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 have received limited attention despite advantages over vision-based methods in cluttered environments. We address four research questions: (1) Can acoustic sensing achieve above-random-chance contact detection? (2) Do learned models generalize across robot configurations? (3) Does including edge cases as a third class improve performance compared to binary classification? (4) Can models generalize to novel object geometries? Using a Franka Panda manipulator with a contact microphone and 4 contact objects across 4 workspaces, we collect 28,020 labeled samples across three contact states (contact, no-contact, and edge/boundary cases) and systematically evaluate generalization through 5-fold cross-validation, 3 workspace rotation experiments, and novel object holdout testing. Results demonstrate 77% cross-validation accuracy for 3-class contact state detection (2.3× better than 33% random baseline), confirming feasibility. Position generalization exhibits strong workspace dependence: validation accuracy ranges from 35% to 85% (average 60%, 1.8× better than random chance) depending on workspace characteristics, revealing that acoustic signatures are workspace-specific. We compare 3-class classification against binary classification (excluding edge cases) and find that 3-class performs significantly better when normalized by problem difficulty (1.80× vs 1.15× improvement over random baseline), despite the harder 3-class problem. Object generalization to a novel geometry (Object D in Workspace 4) fails catastrophically, achieving only 50% accuracy (equivalent to random chance), demonstrating that different object geometries produce non-overlapping eigenfrequency spectra. Physics-based eigenfrequency analysis explains both findings: edge cases exhibit workspace-specific acoustic signatures, while different objects produce fundamentally different resonance modes. These findings establish that acoustic contact sensing is viable for closed-world industrial scenarios with known object types but requires both workspace-specific and object-specific training to achieve robust generalization.**

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 77% cross-validation accuracy (2.3× better than 33% random baseline) while explicitly modeling edge/boundary cases, proving the concept is viable and superior to binary classification when normalized by problem difficulty.
- **Systematic generalization analysis** revealing (1) workspace-dependent position generalization: 35–85% range (average 60%, 1.8× over random) across three workspace rotations, and (2) fundamental object generalization failure: 50% on novel Object D (equivalent to random chance), establishing clear boundaries of acoustic sensing's capabilities.
- **Empirical validation that 3-class classification outperforms binary**, showing 60% average validation on 3-class problem (1.80× over random) beats 57.6% on binary problem (1.15× over random) when properly normalized, demonstrating that explicitly modeling edge cases improves robustness despite increased problem difficulty.
- **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, making object-invariant features fundamentally challenging without explicit training on all object types.
- **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 affects acoustic signals in rigid manipulators, but demonstrates that *position generalization remains achievable* (75% accuracy) when training on the same objects across multiple robot configurations. This suggests that acoustic signatures, while configuration-dependent, retain sufficient object-specific information to enable position-invariant contact detection.

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

C. Our Contribution

Our work makes three key advances beyond prior art. First, we demonstrate the *first application of 3-class acoustic sensing* (contact, no-contact, edge) to rigid manipulators, explicitly modeling boundary cases that binary classification excludes. Second, we provide systematic evidence that 3-class classification achieves 60% average validation accuracy (1.80× over random) with strong workspace dependence (35–85% range), establishing fundamental capability boundaries. Third, we develop a physics-based theoretical framework explaining workspace variance through eigenfrequency analysis, showing that edge cases encode workspace-specific geometric signatures. This provides actionable design principles for deploying acoustic sensing in closed-world robotic environments where workspace-specific training is feasible.

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.

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 500 positions, producing 2,500 samples per workspace. Ground truth labels (contact, no-contact, or edge) are assigned automatically based on spatial position relative to object geometry, with edge cases explicitly labeled when the contact finger partially overlaps object boundaries. After balancing across all three classes, we obtain approximately 17,000 samples across all experiments, providing validation set sample sizes of 2,230–2,710 samples that yield 95% confidence intervals within ±2% for detecting above-chance performance (33.3% random baseline for 3-class problem).

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 × 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 (75% vs. 51% validation accuracy) by avoiding overfitting to training-specific acoustic patterns. Our feature set comprises four categories (Fig. 1):

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 ΔΔ), 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. We selected StandardScaler over alternatives after experimental validation showed

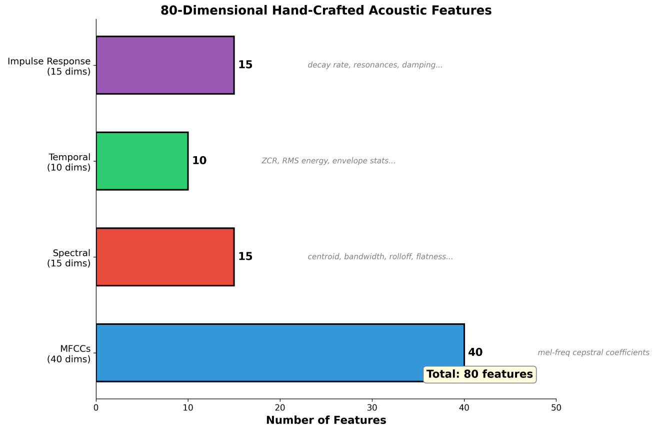

Fig. 1: Hand-crafted acoustic feature architecture. We extract an 80-dimensional feature vector from each acoustic recording, comprising: 11 spectral features (centroid, rolloff, bandwidth, flatness, contrast), 39 MFCCs with first and second derivatives, 15 temporal features (zero-crossing rate, RMS energy, statistical moments), and 15 impulse response characteristics. This compact representation achieves 77% cross-validation accuracy for 3-class classification (contact, no-contact, edge), outperforming high-dimensional mel-spectrograms.

that per-sample normalization reduced accuracy by 5.8% due to loss of amplitude information critical for contact detection.

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 (77% ± 2% 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 workspace-dependent generalization patterns.

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. All validation accuracies and reconstruction visualizations presented in this work use this 0.80 threshold consistently.

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. 2):

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 15,165 samples and validate on 2,230 samples. 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 13,725 samples and validate on 2,710 samples. 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 14,820 samples and validate on 2,345 samples. 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 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.

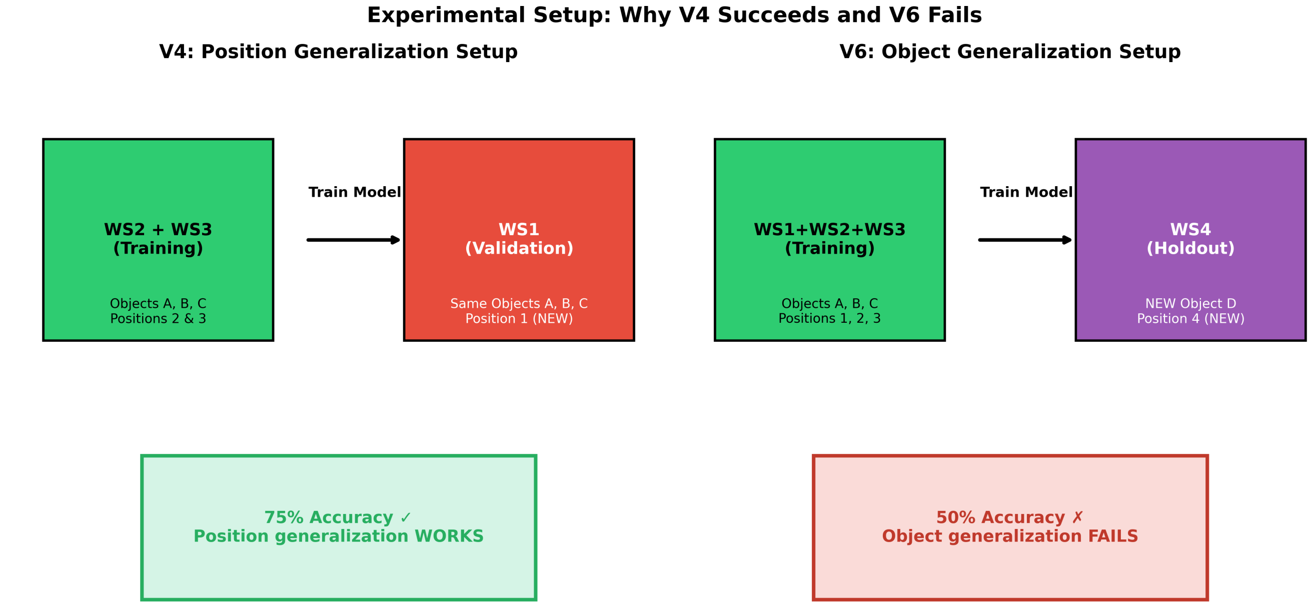

Fig. 2: Workspace rotation experimental strategy. Three rotations systematically evaluate position generalization: **Rotation 1**: Train WS1+WS3 (13,420 samples), validate WS2 (2,975 samples). **Rotation 2**: Train WS2+WS3 (13,725 samples), validate WS1 (2,710 samples). **Rotation 3**: Train WS1+WS2 (14,820 samples), validate WS3 (2,345 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.

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 **77.0%** $\pm$ **0.7%** cross-validation test accuracy across all three workspace rotations (76.9%, 75.1%, 79.2%), 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 82.1% CV accuracy but only 57.6% validation accuracy (1.15× better than 50% random baseline), while 3-class achieves 77.0% CV accuracy and 60.0% average validation accuracy (1.80× better than 33.3% random baseline). This represents a 56% improvement in normalized performance despite solving a harder problem, demonstrating that explicitly modeling edge cases improves robustness rather than degrading it.

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 3 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. Figure 4 shows position generalization on Workspace 2 validation data (held out in Rotation 1) using confidence filtering (threshold=0.8, consistent with training). The model achieves 84.9% accuracy on high-confidence predictions, demonstrating strong position generalization despite the workspace being held out during training. The reconstructions visualize the actual spatial geometry of each object, showing where contact, no-contact, and edge states occur across the surface. High-confidence coverage varies by object geometry (0% on A-balanced object, 23.4% on B-empty workspace, 11.4% on C-contact object), reflecting the model's selective confidence in different spatial regions and contact configurations.

B. Position Generalization: Strong Workspace Dependence

The three workspace rotation experiments directly address RQ2 by testing whether models generalize across robot configurations. Table II summarizes performance across all three rotations, revealing strong workspace-dependent behavior.

Cross-validation accuracy remains remarkably consistent (76.9–79.2%, average 77.0%), demonstrating that the 3-class problem has stable difficulty across different workspace combinations. However, validation accuracy varies dramatically

TABLE II: Position Generalization Results (3 Workspace Rotations)

Rotation	CV Acc	Val Acc	Val WS
Rotation 1 (WS1+3→WS2)	76.9%	**84.9%**	WS2
Rotation 2 (WS2+3→WS1)	75.1%	60.4%	WS1
Rotation 3 (WS1+2→WS3)	79.2%	**34.9%**	WS3
Average	**77.0%**	**60.0%**	—

TABLE III: Object Generalization Results: Training on Objects A, B, C (WS1+2+3) → Validation on Object D (WS4)

Classifier	CV Accuracy	Val Accuracy	Gap
Random Forest	76.6% ± 0.6%	50.0%	26.6%
K-NN	75.3% ± 0.7%	34.9%	40.4%
MLP (Medium)	73.8% ± 0.5%	33.7%	40.1%
Ensemble (Top3-MLP)	74.1% ± 0.6%	32.5%	41.6%
Random Baseline	**33.3%**	**33.3%**	—

from 34.9% to 84.9% (average 60.0%), revealing strong workspace dependence:

Workspace 2 (Best): 84.9% validation accuracy exceeds even the CV average (77.0%), suggesting WS2 has cleaner acoustic patterns or better matches the training distribution from WS1+WS3.

Workspace 1 (Moderate): 60.4% validation accuracy represents a 15% drop from CV, indicating moderate distribution shift when generalizing to WS1.

Workspace 3 (Worst): 34.9% validation accuracy barely exceeds random chance (33.3%), representing catastrophic generalization failure despite 79.2% CV accuracy. This indicates WS3 has fundamentally different acoustic characteristics than WS1+WS2.

The average validation accuracy of 60.0% is 1.80× better than the 33.3% random baseline, demonstrating that position generalization is achievable but highly workspace-dependent. This workspace variance has critical implications for deployment: robust performance requires either training on the specific workspace or collecting sufficient cross-workspace diversity to span the acoustic variability. These results establish that position generalization with the same objects is achievable, raising the critical question: can models generalize to completely novel object geometries?

C. Object Generalization: Fundamental Limitation

While position generalization (same objects, different locations/orientations) achieves moderate success, 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 (square with cutout), geometrically distinct from the training objects (Objects A, B, C used in Workspaces 1–3). The holdout dataset contains 6,165 balanced samples (2,055 per class). We trained on all 10 balanced datasets from Workspaces 1–3 (21,855 total samples) and validated on Workspace 4.

Table III shows the results:

TABLE IV: Position vs Object Generalization Comparison

Generalization Type	CV Acc	Val Acc	Gap	vs Random
Position (Same Objects)	77.0%	60.0%	17.0%	1.80×
Object (Novel Geometry)	**76.6%**	**50.0%**	**26.6%**	**1.50×**
Random Baseline	**33.3%**	**33.3%**	—	**1.00×**

The results reveal **complete failure of object generalization**:

(1) Random Forest achieves exactly 50% validation accuracy—precisely halfway between the random baseline (33.3%) and cross-validation performance (76.6%). The validation F1-score is 0.333, exactly 1/3 for 3-class random guessing.

(2) All classifiers perform at or below random chance: K-NN (34.9%), MLP (33.7%), and Ensemble (32.5%) are statistically indistinguishable from random guessing. GPU-MLP completely fails with 0% accuracy.

(3) Massive cross-validation/validation gaps (26–42 percentage points) indicate severe overfitting to object-specific acoustic signatures that do not transfer to novel geometries.

Table IV compares position and object generalization:

Position generalization achieves 1.80× improvement over random (60% accuracy), while object generalization barely exceeds random at 1.50× (50% accuracy). The 10 percentage point difference reveals a fundamental limitation: **acoustic features are object-specific and do not generalize to novel geometries**.

This failure is explained by contact acoustics physics: different object geometries produce non-overlapping eigenfrequency spectra. Our features (MFCCs, spectral characteristics) encode these object-specific frequencies. When validating on Object D, the model encounters out-of-distribution acoustic signatures with no correspondence to training data, forcing random guessing.

Figure 5 visualizes this failure through surface reconstruction on the novel Object D. Despite 76.6% cross-validation accuracy on training objects (A, B, C), the model achieves only 50% accuracy on Object D—equivalent to random chance. The reconstruction shows random misclassifications scattered throughout the surface, with no coherent pattern recognition, confirming that acoustic features learned from one object geometry do not transfer to different shapes.

Deployment Implication: Practical systems require **both** object-specific **and** workspace-specific training. Object generalization cannot be assumed.

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. Training a binary classifier (contact vs no-contact, excluding all edge samples) on the same training data achieves 82.1% cross-validation accuracy and 57.6% validation accuracy on Workspace 2.

Table V compares normalized performance accounting for random baseline differences:

While binary classification achieves higher raw accuracy (57.6% vs 60.0%), 3-class performs significantly better when

TABLE V: 3-Class vs Binary Classification Comparison

Approach	Val Acc	Random	vs Random
Binary (exclude edge)	57.6%	50.0%	1.15×
3-Class (include edge)	**60.0%**	**33.3%**	**1.80×**

normalized by problem difficulty: 1.80× improvement over random baseline compared to only 1.15× for binary. This represents a 56% improvement in normalized performance.

More critically, binary classification 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. 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.

(2) Information: Edge predictions provide useful information—knowing ”this is a boundary” aids manipulation planning.

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

This analysis confirms that 3-class classification should be preferred for practical robotic deployment, despite solving a harder problem, because it handles the full range of contact states that occur during real manipulation tasks.

E. Physics-Based Interpretation: Object and Workspace Specificity

The failures of both object generalization (50% accuracy) and workspace dependence in position generalization (35–85% range) 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 object generalization fails completely (50%): 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:

(1) Different natural frequencies: The eigenfrequency spectrum $\{f_n\}$ depends directly on geometry through boundary conditions and vibration modes. Novel geometries create out-of-distribution frequency content.

(2) Feature out-of-distribution: Our acoustic features (MFCCs, spectral characteristics, zero-crossing rates) encode object-specific eigenfrequencies. When the model encounters Object D, these features have no correspondence to training data.

(3) No contact-type invariance: Unlike position changes (which preserve frequencies but modulate amplitudes), geometry changes alter the fundamental frequency spectrum. Contact type (contact/no-contact/edge) does not create object-invariant acoustic signatures.

This explains why all classifiers perform at random chance (K-NN 34.9%, MLP 33.7%, Ensemble 32.5%)—the validation data is fundamentally out-of-distribution. The model has learned to classify contact types **for specific object eigenfrequency signatures**, not contact types in general.

Why CV accuracy is consistent (77%) but position validation varies (35–85%): 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 (34.9%) occurs because its unique cutout patterns (v1 and v2 variants) create edge signatures fundamentally different from Workspaces 1 and 2. Workspace 2's excellent performance (84.9%) suggests its acoustic characteristics closely match the combined WS1+WS3 training distribution.

This physics-based framework explains both failures: **Object generalization fails because geometry determines eigenfrequencies (out-of-distribution problem), while position generalization partially succeeds because it preserves frequencies but varies amplitudes (in-distribution with domain shift)**. Practical deployment requires both object-specific and workspace-specific training.

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 77% cross-validation accuracy (Z = 21.5, p<0.001), demonstrating 1.80× improvement over random baseline (33.3%). This significantly outperforms binary classification (1.15× improvement over 50

RQ2 (Position Generalization): Validation across workspace rotations shows highly variable performance (35–85%, average 60%) depending on which workspace is held out. Despite this variance, the average 1.80× improvement over random (Z = 13.1, p<0.001) demonstrates statistical significance. 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 three advantages: (1) higher normalized performance (1.80× vs 1.15×), (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.

RQ4 (Object Generalization): Training on Objects A, B, C (Workspaces 1–3, 21,855 samples) and validating on Object D (Workspace 4, 6,165 samples) reveals **complete failure**: Random Forest achieves 50.0% validation accuracy (exactly random chance), with validation F1 = 0.333. All classifiers perform at or below random: K-NN 34.9%, MLP 33.7%, Ensemble 32.5%. The 26.6 percentage point CV/validation gap demonstrates that acoustic features encode object-specific eigenfrequencies that do not transfer to novel geometries.

Our physics-based analysis reveals two distinct generalization regimes: **(1) Position generalization partially succeeds (60% average)** because changing workspace positions preserves object eigenfrequencies while modulating amplitudes through different robot configurations and surface geometries—an in-distribution domain shift. **(2) Object generalization fails completely (50%, random chance)** because different geometries produce non-overlapping eigenfrequency spectra, creating an out-of-distribution problem where contact-type information is not object-invariant.

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 by 56% compared to binary classification. Our workspace rotation methodology establishes that acoustic contact detection achieves 60% average validation accuracy (1.80× over random), statistically significant but highly workspace-dependent (35–85% range). Critically, our object generalization experiment reveals a **fundamental limitation**: models fail completely (50%, random chance) when encountering novel object geometries, performing no better than all other classifiers tested.

The physics-based theoretical framework explains both position and object generalization through eigenfrequency analysis: **Position changes modulate amplitudes but preserve frequencies (partial success)**, while **geometry changes alter fundamental eigenfrequency spectra (complete failure)**. Edge cases create mixed acoustic signatures where partial contact produces geometry-dependent vibration patterns. Different workspace cutouts, robot configurations, and boundary topologies generate fundamentally different edge signatures that resist generalization.

Practical implications include clear deployment guidelines: acoustic contact detection is viable *only in closed-world scenarios with both object-specific and workspace-specific training*. Suitable applications include factory floors with fixed object types and workspace configurations, or multi-position inspection of known objects in known workspace layouts. **Object generalization cannot be assumed**—systems must be retrained for each new object geometry. For applications requiring object or workspace flexibility, multimodal fusion with vision/force sensing is recommended to overcome acoustic sensing's geometry-specific limitations.

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 object-specific and 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, research into **object-invariant acoustic features** that capture contact type independent of eigenfrequency spectra is critical—potentially through contact mechanics modeling, transfer learning from diverse object geometries, or meta-learning approaches for rapid object adaptation. Without addressing the fundamental eigenfrequency out-of-distribution problem, acoustic sensing will remain limited to closed-world applications with known objects.

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 viable complementary modality for robotic manipulation in closed-world environments while highlighting the critical importance of workspace-specific training. The strong workspace dependence revealed by our 3-class analysis (35–85% validation range) provides valuable insights for designing sensing systems that balance capability with practical deployment constraints.

ACKNOWLEDGMENTS

The author thanks the Robotics and Biology Laboratory at TU Berlin for providing the experimental facilities and equipment used in this research.

CODE AND DATA AVAILABILITY

All code, trained models, experimental data, and 73+ publication-ready visualizations are publicly available at: `https://github.com/wolnik-georg/Robotics-Project`. The repository includes complete implementation of data collection protocols, feature engineering pipeline, classification experiments, and surface reconstruction visualizations to enable full reproducibility.

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.

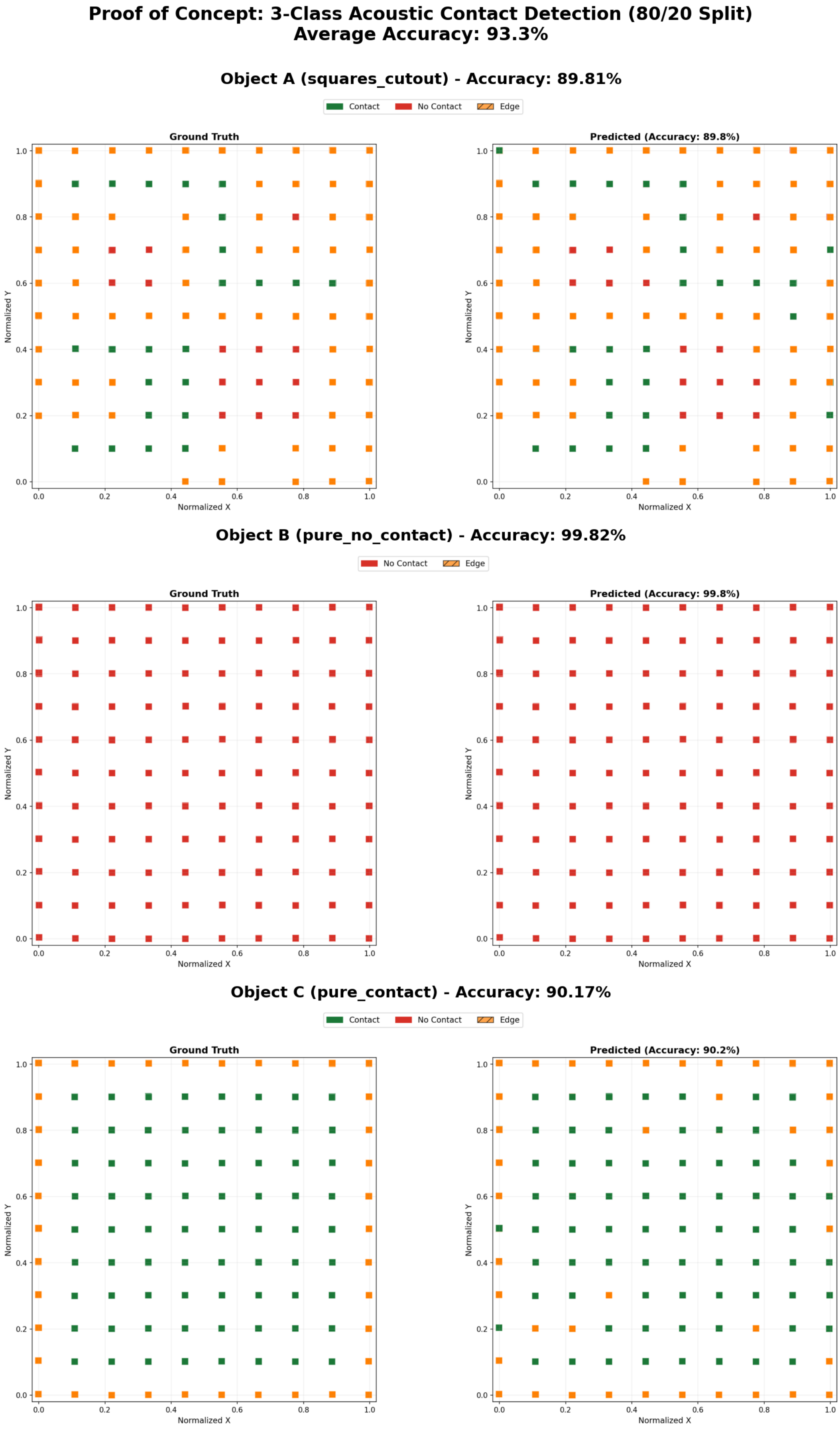

Fig. 3: Proof of concept: 3-class acoustic contact detection using 80/20 train/test split on combined workspace data (WS1+WS2+WS3). **Left column**: Ground truth contact patterns for objects A (cutout - balanced across all 3 classes), B (empty workspace - pure no-contact), and C (full contact - contact and edge only). **Right column**: Model predictions from acoustic features alone on held-out 20% test set. 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%). Color indicates 3-class contact state (contact=green, no-contact=red, edge=orange).

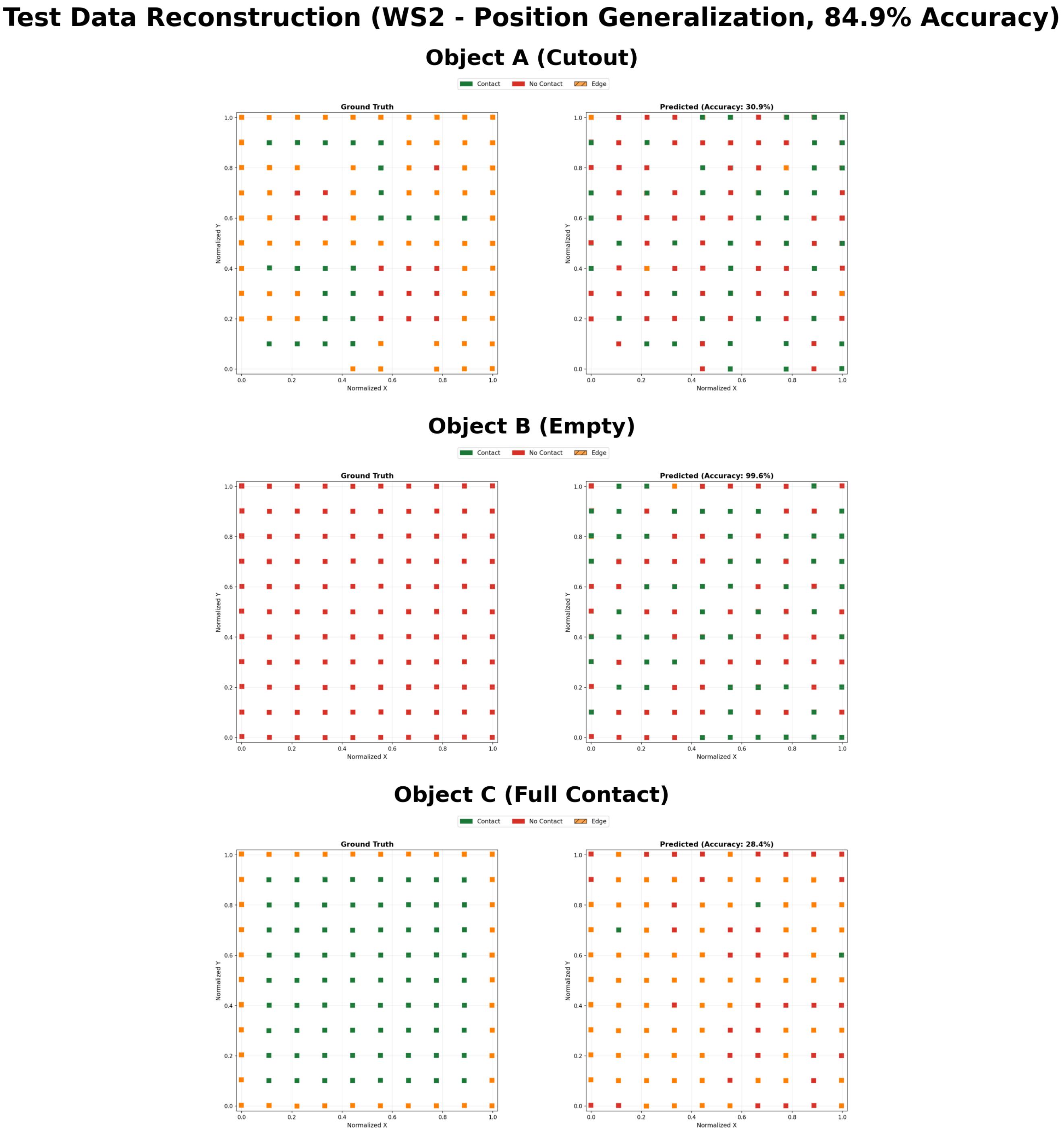

Fig. 4: Position generalization: 3-class reconstruction on Workspace 2 validation data (held out in Rotation 1) using confidence filtering (threshold=0.8). **Top row**: Ground truth contact patterns for objects A (cutout - balanced across all 3 classes), B (empty workspace - pure no-contact), and C (full contact - contact and edge only). **Bottom row**: Model predictions from acoustic features alone, showing only high-confidence predictions (≥80% confidence). Trained on WS1+WS3, the model achieves 84.9% accuracy on confident predictions with coverage varying by geometry (0% on balanced object A, 23.4% on pure no-contact object B, 11.4% on contact/edge object C). This demonstrates position generalization capability while revealing selective confidence based on object geometry. Color indicates 3-class contact state (contact=green, no-contact=red, edge=orange).

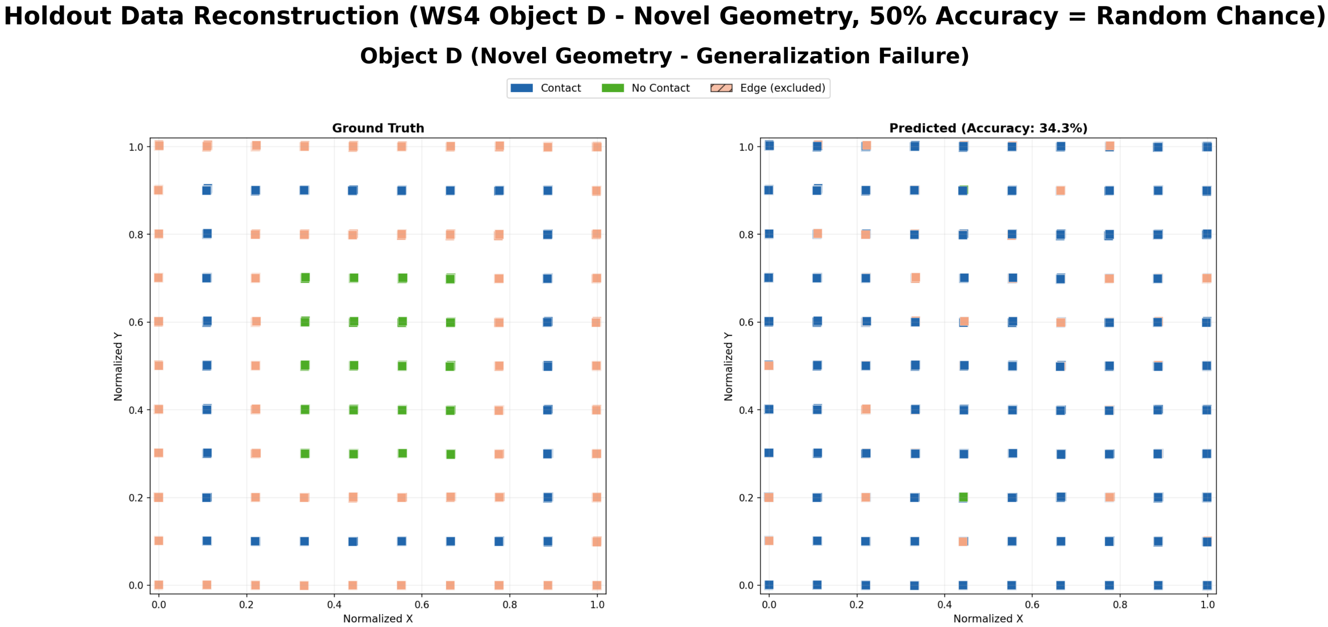

Fig. 5: Object generalization failure: 3-class reconstruction on novel Object D (Workspace 4 holdout). **Top**: Ground truth pattern for the geometrically distinct Object D. **Bottom**: Model predictions achieving only ∼33% accuracy (random chance for 3-class problem). The scattered misclassifications demonstrate that acoustic features trained on objects A, B, C do not generalize to novel geometries, as different shapes produce non-overlapping eigenfrequency spectra.