_{\texttt{safe}}$ denotes the number of test scenarios which is manipulated successfully in a relaxed safe manner, and $\mathbf{Num}_{\texttt{success}}$ denotes the total number of test scenarios which is manipulated successfully. Similarly, a test scenario can be considered manipulated by the model in a sub-safe manner when 80% of its generated states have the harmful force below $\mathbf{f}$. The sub-safety of the state planning model can be computed as,

$$\texttt{SaR-80} = \frac{\mathbf{Num}_{\texttt{sub-safe}}}{\mathbf{Num}_{\texttt{success}}} \tag{8}$$

where $\mathbf{Num}_{\texttt{sub-safe}}$ denotes the number of test scenarios which is manipulated successfully in a sub-safe manner.

Average Harmful Force (AHF) AHF is applied to evaluate the force-wise force safety of the state planning model in robotic manipulation tasks. It is calculated as the average harmful force magnitude $\|\mathbf{F}_{\texttt{harmful}}\|$ applied throughout the robot manipulation process across all test scenarios.

B. Simulation Experiments

Implementation Our proposed SafeDiff model is implemented based on the publicly available Decision Diffuser code base [1]. The training and testing processes are conducted using an NVIDIA A100 Tensor Core GPU. We utilize the training demonstrations provided by our SafeDoorManip50k for safe state planning. The training configuration is as follows: batch size is 256, total training epochs are 500, an initial learning rate of 10^{-4} with a decay rate of 0.985, and the application of an Exponential Moving Average (EMA) with a decay factor of 0.995. During testing, we evaluate the performance of the safe state planning models under $4,580$ seen-door scenarios and $4,438$ unseen-door scenarios in the simulator. Given the scarcity of prior research on door-opening tasks incorporating both tactile and visual information, we compare our method with the transformer-based multi-modal generator [17]. For fairness and practicality, we re-implement the latter without its auditory modality.

Quantitative Results In order to accommodate the limitation of our real experiment, the robot used in our simulated experiment has a fixed base and is stationary. Therefore, the door is considered successfully opened if its angle only surpasses 30°. In addition, we establish 3 levels of force thresholds (i.e. $\mathbf{f} = 5\text{N},\ 10\text{N and } 15\text{N}$) to define SaR-95 and SaR-80 in order to evaluate the force safety performance of such involved states planning models more comprehensively. Tab. I presents the quantitative results of the models in both the seen-door and unseen-door scenarios discussed earlier. As shown, our method outperforms the others across nearly all metrics. This demonstrates that our method effectively ensures force safety during the robotic manipulation process and can generalize robustly to unseen scenarios.

Q1: How does tactile calibration help safe state planning? As tactile calibration plays an essential role in our method, we conduct an ablation study to validate its importance by removing the force feedback input from our method. In the implementation, we directly bypass all operations associated with Eq. 5 during both the training and inference phases. As demonstrated in Tab. I, without tactile calibration, although our method still manages to successfully open doors, it fails to ensure force safety. More importantly, the absence of tactile calibration significantly impairs our method's generalization capabilities, which indicates that vision-based state planning methods are inadequate for modeling the intricate dynamics inherent in robotic manipulation tasks, rendering them incapable of planning robustly in dynamic, unstructured environments.

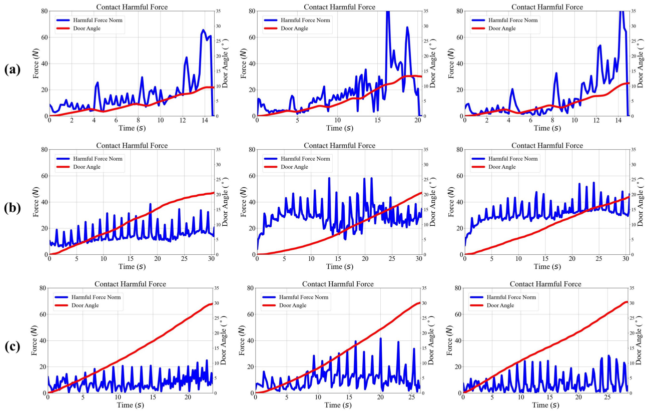

Fig. 3: Qualitative comparisons of different methods on three samples randomly selected from **SafeDoorManip50k unseen-door** scenarios with **disturbance**. Each sub-figure is annotated with two key indicators: the "blue line" represents the magnitude of harmful force applied by the robot's end-effector, demonstrating the safety of the methods, while the "red line" shows the door opening angle, illustrating the effectiveness of the methods. The comparisons include: (a) Li et al. [17], (b) Our method without tactile calibration, and (c) Our method with tactile calibration.

TABLE II: Quantitative evaluation of of different methods on our **SafeDoorManip50k unseen-door** scenarios with **disturbance**, highlighting the anti-disturbance capability of our method.

	SuR (%) ↑	AHF (N) ↓	Threshold - 20 N	
			SaR-95 (%) ↑	SaR-80 (%) ↑
Li et al. [17]	3.83	17.92	0.00	7.83
Ours (V)	68.31	18.37	0.62	45.10
Ours (V+T)	**95.18**	**9.59**	**28.25**	**89.25**

Q2: Does SafeDiff still work under environmental disturbances? The goal of the **disturbance** experiment is to observe whether the state planning methods can counteract the environmental disturbances, preventing their accumulation and ultimately avoiding failure in the robotic manipulation tasks. In the implementation, we tested the involved models using $4,438$ unseen-door scenarios from our SafeDoorManip50k dataset. And during the door-opening process, we applied a periodic impulsive (1.5Hz) disturbance with a positional deviation of 0.03 meters. During evaluation, the force threshold is set to $\mathbf{f} = 20\text{N}$, as external disturbances typically amplify harmful forces. As shown in Fig. 3, trajectories generated by our method and [17] demonstrate that both [17] and our method without tactile calibration fail to resist the disturbance. This leads to an increasing accumulation of positional deviations, eventually causing the robotic gripper to disengage from the door handle. In contrast, our method with tactile calibration responds effectively to the disturbances, maintaining the harmful forces within a relatively small range, and ultimately succeeding in opening the door. Moreover, the quantitative results in Tab. II further confirm that our method can effectively resist disturbances in real time.

C. *Real-world Experiments*

Implementation In the real-world experiments, we constructed three doors with varying colors and radii. One of these doors was utilized for the collection of training data (referred to as the "seen" door), while the remaining two were used for unseen tests. Some door samples are shown in Fig. 4. We deployed our state planning model on the KUKA iiwa14 robot. For input of observation, we obtain visual data from an Intel RealSense D435i camera and force feedback from the robot's interior sensors. Concurrently, we developed a simulated environment within *Isaac Gym* that closely mirrors the actual environment to gather simulation-augmented data for sim2real experiments. The data collection strategies and labeling methods employed in this experiment were broadly consistent with those used in the simulation. Ultimately, we collected 110 real-world demonstrations and 700 simulation demonstrations.

Q1: Can SafeDiff be adapted for real-world robotic manipulation tasks through few-shot fine-tuning? In this experiment, we initially train our model using 700 sampled simulation demonstrations (denoted as Sim), and subsequently fine-tune it with only 20 percent of the 110 real-world demonstrations (denoted as Real (20%)). Fig. 4 demonstrates that our method effectively ensures force safety, even with few-shot fine-tuning.

Q2: How does the generalization performance of SafeDiff in real-world robotic manipulation tasks through few-shot fine-tuning? We continue to employ the few-shot fine-tuned model as the controller for the robot. We then ask the robot to open doors that are unseen during the fine-tuning process. Fig. 4 demonstrates that our method exhibits robust generalization capabilities in real-world robotic manipulations.

Q3: Does SafeDiff still work under real-world environmental disturbances through few-shot fine-tuning? We continue to employ the previously trained model as the robot's controller. However, unlike in the above experiment, we manually introduce external disturbances during the door-opening process. From Fig. 4, it is evident that our method can effectively calibrate real-world disturbances online, maintaining the harmful force at a low level.

VI. Conclusions

In this work, we introduce a novel benchmark dedicated to ensuring force safety in robotic manipulation, focusing specifically on manipulation tasks where the robot's motion trajectory is constrained by the physical properties of the manipulated objects, such as door-opening. Drawing inspiration from bionics, we developed a diffusion-based model named **SafeDiff**, which adeptly integrates real-time tactile feedback to adjust vision-guided planned states, significantly reducing the risk of damage. Additionally, we present the **SafeDoorManip50k** dataset, a pioneering resource that provides a large-scale multimodal environment tailored for safe manipulation. This dataset focuses on the collection of force feedback during robotic manipulation in simulation settings, offering valuable insights that can inspire subsequent tasks. Our experiments demonstrate the robust performance of SafeDiff in ensuring safe robotic manipulation.

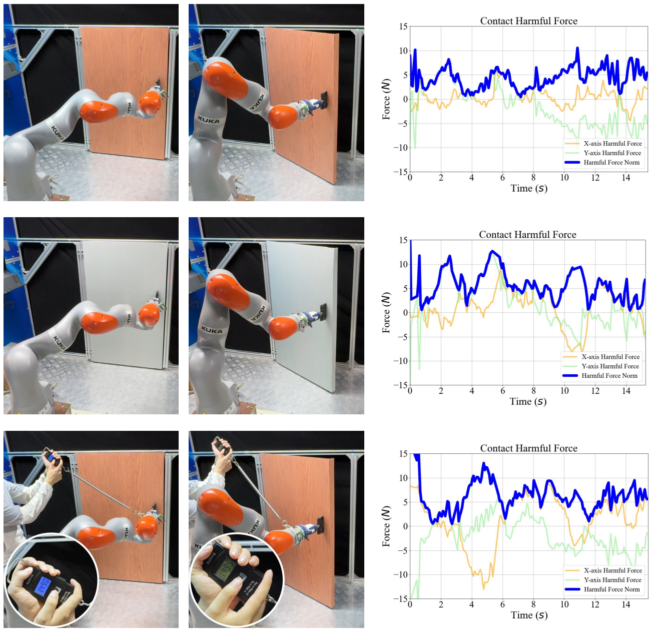

Fig. 4: Qualitative results of our method in real-world scenarios. Each row corresponds to a specific door-opening task: The first row evaluates the effectiveness of our few-shot fine-tuning model in real-world settings (relevant to **Q1**), the second row assesses the model's generalization capabilities (relevant to **Q2**), and the third row examines the model's resistance to disturbances (relevant to **Q3**). Additionally, the first three columns in each row capture two samples from the door-opening process, while the final column quantifies the magnitude of harmful force encountered throughout the entire door-opening. Zoom in 10 times for the better view.

Limitations. Given the cost of data collection for simulation and real-world experiments, our experiments are solely conducted on the door-opening task and have not yet been extended to other manipulation tasks. We only consider a gripper rather than a dexterous hand to manipulate objects. However, we hope that our definition of the evaluation metric, data collection scheme, and model design can stimulate more extensive research in related fields.

Acknowledgements

Special thanks to Prof. Yanding Wei for real-world experiments support and helpful feedback on writing.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

[2] Boshi An, Yiran Geng, Kai Chen, Xiaoqi Li, Qi Dou, and Hao Dong. Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7748–7755. IEEE, 2024.

[3] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[4] Yasemin Bekiroglu, Dan Song, Lu Wang, and Danica Kragic. A probabilistic framework for task-oriented grasp stability assessment. In *2013 IEEE International Conference on Robotics and Automation*, pages 3040–3047. IEEE, 2013.

[5] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[6] Shaowei Cui, Rui Wang, Junhang Wei, Jingyi Hu, and Shuo Wang. Self-attention based visual-tactile fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters*, 5(4):5827–5834, 2020.

[7] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE, 2023.

[8] Hao Dang and Peter K Allen. Learning grasp stability. In *2012 IEEE International Conference on Robotics and Automation*, pages 2392–2397. IEEE, 2012.

[9] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022.

[10] Justin Fu, Sergey Levine, and Pieter Abbeel. One-shot learning of manipulation skills with online dynamics adaptation and neural network priors. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4019–4026. IEEE, 2016.

[11] Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2988, 2023.

[12] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[13] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023.

[14] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-nerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*, 2021.

[15] Mrinal Kalakrishnan, Ludovic Righetti, Peter Pastor, and Stefan Schaal. Learning force control policies for compliant manipulation. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4639–4644. IEEE, 2011.

[16] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[17] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

[18] Yu Li, Xiaojie Zhang, Ruihai Wu, Zilong Zhang, Yiran Geng, Hao Dong, and Zhaofeng He. Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. *arXiv preprint arXiv:2403.02604*, 2024.

[19] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022.

[20] Jun Lv, Qiaojun Yu, Lin Shao, Wenhai Liu, Wenqiang Xu, and Cewu Lu. Sagci-system: Towards sample-efficient, generalizable, compositional, and incremental robot learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 98–105. IEEE, 2022.

[21] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[22] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[23] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In *International Symposium on Experimental Robotics*, pages 375–386. Springer, 2018.

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[25] Zhe Su, Karol Hausman, Yevgen Chebotar, Artem Molchanov, Gerald E Loeb, Gaurav S Sukhatme, and Stefan Schaal. Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 297–303. IEEE, 2015.

[26] Jaeyong Sung, J Kenneth Salisbury, and Ashutosh Saxena. Learning to represent haptic feedback for partially-observable tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2802–2809. IEEE, 2017.

[27] Herke Van Hoof, Nutan Chen, Maximilian Karl, Patrick van der Smagt, and Jan Peters. Stable reinforcement learning with autoencoders for tactile and visual data. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3928–3934. IEEE, 2016.

[28] Herke Van Hoof, Tucker Hermans, Gerhard Neumann, and Jan Peters. Learning robot in-hand manipulation with tactile features. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 121–127. IEEE, 2015.

[29] Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, and Kui Jia. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. *Advances in Neural Information Processing Systems*, 33:13174–13184, 2020.

[30] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *arXiv preprint arXiv:2106.14440*, 2021.

[31] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[32] Zhenjia Xu, Zhanpeng He, and Shuran Song. Universal manipulation policy network for articulated objects. *IEEE robotics and automation letters*, 7(2):2447–2454, 2022.

[33] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.