ACOUSTIGYM: PHYSICS-BASED ACOUSTIC ENVIRONMENT GENERATION FOR ROBOTIC POLICY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken instructions reach a robot through a room and a speech-recognition pipeline, either of which can change the goal selected from the instruction. We present AcoustiGym, which exports generated indoor scenes for acoustic simulation, propagates synthesized speech to a microphone array, and evaluates the top-ranked goal selected from the resulting transcript. We compare written instructions, clean synthesized speech, and room-propagated speech to separate the written-to-spoken grounding gap from the additional effect of room acoustics. In one generated bedroom, a predicate resolver grounds 0.495 of written instructions correctly and 0.305 after clean speech synthesis and recognition across 50 goals with four phrasings. Across four furnished-mesh room variants, the additional room-only regret ranges from 0.022 to 0.068; the hard-floor variant has a 95% confidence interval of $[0.003, 0.130]$. In a separate shoebox-room study, an expert-guided room configuration increases room-only regret to 0.500, and a one-surface replacement reduces it to zero on 24 evaluation items.

1 INTRODUCTION

Language-conditioned robot policies ground natural-language goals in complex scenes (Black et al., 2025; Kim et al., 2024; TRI LBM Team et al., 2025). A spoken instruction reaches a robot through a physical room and a speech-recognition pipeline; each stage can affect the object and destination selected from the instruction. Grounding accuracy across written, clean-spoken, and room-propagated instructions provides a direct way to measure these effects.

Architectural surfaces and furnishings shape reflections, absorption, and occlusion, while source and microphone placement determine which paths reach the receiver. Acoustic simulators provide furnished 3D environments for audio-visual learning (Chen et al., 2020; 2022) and efficient shoebox-room models (Scheibler et al., 2018); scene-generation systems produce furnished environments (Pfaff et al., 2026; Yang et al., 2024). AcoustiGym connects controllable room configurations to speech transcription and goal grounding, allowing material changes to be evaluated by their effect on the selected goal.

We introduce ACOUSTIGYM, a framework that exports a generated furnished scene as an acoustic environment, propagates synthesized instructions to a microphone array, applies far-field speech processing, and evaluates the goal selected by a language-conditioned predicate resolver. The mesh-aware simulator represents interior geometry and material-dependent propagation; a separate shoebox engine is used for environment search. We measure *total Acoustic Regret* as the paired loss in top-ranked goal grounding relative to the written instruction and R_{room} as the incremental difference between clean and room-propagated speech on the same goals. The scoring endpoint is the selected predicate and object binding.

Our evaluation follows the channel from speech to grounding. Controlled speech tests measure recognition and parsed-goal accuracy across reverberation conditions and far-field processing methods. In one generated bedroom, we evaluate 50 goals with four phrasings under furnished-mesh acoustics. Written instructions yield 0.495 top-ranked grounding accuracy versus 0.305 for clean spoken instructions; the additional room-only regret across four material variants is 0.022–0.068, with a hard-floor 95% confidence interval of $[0.003, 0.130]$. In a shoebox-engine search on 24

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

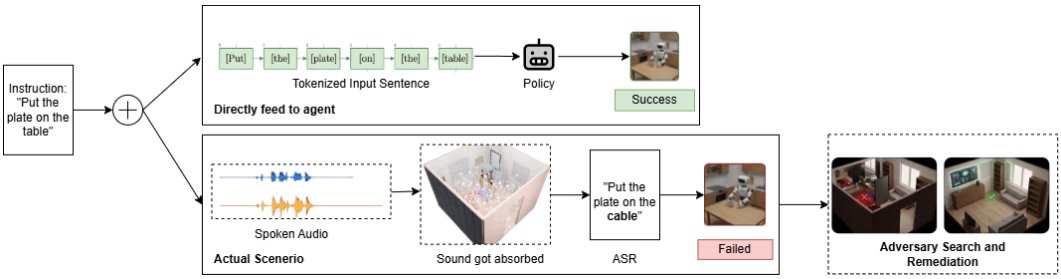

Figure 1: Overview of ACOUSTIGYM. Written instructions provide a room-agnostic reference path, while spoken instructions pass through the acoustic environment and speech-recognition pipeline before policy grounding. Room configurations can then be searched for grounding failures and modified through acoustic remediation.

items, an expert-guided room configuration yields 0.500 room-only regret; changing its ceiling material reduces this to 0.000.

Figure 1 summarizes the written and spoken evaluation paths and the subsequent search for adverse room configurations and acoustic remediation.

Our contributions are:

- We introduce ACOUSTIGYM, an evaluation pipeline that connects generated furnished scenes, geometry- and material-aware sound propagation and language-conditioned goal grounding.
- We measure written-to-spoken grounding loss with *Acoustic Regret* (R) and compare clean and room-propagated speech to estimate the additional room-only loss (R_{room}).
- We evaluate four black-box search methods and an expert-guided method over room materials and source–receiver placement, followed by single-surface remediation at fixed placement in a separate shoebox study.

2 RELATED WORK

Indoor Scene Synthesis for Embodied AI. ProcTHOR (Deitke et al., 2022) generates indoor layouts procedurally, while Holodeck (Yang et al., 2024) combines language models with constraint solvers and SceneSmith (Pfaff et al., 2026) uses multi-agent generation followed by physical settlement; SceneWeaver (Yang et al., 2025) also uses multiple agents for scene synthesis. Their scene geometry and object placements provide an input for acoustic evaluation. AcoustiGym uses a SceneSmith-generated bedroom as input to a material- and geometry-aware acoustic simulator, then tests how the resulting audio changes the goal selected by a predicate resolver.

Acoustic Simulation and Audio-Visual Embodied AI. SoundSpaces (Chen et al., 2020; 2022) provides room impulse responses (RIRs) for audio-visual navigation in scanned environments. PyRoomAcoustics (Scheibler et al., 2018) provides geometric room-acoustics tools, including image-source simulation (Allen & Berkley, 1979), that are useful for controlled room studies. Scanned environments represent furnished geometry, while shoebox models support controlled surface-material changes. Our mesh-aware evaluation uses the generated bedroom's furniture geometry, and the surface search uses a shoebox simulation. Room-conditioned speech is processed with far-field methods (Nakatani et al., 2010; Drude et al., 2018) and recognized (Radford et al., 2023) before measuring goal grounding.

Environment Search and Adversarial Robustness in Robotics. Domain randomization (Tobin et al., 2017) varies simulation parameters during robot training, while POET (Wang et al., 2019) and PAIRED (Dennis et al., 2020) search for challenging environments. Far-field speech processing addresses the incoming audio signal through methods such as multi-channel dereverberation (Nakatani et al., 2010). Our study varies room-surface materials in a shoebox simulator and scores the resulting

change in goal grounding relative to clean speech. Adversarial search varies surface materials and source–receiver placement to maximize this score; remediation varies one surface material at fixed placement to minimize it. The search is scored on 24 goal-grounding items with ten evaluations per method.

3 AcoustiGym Framework

AcoustiGym exports a generated furnished scene for acoustic simulation and compares goal grounding from written instructions, clean synthesized speech, and speech propagated through the simulated room. The measured endpoint is the predicate resolver's selected goal.

3.1 Problem Formulation and Acoustic Regret

Let x be a written instruction specifying a target goal g in a fixed generated scene S. Dry synthesized speech $u_{\text{clean}}(x)$ is propagated from source position $\mathbf{x}_{\text{src}} \in \mathbb{R}^3$ to the receiver array at $\mathbf{x}_{\text{mic}} \in \mathbb{R}^3$, yielding room-conditioned audio $u(S, x)$. Let $T(u)$ be the recognized transcript of audio u, and let $q(z)$ be the predicate resolver's top-ranked goal for text z. We score $A(g, z) = \mathbb{I}[q(z) = g]$, where equality requires the intended predicate and target–reference object binding.

We define total *Acoustic Regret* (R) as the paired difference in grounding accuracy between written instructions and room-conditioned speech:

$$R(S) = \mathbb{E}\left[A(g, x) - A(g, T(u(S, x)))\right]. \tag{1}$$

The expectation is over the evaluated goals, phrasings, and channel realizations; the written and spoken conditions use the same underlying goals and scene.

We separate the written-to-clean-speech gap from the additional effect of the simulated room:

$$R_{\text{clean}} = \mathbb{E}\left[A(g, x) - A(g, T(u_{\text{clean}}(x)))\right], \quad R_{\text{room}}(S) = \mathbb{E}\left[A(g, T(u_{\text{clean}}(x))) - A(g, T(u(S, x)))\right]. \tag{2}$$

Here $u_{\text{clean}}(x)$ is dry synthesized speech. For matched evaluation items, $R(S) = R_{\text{clean}} + R_{\text{room}}(S)$. The two terms quantify the clean-speech gap and the incremental room effect, respectively.

3.2 Generative 3D Acoustic Scenes

The acoustic scene contains architectural geometry G and placed objects $O = \{(A_i, X_i)\}_{i=1}^{N}$, where asset A_i has 6-DoF pose $X_i \in SE(3)$. Thus $S = (G, O)$ includes surfaces that can obstruct or reflect sound between the source and receiver.

For acoustic simulation, surfaces $\mathcal{M}_k \subset (G \cup O)$ are represented by geometry and material parameters in addition to the scene information used for visual and collision modeling (Deitke et al., 2022; Yang et al., 2024):

1. Explicit 3D triangular surface geometry with outward-pointing surface normals.
2. Octave-band acoustic absorption coefficients $\boldsymbol{\alpha}_k = [\alpha_k(f_1), \ldots, \alpha_k(f_B)]^\top \in [0, 1]^B$ mapped from physically based rendering (PBR) material identifiers, governing frequency-dependent energy dissipation across standard acoustic bands from 125 Hz to 4 kHz.
3. Acoustic scattering coefficients $s_k \in [0, 1]$, defining the fraction of reflected energy scattered diffusely versus reflected specularly.

The evaluated bedroom was generated with a hierarchical multi-agent scene-synthesis pipeline adapted from SceneSmith (Pfaff et al., 2026). Construction progresses across successive stages: architectural layout planning, floor-standing furniture placement, wall and ceiling fixtures, and small manipulable tabletop objects. Object translations are adjusted to reduce collisions using constrained optimization in Drake (Tedrake & Drake Development Team, 2019):

$$\min_{\mathbf{p}} \|\mathbf{p} - \mathbf{p}_0\|_2^2 \quad \text{s.t.} \quad d(i, j) \geq \epsilon, \quad \forall i \neq j \tag{3}$$

where $d(i, j)$ is the signed distance between collision bodies i and j, and $\epsilon > 0$ is a separation threshold, followed by dynamic settlement under gravity and contact friction. The resulting scene is exported with object meshes, spatial poses, and per-surface absorption parameters for acoustic simulation.



Figure 2: Progression of stochastic acoustic-ray propagation through the generated furnished room. From left to right, successive snapshots show rays spreading from the source and accumulating reflections throughout the room geometry.

3.3 Mesh-Aware Geometric Acoustic Simulation

For the furnished-scene evaluation, exported triangle meshes provide geometry for occlusion and reflection checks, with material-dependent absorption and scattering. The simulator combines the following propagation components:

Direct Sound and Geometric Occlusion. Let $\mathbf{r}(t) = \mathbf{x}_{\text{src}} + t(\mathbf{x}_{\text{mic}} - \mathbf{x}_{\text{src}})$ denote the direct ray segment between the speaker $\mathbf{x}_{\text{src}}$ and a receiver microphone $\mathbf{x}_{\text{mic}}$. AcoustiGym performs ray-triangle intersection tests across all interior object meshes O. If the direct path is unobstructed, the sound arrives with spherical geometric spreading and time delay $\tau_0 = \|\mathbf{x}_{\text{mic}} - \mathbf{x}_{\text{src}}\|_2/c$, where $c \approx 343\,\text{m/s}$ is the speed of sound. If an object mesh occludes the line of sight, the direct ray is blocked. For an occluded path, AcoustiGym approximates diffraction around silhouette edges using a frequency-dependent loss derived from the Fresnel-Kirchhoff diffraction parameter:

$$v = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} \tag{4}$$

where h is the obstacle clearance height, d_1 and d_2 are distances from the diffracting edge to the source and receiver, and $\lambda = c/f$ is the acoustic wavelength. The diffraction approximation assigns a frequency-dependent attenuated arrival to an occluded source–receiver path.

Mesh-Checked Specular Reflections. Low-order early specular reflections off the room boundaries are computed via the image-source method (Allen & Berkley, 1979). For each candidate reflection from an image source $\mathbf{x}_{\text{img}}$ bouncing off a boundary point $\mathbf{x}_{\text{wall}}$ to $\mathbf{x}_{\text{mic}}$, AcoustiGym validates both path segments $(\mathbf{x}_{\text{src}} \to \mathbf{x}_{\text{wall}})$ and $(\mathbf{x}_{\text{wall}} \to \mathbf{x}_{\text{mic}})$ against all interior object meshes. Furniture occlusion prunes candidate specular paths, such as a wall reflection intercepted by a shelf.

Stochastic Late Reverberant Field. To model late diffuse reverberation, AcoustiGym launches stochastic energy-carrying acoustic rays through the full triangle mesh (architectural shell plus every object mesh). At each ray bounce against surface $\mathcal{M}_k$, ray energy in octave band b decays as $E_b^{(n)} = E_b^{(n-1)}(1-\alpha_{k,b})$, while the reflection direction is sampled as a mixture of specular reflection and Lambertian diffuse scattering weighted by s_k. Rays passing through a reception sphere around $\mathbf{x}_{\text{mic}}$ deposit energy into octave-band energy histograms, from which multi-channel room impulse responses $h_m(t)$ are synthesized for each microphone m.

Figure 2 visualizes the progression of stochastic acoustic-ray propagation through the furnished room.

Multichannel Reception and Goal Grounding. Simulated multi-channel room impulse responses are convolved with dry synthesized speech, with speech-shaped noise added in the room conditions. A four-channel circular microphone array provides audio for Weighted Prediction Error (WPE) dereverberation (Nakatani et al., 2010; Drude et al., 2018) and delay-and-sum beamforming before Whisper transcription (Radford et al., 2023). A predicate resolver then selects a top-ranked goal from the transcript. We score the selected predicate and its target–reference binding against the intended goal.

3.4 Search over Shoebox-Room Configurations

Adversarial search varies floor, wall, and ceiling materials together with the horizontal source and receiver positions. The valid search space is $\Theta \subseteq \mathcal{M}_{\text{floor}} \times \mathcal{M}_{\text{wall}} \times \mathcal{M}_{\text{ceil}} \times \mathcal{X}_{\text{src}} \times \mathcal{X}_{\text{mic}}$, where

$\mathcal{X}_{\text{src}}, \mathcal{X}_{\text{mic}} \subset \mathbb{R}^2$ are positions within the room subject to a minimum source–receiver separation; their heights remain fixed. Candidates are scored by room-only grounding regret on the same 24 items.

Higher-Regret Configurations. To find a room configuration that lowers grounding accuracy relative to clean speech, we maximize room-only regret over this space:

$$\theta^*_{\text{adv}} = \arg\max_{\theta \in \Theta} R_{\text{room}}(\theta) \tag{5}$$

We compare uniform random sampling, coordinate-wise greedy search, (1+1) evolutionary hill-climbing, and Bayesian optimization with a Gaussian Process surrogate and Expected Improvement acquisition, each with ten candidate evaluations. A fifth, expert-guided method selects reflective materials and explores source–receiver placement.

Single-Surface Remediation. Starting from the higher-regret configuration, we hold source and receiver positions fixed and search for a material change that reduces room-only grounding regret. Writing θ_{mat} for the three surface-material assignments, the constrained objective is

$$\begin{aligned} \theta^*_{\text{rem}} &= \arg\min_{\theta \in \Theta} R_{\text{room}}(\theta) \\ \text{s.t.} \quad \|\theta_{\text{mat}} - \theta^*_{\text{adv,mat}}\|_0 &\le B, \\ \theta_{\text{src}} = \theta^*_{\text{adv,src}}, \quad \theta_{\text{mic}} &= \theta^*_{\text{adv,mic}}. \end{aligned} \tag{6}$$

The budget $B = 1$ allows one surface-material change. Each candidate is evaluated through simulated audio and goal grounding on the same search items.

4 Experimental Evaluation

We evaluate three distinct questions: whether simulated room acoustics degrade the spoken instruction channel, whether far-field processing mitigates this degradation, and whether changes to the acoustic environment alter language-conditioned policy grounding. The speech stress tests use controlled reverberation and noise, the grounding study uses the mesh-aware engine on a generated bedroom, and the surface search uses a shoebox engine. Results are reported for each study with its own scene, engine, and evaluation set.

4.1 Experimental Setup and Generated Scene

We generated a bedroom from the prompt "A small bedroom with a bed and a nightstand." The completed room measures $3.5 \times 3.5 \times 2.6$ m, with two windows, one door, and 26 objects: four walls, seven pieces of floor-standing furniture, six wall-mounted objects, one ceiling light, and eight manipulable objects. The exported acoustic representation contains 29 surfaces; the scene exports include the Blender house, the simulation scene and state, and the acoustic scene specification. Its five-stage generation took 5 h 31 min and produced 412 MB of artifacts.

For the speech-channel experiments, we synthesize instructions and retain only items transcribed correctly in the clean condition, so reported errors measure degradation relative to clean speech on the same utterances. The reverberation/noise comparison uses 27 instructions shared by Whisper-small.en and Whisper-medium.en. A separate front-end experiment uses 30 clean-correct instructions and Whisper-medium.en. Both speech experiments use a controlled $4.2 \times 3.6 \times 2.7$ m shoebox room. Word error rate (WER) counts substitutions, deletions, and insertions divided by reference words. Front-end goal accuracy checks preservation of the parsed action, object, relation, target, and polarity.

The policy experiment uses 50 geometrically valid "put object on surface" goals enumerated from the generated scene, each rendered with four phrasings. The same Qwen2.5-14B-based predicate resolver receives either the written instruction (ORACLE), a transcript of clean synthesized speech (CLEAN), or a transcript after propagation through a room variant (ROOM). Room audio is processed by weighted prediction error (WPE) dereverberation followed by delay-and-sum beamforming from a four-microphone circular array of radius 0.05 m, then transcribed with Whisper-medium.en. A fixed 20 dB speech-shaped noise level is used in the room conditions. Grounding success requires

Table 1: Word error rate (WER) under reverberation and speech-shaped noise on 27 clean-correct instructions. Small and Medium denote Whisper-small.en and Whisper-medium.en; RT_{60} is measured, and noise levels are given as signal-to-noise ratios (SNRs).

Condition	RT_{60} (s)	Small WER	Medium WER
Clean	–	0.000	0.000
Reverb target 0.3	0.37	0.006	0.006
Reverb target 0.6	0.61	0.074	0.011
Reverb target 1.0	1.15	0.114	0.040
Reverb target 1.5	1.75	0.189	0.086
Reverb target 2.0	2.30	0.274	0.183
Reverb target 2.5	2.89	0.434	0.251
Noise SNR 20 to 5 dB	–	0.006–0.011	0.006–0.011
Noise SNR 0 dB	–	0.051	0.057
Noise SNR -5 dB	–	0.217	0.086
Noise SNR -10 dB	–	3.57	2.67
Noise SNR -15 dB	–	16.04	6.77
Noise SNR -20 dB	–	6.26	14.78

Table 2: Far-field speech processing on 30 clean-correct instructions with Whisper-medium.en and a four-microphone array ($r = 0.05$ m). Each entry is word error rate (WER) / parsed-goal accuracy; ch1 is an unprocessed microphone, DS is delay-and-sum beamforming, and WPE is weighted prediction error dereverberation.

RT_{60} (s)	ch1	DS	WPE	WPE+DS
0.37	.021/.87	.010/.93	.005/.97	.005/.97
0.61	.015/.90	.015/.93	.010/.97	.010/.93
1.15	.072/.77	.057/.80	.036/.87	.031/.87
1.75	.175/.57	.160/.57	.088/.73	.108/.63
2.30	.258/.50	.247/.40	.170/.57	.149/.53

the resolver's *top-ranked* predicate and target–reference binding to match the intended goal. We compare matched goals and phrasings across conditions and average over channel realizations before estimating uncertainty. The reported 95% intervals use 5,000 paired bootstrap resamples clustered by goal.

4.2 Speech Recognition under Reverberation and Noise

Table 1 isolates the instruction-channel effect before far-field processing or policy grounding. Both recognizers have zero WER on the screened clean set. With increasing measured reverberation time, medium.en WER rises from 0.006 at 0.37 s to 0.251 at 2.89 s; small.en rises from 0.006 to 0.434. At 1.15 s, their WERs are 0.040 and 0.114, respectively, showing different sensitivity to reverberation in the controlled shoebox setting.

At 5–20 dB SNR, WER remains near clean levels. At -10 dB and below, repetition loops produce insertion-heavy transcripts and WER values above 1. At the most severe noise levels, measured error rates for spatial and negation tokens reach 1.00, while content-token errors approach 1.00.

4.3 Far-Field Front-End Ablation

We next measure the effect of multi-channel processing on recognition and parsed-goal accuracy. Table 2 compares one unprocessed microphone (ch1), delay-and-sum beamforming (DS), WPE dereverberation on one microphone (WPE), and WPE followed by DS. The beamformer uses the known source position for steering, and all conditions are level-normalized. Clean audio has WER 0 and goal accuracy 1.00 on the screened 30-item set.

At 1.75 s, WPE lowers WER from 0.175 to 0.088 and raises parsed-goal accuracy from 0.57 to 0.73 relative to ch1. At 2.30 s, WPE+DS has the lowest WER (0.149 versus 0.258 for ch1), while WPE alone has the highest parsed-goal accuracy (0.57 versus 0.53 for WPE+DS). Beamforming alone

Table 3: Acoustic measurements for four material variants of the generated, furnished bedroom. RT_{60}^{500} and RT_{60}^{sp} are the 500-Hz and speech-band reverberation times; C_{50} is clarity and DRR is the direct-to-reverberant ratio.

Variant	RT_{60}^{500} (s)	RT_{60}^{sp} (s)	C_{50} (dB)	DRR (dB)
As built	.39	.41	14.3	-3.2
Hard floor	.46	.43	13.5	-3.2
Ceiling treated	.17	.27	21.8	-1.3
Both treated	.28	.29	24.7	-1.1

Table 4: Top-ranked predicate and object-binding accuracy on 50 goals with four phrasings in the furnished mesh scene. Spoken accuracy is compared against written-oracle accuracy (0.495) for total regret R, and against clean-spoken accuracy (0.305) for room-only regret R_{room}. Brackets give paired, goal-clustered 95% bootstrap confidence intervals (CIs).

Condition	Spoken acc.	R [95% CI]	R_{room} [95% CI]
Clean speech	.305	$+.190$ [.105, .280]	–
As built	.283	$+.212$ [.127, .293]	$+.022$ [$-.047$, .088]
Hard floor	.237	$+.258$ [.180, .338]	$+.068$ [.003, .130]
Ceiling treated	.268	$+.227$ [.145, .307]	$+.037$ [$-.023$, .095]
Both treated	.247	$+.248$ [.167, .332]	$+.058$ [$-.007$, .122]

yields 0.40 parsed-goal accuracy, compared with 0.50 for ch1. The ranking of processing methods varies between WER and parsed-goal accuracy.

4.4 Policy Grounding and Acoustic Regret in the Mesh Scene

We vary the floor and ceiling materials of the generated bedroom while preserving its furnished mesh and policy goals. Table 3 reports the resulting acoustic descriptors. The treated ceiling reduces the measured 500 Hz RT_{60} from 0.39 to 0.17 s, while the hard floor increases it to 0.46 s. Treating both surfaces gives the highest clarity ($C_{50} = 24.7$ dB) and a 500 Hz RT_{60} of 0.28 s. The material variants produce distinct profiles across RT_{60}, clarity, and direct-to-reverberant ratio.

Table 4 separates the loss from converting written text into clean spoken audio from the additional effect of the room. The written oracle grounds 0.495 of the goals correctly, compared with 0.305 after clean speech synthesis and recognition: a baseline gap of 0.190 (95% CI $[0.105, 0.280]$). Across the four mesh rooms, spoken grounding ranges from 0.237 to 0.283, giving total written-to-spoken regret of 0.212–0.258. The corresponding total-regret confidence intervals lie above zero.

Relative to clean spoken speech, *room-only* regret ranges from 0.022 for the as-built room to 0.068 for the hard floor. The hard-floor interval is $[0.003, 0.130]$; the other three intervals cross zero. The ceiling-treated room has the highest grounding accuracy among the treated variants (0.268), compared with 0.283 in the as-built room.

Table 5 reports the five goals grounded correctly under all four written-oracle phrasings. On these oracle-achievable goals, written-to-spoken regret is 0.400 for clean speech and 0.517–0.617 in the mesh rooms.

4.5 Phrasing and Channel Variability

Table 6 breaks the written-to-spoken regret down by four ways of expressing each goal. "Move the X onto the Y" has the largest clean-speech gap (0.26), with room-conditioned gaps of 0.25–0.33 across the material variants. The remaining phrasings change rank across rooms. Across noise realizations in the room conditions, within-item channel-outcome agreement is 0.70–0.78.

4.6 Adversarial Search and Single-Surface Remediation

We evaluate room-configuration search over materials and source–receiver placement in a shoebox simulation, followed by fixed-placement, single-surface remediation. Each candidate is scored on 24

Table 5: Written-to-spoken regret on five goals selected for correct written-oracle grounding on all four phrasings. Confidence intervals (CIs) use a paired, goal-clustered 95% bootstrap.

Condition	Total regret R	95% CI
Clean speech	$+.400$	[.150, .600]
As built	$+.617$	[.483, .750]
Hard floor	$+.533$	[.350, .733]
Ceiling treated	$+.517$	[.433, .600]
Both treated	$+.617$	[.517, .733]

Table 6: Written-to-spoken regret by instruction phrasing on 50 goals in the furnished mesh scene. X and Y stand for the target and reference objects. Columns denote clean speech, the as-built room, a hard floor, a treated ceiling, and both surfaces treated; values are descriptive point estimates.

Phrasing	Clean	As built	Hard floor	Ceiling	Both
Put X on Y	.12	.21	.22	.22	.24
Place X on Y	.16	.18	.23	.15	.26
Move X onto Y	.26	.25	.33	.30	.29
Take X and put on Y	.22	.20	.26	.23	.21

grounding items; the five search methods receive a budget of ten evaluations each per direction. The objective is R_{room}, the difference between clean-spoken and room-spoken grounding on the search items. Clean-spoken grounding accuracy is 0.500, and the as-built shoebox room has $R_{\mathrm{room}} = +0.250$ with measured $\mathrm{RT}_{60} = 1.86\,\mathrm{s}$.

Table 7 shows that an expert-guided reflective configuration raises room regret to $+0.500$, twice the as-built objective on these items. Random search reaches $+0.375$; greedy, evolution, and Bayesian optimization retain the as-built score of $+0.250$ within this budget.

Starting from the high-regret room, four methods find the same ceiling replacement, reducing the measured room-only regret from $+0.500$ to 0.000 on the search items (Table 8); the expert wall replacement leaves $+0.125$. This matches clean-spoken grounding accuracy on the shoebox search items. Across the study there were 101 evaluator calls, 319 model calls, 2,105 grounding-cache hits, and 43 cached configurations; the search took approximately 3.3 h.

5 Limitations

Grounding success measures the resolver's selected predicate and object binding, not completed robot manipulation. The mesh evaluation contains one generated bedroom and 50 goals, and the shoebox search scores 24 items in a separate simulation. The studies therefore do not establish generation reliability across prompts, real-room acoustic fidelity, or comparative performance of the two engines under matched conditions. Source and microphone positions were held fixed during single-surface remediation, while adversarial search also varied their horizontal positions; real-world installation costs and robustness after policy training remain unevaluated.

The written-to-spoken gap includes the effects of speech synthesis and recognition. For room-only regret, three of four mesh-room confidence intervals include zero; the treated variants show no reliable grounding advantage over the as-built room despite higher clarity. The five-goal oracle-achievable subset is selected by written-oracle outcomes and is not representative of all 50 goals. The speech stress tests use clean-correct screened utterances in a shoebox room, and the sparse spatial and negation token classes limit interpretation of their maximum error rates. At severe noise levels, insertion loops produce WER values above 1 with non-monotonic behavior across SNRs. Channel-outcome agreement does not isolate resolver sampling variance because inputs were not independently repeated within the same realization. Search results come from one discrete evaluation with no cross-scene replication; cache hits and repeated evaluations do not add independent scenes or goals.

Table 7: Adversarial search over shoebox-room materials and source–receiver positions with ten evaluations per method and 24 grounding items per configuration. Higher room-only regret R_{room} indicates worse grounding; surface materials are listed as floor, wall, and ceiling, in that order. As built denotes the reference room.

Method	Best R_{room}	Floor / wall / ceiling
Expert	$+.500$	Marble / ceramic tile / concrete
Random	$+.375$	Ceramic tile / ceramic tile / concrete
Greedy	$+.250$	As built
Evolution	$+.250$	As built
BayesOpt	$+.250$	As built

Table 8: Single-surface remediation of the high-regret shoebox room (initial room-only regret $R_{\mathrm{room}} = +0.500$). Regret reduction is measured relative to this starting point; each intervention changes at most one surface.

Method	Best R_{room}	Regret reduction	Surface change
Random	.000	$+.500$	Ceiling $\rightarrow$ cardboard
Greedy	.000	$+.500$	Ceiling $\rightarrow$ cardboard
Evolution	.000	$+.500$	Ceiling $\rightarrow$ cardboard
BayesOpt	.000	$+.500$	Ceiling $\rightarrow$ cardboard
Expert	$+.125$	$+.375$	Wall $\rightarrow$ acoustic foam

6 Conclusion

AcoustiGym connects generated indoor scenes with acoustic propagation and spoken-instruction grounding. In the furnished mesh scene, written instructions reach 0.495 grounding accuracy, clean speech reaches 0.305, and room materials produce room-only regret between 0.022 and 0.068. A separate shoebox study uses goal-grounding regret to find a higher-regret surface configuration and a single-surface change that recovers the clean-spoken score on its search items.

References

Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision (ECCV)*, pp. 17–36, 2020.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 5982–5994, 2022.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 13049–13061, 2020.

Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach. NARA-WPE: A python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing. In *ITG-Symposium on Speech Communication*, pp. 1–5, 2018.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. Scenesmith: Agentic generation of simulation-ready indoor scenes. *arXiv preprint arXiv:2602.09153*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 351–355, 2018.

Russ Tedrake and Drake Development Team. Drake: Model-based design and verification for robotics, 2019. URL `https://drake.mit.edu`.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753*, 2019.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16227–16237, 2024.