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

Visual-auditory Extrinsic Contact Estimation

Xili Yi*[1], Jayjun Lee*[1], Nima Fazeli[1]

***Abstract*— Robust manipulation often hinges on a robot's ability to perceive extrinsic contacts—contacts between a grasped object and its surrounding environment. However, these contacts are difficult to observe through vision alone due to occlusions, limited resolution, and ambiguous near-contact states. In this paper, we propose a visual-auditory method for extrinsic contact estimation that integrates global scene information from vision with local contact cues obtained through active audio sensing. Our approach equips a robotic gripper with contact microphones and conduction speakers, enabling the system to emit and receive acoustic signals through the grasped object to detect external contacts. We train our perception pipeline entirely in simulation and zero-shot transfer to the real-world. To bridge the sim-to-real gap, we introduce a real-to-sim audio hallucination technique, injecting real-world audio samples into simulated scenes with ground-truth contact labels. The resulting multimodal model accurately estimates both the location and size of extrinsic contacts across a range of cluttered and occluded scenarios. We further demonstrate that explicit contact prediction significantly improves policy learning for downstream contact-rich manipulation tasks. Project webpage: va2contact.github.io**

I. Introduction

Extrinsic contact estimation is a crucial capability for robots to accurately understand how tools interact with their environment. Properly perceiving these contacts enables the robot to plan and control its actions effectively. Vision-based methods allow observation of the entire scene, but they often fall short in providing sufficient local information, particularly when contacts are occluded or within the resolution limits of the sensor. While tactile sensors or force/torque sensors offer precise measurements of direct contact surfaces, they struggle to perceive indirect contacts, such as those between a tool and the environment. This challenge creates a sensing gap in extrinsic contact estimation. As illustrated in Fig. 1(d), distinguishing whether the objects are in contact when they are close to each other within the sensor's resolution is difficult. Similarly, in Fig. 1(b), occlusions make it challenging to determine contact status as it is not directly observable.

To address this challenge, we propose a novel visual-auditory extrinsic contact estimation method, **`VA2Contact`**. Our approach integrates global visual feedback with local information obtained through active audio sensing, as illustrated in Fig. 1(a). This enables the estimation of extrinsic contacts as masks in 2D image space, from which both the contact location and type can be inferred. Additionally, we introduce an innovative audio-hallucination technique to overcome the difficulty of scalably obtaining audio feedback in simulation. This technique involves injecting real-world audio data into the simulation dataset with corresponding labels, thereby bridging the sim2real gap.

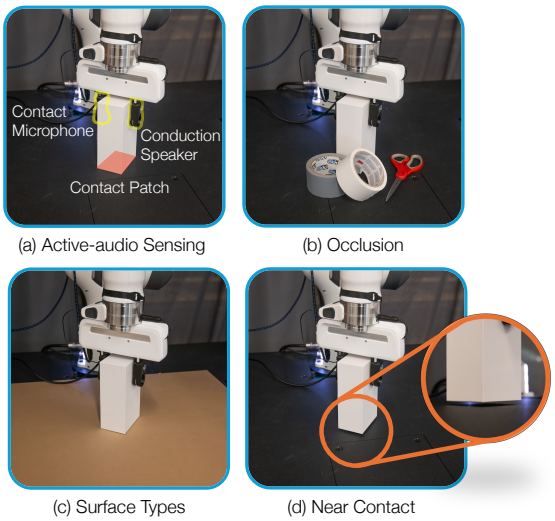

Fig. 1: (a) Our proposed fingers with an active conduction speaker and contact microphone emitting and receiving sound through the object. The absorption and reflection of the audio from the contact between the object and environment enables extrinsic contact estimation despite visual ambiguities. Challenges include (b) where objects occlude the contact between the box and the table, (c) where a different surface type can change the acoustic feedback, and (d) estimating the object's contact status for near-contact scenarios.

Our work builds upon and improves recent advancements in the field. A closely related study is Im2Contact [1], where a single RGB-D camera was used to learn a probability map of contact from depth images, optical flow extracted from RGB images, and proprioception. While they demonstrated a promising ability to estimate contact locations in pixel space, two main issues persist. The first is that the existing method struggles with obtaining the location and shape of the contact patch in static scenes as it relies solely on the optical flow for extracting the temporal information of movement in the scene to infer contacts and such indication of motion is not always correlated to contact events. Second, the existing approach is limited by occlusions and the resolution of the camera. By incorporating local active-audio information, our approach provides more comprehensive contact details, regardless of the contact type, and in spite of heavy occlusion, addressing these limitations and enhancing the performance of binary extrinsic contact detection and the accuracy and robustness of estimating the geometry of the extrinsic contact patches.

* Equal contribution
[1] Robotics Department, University of Michigan, USA <yixili, jayjun, nfz>@umich.edu

II. Related Works

Extrinsic Dexterity and Contact Estimation: In the field of manipulation, external contact sensing plays a crucial role. "Intrinsic contact," which refers to the direct contact between the robot and the environment, has been well studied [2], [3], [4]. However, when manipulating tools, estimating "extrinsic contact," such as sensing the contact between a tool held by the robot and the environment, becomes important. Sensing extrinsic contact is more challenging due to the indirect transmission of contact force/torque and the uncertainties in the object's geometry, stiffness, and pose [5]. [6] combines neural fields and vision-based tactile sensing to estimate the probability of extrinsic contacts between object and environment for any point on the object's surface. Most studies rely on force/torque sensors or tactile sensors, which often involve strong assumptions or are limited to predefined contact configurations [7], [8], [9]. Unlike high-fidelity tactile sensors that provide precise, localized measurements of direct contact forces, our active audio approach offers a low-cost, hardware-efficient alternative. While it may not achieve sub-millimeter force resolution, it uniquely resolves contact ambiguity under occlusion without requiring complex sensorization of the entire tool surface. Some prior works require knowledge about the tool and the environment. For instance, [10] and the subsequent work [11] focused on estimating object pose and contact simultaneously from force/torque feedback. Additionally, data-driven methods have shown potential in estimating extrinsic contacts. [1] demonstrated that with visual data and robot proprioception, their model could predict the contact location in image space without assuming prior knowledge about the object and environment. ViTaSCOPE learns an implicit representation that is trained on simulated point clouds and tactile shear data to integrate vision and touch for simultaneous in-hand object pose and extrinsic contact estimation [5]. [12] uses active tactile feedback to regulate a more consistent contact mode and make contact estimates for peg insertion tasks. Recently, multimodal policies that combine proprioception, vision, and audio have also been explored to decompose tasks into stages [13] and by finetuning vision-language-action models [14] for contact-rich manipulation tasks.

Audio for Robotic Manipulation: Audio is also a widely used modality in the field of robotics. Due to its physical principles, audio signals can provide frequency-related characteristics. Passive acoustic sensing directly uses sound waves from structural vibrations [15]. Some studies have modeled the sounds of object-surface interactions [16]. This approach is also commonly used in soft pneumatic actuators (SPAs) to sense state changes. Previous studies also showed that in end-to-end robot learning algorithms, including audio signals as input improved task performance [17], [18], [19], [20]. SonicSense [21] recently introduced a method to extract object material and shape information from in-hand acoustic vibrations, highlighting how such passive audio cues can enhance object understanding and contact estimation during manipulation. Other works also use active audio sensing methods to gain more information from the environment. Zöller et al. detected contact by embedding a microphone into an SPA to measure the sound induced by contact [22]. Similarly, studies [23], [24], [25], [26] proposed embedding a microphone and speaker in an SPA, playing sweeping sounds, and measuring changes from the microphone to sense deformation or contacts. Multimodal sensing with audio can be used for surface proximity detection with piezoelectric transducers for robot collision avoidance [27]. Other works use active audio on rigid grippers for grasp estimation and object recognition [28]. Additionally, [29] uses a gripper-mounted microphone for audio feedback on contact events under partial observability.

In this work, we leverage acoustic signals that transmit through solid objects by combining this with vision-based methods to estimate extrinsic contacts using a model-free approach. Unlike prior approaches that rely on fixed sensor setups or require assumptions about contact geometry, our method generalizes across contact types and tool configurations by fusing sound and vision in a flexible learning pipeline. Moreover, we introduce active-audio sensing to probe extrinsic contacts.

III. Methodology

Problem Statement: Consider a robot holding an unknown object and making contact with an unstructured environment. The goal of our method is to estimate the location and shape of the *extrinsic contact* between the grasped object and the environment. We assume no prior knowledge about the grasped object or environment. The inputs to our method are a depth map from a statically mounted camera, an optical flow image, fixed-length audio signatures measured at finger tip, and proprioceptive sensing of the robot state.

Method Overview: We train a multimodal model that can predict a per-pixel contact probability map over the scene given the inputs. All the training data except for audio are synthetically generated in simulation where pre-recorded audio data from the real-world are injected. The model is then zero-shot transferred to the real-world as in Figure 2. In the following, we outline the active-audio sensing mechanism that enables extrinsic contact sensing beyond vision, data generation process to train such a model, the model architecture, and training details unique to this problem.

Active-Audio Sensing: Suppose the robot is dragging the object it is grasping across a surface. The physical interaction between the object and environment creates an audio signature that contains a wealth of information about the interaction including its extrinsic contact patch size/shape, object/surface material types, contact location, and movement speed. However, if the object is in a static contact with the environment, there will be no audio feedback from the relative motion of the object, as audio signals require a source, either motion or an active audio source. To overcome this limitation and generalize to such static scenes, we introduce an active-audio system where one finger acts as an actuator (i.e. a speaker) by emitting controlled acoustic

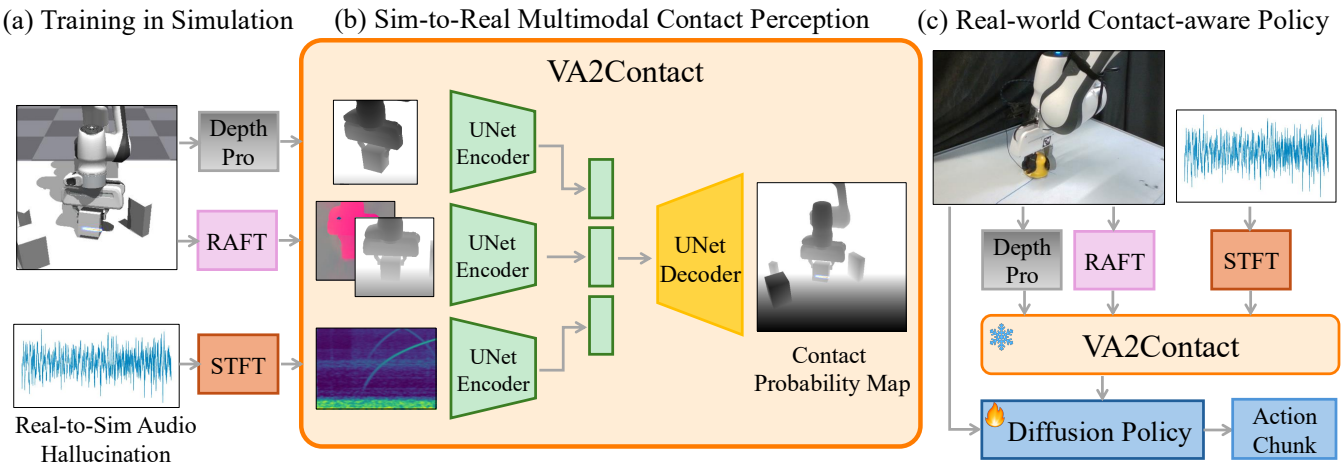

Fig. 2: **System Architecture.** Visual-Auditory Extrinsic Contact Estimation (**`VA2Contact`**) is trained in simulation with real-world audio collected through our active-audio sensing mechanism. Raw audio waveforms are processed with Short-Time Fourier Transform (STFT). **`VA2Contact`** can zero-shot transfer to the real-world for contact prediction tasks. Here, the output contact probability map is overlaid onto the full depth image of the scene. Note the usage of off-the-shelf metric depth estimation models (Depth-Pro [30]) and optical flow estimation model (RAFT [31]) both for scalable sim-based training and real-world inference, to bridge the sim-to-real gap effectively. **`VA2Contact`** unlocks contact perception under occlusions for contact-rich tool manipulation tasks such as wiping, which we demonstrate through real-world policy learning experiments.

signals (i.e. a sweeping impulse signal), while the other finger as a receptor (i.e. a microphone) that receives the acoustic feedback, which changes depending on the grasped object's material properties and the presence of extrinsic contacts. As illustrated in Figure 3, contact with the environment alters the received signal due to energy absorption, enabling active contact perception even in the absence of motion. This mechanism extends the robot's perception to static contact scenarios, complementing vision and proprioception. Even without motion, contact with the environment alters the acoustic response by absorbing energy and modifying the transmission characteristics. This allows the system to distinguish between contact and non-contact scenarios based on the presence and properties of the received audio signal.

Challenges in Data Generation: However, one major challenge in training a multimodal model for the task of visual-auditory extrinsic contact estimation is obtaining a dataset with aligned vision, audio, and ground-truth contact patch

Fig. 3: **Active-audio Sensing.** (1). A sweeping acoustic signal is generated from (2) using conduction speaker finger, where (3) the sound propagates through the object and vibrates with any extrinsic contact it makes, and (4) which is received at the contact microphone finger, and (5) the audio waveform is converted to a spectrogram.

labels. In the real-world, it is relatively easy to obtain the audio information along side with both depth maps and robot proprioception data. However, obtaining correct contact patch masks is not only hard but also inaccurate with human labels, because there is no direct way to observe the exact contact patch shape and size. Meanwhile in simulation, it is simple to generate a large dataset with both correct contact patch masks and correct labels. However, obtaining realistic audio signals from simulation remains a great challenge.

Real-to-Sim Audio Hallucination: To overcome this challenge, we propose an audio hallucination technique. In the real-world, we collect audio signals ($\mathbf{v}_t$) as the robot manipulating various grasped objects by making diverse contact with the environment. During the robot motion, we record contact and contact geometry labels ($\mathbf{g}_t$) from the set `[free, point, line, patch]`. We emphasize that we do not record contact locations, only the simple-to-observe contact labels. Thus the labeled dataset contains audio samples paired with contact types, represented as $\mathcal{D}_{\text{audio}} = \{(\mathbf{v}_0, \mathbf{g}_0), \cdots, (\mathbf{v}_n, \mathbf{g}_n)\}$. Next, in simulation we create scenes with the same robot and table as well as a large distribution of objects that the robot can grasp and interact with. This distribution of objects is not the same as the test objects used in the real-world setup. From the simulated visual scene, we obtain a depth map ($\mathbf{d}_t$) and obtain proprioceptive state ($\mathbf{p}_t$). We also take one depth map when the object is lifted without contact as a reference frame ($\mathbf{r}_t$) of the grasped object. Although the simulator can generate optical flow images, we generate the optical flow ($\mathbf{f}_t$) using an off-the-shelf RAFT model [31] to later facilitate sim-to-real transfer, where we use the same model to generate the optical flow images for the real world test data and similarly use

Depth-Pro [30] for depth. It is noted that we generate optical flows only from depth maps to avoid the influence from shadows that would break the color constancy assumption. Most importantly, simulation also directly provides us with the contact geometry label and its corresponding shape ($\mathbf{s}_t$) when contact occurs. At each contact event, we randomly select a sample of the real-world audio corresponding to the label provided by the simulator and create a labeled dataset of contacts $\mathcal{D} = \{(\mathbf{O}_0, \mathbf{s}_0), \cdots, (\mathbf{O}_m, \mathbf{s}_m)\}$ where the first element is the observation vector $\mathbf{O}_i = (\mathbf{v}_i, \mathbf{d}_i, \mathbf{r}_i, \mathbf{f}_i, \mathbf{p}_i)$ and the second is the label.

The premise of our approach is that, while audio signatures vary across specific object interactions, they share consistent patterns that reflect the underlying contact geometry. A model trained on such data can learn to generalize by ignoring object-specific details and focusing on features indicative of contact type. For instance, point or line contacts tend to produce higher-pressure interactions, resulting in sharper, high-frequency sounds. In contrast, patch contacts generate diffuse sounds resembling white noise, with energy spread across a wider frequency range due to a larger contact area.

Model Architecture: For visual-auditory extrinsic contact estimation, our model architecture builds on the UNet architecture [32] as illustrated in Figure 2(b). The model consists of three streams of UNet encoders, each processing different streams of data. The first stream encodes a cropped depth image of the object at a reference frame. The second stream encodes a depth image and an optical flow image at the current frame, which are cropped around the projected end-effector pixel coordinates and stacked along the channel dimension. The third stream encodes an image of a log-mel spectrogram from a 1s long audio waveform sampled at 44.1 kHz. All input images are of shape 256×256 and the 7-DoF robot pose is fused at the bottleneck of the UNet. While we acknowledge that alternative conditioning mechanisms (e.g., cross-attention or feature-wise linear modulation) might offer a more theoretically principled approach for fusing modalities with distinct dimensional structures like spectrograms and depth maps, our empirical results indicate that channel-wise concatenation within the U-Net effectively learns to correlate spectral features with spatial contact masks.

Additionally, a prior work [1] suggests a few strategies in which we adopt to facilitate better generalization and sim2real transfer. First, cropping the input images of the task scene centered around the end-effector position projected onto the pixel space simplifies the learning problem by removing background distractions and with an improved focus on the tool-environment interaction. For every cropping operation, additional three channels are stacked for coordinate convolution [33] to provide the model with an explicit information about the spatial location of each pixel within the cropped image and help the model maintain the understanding of spatial relationships even after cropping. Another major challenge with the extrinsic contact estimation is when the grasped tool is heavily occluded in a cluttered scene that the parts of the grasped object that make contact with the environment is not visually observable. Thus, we opt to provide the model with a contact-free reference depth image of the object in the gripper. Moreover, visual ambiguities are present when differentiating an in-contact versus a near-contact state of the robot. This can be partially resolved by introducing the optical flow as an input to represent the temporal information of the robot motion.

We further utilize the audio modality to acquire active feedback from extrinsic contact and interaction with the environment from the sine sweep impulse response that changes per contact mode type, grasped object, and robot motion. The output of the model is a contact probability map that can be thresholded to obtain the contact patch as a single channel mask image.

IV. Implementation

Active-Audio Processing: The frequency characteristics of each object and contact vary depending on their physical properties, including stiffness and geometry. To capture these variations comprehensively, we transmit a sweeping impulse sound that spans a broad frequency range (20 to 20k Hz) through the object repeatedly at 1 Hz. We collect the audio feedback from a microphone and then use Short-Time Fourier Transform (STFT) and scaling to generate a log-mel spectrogram using 64 mel frequency bins. In the spectrogram, shown in Figure 3, the x-axis represents time, the y-axis represents frequency, and the color represents energy at that time and frequency. The frequency axis is non-linearly scaled to show more information in the lower frequency range, which typically contains more useful information for contact. Given its ability to display both temporal and frequency information and their relationship, we choose this as the audio representation. We fix the duration of each audio sample to 1 second, ensuring that it captures the full impulse response across the entire sweeping frequency range.

Real World Setup: For collecting and generating such scalable dataset for extrinsic contact perception, we use the following setup. In both simulation and the real-world, we use a 7-DoF Franka Emika Panda robot arm in a tabletop environment. Active-audio fingertips, illustrated in Fig. 3 are mounted onto the gripper, and both input and output audio signals are routed to a computer through a Focusrite Scarlett 4i4 3rd Gen amplifier. We set a single RealSense RGB-D camera pointing to the table for collecting visual feedback. We selected several objects from YCB dataset [34] (a spoon, clamp, tube, apple, lemon, orange, pear, screw driver, scissors, pods) and a few other objects such as a RealSense box and a dustpan as test objects. Note that these are not seen during training.

Real World Audio Collection: To collect the active audio feedback with different types of contact, after the robot grasps an object, we teleoperate the robot to random orientation, and make contact with the environment. During the execution of each motion, we maintain the same contact mode(free, point, line, or patch contact), we also collect the active audio feedback from the scene, with sample rate of 44100 Hz. To provide the active audio signal, we input sweeping sound from 20Hz to 20000Hz in 1 second, and

repeat it at 1 Hz. We also label the audio given the contact mode. We collected 1000 audio samples with variety of objects, contact shape and contact mode.

Simulation Setup and Real-to-Sim Audio Injection: We use the GPU-based Isaac Gym [35] simulation to replicate the real-world scene generating diverse extrinsic contact data at scale. In each episode, the robot grasps random objects and rotate to a random orientation, then lower down to the tabletop, and starts sliding the object on the table. For each environment, we randomly select cubes with random size and objects from ycb dataset[34] that are smaller than 10cm in diameter. We obtain the contact points of the object and the environment, then project them back to the camera frame to obtain the contact masks. We also obtained a contact label from the size and shape of the contact, which are `[free, point, line, patch]`. With these labels, we randomly select real-world audio samples with the same labels and then pair with depth image. The training data contains 6880 episodes where each episode contains around 60 frames, which approximates to a total of 440k samples.

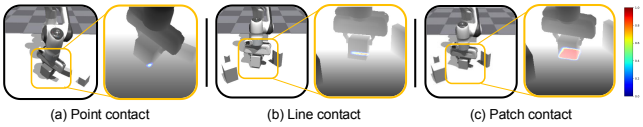

Fig. 4: **Different contact modes learned from simulation data.** The cropped depth is centered around the projected pixel coordinate of the EE pose. Sample predictions from test simulation data are shown as contact probability maps overlaid on top of scene depth.

Training Details: We train our model with pixel-wise binary cross entropy loss with logits using the ground-truth contact masks obtained from simulation. We train for 20 epochs with a learning rate of $5e-4$ on a single A6000 GPU.

Real World Test Data Collection: To evaluate **`VA2Contact`** on extrinsic contact detection, we collect data that consists of RGB-D images of the scene, EE pose, camera extrinsics, a reference RGB-D image, audio feedback, and a manually labelled extrinsic contact mask. We use the same off-the-shelf models (Depth-Pro and RAFT) to obtain consistent depth and optical flow images

Additionally, for collecting the real-world depth images, we primarily use the monocular metric depth estimation model Depth-Pro [30] to generate hole-less depth map from raw RGB images of the scene. We collected data in 4 cases: general, occlusion, near-contact, and different surface cases. We use RAFT [31] to generate optical flow images. We use the same audio collection pipeline as audio collection for test data collection.

For downstream policy learning with **`VA2Contact`** in the perception pipeline, we use an Oculus Quest 2 device to teleoperate and collect 41 demonstrations for the wiping task that amount to 16k timesteps of training data with audio feedback at 15 Hz. We preprocess this data to generate consistent depth, optical flow, and contact maps using Depth-Pro, RAFT, and **`VA2Contact`** to train diffusion policy that takes two images (one for raw RGB and one for contact prediction image) and joint state as input. The diffusion policy is trained on relative joint action space that predicts a action horizon of 16 where we execute the first 8 for receding horizon control at 15 Hz.

V. Experiments

A. Extrinsic Contact Estimation

Baselines and Ablations: We compare our visual-auditory method against a similar model-free and vision-only method Im2Contact [1]. We also evaluate a variant of **`VA2Contact`** trained without optical flow to experiment if audio can fully replace it. **Simulation Test Scenarios:** We first evaluate our methods in simulation

Real World Test Scenarios: To evaluate the sim2real transfer of our models on the extrinsic contact estimation problem, we design the following real world test scenarios. These scenarios highlight the primary challenges, including demonstrating sim2real transfer under substantial sim2real gap (S1, S2), contacts invisible to vision (S3, S4).

- **S1**: Generalization to unseen gripper-held object types and geometry. (Figure 5)
- **S2**: Generalization to unseen table surface to introduce audio variation. (Figure 6)
- **S3**: Near contact cases that are not in contact.
- **S4**: Cases where the contact between gripper-held object and the table is occluded. (Figure 6)

In total, we collected 310 evaluation samples. The number per contact mode is `free:point:line:patch` = {48:37:41:34} for **S1**, {0:15:15:20} for **S2**, {4:14:9:23} for **S3**, and {50:0:0:0} for **S4**.

Experimental Results: The main results of the sim2real transfer of visual-auditory extrinsic contact estimation in the real world setup are shown in Table I and Figure 5. Here we show the performance of our methods across all 4 scenarios. In the *general* case, the model with audio overall exhibits higher recall and F1 score, indicating a superior ability to detect true contacts. This suggests that audio cues help reduce false negatives without substantially increasing false positives. **`VA2Contact`** w/o optical flow scores the highest across all metrics for binary contact detection because audio provides a more direct information about contact whereas optical flow just tells us where things moved, which might not be directly correlated. However, in terms of the actual contact patch prediction, **`VA2Contact`** which uses both active-audio and optical flow performs the best as shown in Figure 5.

We also explore edge cases such as visual occlusions and near-contact scenarios where active-acoustic signals provide critical cues, as shown in Figure 6. We evaluate **`VA2Contact`** under three challenging real-world settings: *different surface types*, *near-contact*, and *occlusions*. On different surfaces (S2, Table II), both vision-only and audio-augmented models achieve strong binary contact detection. While IOU is slightly better without audio, our method generalizes well despite variation in acoustic feedback, showing

Models	Binary Contact Detection (macro)				Contact Location, Shape, & Size (macro)		
	Prsn ↑	Rcll ↑	F1 ↑	Acc ↑	BCE ↓	IOU ↑	CD (px^2) ↓
Im2Contact	0.87	0.87	0.86	0.58	**0.031**	0.092	15.01
VA2Contact w/o flow	**0.99**	0.89	0.94	0.91	0.038	0.096	11.73
VA2Contact	0.94	**0.95**	**0.94**	**0.93**	0.038	**0.240**	**6.44**

TABLE I: **Real world results for sim-to-real extrinsic contact estimation on all 4 scenarios.** Models are evaluated on binary contact detection and the accuracy of contact geometry estimation, which is only provided for the true positive cases where the model predicts a mask above a threshold level and the ground-truth is in-contact. CD: Chamfer Distance. IOU: Intersection over Union. BCE: Binary cross-entropy. Prsn: Precision. Rcll: Recall. Acc: Accuracy.

Subset	Model	Prsn ↑	Rcll ↑	F1 ↑	Acc ↑	BCE ↓	IOU ↑	CD (px^2) ↓
S2: Surface Types	Im2Contact	**1.00**	0.78	0.88	0.78	0.040	0.139	10.52
	VA2Contact w/o flow	**1.00**	0.86	0.92	0.86	0.060	0.116	10.74
	VA2Contact	**1.00**	**0.94**	**0.97**	**0.94**	0.054	**0.221**	**4.36**
S3: Near Contact (no positive GT; Precision/Recall/F1 not defined)								
	Im2Contact	TN=2, FP=48			0.04	–	–	–
	VA2Contact w/o flow	TN=48, FP=2			0.96	–	–	–
	VA2Contact	TN=49, FP=1			**0.98**	–	–	–
S4: Occlusions	Im2Contact	0.91	0.89	0.90	0.82	0.029	0.079	15.30
	VA2Contact w/o flow	**1.00**	0.87	0.93	0.88	**0.025**	0.099	10.84
	VA2Contact	0.96	**1.00**	**0.98**	**0.96**	0.034	**0.220**	**10.16**

TABLE II: **Real-world results for sim-to-real extrinsic contact estimation.** Binary detection metrics (Prsn: Precision, Rcll: Recall, F1, Acc: Accuracy) are reported where positive ground-truth contacts exist (S2, S4). For S3, only near-contact negatives are present; we therefore report true negatives (TN), false positives (FP), and accuracy. Geometry metrics (BCE: Binary Cross-Entropy, IOU: Intersection over Union, CD: Chamfer Distance) are evaluated on true positives only.

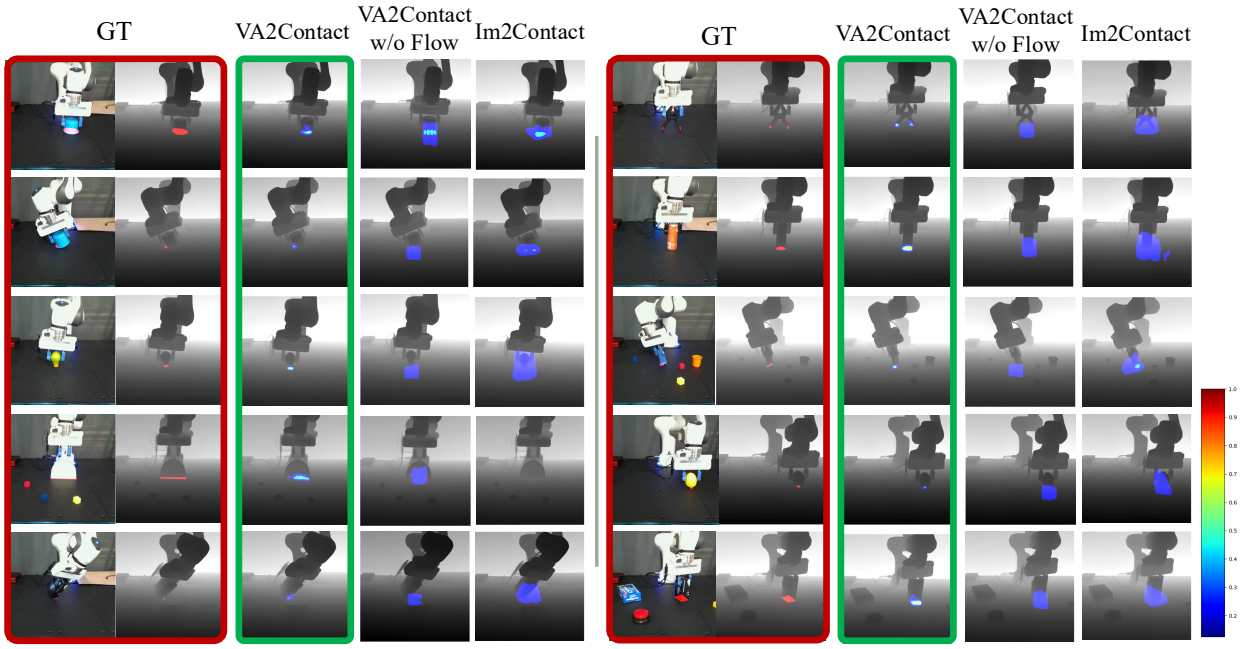

Fig. 5: **Sim-to-Real Transfer and Real-world Extrinsic Contact Predictions for S1.** The RGB-D images with GT contact probability masks are wrapped in red. Three models, **VA2Contact**, **VA2Contact** w/o optical flow, and Im2Contact, are tested where the results are shown per column. **VA2Contact**'s contact probability predictions are overlaid to depth images, wrapped in green. The color bar represents contact probability (0.0 ← | → 1.0). All grasped objects used for real-world testing are unseen geometries (a cup, pear, dustpan, box, blue sponge, lemon, can, clamp). **VA2Contact** is able to zero-shot predict diverse contact types over objects with varying properties.

audio input remains effective under domain shifts. In near-contact scenarios (S3, Table II), **VA2Contact** significantly reduces false positives compared to Im2Contact, indicating its strength in disambiguating ambiguous visual cases via active sensing. Under occlusions (S4, Table II), **VA2Contact** consistently outperforms all baselines in both detection and

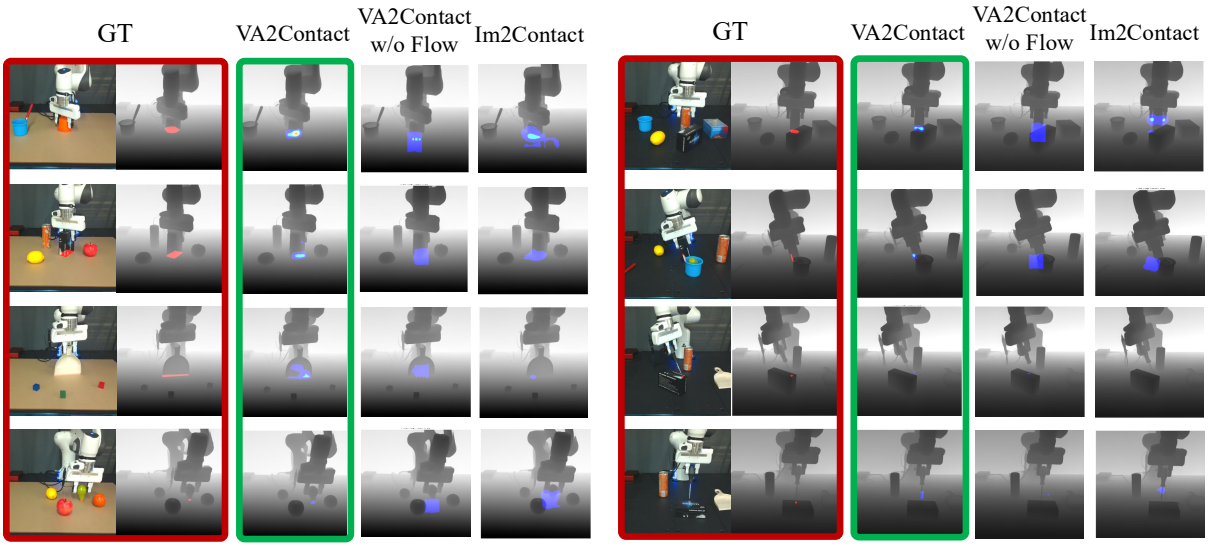

Fig. 6: **Sim-to-Real Transfer and Real-world Extrinsic Contact Predictions for S2 and S4.** The RGB-D images with GT contact probability masks are wrapped in red. Three models, **VA2Contact**, **VA2Contact** w/o optical flow, and Im2Contact, are tested where the results are shown per column. **VA2Contact**'s contact probability predictions are overlaid to depth images, wrapped in green. The color bar represents contact probability ($0.0 \leftarrow | \rightarrow 1.0$). All grasped objects used for real-world testing are unseen geometries (a cup, pear, dustpan, box, blue sponge, lemon, can, clamp). **VA2Contact** is able to zero-shot predict diverse contact types over objects with varying properties.

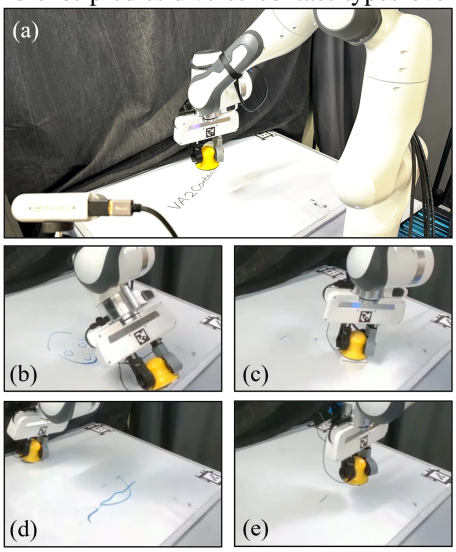

Fig. 7: Real-world wiping task using **VA2Contact** contact-aware diffusion policy. (a) Experimental setup where the robot wipes a whiteboard using our method. Comparison of baseline and **VA2Contact** wiping a drawing of a face in (b)-(c) and wiping a boat in (d)-(e). In both tasks, **VA2Contact** enables more consistent and complete contact with the surface, leading to better performance.

geometry metrics (e.g., Recall: 0.94 vs. 0.78; F1: 0.97 vs. 0.88), demonstrating that audio cues robustly compensate for missing visual information.

B. Contact-aware Policy Learning

We evaluate whether our sim2real visual-auditory extrinsic contact sensing improves downstream policy learning for contact-rich, occlusion-heavy manipulation. We selected a wiping task, which requires the robot to hold a marker eraser and maintain contact to erase the markers(Figure 7). To test this, we train a baseline diffusion policy with just an RGB image stream and a contact-aware diffusion policy as in Figure 2 that takes the contact probability map image in addition to the RGB image stream as inputs. Across 10 rollouts, the baseline achieved 4/10 success due to inconsistent surface contact, whereas our contact-aware policy succeeded 8/10 times.

Overall, integrating audio enhances reliable contact detection and localization in complex real-world environments, effectively bridging the sim-to-real gap. Audio cues effectively complement vision, particularly during common real-world visual ambiguities like occlusions. Furthermore, these multimodal contact-aware representations significantly enhance contact-rich manipulation capabilities.

VI. Conclusion and Discussions

We present **VA2Contact**, a simulation-trained method estimating extrinsic contacts via fingertip audio, vision, and proprioception. Active audio provides local cues complementing vision for robust perception despite occlusions and ambiguity. To bypass audio simulation challenges, our real-to-sim hallucination strategy injects real audio during training, yielding strong zero-shot transfer and outperforming visual-only baselines.

Despite promising results, limitations remain. Internal robot vibrations, especially from the gripper, degrade audio signal quality. External environmental noise in uncontrolled settings warrants future investigation, potentially requiring

active noise cancellation or domain adaptation. Manual contact mask annotations suffer from occlusion biases. Our audio representation may miss rich contact cues; future work could explore pretrained models. Additionally, the 1-second frequency sweep introduces latency, limiting applicability in highly dynamic tasks and motivating shorter waveforms. Reliance on off-the-shelf vision models (Depth-Pro, RAFT) bounds performance by their generalization limits, particularly under extreme lighting or with transparent objects. Finally, generalizing beyond our dataset's rigid objects to deformable or liquid-filled items requires improved sensing and annotation.

REFERENCES

[1] Leon Kim, Yunshuang Li, Michael Posa, and Dinesh Jayaraman. Im2contact: Vision-based contact localization without touch or force sensing. In *Conference on Robot Learning*, pages 1533–1546. PMLR, 2023.

[2] Alessandro De Luca, Alin Albu-Schaffer, Sami Haddadin, and Gerd Hirzinger. Collision detection and safe reaction with the dlr-iii lightweight manipulator arm. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1623–1630. IEEE, 2006.

[3] Lucas Manuelli and Russ Tedrake. Localizing external contact using proprioceptive sensors: The contact particle filter. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5062–5069. IEEE, 2016.

[4] Antonio Bicchi, J Kenneth Salisbury, and David L Brock. Contact sensing from force measurements. *The International Journal of Robotics Research*, 12(3):249–262, 1993.

[5] Jayjun Lee and Nima Fazeli. Vitascope: Visuo-tactile implicit representation for in-hand pose and extrinsic contact estimation. In *Proceedings of Robotics: Science and Systems*, 2025.

[6] Carolina Higuera, Siyuan Dong, Byron Boots, and Mustafa Mukadam. Neural contact fields: Tracking extrinsic contact with tactile sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12576–12582. IEEE, 2023.

[7] Kuan-Ting Yu and Alberto Rodriguez. Realtime state estimation with tactile and visual sensing for inserting a suction-held object. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1628–1635. IEEE, 2018.

[8] Daolin Ma, Siyuan Dong, and Alberto Rodriguez. Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11262–11268. IEEE, 2021.

[9] Sangwoon Kim, Devesh K Jha, Diego Romeres, Parag Patre, and Alberto Rodriguez. Simultaneous tactile estimation and control of extrinsic contact. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12563–12569. IEEE, 2023.

[10] Andrea Sipos and Nima Fazeli. Simultaneous contact location and object pose estimation using proprioception and tactile feedback. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3233–3240. IEEE, 2022.

[11] Andrea Sipos and Nima Fazeli. Multiscope: Disambiguating in-hand object poses with proprioception and tactile feedback. *arXiv preprint arXiv:2305.14204*, 2023.

[12] Sangwoon Kim and Alberto Rodriguez. Active extrinsic contact sensing: Application to general peg-in-hole insertion. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10241–10247. IEEE, 2022.

[13] Ruoxuan Feng, Di Hu, Wenke Ma, and Xuelong Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024.

[14] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, 2025.

[15] Chris Harrison and Scott E Hudson. Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 205–208, 2008.

[16] Shihan Lu, Yang Chen, and Heather Culbertson. Towards multisensory perception: Modeling and rendering sounds of tool-surface interactions. *IEEE transactions on haptics*, 13(1):94–101, 2020.

[17] Abitha Thankaraj and Lerrel Pinto. That sounds right: Auditory self-supervision for dynamic robot manipulation. In *Conference on Robot Learning*, pages 1036–1049. PMLR, 2023.

[18] 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.

[19] Jared Mejia, Victoria Dean, Tess Hellebrekers, and Abhinav Gupta. Hearing touch: Audio-visual pretraining for contact-rich manipulation. *arXiv preprint arXiv:2405.08576*, 2024.

[20] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. *arXiv preprint arXiv:2406.19464*, 2024.

[21] Jiaxun Liu and Boyuan Chen. Sonicsense: Object perception from in-hand acoustic vibration. In *8th Annual Conference on Robot Learning*, 2024.

[22] Gabriel Zöller, Vincent Wall, and Oliver Brock. Acoustic sensing for soft pneumatic actuators. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6986–6991. IEEE, 2018.

[23] Ken Takaki, Yoshitaka Taguchi, Satoshi Nishikawa, Ryuma Niiyama, and Yoshihiro Kawahara. Acoustic length sensor for soft extensible pneumatic actuators with a frequency characteristics model. *IEEE Robotics and Automation Letters*, 4(4):4292–4297, 2019.

[24] Gabriel Zöller, Vincent Wall, and Oliver Brock. Active acoustic contact sensing for soft pneumatic actuators. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7966–7972. IEEE, 2020.

[25] Shinichi Mikogai, BDC Kazumi, and Kentaro Takemura. Contact point estimation along air tube based on acoustic sensing of pneumatic system noise. *IEEE Robotics and Automation Letters*, 5(3):4618–4625, 2020.

[26] Uksang Yoo, Ziven Lopez, Jeffrey Ichnowski, and Jean Oh. Poe: Acoustic soft robotic proprioception for omnidirectional end-effectors. *arXiv preprint arXiv:2401.09382*, 2024.

[27] Xiaoran Fan, Riley Simmons-Edler, Daewon Lee, Larry Jackel, Richard Howard, and Daniel Lee. Aurasense: Robot collision avoidance by full surface proximity detection, 2021.

[28] Shihan Lu and Heather Culbertson. Active acoustic sensing for robot manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3161–3168. IEEE, 2023.

[29] Maximilian Du, Olivia Y Lee, Suraj Nair, and Chelsea Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. *arXiv preprint arXiv:2205.14850*, 2022.

[30] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

[31] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[33] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution, 2018.

[34] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

[35] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance GPU based physics simulation for robot learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.