2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

UNIC: Learning Unified Multimodal Extrinsic Contact Estimation

Zhengtong Xu and Yuki Shirai

Abstract— **Contact-rich manipulation requires reliable estimation of extrinsic contacts—the interactions between a grasped object and its environment—which provide essential contextual information for planning, control, and policy learning. However, existing approaches often rely on restrictive assumptions, such as predefined contact types, fixed grasp configurations, or camera calibration, that hinder generalization to novel objects and deployment in unstructured environments. In this paper, we present UNIC, a unified multimodal framework for extrinsic contact estimation that operates without any prior knowledge or camera calibration. UNIC directly encodes visual observations in the camera frame and integrates them with proprioceptive and tactile modalities in a fully data-driven manner. It introduces a unified contact representation based on scene affordance maps that captures diverse contact formations and employs a multimodal fusion mechanism with random masking, enabling robust multimodal representation learning.**

Extensive experiments demonstrate that UNIC performs reliably. It achieves a 9.6 mm average Chamfer distance error on unseen contact locations, performs well on unseen objects, remains robust under missing modalities, and adapts to dynamic camera viewpoints. These results establish extrinsic contact estimation as a practical and versatile capability for contact-rich manipulation. The overview and hardware experiment videos are here.

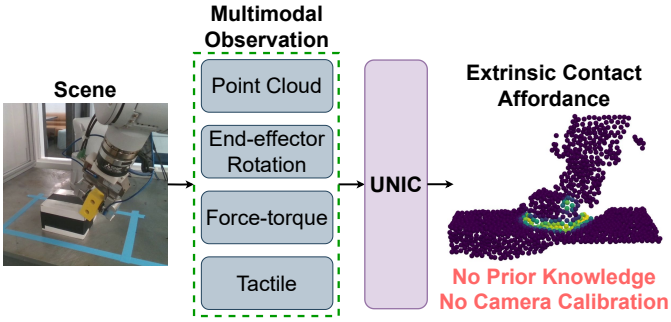

Fig. 1: UNIC leverages multimodal inputs to estimate a unified contact affordance map that captures diverse forms of extrinsic contact, including complex interaction chains such as gripper–object–object–environment. Throughout this process, UNIC does not rely on prior knowledge or camera calibration. At deployment, it remains effective even under missing modalities, adapts to dynamic camera viewpoints, and generalizes well to unseen objects. The overview and hardware experiment videos are here.

I. Introduction

Extrinsic contact [1] refers to contact events involving external objects and the environment, beyond the robot's own body. Estimating extrinsic contact is particularly challenging [2], as it requires reasoning about interactions that are often indirect, involve multiple objects, and occur under uncertain geometry or dynamics. Nevertheless, extrinsic contact estimation is essential for enhancing robotic dexterity, as it enables robots to perceive and reason about diverse interactions with the environment. From tool use [3], [4] and non-prehensile manipulation [5] to precise environment interactions [6], reliable extrinsic contact estimation supports robust performance while ensuring safety and stability in diverse, unstructured settings.

Recent research has investigated tactile-based and multimodal approaches for extrinsic contact estimation. However, these methods often depend on strong priors or restrictive assumptions, such as predefined contact types [1], [7], [8], initial contact conditions [9], fixed grasps without slip [10], object geometries [11], camera calibration [11], [8], or fixed camera placement [12]. These constraints limit their practicality in real-world scenarios, hinder the development of a unified framework, and restrict deployment in diverse, unstructured environments. Moreover, these limitations diminish the feasibility of leveraging extrinsic contact estimation as a general-purpose capability for downstream applications, such as integrating contact estimation with policy learning.

In this paper, we present UNIC, a unified framework for extrinsic contact estimation, as shown in Fig. 1. The main contributions of this work are as follows:

1. **Prior-free framework:** We present a unified framework for extrinsic contact estimation that requires neither prior knowledge nor camera calibration. Visual observations are encoded directly in the camera frame and fused with proprioceptive and tactile inputs in a fully data-driven manner. This design supports flexible deployment under varying camera viewpoints and generalizes effectively to diverse scenarios, including previously unseen objects.

2. **Unified contact representation:** We introduce a unified contact representation based on scene affordance maps, which captures transitions from no-contact to contact states and accommodates diverse contact types, including point, line, and patch. Moreover, this representation explicitly models complex contact chains, extending beyond grasped object–environment interactions to encompass contacts among multiple objects in the scene.

3. **Unified multimodal fusion:** We propose a unified multimodal fusion mechanism that randomly masks feature tokens during training, encouraging the model to learn robust cross-modal representations. Consequently, UNIC can deliver reliable estimation even when one or more modalities are missing at deployment, thereby greatly improving the flexibility and practicality of real-world deployment.

Zhengtong Xu is with the Edwardson School of Industrial Engineering, Purdue University, West Lafayette, USA xu1703@purdue.edu
Yuki Shirai is with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA shirai@merl.com

II. Related Work

In this section, we review prior work on extrinsic contact estimation, which can be broadly categorized into model-based and model-free methods.

Model-based methods rely heavily on underlying physical models [1], [4], [13], [8], [3], [14]. These approaches collect data during manipulation and apply optimization techniques subject to physical constraints to infer unobservable contacts. For example, [4] employs nonlinear optimization with tactile sensing to estimate slip during tool manipulation, while [3] leverages factor graphs with tactile input to estimate contacts.

Model-free methods [10], [12], [11], [7], [15], [16] directly map multimodal sensory inputs to contact states using neural networks. For instance, [11] trains a vision–tactile estimator in simulation and validates it in real settings, and [16] infers contact solely from joint torques and positions.

While these approaches have demonstrated promising results, they often depend on strong priors or restrictive assumptions that limit their practical deployment. Examples include predefined contact types [1], [7], [8], assumptions on initial contact conditions [9], fixed grasp configurations without slip [10], access to precise object geometries [11], strict camera calibration [11], [8], or fixed viewpoints [12].

In contrast, our work introduces a unified, prior-free framework for extrinsic contact estimation that is robust, deployable, and generalizable across diverse settings.

A related direction bypasses explicit contact estimation by learning end-to-end policies that map multimodal inputs directly to actions [17], [18], [19], [20], [21], [22]. While effective, these policies lack interpretability. In addition, they are task-specific, and require retraining for new tasks. In contrast, our task-agnostic estimator generalizes across conditions and provides a reusable capability that can interface with diverse manipulation policies.

III. Method

In this section, we present the details of UNIC.

A. Background: Extrinsic Contact

Extrinsic contact requires reasoning about interactions such as gripper–object, object–object, and object–environment, where object geometry and pose are often unknown [1], [2].

Previous work defined extrinsic contact mainly as gripper–object–environment interactions. Our formulation generalizes this to a unified representation that also captures complex cases such as multi-object manipulation (e.g., gripper–object–object–environment), where contacts may occur simultaneously across multiple objects and the environment.

B. Overview of UNIC

The architecture of UNIC is illustrated in Fig. 2. The objective of UNIC is to estimate extrinsic contact by **(i)** leveraging multimodal inputs, including point clouds, tactile signals, force–torque, and end-effector rotation, and **(ii)** avoiding reliance on additional information such as pre-constructed object geometries or camera calibration.

C. Prior-free Contact Affordance Representation

We introduce our prior-free contact affordance representation, a unified formulation capable of capturing diverse contact types, including complex chains such as gripper–object–object–environment, without requiring priors like camera calibration or pre-defined object geometries.

We represent extrinsic contact as an affordance map over the entire scene. As shown in Fig. 3, UNIC directly uses the camera-frame point cloud as the reference for constructing this affordance map, without relying on ground-truth information such as the object's geometries. Although this design inevitably sacrifices some precision and completeness due to the limited resolution and coverage of point clouds, it eliminates the need for constraints such as camera calibration and ensures that the learning process remains independent of additional priors from multimodal sensory inputs. This paradigm enables flexible deployment, allowing cameras to be positioned freely or moved dynamically during estimation. Notably, the same paradigm of using camera-frame point cloud observations has also demonstrated strong performance and generalization in policy learning [23], [24].

For point clouds, extrinsic contact is annotated by having humans select contact points directly in 3D space. We will provide more details in Section V-A.3 on how this annotation process can be performed both quickly and reliably. Regardless of the contact type, this point-based annotation offers a consistent way to represent contacts, as shown in Fig. 3. We denote the set of N annotated points as $\{p_c^i\}_{i=1}^N$, where $p_c^i \in \mathbb{R}^3$ represent the i-th annotated point. Importantly, N is not fixed; allowing it to vary across point clouds makes the annotation process highly flexible and simple.

However, human annotations often exhibit non-uniform and sparse density across different regions, making them unsuitable as a dense supervision signal. To generate a uniform affordance representation from these annotated points, we adopt the following procedure, as shown in Fig. 3. For each point $p^k \in \mathbb{R}^3$, $k = 1, \ldots, M$, in the point cloud captured by an RGB-D camera, we compute its minimum distance to the annotated set $\{p_c^i\}_{i=1}^N$:

$$d_k = \min_{i=1,\ldots,N} \|p^k - p_c^i\|_2.$$

We then transform this distance into an affordance score using a Gaussian kernel:

$$y_k = \exp\left(-\frac{d_k^2}{2\sigma^2}\right).$$

Here, σ controls the kernel bandwidth. Since the kernel value y_k lies in $(0, 1]$, we further normalize it to match the affordance label range used during training. Specifically, we scale by a factor s, clamp to $[0, 1]$, and linearly rescale to the range $[-1, 1]$:

$$a_k = 2 \cdot \mathrm{clamp}(s \cdot y_k, 0, 1) - 1,$$

where the scaling factor s controls the effective range before clamping. This process yields an affordance value a_k for each point p^k, where points in contact regions receive higher

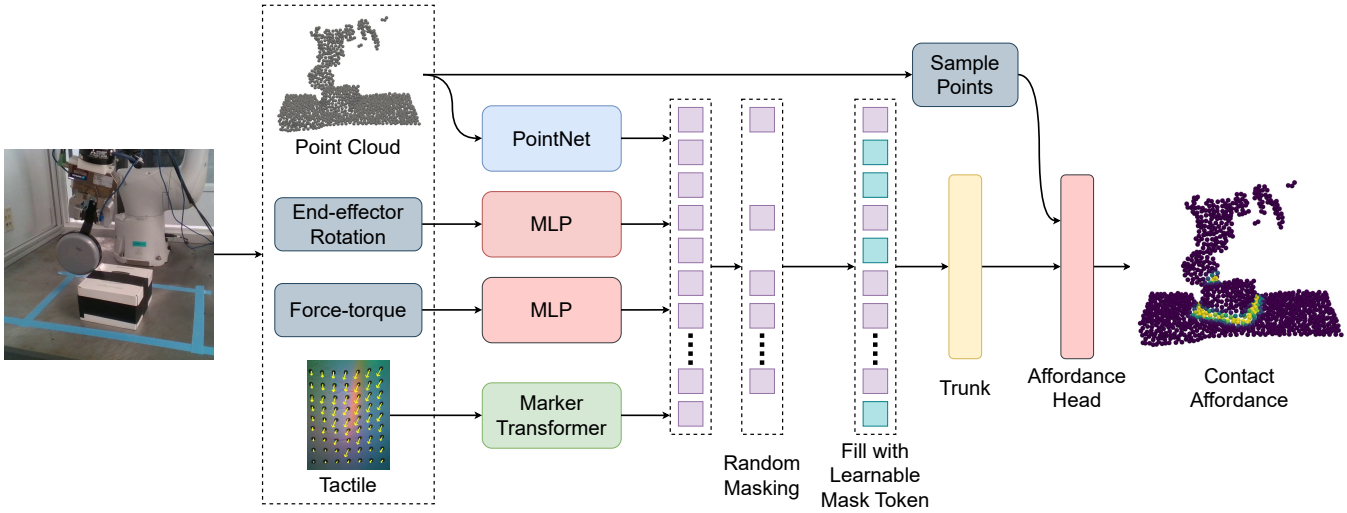

Fig. 2: Pipeline of UNIC. UNIC integrates four sensing modalities as inputs—point clouds, end-effector rotation, force–torque, and tactile marker displacements—and outputs an extrinsic contact affordance map. We adopt a masked multimodal fusion strategy to ensure a robust multimodal representation learning. In addition, UNIC employs a sampling strategy designed to enhance computational efficiency.

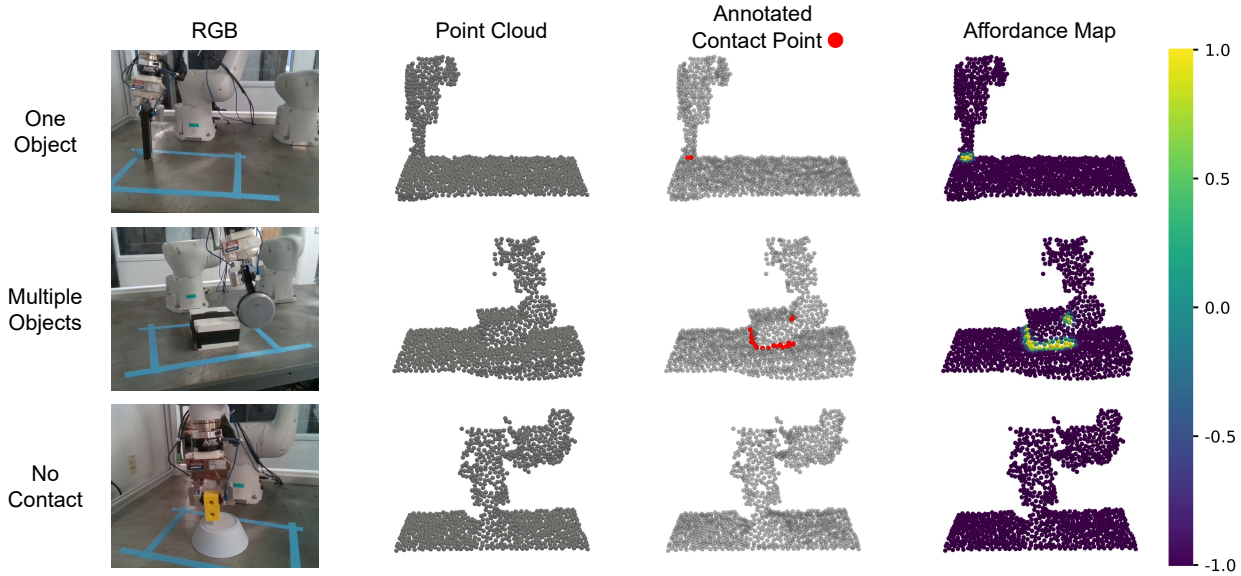

Fig. 3: Illustration of proposed prior-free contact affordance representation over three different cases. For each case, we show the RGB image and the corresponding point cloud. The human operator annotates the contact points. Based on these annotations, the proposed Gaussian kernel-based generation method is applied to produce the contact affordance map. Affordance values range from -1 to 1, with higher values at contact locations and lower values in non-contact regions.

affordance values and those in non-contact regions receive lower values. As shown in Fig. 3, the affordance unifies contact and non-contact regions with minimal annotation effort and without requiring uniform human labeling.

D. Multimodal Encoding

We motivate our modality choices and describe how each input is encoded. In particular, we integrate four input streams: tactile sensing, end-effector rotation, force–torque sensing, and camera-frame point clouds.

1) Selection of Modalities: For tactile signals, we use marker displacement maps to capture the spatial distribution of shear across the finger, as shown in Fig. 2. In our setup, two marker displacement maps are obtained from the parallel gripper's two tactile sensors, which are GelSight Minis [25]. Compared with other tactile representations such as raw images, depth maps, or binarized tactile maps, this representation provides a more unified and information-rich basis for contact estimation [1]. The tactile data are structured in a tensor of size $(H, W, 2)$, where H and W denote the numbers of markers along the two axes, and the last dimension stores the x and y displacement components.

For the end-effector state, we use rotation rather than position or the full 6D pose. This design choice is motivated by the fact that rotation defines the coordinate frames of the force–torque and tactile signals, whereas position mainly introduces fixed geometric priors that are redundant in our setup. We represent rotation using quaternions. The

force–torque modality is a 6D wrench from a wrist-mounted sensor. Point clouds are obtained in the camera frame from an RGB-D camera with variable placement, meaning the camera is not constrained to a single configuration during data collection and model deployment.

2) Encoders: Each modality is processed by a dedicated encoder to the latent space as illustrated in Fig. 2.

Tactile encoder: Tactile inputs are processed by a transformer encoder tailored to marker displacements. Specifically, we patchify the $(H, W, 2)$ map via a convolutional projection with a stride equal to the patch size, flatten the resulting patches into a sequence of tokens, and apply a transformer encoder over this sequence [26]. This yields a set of tactile tokens. In our setup, marker displacement maps from the two fingertips are independently encoded and then aggregated. This design suits marker displacements well: local correlations capture fine-grained contact patterns, while self-attention models long-range effects such as shear fields.

End-effector rotation and force–torque encoders: End-effector rotation and force–torque signals are each passed through lightweight MLPs to produce compact token sets.

Point cloud encoder: Point clouds are processed with a PointNet encoder [27] to extract geometric features. These features are then pooled to a fixed number of tokens by averaging over downsampled subsets.

3) Shared Tokenization: All modality-specific tokens are projected to a common embedding dimension and balanced to the same per-modality token count. For tactile, we flatten all patch tokens and apply a linear projection to a fixed number of tokens. For end-effector rotation and force-torque, linear projections reshape MLP features into the same token layout. For point clouds, the pooled encoder outputs are projected to the shared token dimension. This balancing ensures comparable contribution from each modality regardless of native resolution or feature size.

E. Masked Multimodal Fusion

We propose a masked multimodal fusion strategy to enhance the robustness and flexibility of multimodal representation. The key idea is to randomly mask a subset of latent tokens during training and replace them with a learnable mask token, which is then optimized jointly with the rest of the network, as shown in Fig. 2. Both modality-specific tokens and mask tokens are then fed into the modality-agnostic transformer trunk.

This design brings three main benefits: **(i)** it simulates sensor dropouts or partial observations during training, forcing the model to leverage complementary signals; **(ii)** the learnable mask token provides a modality-agnostic prior that allows the transformer trunk to treat missing information properly rather than noisy or zero-padded inputs; and **(iii)** it ensures that the model naturally adapts to incomplete multimodal observations at deployment.

As described in Section III-D, all modalities are encoded into a balanced tokenized representation. During training, we randomly mask a portion of these tokens according to a specified ratio, and in our experiments we set the masking ratio to 0.5. Masked positions are replaced with the learnable mask token, yielding a consistent placeholder that the modality-agnostic transformer trunk can interpret reliably across training iterations. The learnable mask token is updated through training. The fused token sequence is then processed by the trunk, which performs cross-modal reasoning and produces a compact latent representation.

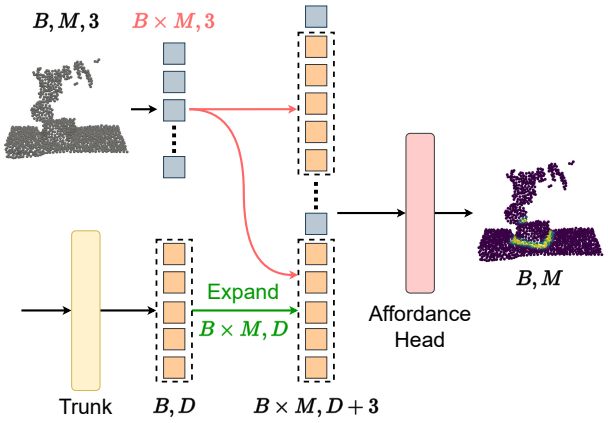

Fig. 4: Illustration of the sampling process. The gray blocks represent flattened sample points obtained from the camera-captured point cloud, while the orange blocks represent the multimodal feature generated by the trunk. Here, B denotes the batch size, M denotes the number of points in point cloud/sample points, and D denotes the dimensionality of the multimodal feature. See Section III-F for details.

At deployment, random masking is disabled: all available modality tokens are passed directly into the transformer trunk. Thanks to this masking design, we can even discard one or more modalities at deployment while the system still functions reliably. If a modality is missing, its tokens are replaced with the learnable mask token, ensuring distributional consistency between training and inference. Consequently, the model remains reliable under sensor failures or missing inputs, without requiring retraining.

This design makes the overall system highly practical for real-world robotics. Since the trunk has been trained to process both modality tokens and mask tokens, it maintains stability under unpredictable sensing missing and supports flexible deployment across diverse hardware configurations.

F. Efficient Sampling

We present an efficient sampling strategy that decouples global multimodal fusion from point-wise affordance generation, thereby accelerating the forward pass. The transformer trunk shown in Fig. 2 and Fig. 4 first produces a global multimodal feature of shape (B, D), where B is the batch size and D the feature dimension. Meanwhile, the point cloud of size M is retained as the set of sample points. The fused feature is then broadcast to these sample points and concatenated with their 3D coordinates, enabling lightweight point-wise inference. As illustrated in Fig. 4, this design pushes only lightweight computation to the point-wise head, resulting in a streamlined data flow: global fused feature (B, D), broadcast to $(B \times M, D)$, concatenated with sample

Fig. 5: Objects used in our experiments. The left set shows seen objects, while the right set shows unseen objects used only during validation.

points $(B \times M, D{+}3)$, processed by the affordance head to $(B \times M, 1)$, and finally reshaped into the affordance map (B, M).

IV. Baselines

In this section, we introduce three baselines which are benchmarked with UNIC in experiments.

A. UNIC *without Masked Fusion*

This baseline uses the same modalities, encoders, and sampling strategy as UNIC model, but removes the masked multimodal fusion. During both training and inference, features from all modalities are directly concatenated and fed into the transformer trunk without any masking.

B. End-to-end Regression

End-to-end regression employs the same modalities and encoders but omits masked fusion, directly mapping the concatenated features to L points as the predicted contact patch. Training is supervised with the Chamfer distance loss between predicted and ground-truth patches. To generate ground-truth patches, we first construct a dense affordance map, select points with positive affordance, and downsample them to L points as ground-truth labels. For non-contact cases, all predicted points are set to zero [16].

C. Visual Estimation

This baseline keeps the same sampling strategy but removes masked fusion and relies only on point cloud and end-effector rotation, excluding force–torque and tactile inputs.

V. Experiments

In this section, we present our experimental setup and results. Our evaluation is designed to address the following key questions:

1) Can UNIC generalize to diverse contact locations, varying object configurations, different camera viewpoints, and even entirely unseen objects?
2) If models are trained with full modality inputs but some modalities are removed only at test time, can different methods still function effectively?
3) What insights can be drawn from the results regarding the relative importance of different modalities?
4) Is UNIC capable of supporting real-time deployment?

In our experiments, to ensure statistical significance of the results, we performed validation of all metrics every 10 epochs during training. For reporting, we averaged the results over the last 10 validations during training. This entire process was repeated with three different random seeds, and the results were obtained by averaging across these seeds.

A. Setup

1) Hardware: For data collection, we use a 6-DoF Mitsubishi MELFA robot equipped with a wrist-mounted force–torque sensor and a WSG-32 gripper, with each finger instrumented with a GelSight Mini. The robot operates under a stiffness controller to enable compliant interactions. An Intel RealSense D435 RGB-D camera is used to capture point cloud observations.

2) Objects: As illustrated in Fig. 5, we employ two sets of objects for data collection. The seen objects are used to gather training data, and based on these objects we also construct a test set consisting of seen objects with unseen contact locations. The unseen objects are used exclusively to build another test set containing entirely novel objects that the model has never encountered. In Fig. 5, the top row shows the grasped objects, while the bottom row shows the objects placed on the table to provide contact surfaces.

3) Data Collection: All data streams were uniformly recorded at 10 Hz. Point clouds are captured with an flexible RGB-D camera, then cropped and downsampled to 1,024 points. The cropping is governed by a fixed set of parameters and applied identically across all data and inference runs.

We use an episode as the basic unit, where each episode is a continuous data chunk lasting a few seconds. For each episode, we consider two cases: **(i)** contact at a fixed location, where the robot varies force and pose (e.g., rotation around the contact point). The contact point remains fixed within an episode but differs across episodes. **(ii)** no contact, where the end-effector moves freely. Although multimodal observations vary frame by frame, annotation is required only once per episode, as the fixed contact points generate affordance maps for all frames. Because contact locations vary across episodes and non-contact episodes involve unconstrained motion, episodes are naturally independent.

To avoid data leakage, we split training and validation sets for seen objects at the episode level, with a ratio of 0.8:0.2. The resulting numbers of episodes and frames are reported in Table I. For unseen objects, we collect both contact and non-contact episodes but use them only for validation. The whole validation set thus consists of: **(i)** unseen contact locations on seen objects, testing generalization to new configurations, and **(ii)** entirely unseen objects, testing generalization to new categories and shapes.

Each validation set (seen and unseen) is divided into all contact and no contact subsets, whose union forms the full set. Within the all contact subset, we further isolate single-object tabletop contacts for controlled evaluation, referred to

TABLE I: Number of Episodes/Frames in Each Set. We split the dataset into training and validation sets with a ratio of 0.8:0.2. See Section V-A.3 for more details about the dataset.

	All	Train	Valid	All Contact	Single Contact	No Contact
Seen	1500/32274	1200/25751	300/6523	272/4476	131/1810	28/2047
Unseen	150/4555	0/0	150/4555	120/3250	60/1574	30/1305

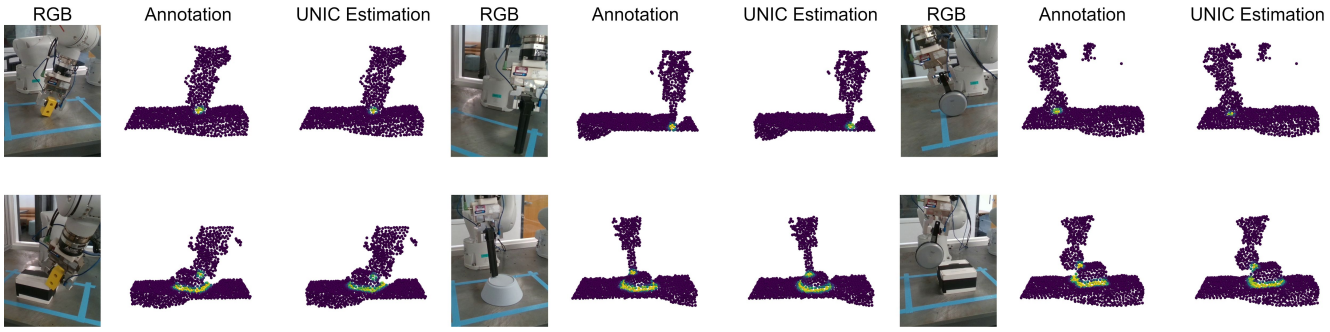

Fig. 6: UNIC inference results on seen objects under unseen contact locations. UNIC shows accurate extrinsic contact estimation across diverse contact configurations.

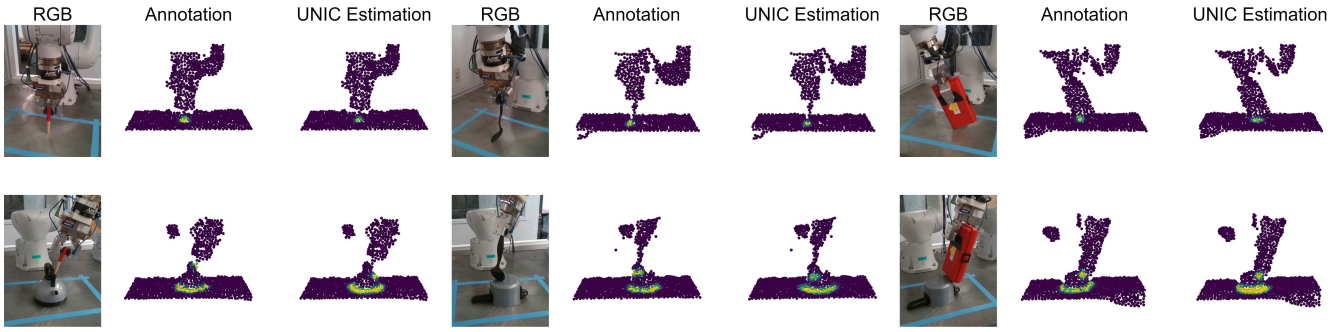

Fig. 7: UNIC inference results on unseen objects, whose data were not used during training, as detailed in Sec. V-A.2. UNIC demonstrates strong generalization to previously unseen objects.

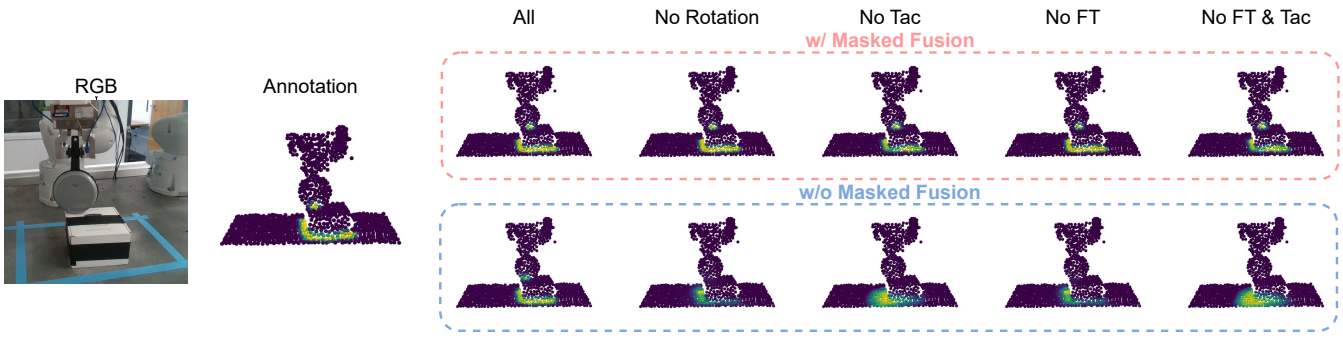

Fig. 8: Inference results with and without masked fusion under test-time modality removal. UNIC with masked fusion demonstrates more robust predictions compared to the variant without masked fusion. This shows that masked fusion enables flexible test-time configurations of modality inputs without retraining the model, while maintaining performance.

as the single contact set. These partitions are constructed alongside annotation, ensuring consistency with semantic labels for analysis. Details of the dataset are provided in Table I.

B. Metrics

We evaluate extrinsic contact estimation using three metrics, corresponding to the all contact, single contact, and no contact validation sets in Table I.

In the all contact setting, contact is uniformly represented as a point cloud, allowing diverse forms of contact to be expressed as patches of points. The predicted patch is compared with the ground truth using Chamfer distance in mean absolute error. The results are summarized in Table II.

The single contact set covers cases where the grasped object touches the table. A single representative point, obtained by averaging the contact patch, is compared with the ground-truth point using mean absolute error. The results are summarized in Table III.

The no contact set includes cases where the grasped object does not touch the environment. The model is expected to output an affordance map of all -1 values, and performance is

TABLE II: Mean Absolute Error (Chamfer Distance, mm) of Contact Patch Estimation on the All Contact Validation Set.

	Seen Objects					Unseen Objects				
	All	Test-time Modality Removal				All	Test-time Modality Removal			
		Rotation	Tac	FT	FT & Tac		Rotation	Tac	FT	FT & Tac
UNIC	9.6	**9.4**	**9.7**	**9.7**	**9.8**	**33.4**	32.5	**33.5**	33.5	**32.9**
UNIC w/o Masked Fusion	**7.9**	15.5	199.2	15.1	210.2	**33.4**	**28.4**	215.4	**27.9**	230.3
E2E Regression	32.8	33.3	36.9	31.8	34.9	45.3	45.1	47.0	44.6	45.2
Visual Estimation	17.9	18.0	-	-	-	54.0	55.4	-	-	-

TABLE III: Mean Absolute Error (Distance, mm) of Contact Point Estimation on the Single Contact Validation Set.

	Seen Objects					Unseen Objects				
	All	Test-time Modality Removal				All	Test-time Modality Removal			
		Rotation	Tac	FT	FT & Tac		Rotation	Tac	FT	FT & Tac
UNIC	16.7	**15.4**	**16.2**	**15.5**	**16.8**	**23.4**	**23.0**	**25.3**	**23.8**	**25.1**
UNIC w/o Masked Fusion	**14.8**	18.8	118.4	18.2	128.0	26.4	27.8	125.7	28.6	132.4
E2E Regression	16.1	16.4	23.1	16.6	23.0	25.3	24.6	26.7	24.7	27.2
Visual Estimation	22.2	22.2	-	-	-	40.0	41.0	-	-	-

TABLE IV: Mean Absolute Error of Non-Contact Affordance Estimation. Since E2E regression does not predict affordance, it is omitted from this metric.

	Seen Objects					Unseen Objects				
	All	Test-time Modality Removal				All	Test-time Modality Removal			
		Rotation	Tac	FT	FT & Tac		Rotation	Tac	FT	FT & Tac
UNIC	0.024	**0.050**	**0.022**	**0.088**	**0.092**	0.018	**0.012**	**0.014**	**0.058**	**0.076**
UNIC w/o Masked Fusion	**0.006**	0.081	0.092	0.106	0.104	**0.004**	0.046	0.086	0.088	0.098
Visual Estimation	0.054	0.054	-	-	-	0.051	0.048	-	-	-

TABLE V: Average inference time over 100 forward passes under different test-time modality drop conditions.

All	**No Rotation**	**No Tac**	**No FT**	**No FT & Tac**
0.0015 s	0.0015 s	0.0011 s	0.0015 s	0.0010 s

measured by the mean absolute error between the predicted and ground-truth affordance maps. The results are summarized in Table IV.

For both the all contact and single contact sets, contact patches are obtained by thresholding the affordance map: points with values greater than zero are selected as the contact point cloud.

C. Results

The quantitative results are summarized in Tables II–IV. From these results, we draw the following conclusions:

1. Performance on seen objects with unseen contact locations: As shown in Tables II and III, UNIC achieves a Chamfer distance error of 9.6 mm for contact patch estimation and a 16.7 mm distance error for single-object contact estimation. Importantly, these results are obtained under randomized camera viewpoints and without relying on any prior knowledge, demonstrating the impressive generalization capability of our approach. Representative visualization results are shown in Fig. 6. In our supplementary video, we present additional real-time inference results, including estimation on unseen contact locations, transitions from non-contact to contact states, and robust estimation under dynamic camera movements.

2. Performance on unseen objects: Even when deployed on entirely unseen objects, UNIC exhibits acceptable performance, as shown in Tables II and III, with corresponding qualitative results illustrated in Fig. 7. These results suggest that UNIC can make reasonable predictions for previously unseen objects. This generalization stems from the fact that the input modalities used by UNIC—point clouds and marker displacement maps—primarily capture geometric and interaction-related information, rather than relying on object-specific features. For objects with significantly different geometries, especially in multi-object contact scenarios, Fig. 7 shows that UNIC's predictions, while not always precise, remain spatially consistent and reasonable.

3. Robustness of masked fusion on UNIC: As shown in Tables II and III, and Fig. 8, we also observe that UNIC without masked fusion achieves higher performance than UNIC when all modalities are present at test time. However, once a modality is removed during inference, its performance degrades substantially, with the drop being especially pronounced for high-dimensional inputs such as tactile marker displacement maps. This indicates that while masked fusion sacrifices some global performance, it greatly enhances the robustness of multimodal representations. Moreover, it introduces a new capability—allowing test-time modalities to be flexibly reconfigured without severely compromising performance—thereby making deployment both more robust and more adaptable.

4. Role of tactile vs. force-torque: From the evaluation on the all contact dataset, tactile marker tracking proves more beneficial for fine-grained estimation, as removing tactile input leads to a larger performance drop compared to removing force–torque, as shown in Table III. In contrast, force–torque signals play a more critical role in distinguishing whether contact occurs: on the no contact validation set, removing force–torque causes a significant degradation in performance, as shown in Table IV. This is intuitive, since tactile marker tracking provides richer and more detailed feedback, implicitly encoding fine-grained interaction and spatial information

that supports accurate estimation. Force–torque, on the other hand, is a low-dimensional signal that is more directly informative for detecting the presence or absence of contact.

5. Comparison with baselines: Both end-to-end regression and vision estimation perform worse than UNIC and its ablations. For end-to-end regression, directly predicting contact patches without sampling makes it more difficult to capture the distribution of contacts, leading to substantially higher learning difficulty. For vision estimation, despite using the same architecture, the absence of force–torque and tactile inputs results in a pronounced drop in accuracy. This highlights the critical importance of multimodal inputs for precise extrinsic contact estimation.

6. Inference time: We measure the inference time of UNIC on a single RTX 3080 GPU, both with all modalities and under different test-time modality drop conditions, as shown in Table V. Each reported value corresponds to generating a full contact affordance map from a point cloud of 1,024 points. UNIC runs at over 600 Hz, demonstrating real-time efficiency.

VI. Discussion

We introduced UNIC, a unified multimodal framework for extrinsic contact estimation that eliminates reliance on prior knowledge and camera calibration. By directly leveraging multimodal inputs, UNIC generalizes across diverse contact formations, remains robust under missing modalities, and demonstrates effectiveness across challenging scenarios.

Looking ahead, UNIC can be advanced in three main directions. First, scalable multimodal data generation through simulation and sim-to-real transfer can overcome the limitations of human annotations and the scarcity of real-world data. Second, integrating vision [28] or vision-language foundation models [29] may strengthen multimodal alignment and contextual reasoning, enabling richer open-world understanding for contact estimation. Finally, coupling UNIC with contact-rich manipulation policies [6] could provide informative conditions for learning and control, paving the way for more adaptive and general-purpose robotic manipulation.

References

[1] D. Ma, S. Dong, and A. Rodriguez, "Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 11 262–11 268.

[2] A. Rodriguez, "The unstable queen: Uncertainty, mechanics, and tactile feedback," *Science Robotics*, vol. 6, no. 54, p. eabi4667, 2021.

[3] S. Kim, A. Bronars, P. Patre, and A. Rodriguez, "Texterity–tactile extrinsic dexterity: Simultaneous tactile estimation and control for extrinsic dexterity," *arXiv preprint arXiv:2403.00049*, 2024.

[4] Y. Shirai, D. K. Jha, A. U. Raghunathan, and D. Hong, "Tactile tool manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 597–12 603.

[5] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[6] C. Higuera, J. Ortiz, H. Qi, L. Pineda, B. Boots, and M. Mukadam, "Perceiving extrinsic contacts from touch improves learning insertion policies," *arXiv preprint arXiv:2309.16652*, 2023.

[7] K. Ota, D. K. Jha, K. M. Jatavallabhula, A. Kanezaki, and J. B. Tenenbaum, "Tactile estimation of extrinsic contact patch for stable placement," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 876–13 882.

[8] M. Oller, D. Berenson, and N. Fazeli, "Tactile-driven non-prehensile object manipulation via extrinsic contact mode control," *arXiv preprint arXiv:2405.18214*, vol. 3, 2024.

[9] S. Kim, D. K. Jha, D. Romeres, P. Patre, and A. Rodriguez, "Simultaneous tactile estimation and control of extrinsic contact," *arXiv preprint arXiv:2303.03385*, 2023.

[10] C. Higuera, S. Dong, B. Boots, and M. Mukadam, "Neural contact fields: Tracking extrinsic contact with tactile sensing," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 576–12 582.

[11] J. Lee and N. Fazeli, "Vitascope: Visuo-tactile implicit representation for in-hand pose and extrinsic contact estimation," *arXiv preprint arXiv:2506.12239*, 2025.

[12] X. Yi, J. Lee, and N. Fazeli, "Visual-auditory extrinsic contact estimation," *arXiv preprint arXiv:2409.14608*, 2024.

[13] O. Taylor, N. Doshi, and A. Rodriguez, "Object manipulation through contact configuration regulation: multiple and intermittent contacts," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 8735–8743.

[14] B. Bianchini, M. Zhu, M. Sun, B. Jiang, C. J. Taylor, and M. Posa, "Vysics: Object reconstruction under occlusion by fusing vision and contact-rich physics," in *Robotics: Science and Systems (RSS)*, june 2025.

[15] M. Y. Aoyama, S. Vijayakumar, and T. Narita, "Few-shot transfer of tool-use skills using human demonstrations with proximity and tactile sensing," *IEEE Robotics and Automation Letters*, 2025.

[16] W. Fu, H. Li, I. X. He, S. Tellex, and S. Sridhar, "Unitac: Whole-robot touch sensing without tactile sensors," *arXiv preprint arXiv:2507.07980*, 2025.

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

[18] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.

[19] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, "Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," in *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[20] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, "Factr: Force-attending curriculum training for contact-rich policy learning," *arXiv preprint arXiv:2502.17432*, 2025.

[21] Q. K. Luu, P. Zhou, Z. Xu, Z. Zhang, Q. Qiu, and Y. She, "Manifeel: Benchmarking and understanding visuotactile manipulation policy learning," 2025. [Online]. Available: https://arxiv.org/abs/2505.18472

[22] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She, "UniT: Data efficient tactile representation with generalization to unseen objects," 2025. [Online]. Available: https://arxiv.org/abs/2408.06481

[23] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with 3d diffusion policies," *arXiv preprint arXiv:2410.10803*, 2024.

[24] Z. Zhang, Z. Xu, J. N. Lakamsani, and Y. She, "Canonical policy: Learning canonical 3d representation for equivariant policy," 2025. [Online]. Available: https://arxiv.org/abs/2505.18474

[25] 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.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[28] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[29] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.