ANAVI: Audio Noise Awareness using Visuals of Indoors for NAVIgation

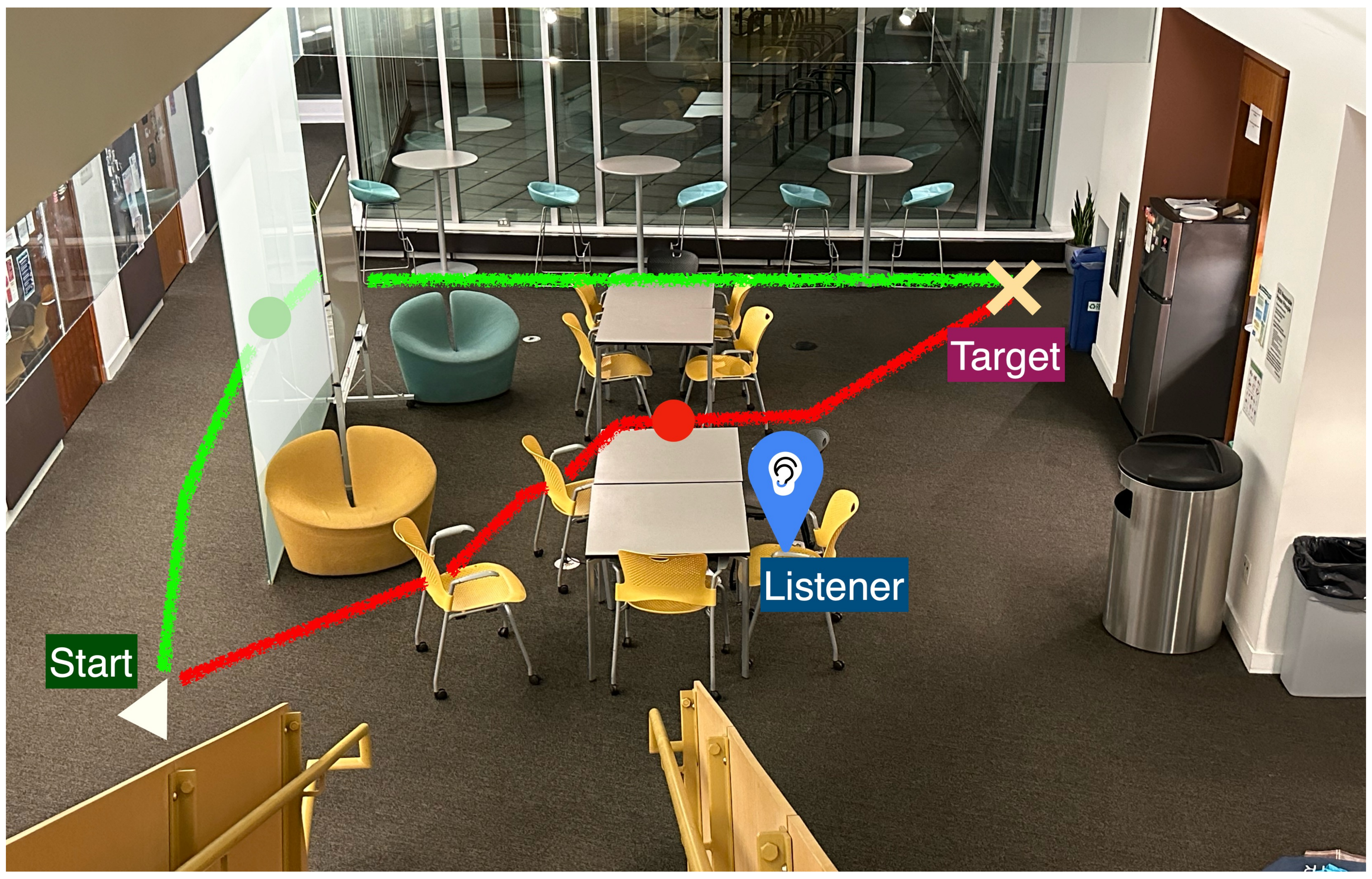
Carnegie Mellon University

Vidhi Jain Rishi Veerapaneni Yonatan Bisk

vidhij@andrew.cmu.edu

ANAVI-CoRL.24

Motivation



Red Trajectory

- ✓ Shortest possible route
- ✗ Very loud for the Listener

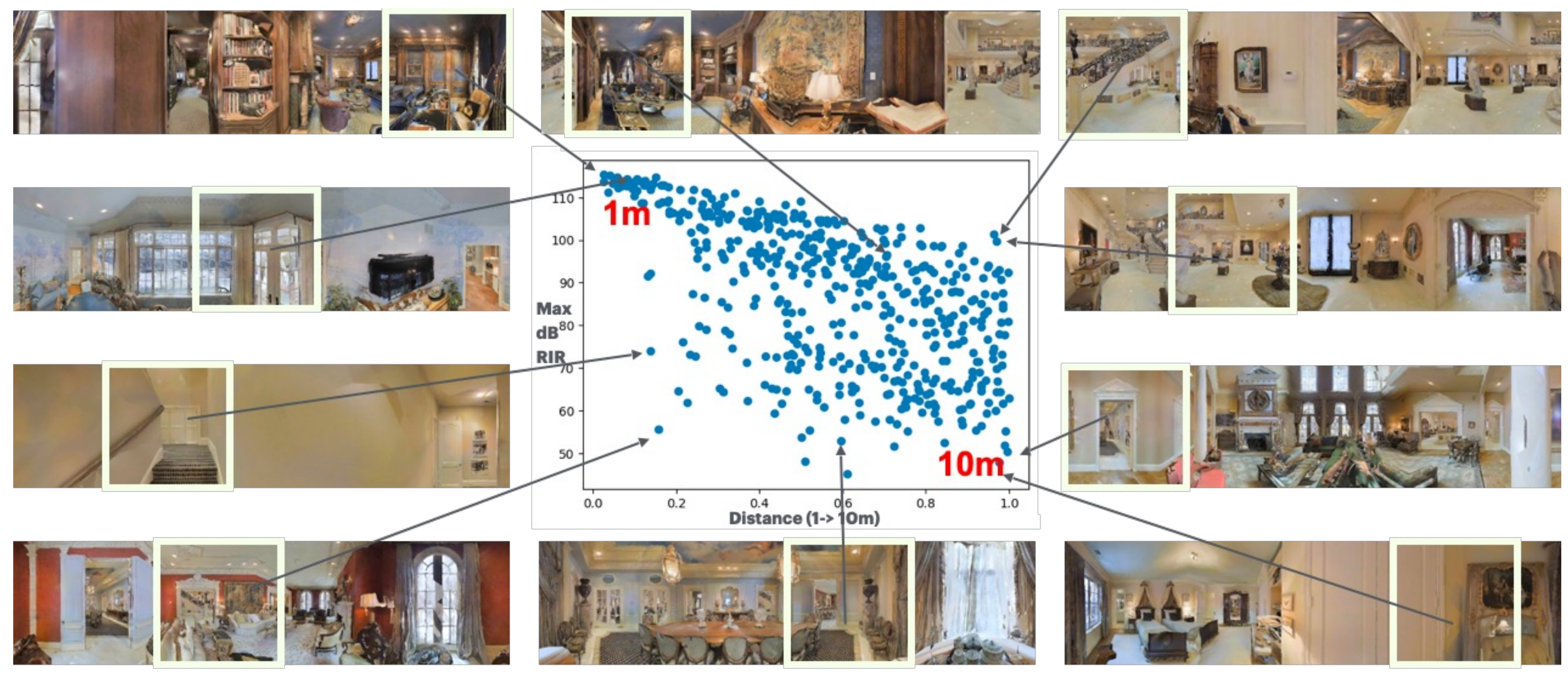
Green Trajectory

- ✗ Slightly longer route
- ✓ Less noisy for the Listener

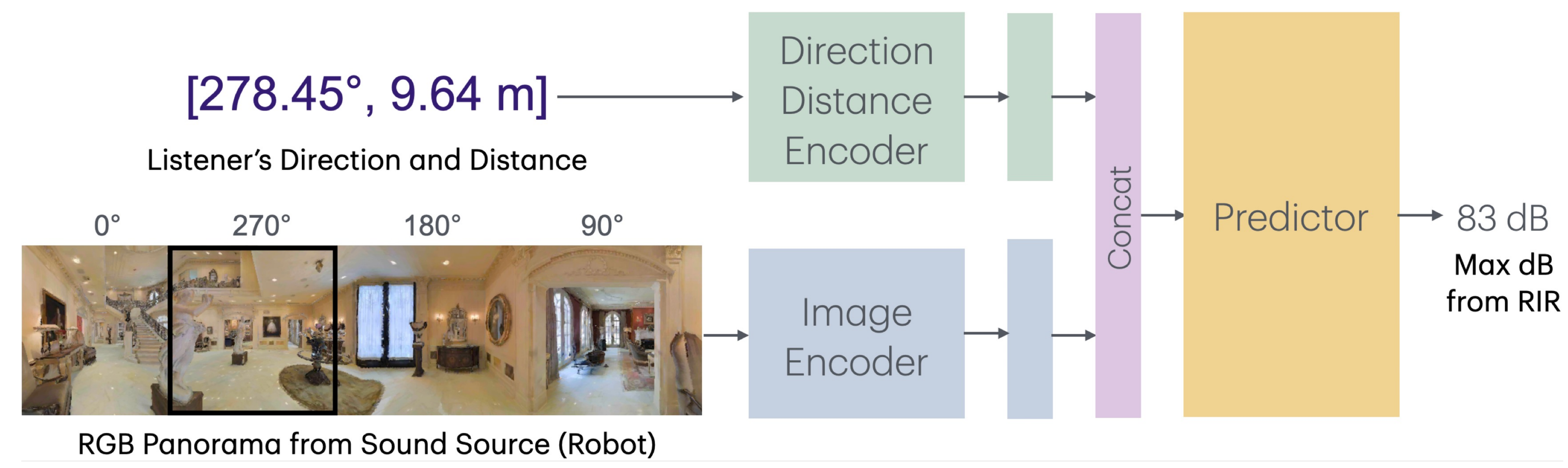
While humans are naturally aware of the noise they make and its impact on those around them, robots currently lack this awareness. In isolation, e.g. without acoustic reflections or echoes, unoccluded sound intensity decays quadratically with distance. A simplistic approach to reducing this noise is for the robot to move slowly and steer clear of people and pets in its environment. But sound depends upon the geometry and material composition of rooms. To this end, we train the robot to passively perceive loudness using visual observations of indoor environments. Existing work on audio for robotics focuses on finding the source of an audio signal, guiding navigation. In contrast, we focus on awareness of self-generated sounds; a related yet orthogonal problem:

How loud will the robot's actions be at a listener's location?

Method

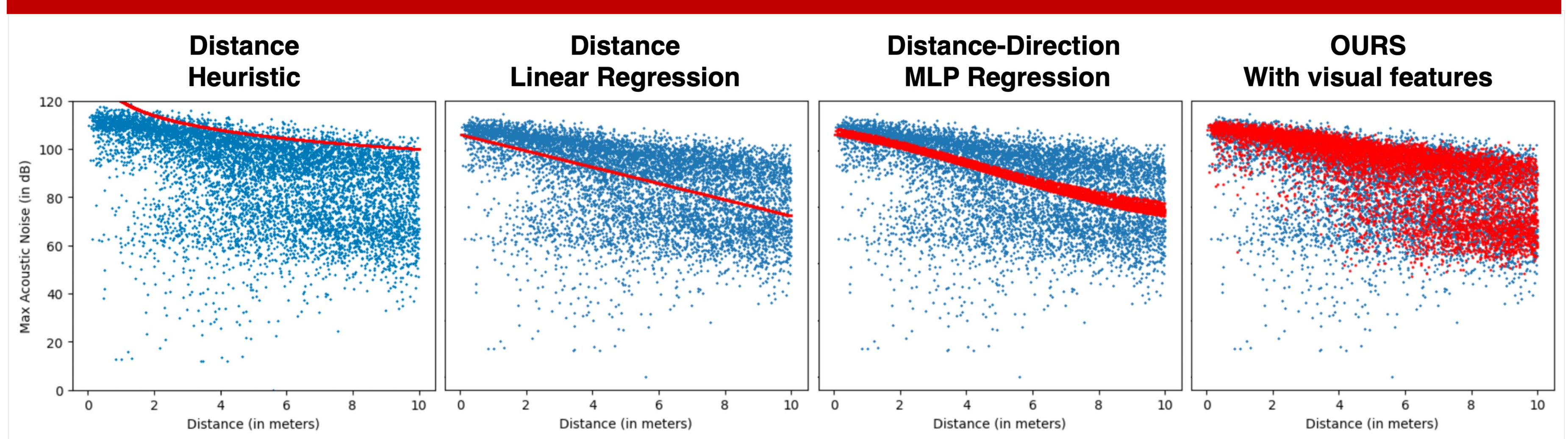


Our objective is to train an Acoustic Noise Predictor (ANP) which tries to estimate how loud an impulse sound will be at different locations. We do this by generating data on how loud an 'impulse' sounds at different listener locations in simulated homes. To render realistic sounds, the simulator uses privileged information about 3D mesh geometry and material sound properties for ray tracing. Here the x-axis shows the relative distance of the listener from the source agent, y-axis shows the max decibels of a simulated Impulse Response (IR) at listener. The box area on the panorama shows the listener's direction relative to the agent. Note the complex response pattern as materials, objects, and geometry have non-linear interactions with the sound.

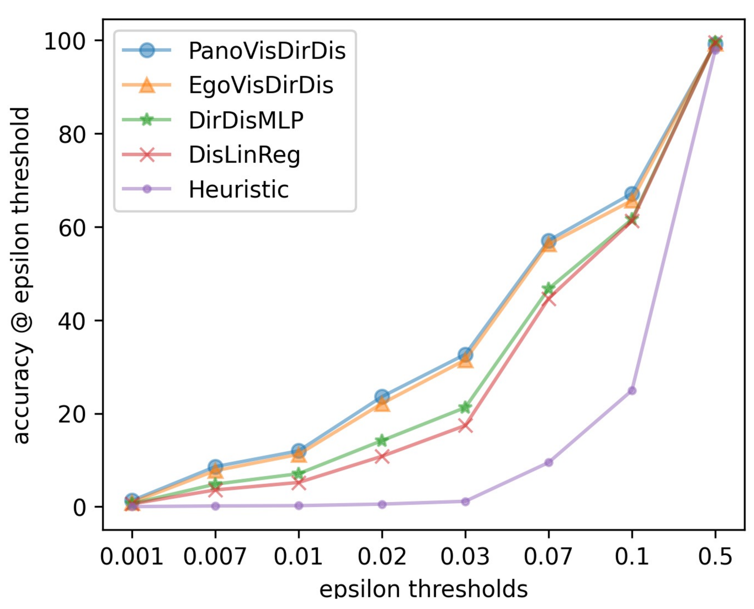


Our Acoustic Noise Prediction (ANP) model consists of Image encoder, Direction-Distance encoder and Predictor modules. The inputs are the 360° RGB panorama view at the robot's location, and the relative polar coordinates of the listener. The square box on the panorama highlights robot's current facing direction, and is drawn for illustration purposes only. The output is the max decibel (dB) of the Room Impulse Response (RIR) at the listener's location.

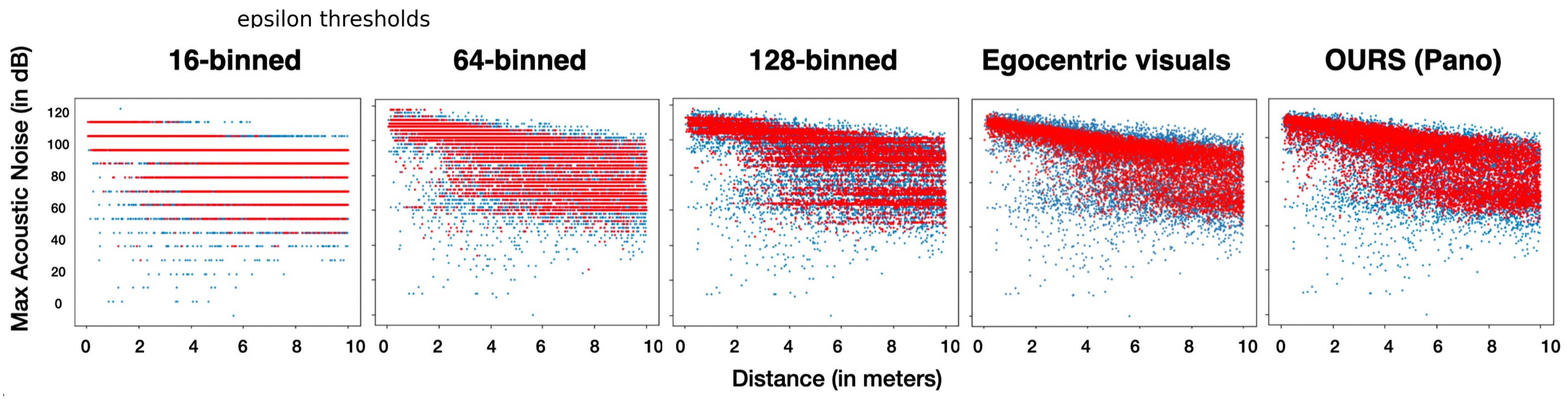
Results



Data distribution coverage plots for the models: Heuristic, DisLinReg, DirDisMLP, (Ours) PanoVisDirDis. The predicted data distribution is in **red**, overlayed on the true test data in **blue**. The ideal model would directly match the blue data distribution, our model most closely aligns with the ground truth.



Here we show ∊ – accuracy for learned models. On the far left, the geometry spreading-based distance heuristic (Heuristic) only predicts correctly for free space and doesn't account for reflections in corridors or absorption by walls in home environments. Others also lack sufficient signal about the spatial context that could inform if there are possible reflections, absorptions, and diffractions of the sound. These observations validate our approach that using visual features along with distance and direction (Pano-VisDirDis) results in the best ∊ – accuracy and data distribution coverage.



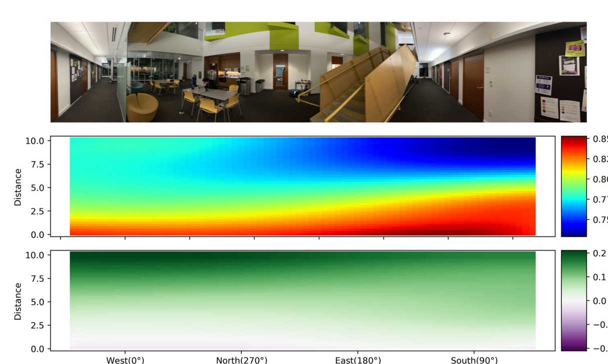
Ablations: The cross-entropy models (1-3) are easier to train and capture outliers but provide prediction at a coarser granularity. The learned distributions with ego-view miss outliers with very high values for the longer distances. Our model trained with regression on panoramic images capture the distribution well.

Takeaway

ANAVI, a framework for Audio Noise Awareness using Visuals of Indoors:

- ANP learns about audio passive perception of indoor visuals.
- Simulated training data for visual and audio RIR + Real Robot Noise Models.
- Real-world experiments for ANP model.

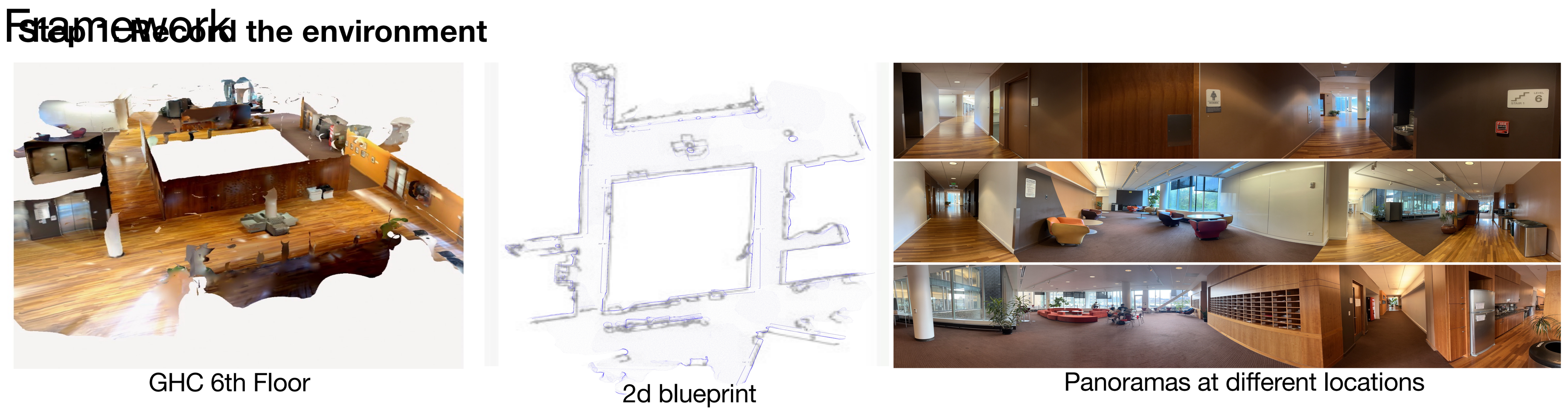
What's next?



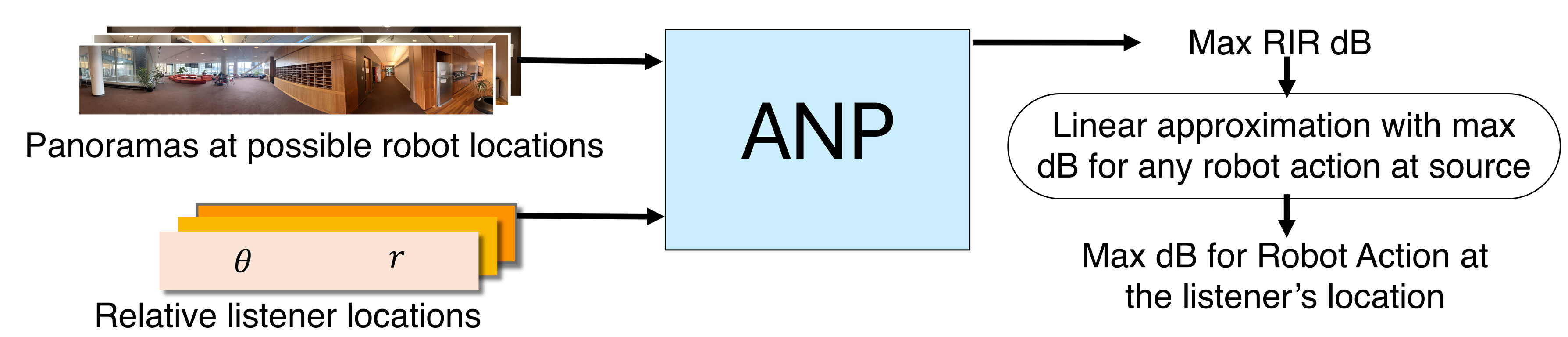
We see a severe sim-to-real gap when applying the ANP model to real world panoramas. For each plot, the top row shows a real-world panorama, second row the ANP predictions depending on the distance and angle to the sensor location, and the third row the difference compared a basic linear regression estimate (green = ANP predicts louder). While the model captures gradual degradation with distance and variation with visual features, it fails to consistently capture meaningful features like walls or corridors that effect dB.

How can we use audio awareness in robot planning? - ANAVI Framework

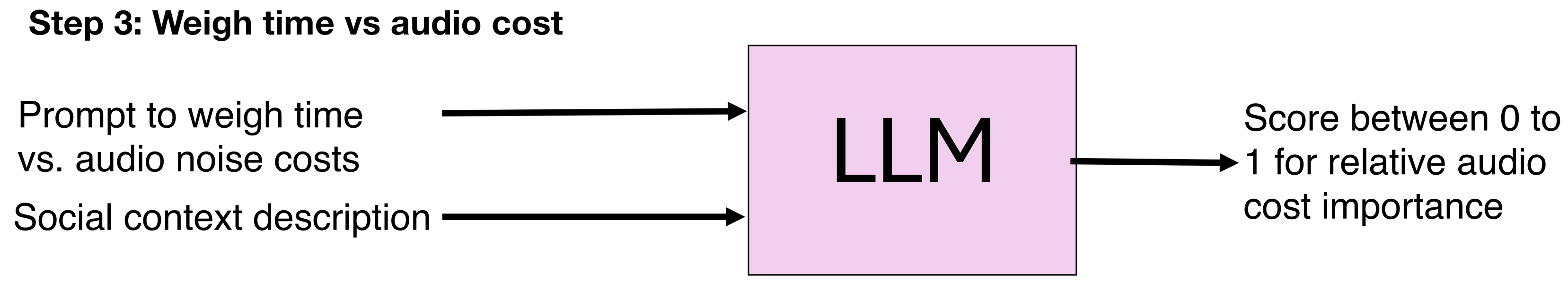
Step 1: Record the environment



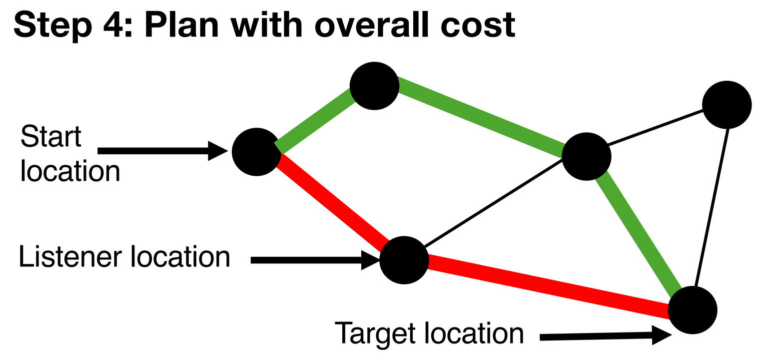
Step 2: ANP inference on set of robot locations



Step 3: Weigh time vs audio cost



Step 4: Plan with overall cost



Questions?

What do you think about robots with audio awareness?

How could this change your daily interactions with them?