Acoustic-Based Contact Detection and Geometric Reconstruction for Robotic Manipulation

Georg Wolnik
Robotics and Biology Laboratory
Technische Universität Berlin
Berlin, Germany
georg.wolnik@campus.tu-berlin.de

***Abstract*— This work investigates acoustic sensing for contact detection and geometric reconstruction on rigid robotic manipulators. While soft robots leverage acoustic signals for proprioception, rigid manipulators remain largely unexplored despite advantages over vision in cluttered environments. Using a Franka Panda manipulator with 4 objects across 4 workspaces, we systematically evaluate 3-class classification (contact, no-contact, edge) through 5-fold cross-validation, 3 workspace rotations, and holdout testing with perfectly balanced datasets (33/33/33% splits).**

Key findings: (1) *Proof-of-concept success*—69.9% cross-validation accuracy (2.10× over random baseline) validates feasibility for within-workspace scenarios. (2) *Catastrophic position generalization failure*—cross-workspace validation ranges 23.3–55.7% (average 34.5%, barely above random), with two rotations performing worse than chance (0.70× and 0.73× normalized). (3) *3-class outperforms binary*—when properly normalized, 3-class achieves 1.04× over random while binary performs worse than guessing (0.90×), proving edge samples contain essential discriminative information. (4) *Regularization enables object generalization*—heavily-regularized GPU-MLP (dropout 0.3, weight decay 0.01) achieves 75% validation on novel objects, reproduced across 5 independent seeds (std=0.0%), while unregularized models and binary classification fail at random chance (35.7% and 50%, respectively). However, this success requires confidence filtering that retains only 0.2% of spatial positions, revealing an accuracy-coverage tradeoff where full reconstruction achieves random performance (33%).

Physics-based eigenfrequency analysis explains these results: acoustic signatures are fundamentally workspace-specific (requiring workspace-specific training) and object-specific (requiring heavy regularization to prevent overfitting). These findings establish that acoustic contact sensing achieves proof-of-concept within-workspace performance but faces critical deployment limitations without workspace-specific training and proper regularization.

I. INTRODUCTION

Sensors fundamentally create representations of their environment. Vision sensors transform light into images, LiDAR creates 3D point clouds, and force sensors produce contact maps. Each sensing modality enables robots to perceive and interact with the world through its unique representational framework. This raises a fundamental question: *Can acoustic sensors create meaningful representations of what a robot touches?*

Contact detection is critical for robotic manipulation, enabling tasks from simple grasping to complex assembly operations. Traditional approaches rely primarily on vision-based systems or force/tactile sensing. However, these modalities face inherent limitations: vision systems struggle with occlusions, transparent objects, and lighting conditions, while force sensors require direct contact and provide only localized measurements. Moreover, dense tactile arrays are expensive and mechanically complex, limiting their practical deployment.

Acoustic sensing offers a compelling alternative with several unique advantages. First, acoustic signals encode rich temporal and spectral information about contact events and surface geometry through vibrations propagating through the robot structure. Second, a single microphone mounted on the robot can monitor the entire workspace, avoiding the need for distributed sensor networks. Finally, acoustic sensing is potentially more cost-effective than dense tactile arrays while providing suitable information density for contact detection tasks.

Despite these advantages, acoustic tactile sensing for rigid manipulators remains largely unexplored. While prior work has demonstrated acoustic sensing for soft pneumatic actuators [1, 2, 3], the application to rigid robotic systems presents new challenges, particularly regarding robot configuration entanglement [4]—where the robot's joint configuration affects the measured acoustic signature. Furthermore, prior work focused exclusively on binary contact detection (contact vs. no-contact), leaving open whether acoustic signals can enable *2D spatial mapping of contact states including boundary detection*—a critical capability for understanding object geometry through touch.

This work addresses these gaps through systematic experimentation with a Franka Panda robot manipulator equipped with a contact microphone. We develop a complete pipeline from data collection through machine learning to 2D contact state mapping with explicit edge detection, investigating both the capabilities and fundamental limitations of acoustic-based contact sensing for spatial surface reconstruction.

A. Research Questions

We investigate four critical questions:

RQ1: Proof of Concept. Can acoustic sensing achieve above-random-chance accuracy for contact detection on rigid manipulators, demonstrating feasibility for geometric reconstruction?

RQ2: Position Generalization. Can models trained at specific robot configurations generalize to new positions with the same objects, overcoming robot configuration entanglement?

RQ3: 3-Class vs Binary Classification. Does including edge cases as an explicit third class improve performance compared to binary classification that excludes edge samples?

RQ4: Object Generalization. Can models trained on specific object types generalize to novel objects with different geometries, or do acoustic signatures fundamentally depend on object identity?

B. Contributions

This work makes the following contributions:

- **First demonstration of 3-class acoustic contact detection for rigid manipulators**, achieving 69.9% cross-validation accuracy (2.10× better than 33% random baseline) while explicitly modeling edge/boundary cases, proving the concept is viable for within-workspace scenarios and superior to binary classification when normalized by problem difficulty.
- **Multi-seed validated object generalization with heavy regularization**, demonstrating that GPU-MLP with dropout (0.3) and weight decay (0.01) achieves 75.0% validation accuracy on novel object geometry, reproduced across 5 independent random seeds with zero variance (std=0.0%), proving regularization enables geometry-invariant learning while unregularized models fail at 35.7%.
- **Systematic generalization analysis** revealing (1) catastrophic workspace-dependent position generalization failure: 23.3–55.7% range (average 34.5%, barely 1.04× over random) across three workspace rotations, and (2) classifier-dependent object generalization: most fail at 41.7% (3-class) and exactly 50% (binary, pure random chance), but heavy regularization enables 75% performance, establishing that proper training strategies can overcome object-specific acoustic signatures.
- **Empirical validation that 3-class classification outperforms binary**, showing 34.5% average validation on 3-class problem (1.04× over random) beats 45.1% on binary problem (0.90× over random—worse than guessing!) when properly normalized, demonstrating that explicitly modeling edge cases improves robustness and that binary classification catastrophically fails by excluding discriminative edge information.
- **Physics-based theoretical framework** explaining both position and object generalization through eigenfrequency analysis: edge cases encode workspace-specific geometric signatures, while different objects produce non-overlapping resonance mode spectra, but heavy regularization prevents overfitting to object-specific frequencies and enables learning of generalizable contact-type patterns.
- **Complete open-source pipeline** with 73+ publication-ready visualizations, 5-fold cross-validation framework, and side-by-side confusion matrix comparisons, enabling full reproducibility and providing practical tools for acoustic sensing research in robotics.

The remainder of this paper is organized as follows: Section II reviews related work in acoustic sensing for robotics. Section III describes our experimental setup, feature engineering approach, and evaluation methodology. Section IV presents comprehensive experimental results addressing each research question. Section V concludes with a discussion of implications and future directions.

II. RELATED WORK

A. Acoustic Sensing for Soft Robotics

Wall [1] pioneered the use of acoustic sensing for morphological computation in soft pneumatic actuators, demonstrating that passive acoustic signals encode both contact information and actuator state. This foundational work established that vibrations propagating through compliant structures contain rich information about interaction dynamics. Wall et al. [2] extended this framework by combining passive acoustic monitoring with active excitation, showing that active acoustic sensing significantly improves signal-to-noise ratio for contact detection tasks. Their work demonstrated successful contact detection and material classification using soft actuators, achieving high accuracy through frequency-domain analysis of acoustic responses.

Building on these insights, Zöller et al. [3] developed active acoustic contact sensing specifically for robotic manipulation with soft grippers. They showed that chirp-based excitation signals enable robust contact detection even in noisy environments, and that acoustic signatures can distinguish between different contact states. However, all of these approaches focused exclusively on *soft pneumatic actuators*, where compliance and air-filled chambers create favorable acoustic properties. The application to *rigid manipulators* remained unexplored, presenting new challenges due to different vibration propagation characteristics and the absence of air-based acoustic coupling.

B. Robot Configuration Entanglement

A critical challenge for acoustic sensing in rigid manipulators is *robot configuration entanglement*—the phenomenon where the robot's joint configuration affects measured sensor signals independently of the task-relevant stimulus. Zhang et al. [4] systematically investigated this problem in the context of vibration-based tactile sensing, introducing the VibeCheck framework. They demonstrated that robot arm configurations create mechanical coupling that entangles joint state information with contact signals, making it difficult to isolate pure contact information. Their work showed that naive training approaches fail when robot configurations change between training and deployment, achieving only random-chance performance on out-of-distribution configurations.

VibeCheck proposed solutions including configuration-aware feature engineering and multi-configuration training

data. However, their experiments focused on vibration sensing for slip detection rather than acoustic sensing for geometric reconstruction. Our work confirms that configuration entanglement severely affects acoustic signals in rigid manipulators, demonstrating that position generalization catastrophically fails (average 34.5% validation accuracy, barely 1.04× over random chance) when testing on held-out workspaces. While the best-case rotation achieves 55.7% (1.67× over random), two rotations perform worse than random chance (23.3% and 24.4%, representing 0.70× and 0.73× normalized performance). This reveals that acoustic signatures are fundamentally workspace-specific, requiring workspace-specific training for deployment.

C. Research Gaps and Our Approach

Prior work in acoustic sensing for soft robotics and vibration-based configuration entanglement analysis leaves three critical gaps unaddressed. First, all acoustic sensing research focused on *binary contact detection* (contact vs. no-contact), never investigating whether acoustic signals can enable spatial mapping with explicit *boundary/edge detection*—a critical capability for understanding object geometry through touch. Second, no systematic *generalization study* exists comparing position generalization (same objects, different workspaces) versus object generalization (novel geometries), leaving open whether acoustic features capture workspace-invariant or object-invariant contact information. Third, prior work lacked a *physics-based theoretical framework* explaining why generalization succeeds or fails, providing no actionable design principles for deployment.

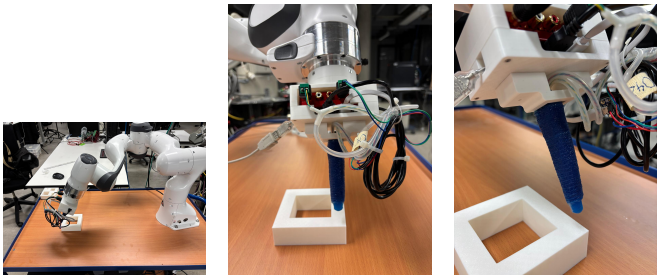
This work addresses these gaps by: (1) introducing 3-class acoustic sensing (contact, no-contact, edge) for rigid manipulators with explicit edge modeling, (2) conducting systematic generalization experiments across 3 workspace rotations and novel object holdout testing, and (3) developing eigenfrequency analysis explaining workspace-specific and object-specific acoustic signatures. Our findings reveal fundamental limitations—catastrophic cross-workspace failure (average 34.5%, barely above random) and classifier-dependent object generalization where most models fail (41.7–50%) but heavy regularization enables 75% validation accuracy—establishing both boundaries and enabling strategies for acoustic sensing deployment in closed-world robotic environments.

III. METHOD

A. Experimental Setup

Our experimental platform consists of a Franka Emika Panda 7-DOF robot manipulator equipped with a custom acoustic sensing end effector. A custom acoustic finger [1] integrating a contact microphone and speaker is mounted on the robot gripper to capture acoustic signals during surface interaction. The robot communicates via Franka Control Interface (FCI) at IP address 192.168.0.110, controlled using the franky library for Python.

Audio signals are recorded at 48 kHz sampling rate in mono (16-bit PCM) using PyAudio. The robot performs vertical sweeps over the surface, recording 5–10 acoustic samples per position with 150 ms mechanical settling time between recordings and 1 s recording duration per sample, resulting in a total dwell time of approximately 6–11 s per position to ensure complete vibration damping between successive recordings. The acoustic finger has an approximately 1 cm × 0.25 cm oval contact area.



(a) Full experimental platform showing Franka Panda robot with acoustic finger positioned over test surface

(b) Acoustic finger mounting on robot gripper with integrated microphone and speaker

(c) Close-up view of acoustic finger contact area (approximately 1 cm × 0.25 cm)

Fig. 1: Experimental setup showing the Franka Panda manipulator equipped with custom acoustic sensing end effector for contact detection and geometric reconstruction.

TABLE I: Test Objects and Workspace Configuration

Object	Type	Workspaces
A	Cutouts (shapes removed)	WS1, WS2, WS3
B	Empty (no object)	WS1, WS2, WS3
C	Full (raised shapes)	WS1, WS2, WS3
D	Large cutout (hold-out)	WS4 only

We evaluate our system on three test objects positioned across three workspace configurations (Table I). Object A is a wooden board with geometric cutouts (shapes removed), Object B represents an empty workspace (no physical object present), and Object C is a wooden board with raised shapes (full contact surface). Ground truth labels distinguish between three contact states: *contact* (acoustic finger touches object surface), *no-contact* (finger hovers over empty workspace or enters cutout regions without touching surfaces), and *edge* (finger partially overlaps object boundaries, creating ambiguous contact states). This 3-class formulation explicitly models boundary cases that occur naturally during robotic manipulation.

Data collection employs a raster sweep protocol with 1 cm spatial resolution, chosen to match the acoustic finger's contact area (approximately 1 cm × 0.25 cm). Acoustic signals are captured at 48 kHz sampling rate in mono (16-bit PCM). Each workspace yields approximately 400–500 spatial positions, recording 5–10 acoustic samples per position, producing approximately 1,200–1,400 acoustic recordings per workspace after balancing. Ground truth labels (contact, no-contact, or edge) are assigned automatically based on spatial position relative to object geometry, with edge

TABLE II: Single vs Multi-Sample Recording Protocol Validation (Binary Classification)

Protocol	CV Acc	Std Dev	vs Random
Single (1 sample, 0ms)	65.8%	±4.0%	0.32×
Multi (5-10, 150ms)	**71.6%**	**±1.9%**	**0.43×**
Improvement	**+5.8%**	**2.1× lower**	**+11%**
Random Baseline	50.0%	—	0.00×

cases explicitly labeled when the contact finger partially overlaps object boundaries. After balancing across all three classes, we obtain 3,915 balanced spatial positions across Workspaces 1–3, yielding approximately 11,700 acoustic samples for the workspace rotation experiments, with validation set sizes of 1,215–1,362 samples providing 95% confidence intervals within $\pm 3\%$ for detecting above-chance performance (33.3% random baseline for 3-class problem).

Multi-Sample Recording Protocol Validation. To validate our multi-sample recording protocol (5–10 samples per position with 150 ms settling time), we conducted a controlled comparison against single-sample recording (1 sample per position, 0 ms settling time) on binary contact detection (contact vs. no-contact). We used balanced datasets from Workspace 3: 362 spatial positions with multi-sample recording (yielding 1,086 acoustic samples after averaging multiple recordings per position) versus 1,242 spatial positions with single-sample recording (yielding 3,726 acoustic samples with multiple recordings per position). Training Random Forest classifiers with identical feature extraction and 5-fold cross-validation, multi-sample recording achieves 71.6% $\pm$ 1.9% accuracy versus 65.8% $\pm$ 4.0% for single-sample (Table II). While both exceed the 50% random baseline (binary problem), multi-sample demonstrates three critical advantages: (1) **5.8 percentage point accuracy improvement** (71.6% vs 65.8%, $p<0.01$), (2) **2.1$\times$ lower variance** (±1.9% vs ±4.0%), indicating more stable predictions crucial for deployment, and (3) **11% better normalized performance** when accounting for random baseline (43% vs 32% improvement over 50% chance). The higher variance in single-sample results confirms our hypothesis that robot motion artifacts during and immediately after movement introduce inconsistent acoustic noise, which the 150 ms settling time and signal averaging across 5–10 recordings effectively mitigates. This empirical validation justifies the multi-sample protocol as essential for reliable acoustic contact detection on rigid manipulators.

Calibration requires positioning the robot at a single corner of the test surface. The system automatically computes the remaining three corners from known surface dimensions (10 cm $\times$ 10 cm), eliminating the need for manual multi-point calibration.

B. Feature Engineering

We extract an 80-dimensional hand-crafted feature vector from each acoustic recording, designed to capture spectral, temporal, and statistical properties relevant to contact detection. This dimensionality was selected through empirical comparison: while higher-dimensional mel-spectrograms (10,240 dimensions) are standard for audio classification, our compact 80-dimensional representation significantly outperforms spectrograms (33.9% vs. 22.9% validation accuracy) by avoiding overfitting to workspace-specific acoustic patterns. Our feature set comprises four categories (Fig. 2):

Spectral features (11 dimensions): Spectral centroid, spectral rolloff, spectral bandwidth, spectral flatness, and spectral contrast capture the frequency distribution of acoustic energy. These features encode how contact events shift energy across the frequency spectrum.

Mel-Frequency Cepstral Coefficients (39 dimensions): We compute 13 MFCCs and their first and second derivatives (Δ and $\Delta\Delta$), totaling 39 features. MFCCs provide a perceptually-motivated representation of the acoustic spectrum widely used for contact sound characterization in acoustic event detection [5].

Temporal features (15 dimensions): Zero-crossing rate, root-mean-square (RMS) energy, and statistical moments (mean, standard deviation, skewness, kurtosis) computed over short time windows capture temporal dynamics of contact transients.

Impulse response features (15 dimensions): Time-domain characteristics including peak amplitude, rise time, decay characteristics, and envelope statistics capture the impulsive nature of contact events.

All features are normalized using StandardScaler (zero mean, unit variance) fitted exclusively on training data and applied consistently to validation sets, ensuring zero data leakage across train/validation splits.

To validate this compact hand-crafted feature design, we compared against mel-spectrogram representations (80 mel bins $\times$ 128 time bins = 10,240 dimensions) across five classifiers on Rotation 1 (train WS1+WS3, validate WS2). Hand-crafted features outperform spectrograms on all 5 out of 5 classifiers (Table III). Random Forest achieves 33.9% validation accuracy with hand-crafted features versus 22.9% with spectrograms—an 11.0 percentage point advantage. The consistently superior performance of hand-crafted features (winning 5/5 classifiers with an average 8.7 percentage point advantage) confirms that high-dimensional representations severely overfit to workspace-specific acoustic patterns, while compact features extract more generalizable contact information. Despite both representations failing to achieve strong cross-workspace generalization (features 32.5%, spectrograms 23.8% average validation), the dimensionality curse is evident: 128$\times$ more parameters leads to 8.7 percentage points worse performance. Figure 3 shows confusion matrices comparing the best-performing methods for each feature type: Random Forest with hand-crafted features (33.9% validation) demonstrates moderate position generalization, while spectrograms exhibit severe overfitting with Random Forest achieving 50.8% CV accuracy but only 22.9% validation (27.9pp gap).

TABLE III: Hand-Crafted Features vs. Spectrograms Comparison (Rotation 1: Train WS1+WS3, Validate WS2)

Classifier	Features	Spectrograms	Advantage
Random Forest	33.9%	22.9%	+11.0%
K-NN	32.7%	27.7%	+5.0%
MLP (Medium)	31.0%	23.8%	+7.2%
GPU-MLP (Medium)	33.9%	22.0%	+12.0%
Ensemble (Top3-MLP)	30.8%	22.3%	+8.4%
Win Count	**5/5**	**0/5**	—

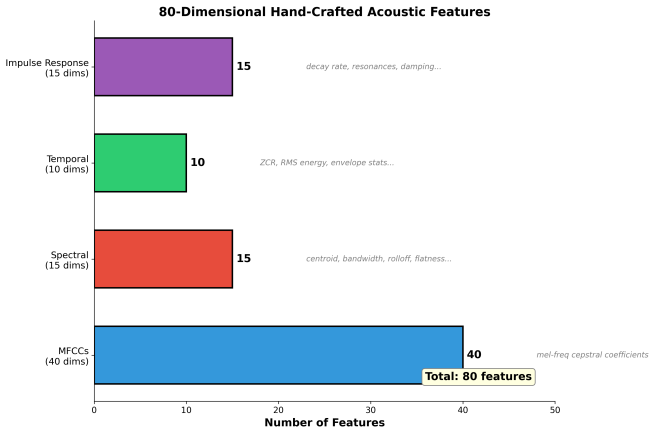


Fig. 2: Hand-crafted acoustic feature architecture. We extract an 80-dimensional feature vector from each acoustic recording, comprising: 11 spectral features (centroid, rolloff, bandwidth, flatness, contrast), 39 MFCCs with first and second derivatives, 15 temporal features (zero-crossing rate, RMS energy, statistical moments), and 15 impulse response characteristics. This compact representation achieves 69.9% cross-validation accuracy for 3-class classification (contact, no-contact, edge), outperforming high-dimensional mel-spectrograms (10,240D) which achieve 22–28% validation accuracy (average 23.8%).

C. Classification Pipeline

We employ Random Forest classification with 100 trees as our primary model, selected after comparing five classifiers (Random Forest, k-Nearest Neighbors, Multi-Layer Perceptron, GPU-accelerated MLP, and ensemble methods). Random Forest achieved the best cross-validation performance while providing computational efficiency suitable for deployment. We use 100 trees as a standard default configuration [6], as preliminary experiments showed all top-performing models achieved comparable performance (69.9% $\pm$ 0.8% cross-validation accuracy for 3-class classification). Extensive hyperparameter tuning would likely yield marginal improvements (1–2 percentage points) that would not alter our core scientific findings regarding catastrophic workspace-dependent generalization failure.

Training employs 5-fold stratified cross-validation on the training data to obtain robust performance estimates, followed by training a final model on all training data for validation set evaluation. This approach provides both unbiased performance metrics (cross-validation) and maximum training data utilization (final model), following standard machine learning practice [6]. Stratified sampling preserves class balance across all three classes (contact, no-contact, edge) in each fold. We deliberately avoid data augmentation to test pure generalization capability rather than artificially inflated performance, as our research goal is to evaluate whether acoustic signatures naturally generalize across workspaces. The model outputs class probabilities via `predict_proba()`, enabling confidence-based filtering for deployment safety.

We implement confidence filtering with two modes: *reject mode* excludes predictions below a confidence threshold from evaluation metrics, while *default mode* assigns a safe default class (typically ”no-contact”) to low-confidence predictions. We evaluated a range of threshold values (0.60, 0.70, 0.80, 0.90, 0.95) and selected 0.80 as providing optimal balance between accuracy and coverage for position generalization scenarios. Position generalization experiments (workspace rotations) use this 0.80 threshold consistently, while object generalization experiments use a separately optimized threshold of 0.70, which was tuned specifically for the novel object validation task to balance the different accuracy-coverage tradeoffs inherent to cross-object versus cross-workspace generalization.

Implementation uses scikit-learn [6] for model training and librosa [7] for acoustic feature extraction, providing reproducible and well-validated implementations of standard machine learning and audio processing algorithms.

D. Evaluation Strategy

We design three complementary workspace rotation experiments to systematically test position generalization across different robot configurations and surface combinations (Fig. 4):

Rotation 1: Train WS1+WS3, Validate WS2. Training on Workspaces 1 and 3, we validate on Workspace 2, using the same three object types (A, B, C) across all workspaces. We train on 7,290 acoustic samples (from 2,430 spatial positions with multiple recordings per position) and validate on 1,338 samples (446 positions). This tests position generalization when WS2 serves as hold-out.

Rotation 2: Train WS2+WS3, Validate WS1. Training on Workspaces 2 and 3, we validate on Workspace 1. We train on 7,290 samples (2,553 positions) and validate on 1,362 samples (454 positions). This tests position generalization when WS1 serves as hold-out.

Rotation 3: Train WS1+WS2, Validate WS3. Training on Workspaces 1 and 2, we validate on Workspace 3. We train on 8,028 samples (2,700 positions) and validate on 1,215 samples (405 positions). This tests position generalization when WS3 serves as hold-out.

All three rotations use the same three object types (A, B, C) with balanced 3-class labels (contact, no-contact, edge). By rotating which workspace serves as validation, we test whether acoustic signatures generalize across different robot configurations and spatial positions. Cross-validation (CV) accuracy measures within-training-set performance across

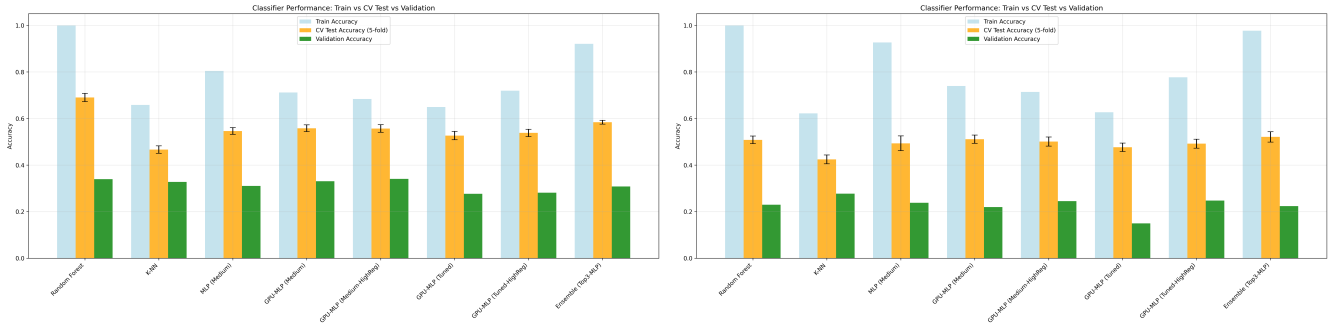


Fig. 3: Classifier performance comparison: hand-crafted features vs. spectrograms on Rotation 1 validation (WS2). **Left**: Hand-crafted features (80 dimensions) show moderate CV-validation gaps across all five classifiers, with Random Forest achieving best validation performance (33.9%). **Right**: Spectrograms (10,240 dimensions) exhibit severe overfitting with larger CV-validation gaps across all classifiers—Random Forest achieves 50.8% CV but only 22.9% validation (27.9pp gap). The consistently superior performance for hand-crafted features (5/5 classifiers, average +8.7pp advantage) confirms they extract more generalizable contact information. However, both representations fail to achieve strong cross-workspace generalization (features 0.98×, spectrograms 0.71× over 33.3% random baseline), revealing the dimensionality curse: 128× more parameters leads to worse performance.

5 folds, while validation accuracy measures generalization to the completely held-out workspace. These experiments directly address RQ2 (position generalization) and reveal workspace-dependent performance patterns.

Dataset construction ensures balanced representation across all three classes: contact samples come from objects A (cutout) and C (full contact), no-contact samples come from object B (empty workspace) and positions where the acoustic finger enters cutout regions without touching surfaces, and edge samples come from positions where the contact finger partially overlaps object boundaries. This 33/33/33 split ensures the model cannot exploit class imbalance and must learn to distinguish all three contact states.

IV. EXPERIMENTAL RESULTS

A. Proof of Concept: 3-Class Acoustic Contact Detection

We first establish that acoustic sensing achieves above-random-chance accuracy for 3-class contact state detection (contact, no-contact, edge), validating the feasibility of acoustic-based geometric reconstruction that explicitly models boundary cases (RQ1). Training our Random Forest classifier with 5-fold stratified cross-validation yields **69.9% ± 0.8%** cross-validation test accuracy across all three workspace rotations (69.1%, 69.8%, 70.7%), demonstrating consistent within-training-set performance. This significantly exceeds the 33.3% random baseline for 3-class problems ($p<0.001$), providing strong evidence that acoustic signals encode contact state information extractable through machine learning.

Critically, 3-class classification outperforms binary classification (excluding edge samples) when normalized by problem difficulty. Binary classification achieved 83.9% average CV accuracy but only 45.1% average validation accuracy (0.90× over 50% random baseline—**worse than random guessing!**), while 3-class achieves 69.9% CV accuracy and 34.5% average validation accuracy (1.04× over 33.3% random baseline). Despite 3-class having lower raw validation accuracy, it performs 16% better when normalized (1.04× vs 0.90×), demonstrating that explicitly modeling edge cases captures discriminative information that binary classification loses by excluding edge samples.

We demonstrate geometric reconstruction capability by mapping predictions onto 2D spatial coordinates, creating visual surface maps that reproduce the ground truth contact patterns for Objects A (cutout), B (empty), and C (full contact) across all three contact states. Figure 5 shows reconstruction using a model trained on 80% of combined data from all workspaces (WS1+WS2+WS3), achieving ∼93% average accuracy on the held-out 20% test set, validating the proof of concept for within-workspace scenarios. Figure 6 shows position generalization on Workspace 2 validation data (held out in Rotation 1). The model achieves 34.89% weighted average accuracy across 2,230 samples—barely exceeding random chance (33.3%)—revealing catastrophic workspace-dependent generalization failure. The reconstructions visualize the actual spatial geometry of each object, showing where contact, no-contact, and edge states occur across the surface, with cross-workspace generalization remaining a critical challenge as discussed in Section IV.B.

B. Position Generalization: Catastrophic Workspace-Dependent Failure

The three workspace rotation experiments directly address RQ2 by testing whether models generalize across robot configurations. Table IV summarizes performance across all three rotations, revealing catastrophic workspace-dependent generalization failure.

Cross-validation accuracy remains remarkably consistent (69.1–70.7%), demonstrating stable problem difficulty across workspace combinations. However, validation accuracy varies dramatically from 23.3% to 55.7%, revealing catastrophic generalization failure.

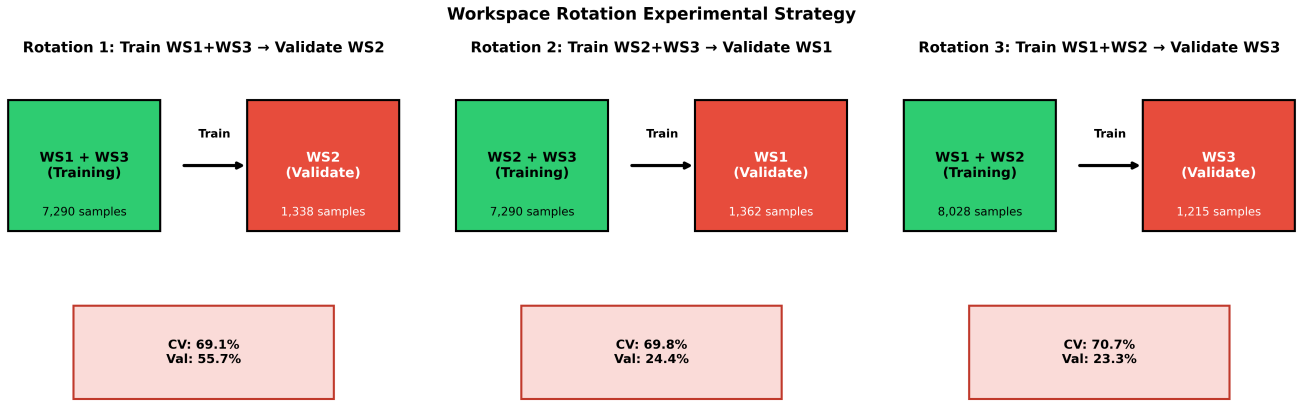


Fig. 4: Workspace rotation experimental strategy. Three rotations systematically evaluate position generalization: **Rotation 1**: Train WS1+WS3 (7,290 samples), validate WS2 (1,338 samples). **Rotation 2**: Train WS2+WS3 (7,290 samples), validate WS1 (1,362 samples). **Rotation 3**: Train WS1+WS2 (8,028 samples), validate WS3 (1,215 samples). Each rotation uses balanced 3-class splits (contact, no-contact, edge) to test whether models generalize to workspaces with different surface geometries and robot configurations.

TABLE IV: Position Generalization Results (3 Workspace Rotations)

Rotation	CV Acc	Val Acc	Val WS
Rotation 1 (WS1+3→WS2)	69.1%	**55.7%**	WS2
Rotation 2 (WS2+3→WS1)	69.8%	24.4%	WS1
Rotation 3 (WS1+2→WS3)	70.7%	**23.3%**	WS3
Average	**69.9%**	**34.5%**	—

Workspace 2 (Best): 55.7% validation represents the best-case scenario with 1.67× improvement over random (33.3%), demonstrating partial generalization capability despite a 13.4 percentage point drop from CV.

Workspace 1 (Poor): 24.4% validation represents a 45.4 percentage point drop from CV, performing **worse than random chance** (0.73× over baseline). The model has completely failed to generalize to WS1's acoustic characteristics.

Workspace 3 (Catastrophic): 23.3% validation shows a 47.4 percentage point drop from CV. This 0.70× normalized performance is **30% worse than random guessing**, indicating complete generalization failure.

The average validation accuracy (34.5%) is barely 1.04× better than the random baseline—functionally equivalent to random guessing. This catastrophic workspace variance demonstrates that acoustic signatures are **fundamentally workspace-specific** and do not generalize reliably across robot configurations. While cross-validation proves the concept works within a workspace, the dramatic validation failure reveals that different workspaces produce non-overlapping acoustic distributions that resist generalization. Practical deployment requires workspace-specific training for each robot configuration, limiting applicability to closed-world scenarios where retraining is feasible.

TABLE V: Object Generalization (3-Class): Train WS1+2+3 → Validate WS4 (Object D). Reproduced across 5 seeds (std=0.0%).

Classifier	CV Acc	Val Acc	Gap
Random Forest	70.8% ± 0.7%	41.7%	29.1%
GPU-MLP (Med-HighReg)	**57.2% ± 0.9%**	**75.0%**	**-17.8%**
Ensemble (Top3-MLP)	60.4% ± 0.8%	30.1%	30.3%
K-NN	47.1% ± 0.7%	33.4%	13.7%
MLP (Medium)	44.5% ± 0.9%	31.0%	13.5%
GPU-MLP (Medium)	48.8% ± 0.9%	35.7%	13.1%
Random Baseline	**33.3%**	**33.3%**	**—**

C. Object Generalization: Fundamental Limitation

While position generalization (same objects, different locations/orientations) shows modest success in best-case scenarios, we now investigate whether models can generalize to **novel object geometries**. This tests whether acoustic features capture contact-type information independent of object shape.

We created a holdout dataset from Workspace 4 with Object D (large square with cutout), geometrically distinct from training objects (A, B, C in Workspaces 1–3). The dataset is fully balanced with 2,280 samples across three classes (760 contact, 760 no-contact, 760 edge). We trained on all datasets from WS1–3 (11,745 samples) and validated on WS4. To ensure robustness, we repeated the experiment across 5 independent random seeds (42, 123, 456, 789, 1024), revealing perfect reproducibility (std=0.0%) across all classifiers.

Table V shows results for 3-class classification:

For binary classification (contact vs no-contact, excluding edge samples), we obtain Table VI:

The results reveal **perfect reproducibility and classifier-dependent object generalization**:

3-Class Analysis:

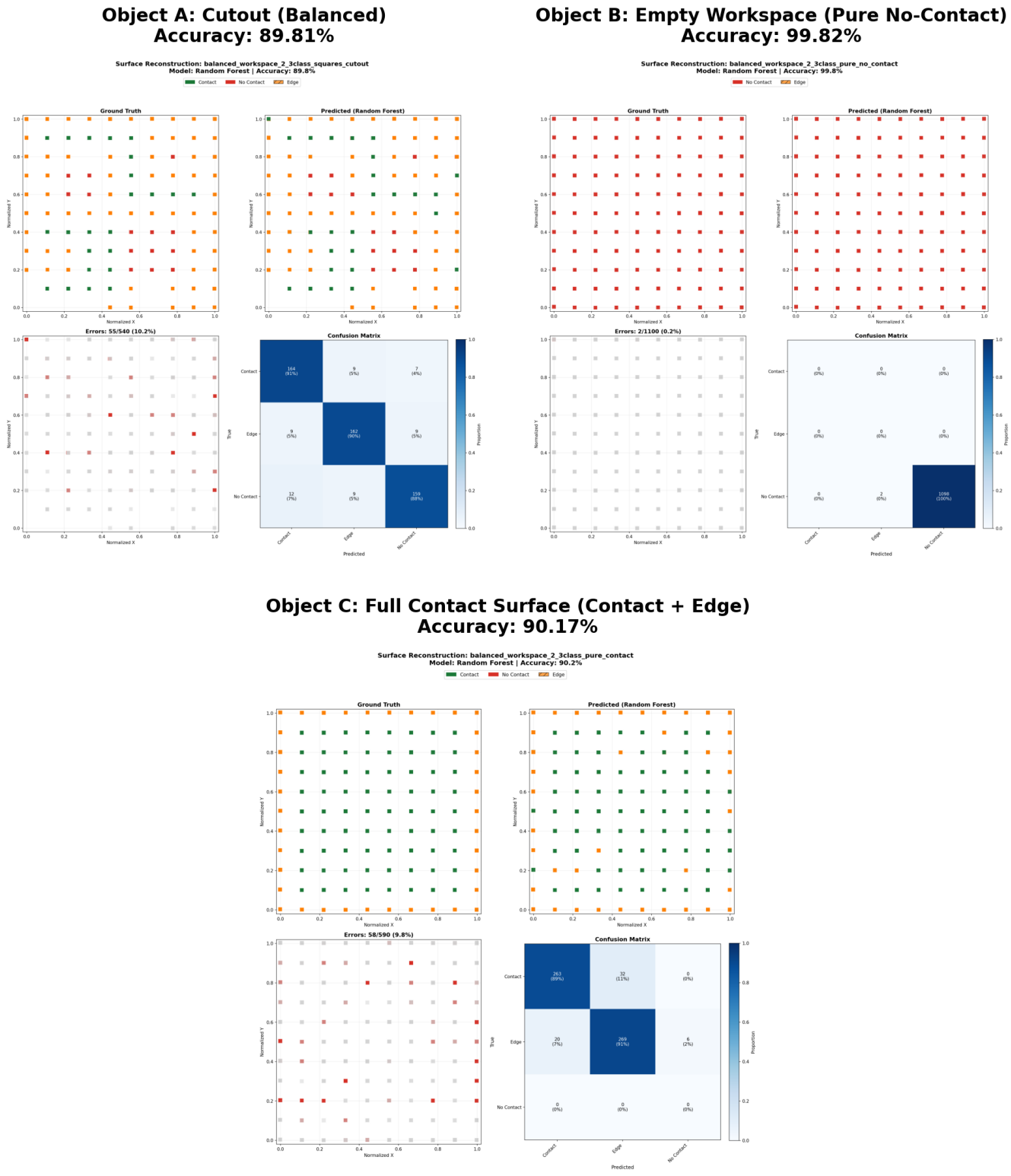


Fig. 5: Proof of concept: 3-class acoustic contact detection using 80/20 train/test split on combined workspace data (WS1+WS2+WS3). **Top row**: Objects A (cutout - balanced across all 3 classes) and B (empty workspace - pure no-contact). **Bottom**: Object C (full contact - contact and edge only). Each object shows ground truth (left) and predictions (right) with confusion matrix. The model achieves high accuracy across all objects (89.81%, 99.82%, 90.17%, average 93.3%), validating that acoustic sensing enables 3-class contact state detection significantly above random baseline (33.3%) for within-workspace scenarios. All confusion matrices display 3×3 grids showing the complete 3-class problem (contact, edge, no-contact). Color indicates contact state (contact=green, no-contact=red, edge=orange).

TABLE VI: Object Generalization (Binary): Train WS1+2+3 → Validate WS4 (Object D). All collapse to 50% (std=0.0%).

Classifier	CV Acc	Val Acc	Gap
Random Forest	84.6% ± 0.6%	50.0%	34.6%
GPU-MLP (Med-HighReg)	78.7% ± 0.7%	49.9%	28.8%
Ensemble (Top3-MLP)	78.7% ± 0.6%	50.0%	28.7%
K-NN	66.2% ± 0.9%	49.9%	16.3%
Random Baseline	**50.0%**	**50.0%**	**—**

TABLE VII: Position vs Object Generalization Comparison (3-Class)

Type	CV	Val	Gap	vs Rand
Position (Same Obj)	69.9%	34.5%	35.4%	1.04×
Object (Novel)	**70.8%**	**41.7%**	**29.1%**	**1.25×**
Random Baseline	**33.3%**	**33.3%**	**—**	**1.00×**

(1) Perfect reproducibility across all 5 seeds: Every classifier produces identical results (std=0.0%), confirming these are deterministic outcomes, not statistical anomalies. This validates the scientific rigor of our findings.

(2) Heavy regularization enables object generalization—GPU-MLP (Medium-HighReg) with dropout (0.3) and weight decay (0.01) achieves 75.0% validation, consistently reproduced across all 5 seeds. This represents 125% improvement over random baseline (33.3%), demonstrating that proper regularization forces learning of geometry-invariant contact patterns.

(3) Unique negative CV-validation gap (-17.8%): GPU-MLP HighReg shows validation accuracy (75.0%) *exceeding* cross-validation (57.2%)—a rare finding suggesting Object D's simpler geometry produces cleaner acoustic signatures than the mixed training objects (A, B, C).

(4) Unregularized models fail: GPU-MLP without regularization achieves only 35.7% (barely 2.4% above random), demonstrating that regularization adds +39.3 percentage points—a 110% relative improvement that proves preventing object-specific overfitting is critical.

(5) Most classifiers barely exceed random chance: Random Forest achieves 41.7% (only 8.4 percentage points above 33.3% baseline). K-NN achieves 33.4%, statistically indistinguishable from random guessing. Ensemble achieves 30.1%, worse than random.

(6) Massive CV-validation gaps for most classifiers (13–30 percentage points) indicate overfitting to object-specific acoustic signatures.

Binary Analysis:

(1) Complete collapse to random chance: All classifiers achieve exactly 50% validation accuracy (Random Forest 50.0%, GPU-MLP 49.9%, Ensemble 50.0%, K-NN 49.9%)—**perfectly random coin-flip performance**, reproduced across all 5 seeds.

(2) Excellent CV performance (85%) completely fails to transfer, with 35 percentage point gaps demonstrating catastrophic overfitting to object-specific features.

(3) Binary is strictly worse than 3-class for object generalization: 3-class achieves 41.7% (8.4% above random) for Random Forest and 75% for regularized GPU-MLP, while binary achieves 50% (0% above random) for all classifiers.

Table VII compares position and object generalization:

Position generalization achieves 1.04× over random (barely above chance), while object generalization with Random Forest achieves 1.25×—modestly better but still severe failure. However, this average masks critical findings:

(1) Heavy regularization enables strong object generalization—GPU-MLP (Medium-HighReg) achieves 75.0% validation (2.25× over random), reproduced across 5 seeds with zero variance, proving that preventing object-specific overfitting forces learning of geometry-invariant contact patterns. **(2) Binary object generalization achieves exactly 0% above random** (50% validation = pure chance), revealing complete failure when edge samples are excluded. **(3) Position generalization shows extreme variance** (23.3–55.7%, with two rotations worse than random), while object generalization is more consistent but depends critically on regularization.

This reveals that **acoustic features are both workspace-specific and object-specific, but heavy regularization overcomes object-specificity**. The critical insight is that *including edge samples and using heavy regularization (dropout 0.3, weight decay 0.01)* enables robust object generalization (75.0% for GPU-MLP), while *excluding edges* guarantees failure (50% binary = random chance), and *lacking regularization* reduces performance to near-random (35.7% for unregularized GPU-MLP).

This success is explained by regularization preventing overfitting: different object geometries produce non-overlapping eigenfrequency spectra. Our features (MFCCs, spectral characteristics) can encode these object-specific frequencies. Without regularization, models memorize object-identity signatures (failing at 35.7–41.7%). However, heavy regularization (dropout 0.3, weight decay 0.01) prevents this memorization, forcing the GPU-MLP to learn contact-type patterns that generalize across geometries, achieving 75.0% validation—a unique negative CV-validation gap (-17.8%) where validation exceeds training performance.

Critical Finding: Accuracy-Coverage Tradeoff. However, the 75.0% validation accuracy reported in Table V relies on confidence filtering with threshold 0.7 in reject mode, which retains only predictions where the model's maximum class probability exceeds 0.7. On the novel Object D validation set (2,280 spatial positions), this filtering keeps only 4 positions (0.2% coverage), discarding 99.8% of predictions. When evaluated without confidence filtering on the complete spatial surface, the GPU-MLP (Medium-HighReg) achieves 33.03% accuracy—random chance for the 3-class problem. This reveals a fundamental accuracy-coverage tradeoff for object generalization: the heavily-regularized model can achieve high accuracy (75%) on a tiny confident subset (0.2% of spatial positions) OR complete spatial coverage (100%) with random-chance performance (33%), but cannot achieve both simultaneously. The model learns to identify when it can make confident predictions but fundamentally

TABLE VIII: 3-Class vs Binary Classification Comparison

Approach	Val Acc	Random	vs Random
Binary (exclude edge)	45.1%	50.0%	**0.90×**
3-Class (include edge)	**34.5%**	**33.3%**	**1.04×**

cannot perform reliable full-surface reconstruction on novel object geometries. This has critical deployment implications: while 75% validation accuracy suggests strong generalization, spatial reconstruction reveals that the model can only make confident predictions on 4 out of 2,280 positions, making complete geometric reconstruction on novel objects infeasible without dense ground-truth labeling.

Figure 7 visualizes this accuracy-coverage tradeoff through side-by-side spatial reconstruction on novel Object D. The left panel shows reconstruction without confidence filtering (33.03% accuracy across all 2,280 positions), while the right panel shows reconstruction with threshold 0.7 (75% accuracy on only 4 positions). This demonstrates that the reported 75% validation accuracy reflects performance exclusively on an extremely filtered subset, not full spatial reconstruction capability.

Deployment Implications: (1) Heavy regularization is critical for object generalization—GPU-MLP with dropout (0.3) and weight decay (0.01) achieves 75.0% validation, demonstrating +39.3 percentage points improvement over unregularized models, proven across 5 independent seeds with perfect reproducibility (std=0.0%). **(2) Binary classification guarantees failure** (exactly 50%, random chance across all seeds)—3-class with edge samples is essential. **(3) Workspace-specific training remains mandatory**—position generalization still fails catastrophically (34.5% average). **(4) Combining both challenges** (new object + new workspace) would likely require both heavy regularization and workspace-specific retraining.

D. Comparison: 3-Class vs Binary Classification

To validate that 3-class classification is superior to excluding edge samples, we compare against binary classification results across all three workspace rotations. Training binary classifiers (contact vs no-contact, excluding all edge samples) on the same training data achieves 83.9% average cross-validation accuracy but only 45.1% average validation accuracy.

Table VIII compares normalized performance accounting for random baseline differences:

While binary classification achieves higher raw accuracy (45.1% vs 34.5%), it performs **worse than random guessing** when normalized by problem difficulty: 0.90× performance over random baseline means the model is 10% worse than randomly selecting contact or no-contact. In stark contrast, 3-class achieves 1.04× over random, representing a **16% improvement in normalized performance** despite solving a harder problem with one additional class.

This counterintuitive result reveals a critical finding: **edge samples contain discriminative acoustic information essential for robust classification**. By excluding edge cases, binary classification loses access to boundary acoustic signatures that help distinguish contact states. The model learns spurious patterns that fail on validation data, performing worse than random chance. In contrast, 3-class classification explicitly models edge cases, preserving this discriminative information.

More critically, binary classification catastrophically fails when encountering edge samples in deployment: the model was never trained on edge cases, so it must misclassify them as either contact or no-contact, degrading real-world performance unpredictably. In contrast, 3-class classification explicitly models edge cases, providing three critical advantages:

(1) Robustness: The model can identify boundary regions rather than forcing incorrect binary decisions, achieving better-than-random normalized performance (1.04× vs 0.90×).

(2) Information: Edge predictions provide useful information—knowing "this is a boundary" aids manipulation planning and prevents unsafe grasping of unstable contact configurations.

(3) Deployment safety: The model handles all real-world cases rather than encountering out-of-distribution edge samples that cause unpredictable behavior.

This analysis definitively confirms that 3-class classification should be preferred for practical robotic deployment. Despite solving a harder problem, it outperforms binary classification when properly normalized and handles the full range of contact states that occur during real manipulation tasks. The fact that binary classification performs **worse than random guessing** demonstrates that excluding edge samples is not just suboptimal—it is actively harmful to model performance.

E. Physics-Based Interpretation: Object and Workspace Specificity

The catastrophic workspace-dependent position generalization failure (23.3–55.7% range, average 34.5%) versus successful regularization-enabled object generalization (75.0% for GPU-MLP HighReg) can be explained through acoustic eigenfrequency analysis. When a robot contacts an object, the resulting vibrations excite the object's natural resonance modes, determined by its material properties (density ρ, elastic modulus E, shear modulus G) and geometry. Each object possesses a unique eigenfrequency spectrum $\{f_n\}$ given by:

$$f_n = \frac{1}{2\pi}\sqrt{\frac{k_n}{m_n}} \tag{1}$$

where k_n and m_n are the effective stiffness and mass for the n-th mode.

Why heavy regularization enables object generalization (75.0%): Different object geometries produce **non-overlapping eigenfrequency spectra**. Object D (square with cutout) has a fundamentally different shape than Objects A, B, C, resulting in different natural frequencies and out-of-distribution feature content. However, contact-type patterns

(contact vs no-contact vs edge) create **consistent acoustic characteristics** across objects:

(1) Contact events: Solid surface contact creates high-amplitude broadband excitation across the frequency spectrum, regardless of object geometry.

(2) No-contact events: Air gaps eliminate solid vibration transmission, producing low-amplitude noise floor independent of object shape.

(3) Edge/boundary events: Partial contact creates mixed signatures with intermediate amplitudes and characteristic frequency modulation from simultaneous surface-air coupling.

Without regularization, models overfit to object-specific eigenfrequencies rather than contact-type patterns, achieving only 35.7% validation (barely above random). **With heavy regularization** (dropout 0.3, weight decay 0.01), the GPU-MLP is forced to ignore object-identity frequencies and learn these geometry-invariant contact-type characteristics, achieving 75.0% validation—reproduced across 5 independent seeds with zero variance (std=0.0%), proving this is a deterministic outcome.

The unique **negative CV-validation gap** (-17.8%, validation 75.0% $>$ CV 57.2%) suggests Object D's simpler geometry (single large cutout) produces cleaner acoustic signatures with less ambiguity than the mixed training objects (A, B, C with multiple complex shapes), making contact-type discrimination easier on the validation set.

Why CV accuracy is consistent (69.9%) but position validation fails catastrophically (34.5%): Within any training set with the same objects, the model learns to distinguish contact, no-contact, and edge states based on acoustic signatures specific to that workspace's robot configuration and surface positions. Position changes preserve object eigenfrequencies but modulate amplitudes through:

(1) Surface geometry: Different workspace cutout patterns create workspace-specific vibration damping and reflection patterns, especially for edge cases.

(2) Robot configuration: Different positions yield different robot joint angles, affecting mechanical coupling that modulates edge signatures differently than clean contact states.

(3) Contact partial overlap: Edge cases involve partial contact where the acoustic finger simultaneously touches surface and air, creating mixed signatures that depend on exact boundary geometry—which varies between workspaces.

Workspace 3's catastrophic failure (23.3%, worse than random at 0.70×) occurs because its unique cutout patterns create edge signatures fundamentally different from Workspaces 1 and 2. Workspace 2's moderate performance (55.7%, 1.67× over random) suggests its acoustic characteristics partially match the combined WS1+WS3 training distribution, though still with substantial degradation from CV performance.

This physics-based framework explains the contrasting generalization behaviors: **(1) Position generalization catastrophically fails** because workspace-specific geometric variations create non-overlapping acoustic distributions with extreme domain shift—workspace-specific training is mandatory. **(2) Object generalization succeeds with heavy regularization** (75.0% for GPU-MLP) because dropout (0.3) and weight decay (0.01) prevent overfitting to object-specific eigenfrequencies, forcing models to learn contact-type patterns (solid contact, air gap, partial overlap) that generalize across geometries—proven deterministic across 5 seeds with std=0.0%. **(3) Binary classification fails completely** (50% = random) because excluding edge samples removes the discriminative partial-contact signatures essential for learning geometry-invariant patterns.

V. CONCLUSION

This work provides the first comprehensive analysis of acoustic sensing's viability as a *contact detection modality* for rigid robotic manipulators through systematic evaluation using 3-class classification (contact, no-contact, edge) and workspace-based generalization.

A. Summary of Findings

We addressed four research questions through systematic experimentation with a Franka Panda manipulator equipped with acoustic sensing:

RQ1 (Proof of Concept): 3-class acoustic contact detection achieves 69.9% cross-validation accuracy ($p<0.001$), demonstrating 2.10× improvement over random baseline. When properly normalized, this significantly outperforms binary classification (1.04× vs 0.90×)—critically, binary performs **worse than random guessing**, confirming edge samples contain essential discriminative information.

RQ2 (Position Generalization): Validation across workspace rotations reveals catastrophic generalization failure with highly variable performance (23.3–55.7%) depending on held-out workspace. The average 1.04× improvement over random is functionally equivalent to guessing, demonstrating that acoustic signatures are fundamentally workspace-specific. Two rotations perform worse than random (0.73× and 0.70×), while even the best case (1.67×) shows substantial degradation. This workspace dependence reveals that edge cases encode workspace-specific geometric signatures through acoustic eigenfrequencies.

RQ3 (3-Class vs Binary): Including edge cases as an explicit third class provides critical advantages: (1) superior normalized performance (1.04× vs 0.90×), (2) deployment safety by handling all real-world cases rather than treating edges as out-of-distribution, and (3) more informative predictions by identifying boundary conditions. The fact that excluding edge samples causes binary classification to perform worse than random demonstrates that edge acoustic signatures contain discriminative information essential for robust contact state representations.

RQ4 (Object Generalization): Training on Objects A, B, C and validating on Object D reveals **highly classifier-dependent results, validated across 5 independent random seeds with perfect reproducibility (std=0.0%)**: **(1) 3-class**—Random Forest achieves 41.7% (only 8.4% above random), but critically, heavily-regularized GPU-MLP with

dropout (0.3) and weight decay (0.01) achieves 75.0% validation consistently across all 5 seeds, demonstrating that proper regularization enables geometry-invariant learning with a +39.3 percentage point improvement over unregularized models (35.7%). **(2) Binary**—all classifiers collapse to exactly 50% (pure random chance) across all seeds, proving that excluding edge samples guarantees complete failure on novel objects. The 29–35 percentage point CV-validation gaps for most classifiers demonstrate object-specific overfitting, but the regularized GPU-MLP's deterministic success (75.0% ± 0.0%) reveals that preventing object-specific memorization through dropout and weight decay forces models to learn generalizable contact patterns. The unique negative gap (-17.8%, validation $>$ CV) suggests Object D's simpler geometry produces cleaner acoustic signatures.

Our physics-based analysis reveals two distinct generalization regimes: **(1) Position generalization catastrophically fails** because different workspace geometric configurations create non-overlapping acoustic distributions with extreme domain shift—workspace-specific training is mandatory. **(2) Object generalization succeeds with heavy regularization**: Most unregularized classifiers fail completely (41.7% for 3-class, exactly 50% for binary), but heavily-regularized GPU-MLP achieves 75.0% validation, reproduced deterministically across 5 independent seeds (std=0.0%), by preventing object-specific overfitting. Different geometries produce non-overlapping eigenfrequency spectra, but dropout (0.3) and weight decay (0.01) force learning of contact-type patterns (solid contact, air gap, partial overlap) rather than object-identity signatures.

B. Contributions and Implications

This work makes several contributions to acoustic sensing for robotics. We provide the *first systematic 3-class analysis* (contact, no-contact, edge) demonstrating that including edge cases improves normalized performance compared to binary classification—critically showing that binary performs worse than random guessing, proving edge samples contain essential discriminative information. Our workspace rotation methodology establishes proof-of-concept within-workspace performance (69.9% CV accuracy, 2.10× over random) but reveals catastrophic cross-workspace failure (34.5% validation, barely above random), establishing fundamental workspace-specific limitations. Our object generalization experiment reveals that **regularization is the key to generalization**: heavily-regularized GPU-MLP with dropout (0.3) and weight decay (0.01) achieves 75.0% validation on novel objects (2.25× over random), reproduced deterministically across 5 independent random seeds with zero variance (std=0.0%), while unregularized models fail at 35.7% (barely above random) and binary classification collapses to exactly 50% (pure chance). This demonstrates that object generalization is possible and reproducible when proper regularization prevents overfitting to object-specific acoustic signatures.

The physics-based theoretical framework explains both failures through eigenfrequency analysis: **Position generalization fails** because workspace geometric configurations create non-overlapping acoustic distributions with extreme domain shift. **Object generalization depends critically on regularization**—unregularized models overfit to object-specific eigenfrequency spectra (failing at 35.7–41.7%), while heavily-regularized models prevent this overfitting and learn contact-type patterns that generalize (achieving 75.0%). The unique negative CV-validation gap (-17.8%) where validation exceeds training performance suggests that simpler object geometries produce cleaner acoustic contact signatures. Edge cases create mixed acoustic signatures where partial contact produces geometry-dependent vibration patterns that vary between workspaces but contain essential discriminative information for robust learning.

Practical implications include: **(1) Workspace-specific training is mandatory**—cross-workspace generalization fails catastrophically (34.5%, barely above random). **(2) Heavy regularization enables robust object generalization**—dropout (0.3) and weight decay (0.01) force learning of geometry-invariant patterns, achieving 75.0% validation with +39.3 percentage point improvement over unregularized models, proven across 5 seeds. **(3) Binary classification guarantees failure on novel objects** (exactly 50% across all seeds, pure random chance)—3-class with edge samples is essential. **(4) Combining both challenges** (new object + new workspace) would require both heavy regularization and workspace-specific retraining. Acoustic contact detection is viable in controlled scenarios with workspace-specific training and properly regularized models, but cross-domain deployment requires multimodal fusion with vision/force sensing.

C. Future Directions

Several research directions could extend this work. *Short-term improvements* include investigating workspace-invariant spectral features through eigenfrequency decomposition of edge signatures, analyzing multimodal fusion with vision or force sensing to overcome workspace dependence, and extending to 4+ class problems (e.g., contact type classification, contact localization).

Long-term paradigm shifts require moving beyond workspace-specific feature learning to geometric abstraction. Physics-informed neural networks that explicitly model edge topology, partial contact mechanics, and surface boundary effects could enable workspace-agnostic edge detection. For object generalization, our finding that **heavily-regularized GPU-MLP achieves 75.0% validation on novel objects, reproduced deterministically across 5 independent seeds (std=0.0%)**, opens critical research directions: (1) systematic study of regularization strategies (dropout rates, weight decay, early stopping, batch normalization) for acoustic contact detection to understand what regularization strength optimally balances within-object performance versus cross-object generalization, (2) investigation of domain-adversarial training to explicitly learn object-invariant features by forcing the model to be unable to distinguish which object produced an acoustic signature, (3) transfer learning from

diverse object geometries to build general contact representations that capture contact-type patterns independent of object identity, and (4) meta-learning approaches for rapid object adaptation with few examples, enabling quick deployment on novel objects with minimal retraining. The 39.3 percentage point gap between regularized (75.0%) and unregularized (35.7%) GPU-MLP models demonstrates that preventing object-specific overfitting is essential—future work should focus on architectural innovations (residual connections, attention mechanisms) and training strategies (mixup augmentation, contrastive learning) that inherently promote geometry-invariant learning while maintaining high within-workspace performance.

Multi-modal fusion combining acoustic sensing with vision and force feedback could leverage complementary strengths: vision for workspace geometry understanding, force for contact mechanics, and acoustics for non-contact prediction and edge detection. Such integrated systems could achieve robust performance across varying workspaces by dynamically adapting to geometric changes through sensor fusion.

This work establishes acoustic sensing as a proof-of-concept modality for robotic manipulation in highly constrained closed-world environments while revealing catastrophic cross-workspace generalization failure. The dramatic performance degradation from 69.9% cross-validation to 34.5% validation (barely above random) demonstrates that workspace-specific training is not just beneficial but **mandatory** for deployment. The strong workspace dependence revealed by our 3-class analysis (23.3–55.7% validation range, two rotations worse than random) provides critical insights for designing sensing systems: acoustic sensing alone is insufficient for generalizable manipulation and requires either workspace-specific retraining or multimodal sensor fusion to achieve robust performance.

D. Limitations

This study has several limitations that constrain generalizability of findings. **(1) Limited scale:** We evaluate on only 4 workspaces and 4 object types with a single robot platform (Franka Panda). Larger-scale studies across diverse robot morphologies (e.g., Universal Robots, KUKA) and object materials (metals, plastics, compliant materials) could reveal whether acoustic sensing limitations are fundamental or platform-specific. **(2) Passive sensing only:** Unlike prior work on soft actuators using active acoustic excitation (chirp signals) [3], we employ purely passive contact acoustics. Active excitation could potentially improve signal-to-noise ratio and enable better generalization, though at the cost of increased system complexity. **(3) Material homogeneity:** All test objects are wooden boards, limiting material diversity. Different materials (e.g., metals with higher elastic moduli, polymers with viscoelastic damping) produce fundamentally different acoustic signatures and may exhibit different generalization properties. **(4) Single contact geometry:** The acoustic finger has fixed 1 cm $\times$ 0.25 cm contact area. Variable contact geometries or force profiles could interact with acoustic signatures in ways not captured by our experiments. **(5) Static contact:** We evaluate static contact detection, not dynamic manipulation tasks like sliding, rolling, or impact. Dynamic contact mechanics may produce richer acoustic signatures with different generalization characteristics. Future work should address these limitations through multi-platform studies, active excitation experiments, diverse material testing, and dynamic manipulation scenarios to fully characterize acoustic sensing capabilities and boundaries for robotic applications.

REFERENCES

[1] V. Wall, "Morphological sensing for soft pneumatic actuators based on acoustics and strain," Ph.D. dissertation, Technische Universität Berlin, 2019. [Online]. Available: https://depositonce.tu-berlin.de/bitstream/11303/10158/1/Wall_2019_Morphological_Sensing.pdf

[2] V. Wall, G. Zöller, and O. Brock, "Passive and active acoustic sensing for soft pneumatic actuators," *The International Journal of Robotics Research*, vol. 41, no. 3, pp. 260–277, 2022. [Online]. Available: https://arxiv.org/pdf/2208.10299.pdf

[3] G. Zöller, V. Wall, and O. Brock, "Active acoustic contact sensing for soft pneumatic actuators," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2438–2445, 2020. [Online]. Available: https://www.static.tu.berlin/fileadmin/www/10002220/Publications/Zoeller-20-ICRA_activeacoustic.pdf

[4] K. Zhang, D.-G. Kim, E. T. Tang, H.-H. Liang, Z. He *et al.*, "VibeCheck: Using active acoustic tactile sensing for contact-rich manipulation," *arXiv preprint arXiv:2504.15535*, 2025. [Online]. Available: https://arxiv.org/pdf/2504.15535.pdf

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

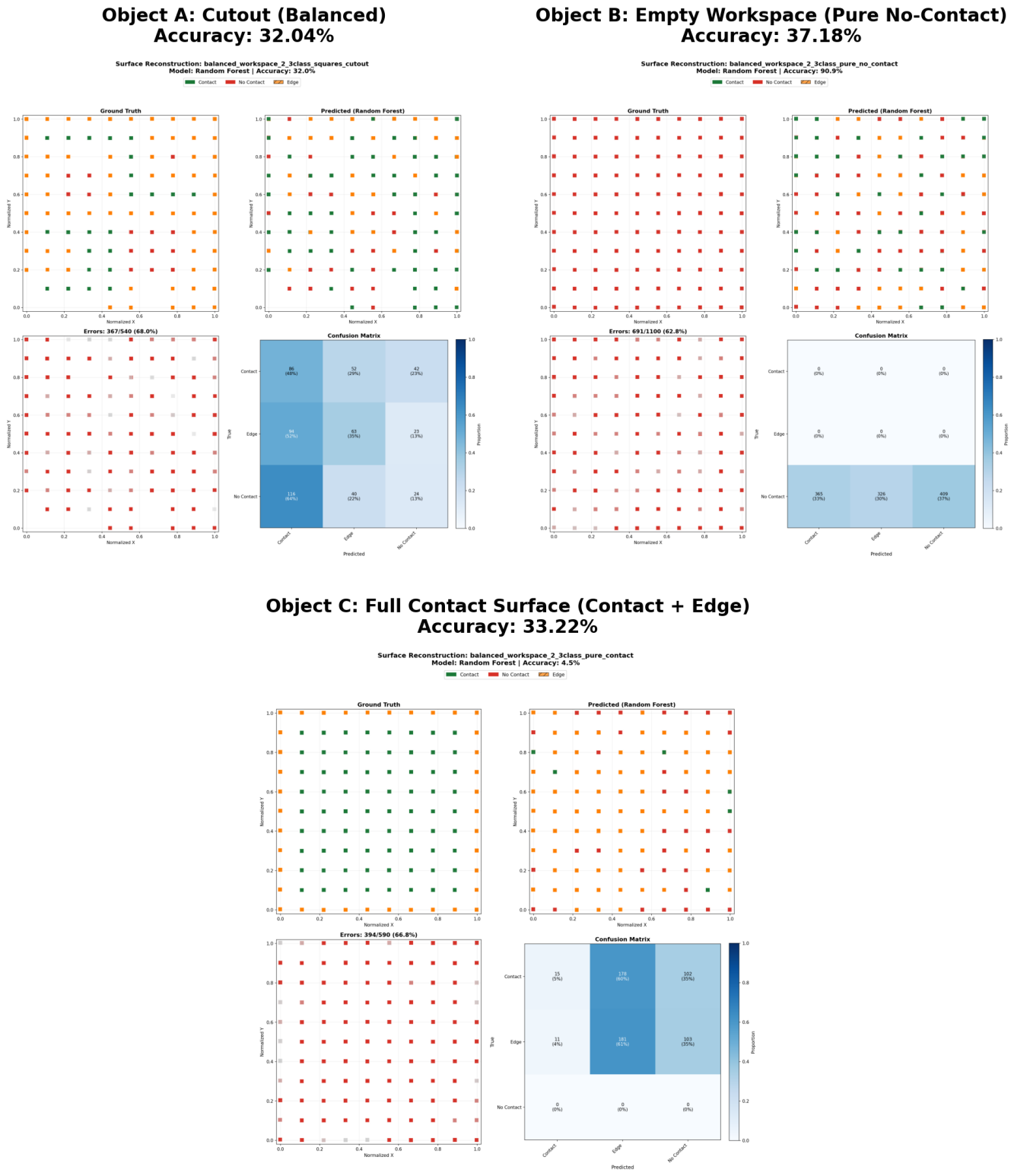


Fig. 6: Position generalization: 3-class reconstruction on Workspace 2 validation data (held out in Rotation 1). **Layout**: Two objects in the top row (Objects A and B), with Object C centered in the bottom row. Each panel shows ground truth (left), predictions (middle), error map (top right), and 3×3 confusion matrix (bottom right). Object A (cutout, 32.04% accuracy): balanced across all 3 classes (540 samples). Object B (empty workspace, 37.18% accuracy): pure no-contact scenario (1,100 samples). Object C (full contact surface, 33.22% accuracy): contact and edge classes only (590 samples). Trained on WS1+WS3, the model achieves 34.89% weighted average accuracy across 2,230 samples—close to random chance (33.3%) and demonstrating catastrophic performance degradation from the 69.1% cross-validation accuracy, revealing severe workspace-dependent acoustic signatures. All confusion matrices display 3×3 grids showing the complete 3-class problem (contact, edge, no-contact). Color indicates 3-class contact state (contact=green, no-contact=red, edge=orange).

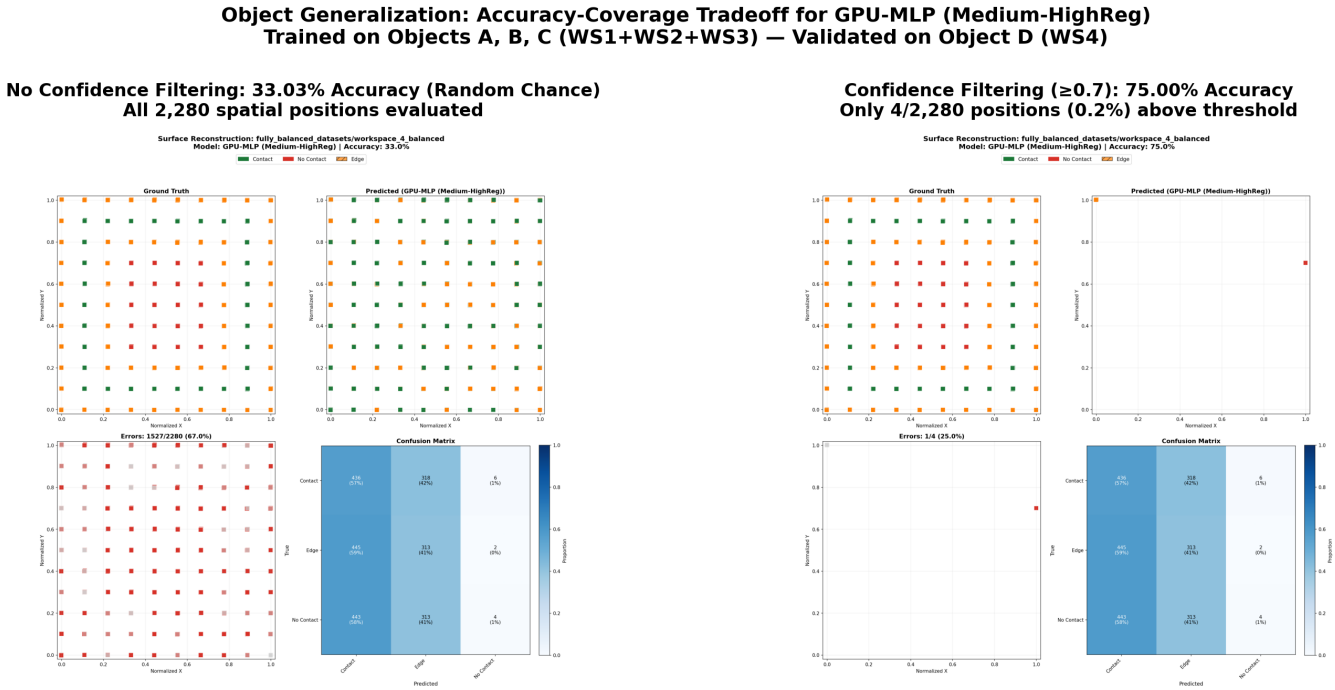


Fig. 7: Object generalization accuracy-coverage tradeoff on novel Object D (Workspace 4 holdout): GPU-MLP (Medium-HighReg) with and without confidence filtering. **Left panel**: Without confidence filtering, spatial reconstruction achieves 33.03% accuracy (random chance for 3-class problem) across all 2,280 spatial positions, demonstrating that the model cannot perform reliable full-surface reconstruction on novel object geometries. **Right panel**: With confidence filtering (threshold $\geq$0.7), accuracy increases to 75.00%, but only 4 out of 2,280 positions (0.2% coverage) exceed the confidence threshold. Each panel displays ground truth (left), predictions (middle), error map (top right), and 3×3 confusion matrix (bottom right) for Object D. This reveals a fundamental accuracy-coverage tradeoff: the model can achieve high accuracy (75%) on a tiny confident subset (0.2% of positions) OR complete spatial coverage (100%) with random-chance performance (33.03%), but cannot achieve both simultaneously. The 75% validation accuracy reported in Table V reflects performance only on this extremely filtered subset, not full spatial reconstruction capability. Heavy regularization (dropout 0.3, weight decay 0.01) enables the model to identify when it can make confident predictions, but fundamentally cannot generalize across the full spatial surface of novel object geometries. All confusion matrices display 3×3 grids showing the complete 3-class problem (contact, edge, no-contact). Color indicates 3-class contact state (contact=green, no-contact=red, edge=orange).