You're Pushing My Buttons: Instrumented Learning of Gentle Button Presses

Raman Talwar[1], Remko Proesmans[1], Thomas Lips[1], Andreas Verleysen[1] and Francis wyffels[1]

***Abstract*—Learning contact-rich manipulation is difficult from cameras and proprioception alone because contact events are only partially observed. We test whether training-time instrumentation, i.e., object sensorisation, can improve policy performance without creating deployment-time dependencies. Specifically, we study button pressing as a testbed and use a microphone fingertip to capture contact-relevant audio. We use an instrumented button-state signal as privileged supervision to fine-tune an audio encoder into a contact event detector. We combine the resulting representation with imitation learning using three strategies, such that the policy only uses vision and audio during inference. Button press success rates are similar across methods, but instrumentation-guided audio representations consistently reduce contact force. These results support instrumentation as a practical training-time auxiliary objective for learning contact-rich manipulation policies.**

***Index Terms*—Instrumentation, Diffusion Policy, Multimodal Learning, Audio-Tactile Integration, Robotic Manipulation**

I. Introduction and Related Work

Imitation learning (IL) of contact-rich manipulation is difficult from cameras and proprioception alone because contact events are only partially observed. Instrumentation, i.e., object sensorisation, has been identified as a useful robot learning tool: it can automate high-quality demonstrations and provide privileged signals to improve policy performance [1–4]. The key research question of instrumentation is: *Can training-time instrumentation improve performace in manipulation policies without creating inference-time dependence?* We investigate this question for button pressing, where performance is measured not only by task success but also by contact quality: the robot should use minimal force for a successful button press. In this paper, the robot is provided with two sensing modalities for the button-pressing task: a wrist camera for localisation, and a microphone in the fingertip for capturing contact-related audio. Prior studies demonstrate that audio signals offer valuable cues for contact-rich manipulation [5]. In the next section, we detail the instrumented setup, data-collection pipeline, and multimodal observation/action space used to study this question. We then describe how instrumentation-supervised audio representations are integrated into the policy through the evaluated training strategies.

II. Methods and Materials

Experimental Setup: We use an UR3e arm with a Robotiq 2F-85 gripper, wrist RGB camera (Intel RealSense D435), a Bota Systems F/T sensor (BFT-DENS-SER-M8), and a custom 3D-printed fingertip with an ICS-43434 microphone (see Fig. 1). A push button is connected to a microcontroller that reads the binary button state (pressed or unpressed). This button-state signal is *privileged information*: it is available during data collection and training, but not during inference.

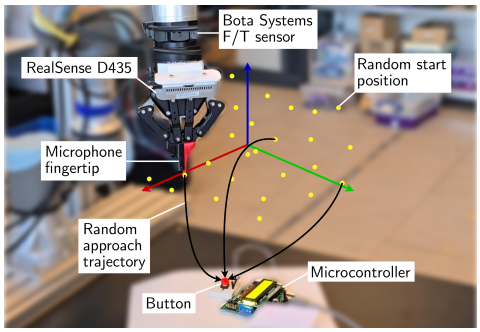

Fig. 1. Overview of the hardware setup. In addition, some of the randomised end-effector start positions and approach trajectories are indicated.

Task Definition and Automated Data Collection: The task is to approach the button, press, and retract. During setup, the button location is measured once and subsequently used as a fixed reference target for all data collection trials. Demonstrations are collected automatically by randomising the end-effector (EEF) start poses and approach trajectories (Fig. 1); once a press event is detected from instrumentation, the robot retracts. This avoids haptic teleoperation and yields high-quality demonstrations [1].

Observation and Action Spaces: For imitation learning, we use Diffusion Policy [6]. To prioritise reactivity, we set the action horizon to 1. Each observation at time t consists of (policy control at 10 Hz, i.e., 100 ms between observations):

- **Wrist RGB image ($\mathbf{I}_t \in \mathbb{R}^{240\times320\times3}$):** encoded with ResNet-18. Images are randomly cropped to 288×216 during training, and center-cropped during inference.
- **Audio spectrogram ($\mathbf{A}_t \in \mathbb{R}^{298\times128}$):** 3-second log-Mel window.
- **Instrumented Button State ($b_t \in \{0, 1\}$):** Binary pressed/unpressed signal, only used in strategies that include privileged instrumentation during training.

The robot's orientation is fixed, and the policy controls only its position. Actions are interpreted as relative displacements expressed in the current EEF frame (i.e., the frame associated with the latest observation).

Audio Processing and Audio Encoder Fine-tuning: To address visual occlusions and the lack of force feedback in image observations, we use the fingertip microphone as

*This work was supported by Research Foundation Flanders (grant no. 1S15925N), and by euROBIN (grant no. 101070596).

[1]Raman Talwar, Remko Proesmans, Thomas Lips, Andreas Verleysen and Francis wyffels are with the AI and Robotics Lab (IDLab-AIRO), Ghent University—imec, Ghent, Belgium raman.talwar@ugent.be

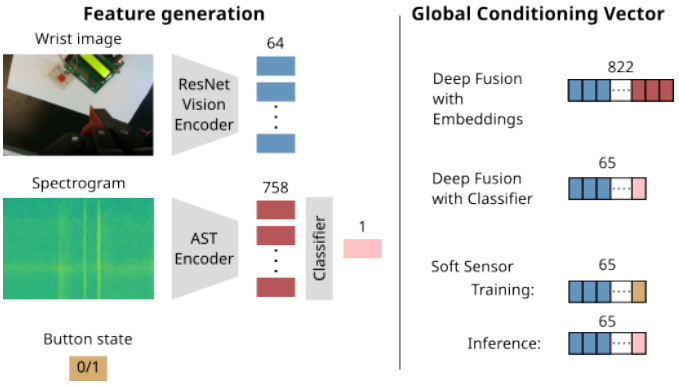

Fig. 2. Overview of the audio integration strategies evaluated in this work: Deep Fusion (direct conditioning on AST logits or embeddings) and Soft Sensor (training with true button state, replaced by AST prediction at test time).

an additional (contact)-sensing modality. Audio is processed in 3-second windows, resampled to 16 kHz, and converted to log-Mel spectrograms. Spectrograms are encoded with an Audio Spectrogram Transformer (AST) [7] pretrained on AudioSet [8].

To adapt this generic encoder to the task, we fine-tune AST as a binary click detector using labels from the instrumented button signal (10 epochs, learning rate 10^{-5}). The model reaches F1 $= 0.988$ with 1.2% false negatives on validation, and its representation is used for policy learning.

Integration Strategies: Fig. 2 summarises the evaluated policy variants:

- **Soft Sensor (Pseudo-Instrumentation):** Train the policy with privileged button state in the observation, then replace it at inference with the button-state prediction from the fine-tuned AST click detector.
- **Deep Fusion:** Inject audio during policy training using either (i) the output logits of the fine-tuned AST click detector or (ii) intermediate AST embeddings (penultimate-layer features) from the same fine-tuned model.

We compare these variants against a baseline that uses embeddings from a generic AudioSet-pretrained AST without task-specific fine-tuning.

All policies are trained with Adam (learning rate 10^{-4}, weight decay 10^{-6}) and a cosine schedule with 500 warmup steps. We apply standard image augmentations (brightness, contrast, saturation, hue, and sharpness jitter). Actions and states are normalised to $[-1, 1]$, and spectrograms are normalised following [7].

Evaluation Protocol: Each policy is evaluated using 40 rollouts. Initial EEF poses are uniformly randomised within the same space used for data collection. A rollout is considered "successful" if the button is pressed and the robot retracts. Wrist force-torque signals are logged to assess interaction quality, where lower contact force is preferred.

III. Results and Discussion

As reported alongside the model labels in Fig. 3, success rates fall within a narrow range of 45% to 55% across models and are statistically indistinguishable over 40 rollouts per

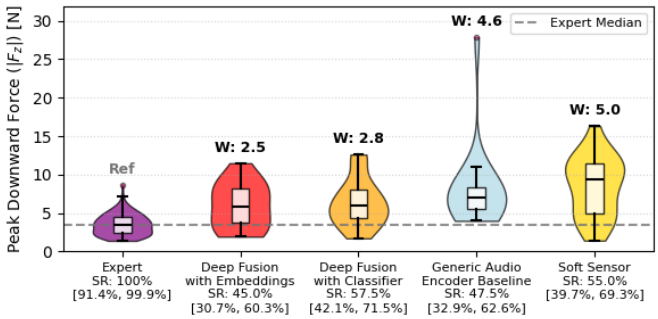

Fig. 3. Force distributions ranked by Wasserstein distance (W [N]) to the expert demonstrations. Success rates (SR) and Bayesian 95% credible intervals are provided for each model.

model, with largely overlapping the Bayesian 95% credible intervals. We therefore focus on contact-force metrics to differentiate policy quality.

We analysed peak vertical force (F_z) during successful rollouts. Compared with expert demonstrations (median 3.41 N), the generic Audio Encoder baseline is much harsher (median 6.98 N), while Deep Fusion improves contact behavior (median 5.87 N with embeddings; 5.95 N with classifier). Soft Sensor performs worst (median $9.37N$), likely due to train–test mismatch: the policy trains on perfect button-state signals but receives predicted states at inference, with additional maximum latency of 50 ms from spectrogram creation.

Computing the Wasserstein distance to the expert force distribution further supports this ranking: Soft Sensor is least similar (W=5.0 N), generic AST remains distant (W=4.6 N), and Deep Fusion with fine-tuned embeddings (W=2.5 N) or fine-tuned classifier (W=2.8 N) is closest to expert behavior. These results suggest that conditioning the policy on acoustic representations fine-tuned using instrumentation can effectively regularise contact forces in this task setting.

IV. Conclusion and Future Work

This work suggests that instrumentation can be used during training to improve interaction quality in contact-rich manipulation policies, while inference remains free of privileged sensors. In our experiments, instrumentation-guided audio representation learning does not significantly change success rate, but consistently reduces excessive contact forces.

The main implication is methodological: instrumentation can be treated as a temporary supervision channel for learning deployable representations. While shown here on button pressing, the approach is promising for broader tasks where safe, compliant contact matters, though broader validation across tasks, hardware, and acoustic shifts is still needed.

Next steps include validating this strategy on broader contact-rich tasks (e.g., insertion, latching, and tool use), testing generalisation across different buttons and button types, testing transfer across objects and mechanisms with different contact signatures, and studying tighter integration schemes such as end-to-end audio-visual fusion and partial AST unfreezing during policy training. Together, these directions can help establish instrumentation-guided learning as a general recipe for robust and gentle robotic manipulation.

REFERENCES

[1] R. Proesmans, T. Lips, and F. wyffels, "Instrumentation for better demonstrations: A case study."

[2] ——, "Instrumentation for imitation learning: Enhancing training datasets for clothes hanger insertion," in *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[3] K. Junge, C. Pires, and J. Hughes, "Lab2field transfer of a robotic raspberry harvester enabled by a soft sensorized physical twin," *Communications Engineering*, vol. 2, no. 1, p. 40, Jun 2023.

[4] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang, "Solving rubik's cube with a robot hand," *CoRR*, vol. abs/1910.07113, 2019.

[5] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwav: Learning robot manipulation from in-the-wild audio-visual data," *arXiv preprint arXiv:2406.19464*, 2024.

[6] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024.

[7] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021.

[8] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.