AVLOC: ROBUST AUDIO-VISUAL FLOORPLAN LOCALIZATION UNDER SENSOR NOISE AND ENVIRONMENTAL CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Floorplan localization estimates an agent's pose within a 2D floorplan from egocentric observations. Existing methods draw every prediction from a single sensor, a camera, whose limited field of view and vulnerability to dim corridors, motion blur, sensor noise, and occluded walls mean that one corrupted image corrupts every cue the model has. We present AVLoc, which pairs vision with binaural acoustic echoes. Echoes are governed by geometry rather than appearance, they cover the full 360°, and are largely unaffected by the conditions that disable a camera. Complementarity alone does not give robustness, since a joint encoder can propagate corruption from one modality into the other. AVLoc therefore combines unimodal and multimodal experts, each predicting the surrounding layout with its uncertainty. The router weights the experts by these uncertainty estimates, so the model dynamically shifts its reliance toward whichever modality the current observation supports, and the fused uncertainty carries through to floorplan matching. Under maximum visual corruption, AVLoc retains 53.0% recall@1m against 7.5% for the strongest vision-only baseline, and 30.1% when vision and audio are degraded at once. Because echoes also constrain geometry beyond the field of view, AVLoc improves clean-condition accuracy as well, raising recall@1m from 56.7% to 74.7% on Gibson(f) and from 63.3% to 75.6% on HM3D. Code and benchmarks will be released upon publication.

1 INTRODUCTION

A floorplan is often the only map of a building that already exists. Localizing an agent against one is therefore more practical than requiring a 3D reconstruction Micusik & Wildenauer (2015) or a database of posed images Taira et al. (2018), which must be rebuilt for every scene, and a model that reads floorplans transfers to buildings it has never seen.. The prevailing formulation predicts from an egocentric image the distances to the surrounding walls along a set of directions (rays), and matches this ray profile against rays rendered from candidate poses on the floorplan Chen et al. (2024).

Every prediction such a model makes comes from a single sensor, which is fragile in exactly the settings where localization is needed: dim corridors, motion blur while turning, sensor noise from inexpensive cameras, and furniture or people occluding the walls that carry the geometric signal. Recent work adds semantics Grader & Averbuch-Elor (2025), room-style cues Chen et al. (2026), top-down floorplan patches Meng et al. (2026), or a pretrained monocular depth prior Wueest et al. (2026), each raising accuracy in clean conditions, but none changes where the evidence comes from: a corrupted image corrupts every one of these cues at once, so they inherit the same failure mode. The limitation is a property of the sensing modality, not of model capacity.

Robustness in such settings requires a second channel that fails under different conditions than the first. Audio has this property. An active acoustic probe, such as a short sweep emitted in the room, returns reflections from the surrounding surfaces, and those reflections are governed by geometry rather than appearance: darkness, blur, and camera noise leave them untouched, and an object that hides a wall from the camera does not remove the wall from the response. The converse holds as well, since ambient noise that masks an echo leaves the image intact. Echoes also return from all directions at once, so a single binaural response constrains geometry outside the camera's

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.

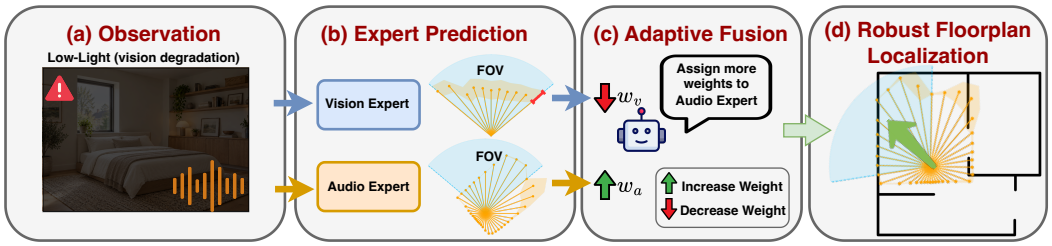


Figure 1: **Conceptual overview of AVLoc.** Given (a) an audio-visual observation, (b) each modality-specific expert estimates per-ray distances along with uncertainty values. (c) A ray-wise router leverages these uncertainties to adjust the weighting of experts for each direction, prioritizing audio when low-light conditions impair vision. (d) Consequently, AVLoc can robustly localize the agent on the floorplan even if one modality degrades. The audio-visual expert is omitted for simplicity.

field of view (FOV), where a 106° view leaves 25 of the 36 directions sampled at 10° intervals unconstrained.

Complementary failure modes are a property of the modalities, not of a model that consumes them. A network trained on both streams learns to depend on both, so corrupting either moves the joint representation off the distribution it was trained on. Fixing the fusion weights does not help either, since any fixed rule assumes a reliability ordering that is violated exactly in the conditions that motivate using two modalities. Reliability is also not observable at test time, and it is not uniform across directions, since visual corruption affects only the FOV while acoustic interference affects the full 360°. Turning two complementary modalities into a robust estimator therefore requires keeping a prediction path that no single corrupted input can contaminate, inferring reliability from the model's own behavior, and resolving it separately for each direction.

We introduce **AVLoc**, an **A**udio-**V**isual **Loc**alization framework built on these three requirements. AVLoc runs an audio-visual, an audio-only, and a vision-only expert, each predicting for every direction a wall distance and its uncertainty. The unimodal experts preserve prediction paths that do not pass through the other modality, so a corrupted stream cannot contaminate the entire estimate, while the joint expert supplies the cross-modal interactions the unimodal experts cannot express. A ray-wise router maps the experts' predicted uncertainties to per-direction mixture weights, and the fused prediction and uncertainty are used directly as the observation likelihood during floorplan matching, so what the model does not know propagates into the pose estimate instead of being discarded after fusion (Figure 1).

Across a grid of visual and acoustic corruptions on Gibson(f), AVLoc degrades gradually where vision-only methods collapse: under maximum visual corruption it retains 53.0% recall@1m against 7.5% for the strongest vision-only baseline, and 30.1% remains when vision and audio are degraded at once. The same behavior holds over trajectories, where the gap between clean and noisy sequences is markedly smaller than for a vision-only model. Our ablations show that replacing uncertainty-guided routing with equal weights costs 11.9 points of recall@1m, and removing either the unimodal experts or the joint expert costs further accuracy. The same design also improves clean-condition accuracy, raising recall@1m from 56.7% to 74.7% on Gibson(f) and from 63.3% to 75.6% on HM3D, more than doubling recall@0.1m, and lifting floor accuracy from 73.4% to 83.8% in multi-floor buildings, where layouts may repeat across floors. The gains also carry over to real-world captures.

Our main contributions:

- We show that pairing vision with binaural echoes makes floorplan localization hold up under lighting changes, motion blur, sensor noise, and acoustic interference, and we introduce the architecture that produces this behavior: unimodal experts that preserve uncontaminated prediction paths, a ray-wise router that infers each expert's reliability from its own predicted uncertainty, and propagation of the fused uncertainty into floorplan matching.
- We show that echoes also constrain geometry outside the FOV, reducing the pose ambiguity caused by repetitive indoor layouts and improving accuracy in the clean conditions studied by prior work, including across floors of the same building.

- We introduce benchmarks for audio-visual floorplan localization: Gibson augmented with binaural echo, an HM3D benchmark with 278 manually annotated floorplans, and HM3D-Multi, which requires localizing within a multi-floor building rather than a known floor.

2 Related Work

Visual Floorplan Localization Early learning-based methods match RGB observations to floorplan representations through shared latent embeddings, pose-dependent map encodings, or differentiable filtering Howard-Jenkins et al. (2021); Howard-Jenkins & Prisacariu (2022); Karkus et al. (2018); Min et al. (2022), but assume constrained input such as upright or panoramic images. F^3Loc Chen et al. (2024) relaxes this by predicting a 1D depth ray from a non-upright image and matching it against rays rendered from the floorplan. While effective, this formulation remains vulnerable to ambiguity: repetitive structures and visually similar layouts in floor plans can produce multiple locations with similar depth-ray observations.

Subsequent methods address this ambiguity by enriching the visual cues read from the image: semantics Grader & Averbuch-Elor (2025), room style Chen et al. (2026), top-down floorplan patches for re-ranking Meng et al. (2026), 3D geometric priors distilled into the encoder Chen et al. (2025), and depth rays predicted with explicit uncertainty from a monocular depth model Wueest et al. (2026); Yang et al. (2024). Two of these target reliability directly, 3DP Chen et al. (2025) through invariance to appearance change, UnLoc Wueest et al. (2026) by down-weighting rays its predictor is unsure about, yet all read the same sensor. The observed region therefore remains bounded by the field of view, and a corrupted image corrupts every cue at once, so modeling uncertainty can report that the evidence is unreliable without supplying evidence to replace it. We add a second sensing channel whose failure modes are largely independent of the camera's, and use predicted uncertainty to choose among sources in a dynamic manner rather than only to discount them.

Audio for Spatial Understanding Multi-channel audio carries spatial information through interaural time and phase differences, since sound propagation reflects the surrounding geometry. It has been used to estimate room geometry from echoes and impulse responses Dokmanić et al. (2013); Yeon et al. (2024); Bicer et al. (2024), to improve monocular depth and spatial representations by relating images to their acoustic response Gao et al. (2020); Christensen et al. (2020); Brunetto et al. (2023), to support embodied navigation, floorplan reconstruction, and moving toward a sound source Chen et al. (2022); Purushwalkam et al. (2021); Majumder et al. (2021), and more recently for higher-level spatial reasoning in audio language models Biswas et al. (2025); Zheng et al. (2024); Sakshi et al. (2025). In all of these, audio describes the scene relative to the listener or indicates where to go next. Acoustic indoor localization is a long-standing line of work: ambient-sound or impulse-response fingerprints are collected at known positions and a query is matched against that database Tarzia et al. (2011); Rossi et al. (2013). Like posed-image databases, these methods require per-environment collection and do not transfer to a building that was never surveyed, whereas we match acoustic observations against a floorplan that already exists. We use acoustic echoes for that purpose, and because they are measured in every direction at once through a channel indifferent to illumination and image noise.

Fusion under Degraded Modalities Multimodal models are typically trained with all streams intact and fuse them with a fixed operation, which degrades sharply when a stream is missing or corrupted at test time Ma et al. (2022; 2021). A separate line predicts heteroscedastic uncertainty Kendall & Gal (2017) and uses it to weight sources so that low-quality views contribute less Han et al. (2021); Zhang et al. (2023). These methods arbitrate between redundant sources: each view predicts the same target, and fusion decides how much each is trusted. Vision and audio are not redundant here, since the camera constrains only the directions inside its field of view and is silent about the rest, so the set of available experts changes with direction and fusion must handle the absence of a source in addition to its quality. Our design also improves robustness through isolation rather than repair, the unimodal experts keep a prediction path that corruption cannot reach, instead of restoring a joint representation after one stream has already contaminated it.

3 Method

AVLoc estimates the layout surrounding the agent from an image and a binaural echo response, and then matches that layout against the floorplan. Based on predicted uncertainties, the model learns to weight the evidence from each modality separately for each direction, according to the quality of

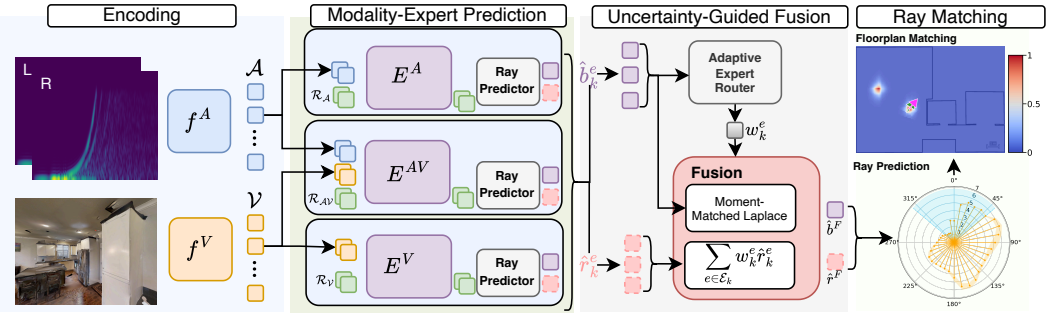


Figure 2: Overview of AVLoc. Given an RGB image and a binaural echo response, Audio, Vision, Audio-Visual Experts predict ray distances $\hat{r}_k^e$ and uncertainties $\hat{b}_k^e$. Then, the uncertainty-guided fusion module dynamically fuses the outputs of three experts based on their uncertainty to obtain the final rays $\hat{r}^F$ and uncertainty estimate $\hat{b}^F$. The predicted rays and uncertainty are then matched to the floorplan to estimate the agent pose.

the input, which varies with sensor noise and environmental conditions, and to the scene structure, which can make a direction hard to predict even from a clean observation. In the following, we state the estimation problem and the ray representation in Sec. 3.1, describe the experts in Sec. 3.2, the router and fusion rule in Sec. 3.3, and floorplan matching and training in Sec. 3.4.

3.1 Problem Formulation

Given an RGB image V, a binaural echo response A, and a 2D floorplan F, we estimate the agent pose $S = (x, y, \theta)$, where (x, y) is the position on the floorplan and θ the heading.

The pose space is discretized into a set of candidate poses $\mathcal{S} = \{S_i\}$, which is sampled on a 10 cm grid over F Chen et al. (2024), and localization is formulated as a maximum a posteriori estimation:

$$\hat{S} = \arg\max_{S_i \in \mathcal{S}} p(S_i \mid V, A, F)$$

We evaluate this likelihood through a ray profile matching process. The surrounding geometry is described by the distances to the nearest wall along $K = 36$ directions, uniformly spaced at 10° over 360° in the agent's frame. The observation is mapped to a distribution over these distances, and a candidate pose is scored by how well the rays rendered at S_i agree with it Chen et al. (2024). Localization accuracy therefore depends not only on the predicted distances but also on the uncertainty estimates associated with each direction.

3.2 Multimodal Multi-Expert Ray Distribution Prediction

AVLoc uses three ray-prediction experts: audio-visual, audio-only, and vision-only (Fig. 2). The audio-visual expert captures cross-modal interactions the unimodal experts cannot express, but its joint representation becomes unreliable once either modality degrades. The unimodal experts keep prediction paths that the other modality cannot reach, and therefore remain informative when the joint representation is compromised. Each expert predicts, for every direction it covers, a ray distance and an uncertainty estimate, which the fusion module uses as a reliability signal.

Modality Encoders The vision encoder f^V maps the RGB image $\mathbf{V}$ to vision tokens $\mathcal{V} = \{\mathbf{v}_i\}$ and the audio encoder f^A maps the binaural echo response $\mathbf{A}$ to audio tokens $\mathcal{A} = \{\mathbf{a}_j\}$. Each encoder uses a convolutional backbone to extract local features, which are augmented with positional and modality embeddings, then projected with two layers of self-attention blocks that capture the global context within each modality into a set of tokens of a shared embedding dimension d.

Ray Tokens Each expert maintains its own learnable ray tokens $\{\mathcal{R}_{AV}, \mathcal{R}_A, \mathcal{R}_V\}$ with learnable directional positional embeddings. The audio-visual E^{AV} and audio-only E^A experts use $K_{AV} = K_A = K$ omnidirectional ray tokens. The vision-only E^V expert uses $K_V = 11$ tokens, covering the 106° FOV, since the image does not constrain the other directions. Following prior work Chen et al. (2024); Wueest et al. (2026); Meng et al. (2026), the in-FOV ray tokens of the vision-only and audio-visual experts are seeded with features sampled from the image columns geometrically

aligned with each ray's angle, anchoring them to direct image evidence. The audio-only expert receives no visual seeding, and its predictions are independent of the image in every direction.

Modality Expert Encoders The audio-visual expert E^{AV} concatenates its ray tokens with $\mathcal{V}$ and $\mathcal{A}$ and processes the joint sequence with Transformer encoder blocks, letting each ray token attend to visual and acoustic evidence simultaneously. Then E^{AV} predicts the distance $\hat{r}_k^{AV}$ and uncertainty estimate $\hat{b}_k^{AV}$ per ray.

The unimodal experts (E^A and E^V) process one modality each. Each concatenates its learned ray tokens with the corresponding modality tokens, applies its own Transformer encoder, and predicts a unimodal estimate for the ray distances and uncertainty, $(\hat{r}_k^V, \hat{b}_k^V)$ and $(\hat{r}_k^A, \hat{b}_k^A)$. Because neither sees the other modality, corruption of one stream leaves the other expert's prediction unaffected.

3.3 Uncertainty-Guided Expert Fusion

Adaptive Expert Router Which expert to trust depends on the quality of the current input, which is not given to the model, so we use each expert's predicted uncertainty as a proxy for its reliability. Reliability also varies across directions, so routing is performed independently per ray: we learn K separate two-layer MLPs, one per direction, each mapping the uncertainty estimate of the experts available at that direction to softmax-normalized weights w_k^e. All three experts participate within the FOV. Outside it the vision weight is zero, since E^V makes no prediction there. We detach the uncertainty estimates before the router to keep them independent and prevent the fusion loss from altering expert uncertainties through the routing path.

Fusion Module The fused ray distance $\hat{r}_k^F$ is the weighted mean of the expert predictions, $\hat{r}_k^F = \sum_{e \in \mathcal{E}_k} w_k^e \hat{r}_k^e$, where $\mathcal{E}_k$ is the set of experts available at direction k.

We obtain the fused uncertainty $\hat{b}_k^F$ by applying the law of total variance to this mixture and moment-matching the result to a single Laplace distribution, which keeps the observation model in the same form the matching step consumes:

$$(\hat{\sigma}_k^F)^2 = \sum_{e \in \mathcal{E}_k} w_k^e \left[2(\hat{b}_k^e)^2 + (\hat{r}_k^e - \hat{r}_k^F)^2 \right], \qquad \hat{b}_k^F = \frac{\hat{\sigma}_k^F}{\sqrt{2}}. \tag{1}$$

The two terms capture within-expert uncertainty and the spread of the expert predictions around the fused mean.

3.4 Localization and Training

Uncertainty-Aware Floorplan Matching Given the fused ray distribution $\{(\hat{r}_k^F, \hat{b}_k^F)\}_{k=1}^K$, we score each candidate pose by its likelihood under that distribution, treating directions as conditionally independent given the pose:

$$\hat{S} = \arg\max_{S_i \in \mathcal{S}} \prod_{k=1}^{K} \frac{1}{2\hat{b}_k^F} \exp\left(-\frac{|r_k(S_i) - \hat{r}_k^F|}{\hat{b}_k^F} \right), \tag{2}$$

where $r_k(S_i)$ is the floorplan ray distance rendered from S_i. Candidate headings are discretized at the same 10° spacing as the rays, giving 36 headings per location. Uncertainty acts directly on the pose estimate, e.g., a direction with a large $\hat{b}_k^F$ constrains the pose weakly.

Learning Objectives The model output is supervised with the negative log-likelihood of the Laplacian distribution Wueest et al. (2026):

$$\mathcal{L}_o = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{|r_k - \hat{r}_k^o|}{\hat{b}_k^o} + \log\big(2\hat{b}_k^o\big) \right), \text{where } o \in \{F, AV, A, V\} \tag{3}$$

In this objective, the first term is small when the uncertainty is large and the second penalizes large uncertainties. The predicted uncertainty is thus tied to the difficulty of the input rather than being free, which is what makes it usable as the router's input signal. We apply the same objective to each expert as well as the final fused result, giving $\mathcal{L}_F$, $\mathcal{L}_{AV}$, $\mathcal{L}_A$, and $\mathcal{L}_V$, with the vision term evaluated only over the K_V in-FOV directions. The overall objective is: $\mathcal{L} = \mathcal{L}_F + \lambda_{AV}\mathcal{L}_{AV} + \lambda_V \mathcal{L}_V + \lambda_A \mathcal{L}_A$, with $\lambda_{AV} = 0.1$ and $\lambda_V = \lambda_A = 0.5$ selected based on the validation loss of each individual expert and used unchanged in all experiments.

4 EXPERIMENTS

Dataset We evaluate on Gibson Xia et al. (2018) and introduce a new benchmark on HM3D Yadav et al. (2023). We also provide results on a real-world dataset to show feasibility of our approach in practice (Sec. 4.4). Following prior work Chen et al. (2024); Meng et al. (2026), we use the Gibson(f), Gibson(g), and Gibson(t) splits, covering 118 scenes (100 training, 9 validation, 9 test). Gibson(g) contains 49,558 samples of four sequential views and Gibson(f) 24,779 samples of four forward-facing sequential views; Gibson(t) contains 118 trajectories of 280 to 5,152 steps. We additionally introduce a floorplan localization benchmark on HM3D Yadav et al. (2023), for which we manually annotate 278 floorplans and generate 119,818 samples. We extend the previous datasets with binaural echo responses simulated in SoundSpaces 2 Chen et al. (2022). See Sec. D for details.

Implementation Details Vision is represented as RGB of size 640x480 and audio as four-channel spectrograms containing left/right log-magnitude and sine/cosine inter-channel phase differences Zheng et al. (2024). The vision encoder is an ImageNet-pretrained ResNet-50, and the audio encoder is a four-channel ResNet-18. The audio-visual and audio-only experts use four Transformer layers with 36 ray tokens. The vision-only expert uses two layers with 11 visually seeded ray tokens. During training, we apply visual corruptions at varying severities and ambient audio noise at varying SNRs. See Sec. C and Sec. G for details.

Baselines F^3Loc Chen et al. (2024) introduces the ray-based observation model coupled with a histogram filter that most subsequent methods build on. 3DP Chen et al. (2025) distills 3D geometric priors from RGB-D scans into the visual encoder through contrastive pre-training, targeting robustness to appearance change and object occlusion. RSK Chen et al. (2026) injects room-style knowledge into the visual encoder through unsupervised pre-training to reduce ambiguity between visually similar rooms. Disco-FLoc Meng et al. (2026) generates pose candidates with a depth-aware ray predictor and re-ranks them by contrastively matching the RGB observation against local top-down floorplan patches. UnLoc Wueest et al. (2026) predicts floorplan depth as an explicit uncertainty distribution from a pretrained monocular depth encoder and fuses uncertainty-weighted predictions over time. On HM3D, we train and evaluate the baselines F^3Loc, Disco-FLoc, and UnLoc with publicly available code using their published hyperparameters. We also train and evaluate UnLoc on Gibson(f) and Gibson(g).

Evaluation Metrics We report recall at three distance thresholds, the percentage of predictions whose estimated position lies within 0.1 m, 0.5 m, or 1 m of the ground truth. To measure position and orientation jointly, we also report recall at 1 m 30°, which requires translation error below 1 m and orientation error below 30°.

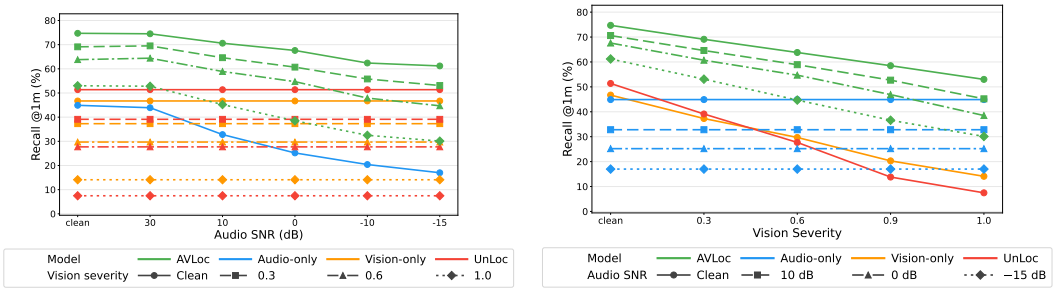
Method (Venue)	**Gibson(f)**				**Gibson(g)**				**HM3D**			
	0.1 m	**0.5 m**	**1 m**	**1 m 30°**	**0.1 m**	**0.5 m**	**1 m**	**1 m 30°**	**0.1 m**	**0.5 m**	**1 m**	**1 m 30°**
F^3Loc (CVPR 2024)	4.7	28.6	36.6	35.1	4.3	26.7	33.7	32.3	7.5	37.2	43.5	42.1
3DP (ACM MM 2025)	5.3	33.2	39.8	38.4	9.4	37.4	43.1	41.5	-	-	-	
RSK (AAAI 2026)	8.3	38.5	45.3	43.6	8.7	36.4	42.3	40.4	-	-	-	
DisCo-FLoc (arXiv 2026)	13.1	50.9	56.7	55.4	12.4	48.0	55.6	54.0	12.1	56.5	63.3	62.1
UnLoc (ICLR 2026)	14.6	47.2	50.2	49	13.0	41.7	45.4	44.0	12.3	45.3	48.8	47.7
AVLoc	**35.2**	**72.4**	**74.7**	**74**	**37.0**	**72.0**	**73.9**	**73.0**	**36.7**	**73.7**	**75.6**	**74.8**

Table 1: Model's performance in single view setting.

4.1 MAIN RESULTS

Table 1 reports single-view localization on Gibson and HM3D under the common evaluation protocol (ideal/clean settings). AVLoc improves recall@1m from 56.7% to 74.7% on Gibson(f), from 55.6% to 73.9% on Gibson(g), and from 63.3% to 75.6% on HM3D over the strongest baseline in each case. The margin widens at the tightest threshold: recall@0.1m rises from 14.6% to 35.2% on Gibson(f) and from 12.3% to 36.7% on HM3D. Coarse thresholds are satisfied by any pose in the right part of a room, whereas 0.1 m requires the predicted rays to pin down the position within it, which is what the 25 directions outside the camera's field of view supply.

Part of this gap could in principle come from architecture rather than from audio. The expert ablation in Table 2 shows that our vision-only expert, which uses the same encoder and ray-token design but no audio, reaches 48.2% recall@1m, above RSK (45.3%) but below Disco-FLoc (56.7%). The architecture alone therefore does not account for the result. The gain follows from adding the audio modality and from how the two are combined. For multi-floor evaluation see Sec. H.



(a) Varying audio degradation with fixed vision

(b) Varying vision degradation with fixed audio

Figure 4: **Robustness to modality degradation on Gibson(f).** Recall@1m as one modality is progressively degraded, while the other is fixed at different noise levels. (a) Audio SNR is varied under fixed vision severity. (b) Vision severity is varied under fixed audio SNR. Colors denote models, and line styles and markers denote the fixed modality conditions.

Fig. 3 shows the effect qualitatively. Where UnLoc Wueest et al. (2026) spreads probability over several candidates in repetitive layouts, AVLoc concentrates it to fewer positions. In the first real-world example, a tree occludes a corner, so the RGB observation resembles a flat wall and UnLoc places the pose along similar walls elsewhere. The echo response is unaffected by the occlusion and recovers the corner geometry, leading to accurate localization.

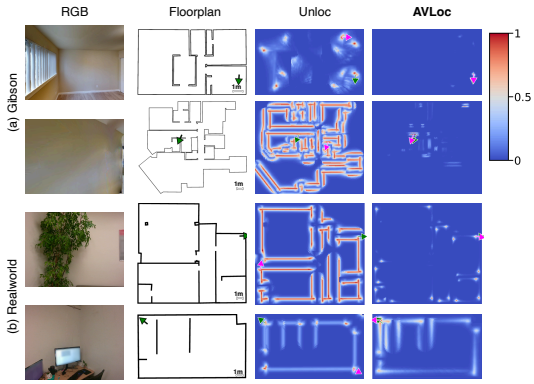


Figure 3: Qualitative results in simulation (top two rows) and real-world environments (bottom two rows). Green and magenta arrows denote ground-truth and predicted poses, respectively.

4.2 Robustness to Modality Degradation

In this experiment, we test what happens in non-ideal conditions when the input modalities are corrupted at test time, which is a common real-world scenario. We evaluate on the Gibson(f) test set over a grid of audio SNRs and visual corruption severities, reporting recall@1m. All models, including the baselines, are retrained with the noise-augmentation protocol used for AVLoc, so the comparison isolates the sensing design rather than exposure to corruption during training, because of this retraining, UnLoc's clean-condition number here (51.4%) differs from Table 1. Test-time ambient noise is drawn from 10 ESC-50 categories disjoint from those seen in training, so the acoustic corruptions are not memorized.

Fig. 4 reports two slices of the grid for AVLoc, UnLoc, and the audio-only and vision-only baselines. AVLoc exceeds UnLoc in every condition, by between 9.8 and 45.5 points. The gap grows with severity rather than staying constant: at maximum visual degradation with clean audio, AVLoc retains 53.0% recall while UnLoc falls to 7.5% (Fig. 4b), and at maximum audio degradation with clean vision AVLoc leads by 9.8 points (Fig. 4a). The asymmetry is expected: visual corruption removes evidence only inside the field of view, whereas acoustic interference affects all 36 directions, so the two modalities are not interchangeable as fallbacks. With both modalities at maximum degradation AVLoc still reaches 30.1%, indicating that the two corruption processes do not disable the same directions at the same time. See Sec. C for the modality degradation protocol.

4.3 Sequential Localization

Single-view localization is the harder setting, but sequential evidence is an option a deployed agent actually has, so we also evaluate on trajectories. We follow the protocol, metrics, and trajectory lengths of UnLoc Wueest et al. (2026) on Gibson(t), measuring pose recall within 1 m over the final 10 observations. For the noisy condition, UnLoc is retrained with the same noise augmentation as AVLoc, and both models receive identical vision-degraded frames. AVLoc additionally receives randomly degraded audio on those frames, so it experiences degradation in both modalities. Specif-

ically, for both models, we add noisy vision inputs at each timestep with a probability of 0.5 at a severity of 0.5.

Additionally, AVLoc receives audio noise at SNR of 0 dB with probability 0.5, but only on frames where vision is also degraded, so noise in the two modalities always co-occurs, and both models always observe the same set of degraded frames.

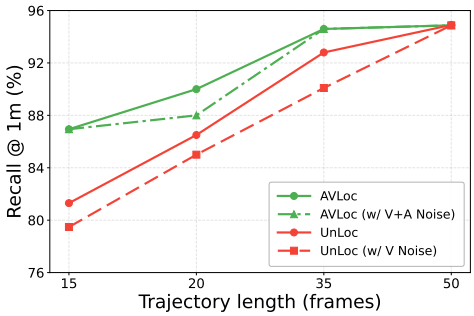


Figure 5: Performance across trajectory lengths on Gibson(t). Solid and dashed lines denote clean and noisy conditions, respectively.

Fig. 5 shows AVLoc leading at most trajectory lengths, most clearly at 15 and 20 observations, where the agent has accumulated the least evidence. The clean-to-noisy gap is also smaller for AVLoc, so the robustness observed on single views is not washed out by temporal filtering. After observing 50 frames, both models reach a similar performance, and we don't see further improvement, which is expected because the agent has already seen a large fraction of the scene with 50% of the observations under ideal (clean) settings.

4.4 Real-World Generalization

No real-world audio-visual benchmark exists. To show evidence of feasibility, we use a RealSense camera and a 3Dio binaural microphone and collected 44 samples from seven distinct floor maps. The floor maps cover a total of 724.9 m^2, with an average area of 103.6 ± 75.7 m^2 per floor map. On this dataset, our AVLoc achieves a recall@1m of 15.9% against 6.8% for UnLoc in the single-view setting. Our model maintains its advantage over the best competitor, however both numbers are below the simulated results. The sim-to-real gap remains a challenge, but the qualitative results in Fig. 3 (rows 3 and 4) show that AVLoc is able to reduce ambiguity in real-world settings as well.

4.5 Ablation Studies

Model	0.1 m	0.5 m	1 m	1 m 30°
AVLoc	**35.2**	**72.4**	**74.7**	**74**
Expert Ablations				
E^{AV}	32.5	70	72.3	71.3
E^V & E^A	30	67.1	70	69.1
E^V	14.3	44.1	48.2	46.6
E^A	13.0	42.3	46.7	45.4
Router Ablations				
Global Modality Router	33.9	71.2	73.5	72.5
Direct Confidence Router	31.0	64.5	67.8	66.8
Uniform Router	25.6	58.7	62.8	61.5
FOV-Based Hard Router	19.3	57.8	63.3	61.5

Table 2: Ablations of AVLoc on Gibson(f).

Expert Ablation. Table 2 tests whether both the unimodal and the joint experts are needed. Either unimodal expert alone is weak (48.2% for E^V, 46.7% for E^A at 1 m), and routing between the two reaches 70.0%, so most of the gain comes from having two modalities rather than from any single stream. The joint expert alone reaches 72.3%, showing that cross-modal interaction adds beyond what routing between separate predictions provides. The full model reaches 74.7%, so the two are not substitutes: the joint expert contributes interaction the unimodal experts cannot express, and the unimodal experts contribute paths the joint expert cannot preserve. Furthermore, our full model is more robust to modality degradation than the unimodal experts (Fig. 4) or the joint expert (Sec. I, Fig. 8).

Router Ablation. In the lower part of Table 2, we compare our *Adaptive Expert Router* (recall@1m: 74.7%), which learns an independent uncertainty-conditioned fusion function per ray to the following alternatives. The *Global Modality Router* uses a single MLP to produce modality weights shared across all rays, testing whether routing must be resolved per direction: performance drops to 73.5%. The *Direct Confidence Router* assigns each expert a confidence which is inversely proportional to the uncertainty scale. Then, weights are obtained by normalizing the experts' confidence for each ray: recall drops to 67.8% indicating that raw confidence alone is not sufficient to capture expert reliability.

The *Uniform Router* assigns equal weights to all experts (62.8%) and the *FOV-Based Hard Router* (63.3%) selects vision inside the field of view and audio outside it, test whether input-conditioned weighting is needed at all. Both lose more than 11 points, the largest effect in the table. Adapting to the input therefore matters most, learning the uncertainty-to-weight mapping second, and per-ray

resolution is third on clean data. However, our analysis in Sec. J shows that the per-ray design is more important under degradation, where the two modalities are not interchangeable.

4.6 Adaptive Expert Router Analysis

In Fig. 6, we examine routing weights over a controlled grid of audio (clean, 0 dB, -15 dB) and visual (clean, 0.6, 1.0 severity) conditions on a fixed set of 500 samples; Sec. J gives a fuller analysis.

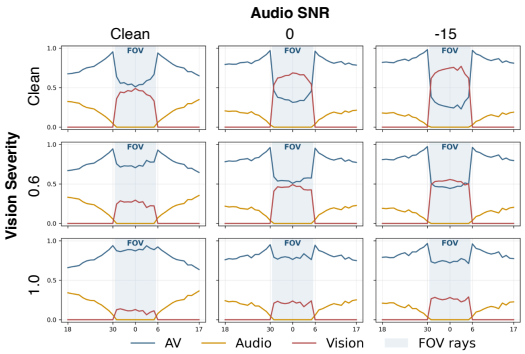


Figure 6: Ray-wise modality weights under different noise settings.

Adaptation to audio noise. Out of FOV, decreasing audio SNR lowers the audio-only weight and raises the audio-visual weight. Inside it, added audio noise shifts weight from the audio-visual expert toward the vision-only expert. In both regions the router moves toward the expert least exposed to the corruption that was applied.

Adaptation to visual noise. At fixed audio quality, increasing visual corruption lowers the vision-only weight within the FOV and shifts it to the audio-visual expert, while weights outside the FOV are largely unchanged. This is the behavior the design predicts, since visual corruption cannot affect directions the camera never observed, and it is what distinguishes uncertainty-driven routing from a global quality estimate.

Directional structure at fixed noise. Weights also vary across rays under fixed conditions. Outside the FOV the audio weight is lowest near the boundaries and rises toward rear-facing rays, with the audio-visual weight following the complementary pattern. Inside FoV, the vision expert receives more weight near the center than at the edges when vision is not severely degraded. Reliability is therefore direction-dependent even when input quality is held constant, which is what per-ray routing is for. Fig. 10 and Sec. J provide further analysis of per-ray routing in relation to scene structure as well as input quality. For example, we notice that for rays just outside the FOV, an audio-visual expert gets more weight where vision context supports extrapolating out of the FOV region, such as a wall that continues past the FOV boundary.

4.7 Failure Cases

Fig. 7 shows two failure modes. In the first, AVLoc places the pose in the correct corridor-like region but cannot resolve two door openings behind the agent, an error attributable to the angular resolution of the echo response, which separates nearby openings less sharply than an image would. In the second, several locations share nearly identical local geometry, and geometry alone does not disambiguate them. Resolving these would require global layout context or semantics.

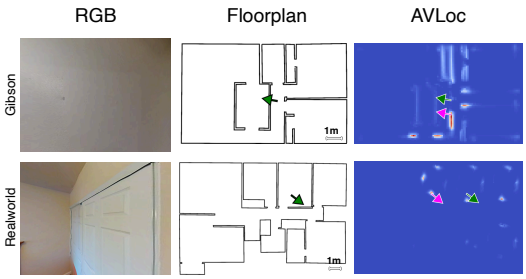


Figure 7: Failure cases caused by low angular resolution of audio (top) and repetitive room layouts (bottom).

5 Conclusion

We presented AVLoc, which localizes an agent against a floorplan from vision and binaural echoes together. Because the two modalities fail under different conditions, and because AVLoc keeps unimodal prediction paths and routes between experts by their own predicted uncertainty, localization degrades gradually where vision-only methods collapse: under maximum visual corruption AVLoc retains 53.0% recall@1m against 7.5% for the strongest vision-only baseline. The same design improves clean-condition accuracy, since echoes constrain the 25 directions the camera cannot see. Analysis of the router shows weights shifting away from each expert as its input degrades, and varying with direction even at fixed input quality.

REFERENCES

H Nazim Bicer, Cagdas Tuna, Andreas Walther, and Emanuël AP Habets. Evaluation of data-driven room geometry inference methods using a smart speaker prototype. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 274–278. IEEE, 2024.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. Owl: Geometry-aware spatial reasoning for audio large language models. *arXiv preprint arXiv:2509.26140*, 2025.

Amandine Brunetto, Sascha Hornauer, X Yu Stella, and Fabien Moutarde. The audio-visual batvision dataset for research on sight and sound. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8. IEEE, 2023.

Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, and Jianxin Wang. Perspective from a higher dimension: Can 3d geometric priors help visual floorplan localization? In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7181–7190, 2025.

Bolei Chen, Shengsheng Yan, Yongzheng Cui, Jiaxu Kang, Ping Zhong, and Jianxin Wang. Perspective from a broader context: Can room style knowledge help visual floorplan localization? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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 *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1581–1587. IEEE, 2020.

Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pp. 658–676. Springer, 2020.

Yuval Grader and Hadar Averbuch-Elor. Supercharging floorplan localization with semantic rays. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Henry Howard-Jenkins and Victor Adrian Prisacariu. Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments. In *European Conference on Computer Vision*. Springer, 2022.

Henry Howard-Jenkins, Jose-Raul Ruiz-Sarmiento, and Victor Adrian Prisacariu. Lalaloc: Latent layout localisation in dynamic, unvisited environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto (eds.), *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 169–178. PMLR, 29–31 Oct 2018. URL `https://proceedings.mlr.press/v87/karkus18a.html`.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Chen Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 275–285, 2021.

Shiyong Meng, Tao Zou, Bolei Chen, Chaoxu Mu, and Jianxin Wang. Disco-floc: Using dual-level visual-geometric contrasts to disambiguate depth-aware visual floorplan localization. *arXiv preprint arXiv:2601.01822*, 2026.

Branislav Micusik and Horst Wildenauer. Descriptor free visual indoor localization with line segments. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Zhixiang Min, Naji Khosravan, Zachary Bessinger, Manjunath Narayana, Sing Bing Kang, Enrique Dunn, and Ivaylo Boyadzhiev. Laser: Latent space rendering for 2d visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, 2015.

Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1183–1192, 2021.

Mirco Rossi, Julia Seiter, Oliver Amft, Seraina Buchmeier, and Gerhard Tröster. Roomsense: An indoor positioning system for smartphones using active sound probing. In *Proceedings of the 4th Augmented Human International Conference*, pp. 89–95, 2013.

S Sakshi, Vaibhavi Lokegaonkar, Neil Zhang, Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. Spur: A plug-and-play framework for integrating spatial audio understanding and reasoning into large audio-language models. *arXiv preprint arXiv:2511.06606*, 2025.

Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Stephen P. Tarzia, Peter A. Dinda, Robert P. Dick, and Gokhan Memik. Indoor localization without infrastructure using the acoustic background spectrum. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pp. 155–168, 2011.

Matthias Wueest, Francis Engelmann, Ondrej Miksik, Marc Pollefeys, and Daniel Barath. Unloc: Leveraging depth uncertainties for floorplan localization. In *Proc. ICLR*, 2026.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9068–9079, 2018.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4927–4936, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Inmo Yeon, Iljoo Jeong, Seungchul Lee, and Jung-Woo Choi. Echoscan: Scanning complex room geometries via acoustic echoes. 32:4768–4782, October 2024. ISSN 2329-9290. doi: 10.1109/TASLP.2024.3485516. URL https://doi.org/10.1109/TASLP.2024.3485516.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning*, 2023.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*, 2024.

In this supplementary material, we provide more details about our approach, ablations, and additional analysis.

- Supplementary video A
- Training Details B
- Input Preprocessing & Augmentation C
- Dataset Details D
- Model Capacity E
- Sound Source Ablation F
- Main Components of AVLoc G
- Multi-Floorplan Localization H
- Contribution of the E^{AV} Expert I
- Qualitative Results J
- Noise Visualization K

A Supplementary Video

We make a supplementary video to help understand the AVLoc and present qualitative analysis. Initially, the video demonstrates the qualitative results, comparing AVLoc to the SOTA model (UnLoc) Wueest et al. (2026) on Gibson and real-world data samples with elaborations. Then, the video shows AVLoc failure cases, followed by an adaptive routing analysis using a specific data sample under varying levels of degradation.

B Training Details

The model is trained with Laplace negative log-likelihood loss for up to 100 epochs, with early stopping based on validation L1 ray error, using a patience of 20 epochs. We select the checkpoint with the lowest validation L1 ray error. Optimization uses AdamW with learning rate 5×10^{-4} with a cosine annealing schedule decaying to 10^{-6}.

C Input Preprocessing & Augmentation

Visual Data RGB images are normalized with ImageNet mean and standard deviation. As *vision augmentation*, each RGB observation is degraded with probability 0.5. Conditioned on degradation, we independently apply low-light brightness changes ($p = 0.6$), Gaussian read noise ($p = 0.6$), random blackout regions ($p = 0.3$), and directional motion blur ($p = 0.3$). For each chosen transformation, a severity value is sampled uniformly and independently from 0 to 1, where 0 leaves the image intact and 1 produces maximum corruption. Values in between smoothly interpolate between these endpoints, allowing training to span the entire spectrum from slight to heavy degradation.

Figure 12 shows each corruption type at four discrete severity levels $\lambda \in \{0.25, 0.50, 0.75, 1.00\}$ for illustration purposes.

Audio Data We sample audio at 48 kHz and obtain the echo response by emitting a 50 ms logarithmic sine sweep from 20 Hz to 20 kHz. The first 100 ms of the recorded response is used as input, covering the full sweep and reflections arriving up to 50 ms after it ends. Then, we compute the STFT with $n_{\text{fft}} = 512$, window length 64, and hop length 16. Audio is represented as a four-channel spectrogram containing left/right log-magnitude and sine/cosine inter-channel phase-difference channels.

As *audio augmentation*, we apply time-frequency masking to the four-channel binaural spectrogram with probability 0.7. Specifically, we use two time masks and two frequency masks, each with a masking parameter of 12. Independently, ambient noise is injected with probability 0.5 during training. We split the ESC-50 Piczak (2015) categories into disjoint training and evaluation sets,

Dataset	Split	Scenes	Floors	DPs	Avg DP/Floor ($\pm$ std)
Gibson(f)	All	78	94	20,305	216.0 $\pm$ 122.0
	Train	61	77	16,724	217.2 $\pm$ 127.0
	Val	8	8	1,574	196.8 $\pm$ 88.6
	Test	9	9	2,007	223.0 $\pm$ 100.2
Gibson(g)	All	78	94	40,610	432.0 $\pm$ 244.0
	Train	61	77	33,448	434.4 $\pm$ 254.0
	Val	8	8	3,148	393.5 $\pm$ 177.3
	Test	9	9	4,014	446.0 $\pm$ 200.3
HM3D	All	155	278	119,818	431.0 $\pm$ 316.4
	Train	123	213	94,927	445.7 $\pm$ 330.1
	Val	32	65	24,891	382.9 $\pm$ 260.8
	Val*	15	35	13,893	396.9 $\pm$ 272.9
	Test*	17	30	10,998	366.6 $\pm$ 245.0

Table 3: Dataset statistics. *DP* = data point.

ensuring the model encounters unseen ambient noise categories during training. During training, we sample either Gaussian noise or a recording from the training categories and scaling it to an SNR selected uniformly from

$$\{-15, -13, \ldots, 9, 10, 20, 30, 60\}\ \mathrm{dB}.$$

This exposes the acoustic encoder to both severe interference and nearly clean observations.

Figure 13 illustrates the effect of five representative held-out ESC-50 categories, *laughing*, *dog barking*, *rain*, *siren*, and *clock alarm*, alongside Gaussian noise, each at four SNR levels used during evaluation: 10, 0, -10, and -15 dB.

Audio-visual horizontal flipping augmentation With probability 0.5, we jointly mirror the complete audio-visual sample. The RGB image is flipped horizontally, while the left and right log-magnitude spectrogram channels are exchanged. For the interaural phase-difference representation, the sine component is negated, and the cosine component is retained. We also mirror the ordering of the target rays. Applying these operations jointly preserves the geometric correspondence among the image, binaural observation, and the localization target.

D Dataset Details

D.1 Dataset Statistics

Table 3 summarizes the scene, floor, and data-point statistics for each dataset split. Gibson(f) and Gibson(g) share the same 78 scenes and 94 floors but use different sampling protocols; Gibson(g) contains twice as many data points as Gibson(f). HM3D provides a larger and more diverse benchmark, comprising 155 scenes and 278 floors. For HM3D, Val* and Test* denote the disjoint subsets of the original validation scenes used for model selection and final evaluation, respectively. The last column reports the mean number of data points per floor and its standard deviation, highlighting the variation in floor size and sampling density.

D.2 Audio-Visual Data Generation

We construct audio-visual observations for Gibson(f), Gibson(g), and HM3D using SoundSpaces 2 (SS2), the state-of-the-art audio simulator. As the result, each data point contains a perspective RGB image, a binaural room impulse response, the camera pose, and a floorplan-derived ray supervision.

Gibson(f) and Gibson(g). We adopt the floorplans, poses, and scene-disjoint splits provided by the F^{3}Loc Chen et al. (2024) Gibson Floorplan Localization Dataset.

At every pose, we place a binaural listener and an acoustic source at the same position and height through SS2 Chen et al. (2022) on a corresponding Gibson scene. The resulting two-channel echo response is simulated at $48\,\mathrm{kHz}$, with indirect sound propagation enabled. This preserves binaural directional and reflection cues. However, SS2 does not support all Gibson scenes, so we exclude 24 unsupported scenes (23 train / 1 val), yielding 20,305 audio-visual samples for Gibson(f) and 40,610 for Gibson(g) across 94 scenes. All 9 test scenes are fully supported, ensuring a fair comparison with prior works on the complete test set.

HM3D. For HM3D, we generate candidate positions by sampling from a $0.35\,\mathrm{m}$ grid and independently perturbing them by up to $0.05\,\mathrm{m}$ along both horizontal axes. Camera orientations are sampled uniformly from $[0, 2\pi)$. We use the same camera configuration and binaural audio collection protocol as for Gibson.

Ray supervision. For every Gibson and HM3D observation, we ray-cast the corresponding binary floorplan to generate $K = 36$ omnidirectional ray distances at 10° intervals.

D.3 Realworld Dataset

For the realworld dataset, we collect 44 samples across seven distinct floor maps. The floor maps cover a total of $724.88\,\mathrm{m}^2$, with an average area of $103.55 \pm 75.65\,\mathrm{m}^2$ per floor map.

E Model Capacity

Table 4 compares the parameter sizes of AVLoc with those of the baselines. UnLoc is the largest model at 336.56M parameters, while its trainable parameters are only 1.24M, since it relies on a pretrained ViT-L as a frozen visual backbone. AVLoc contains 40.19M parameters overall, roughly $8\times$ fewer compared to UnLoc. The three experts are small modules made up only of a transformer and MLP layers. This analysis supports that AVLoc's improvements stem from the added modality and the fusion strategy, not from the increased model capacity.

Model	**# Params**
UnLoc	336.56M
DisCo-FLoc	56.83M
AVLoc	40.19M
E^V	0.42M
E^A	0.82M
E^{AV}	0.82M

Table 4: Number of parameters. Expert rows are included in the AVLoc total.

F Sound Source Ablation

In this section, we explore whether AVLoc is compatible with other audio sources beyond the chirp-sweep signal. We keep the model architecture fixed and only replace the chirp sound with speech and music. For both speech and music setups, we sample five representative clips from LibriSpeech Panayotov et al. (2015) and GTZAN Sturm (2013), respectively, to replace the chirp signal during training and evaluation. As shown in Table 5, replacing chirp with speech decreases performance, with Recall@1m dropping from 74.7% to 60.9% and to 61.6% for music. We believe non-chirp sounds, which are less deterministic and span a narrower frequency range than chirp signals, make it harder for the model to capture distance signals from reflected sound. Importantly, AVLoc trained on speech or music still outperforms the previous state of the art on Gibson(f). This suggests that AVLoc can also be adapted to other sound sources, such as speech or music.

G Main Components of AVLoc

Visual encoder. The visual encoder f^V uses an ImageNet-pretrained ResNet-50. We extract its final convolutional feature map, project each spatial feature to the shared token dimension

Source	0.1m↑	0.5m↑	1m↑	1m30°↑
AVLoc	**35.2**	**72.4**	**74.7**	**74**
AVLoc-Speech	20.1	56.0	60.9	59.0
AVLoc-Music	21.5	57.5	61.6	60.0

Table 5: Sound-source ablation on Gibson(f).

$d = 128$, and add learned positional and modality embeddings. The resulting visual tokens $\mathcal{V} = \{\mathbf{v}_i\}$ are refined using two intra-modality self-attention layers before being passed to the audio-visual and vision-only experts.

Audio encoder. The audio encoder f^A is a ResNet-18 adapted to accept the four-channel binaural spectrogram. The input channels contain the left and right log-magnitude spectrograms and the sine and cosine of the interaural phase difference. We remove the global pooling and classification layers and retain the spatial time–frequency feature map. These features are projected to $d = 128$, augmented with time–frequency positional and audio-modality embeddings, and refined by two intra-modality self-attention layers to obtain $\mathcal{A} = \{\mathbf{a}_j\}$.

Expert architectures. All experts use Transformer encoders with eight-head self-attention and maintain separate learnable ray tokens. The *audio-visual expert* concatenates its $K = 36$ ray tokens with $\mathcal{V}$ and $\mathcal{A}$, and processes the joint sequence using four Transformer encoder layers. This branch predicts an omnidirectional ray distribution by exploiting interactions between the two modalities.

The *audio-only expert* concatenates $K_A = 36$ ray tokens with $\mathcal{A}$ and applies a four-layer Transformer encoder. It therefore predicts distances and uncertainty over the complete 360° range. The *vision-only expert* instead combines $K_V = 11$ visually seeded ray tokens with $\mathcal{V}$ and applies a two-layer Transformer encoder. Its predictions are restricted to the camera field of view.

In every branch, only the resulting ray tokens are forwarded to the prediction heads to estimate $(\hat{r}_k^e, \hat{b}_k^e)$. The three expert outputs are then combined using the uncertainty-guided routing mechanism described in Section 3.3.

H Multi-Floorplan Localization

Previous work focused on single-floor settings, but many buildings have multiple floors which creates additional ambiguity for models but captures real-world scenarios not covered by single-floor benchmarks. We construct HM3D-Multi from the HM3D test set: 5,839 samples across 9 buildings totaling 22 floors, with the candidate pose set spanning all floors of a building rather than a known one. Table 6 shows AVLoc achieving a recall@0.1m of 35.7% against 11.5% of UnLoc, and our model is able to identify the correct floor for 83.8% of samples against 73.4% for Disco-FLoc. This improvement indicates that the ability of AVLoc to utilize information from the rays behind the agent is very useful in this challenging setting as well.

	Floor Acc.↑	**0.1 m↑**	**1 m↑**	**1 m 30° ↑**
F^3Loc	62.5	6.7	34.7	33.8
Disco-FLoc	73.4	11.3	54.1	53.4
Unloc	64.8	11.5	38.6	37.8
AVLoc	83.8	35.7	67.2	66.4

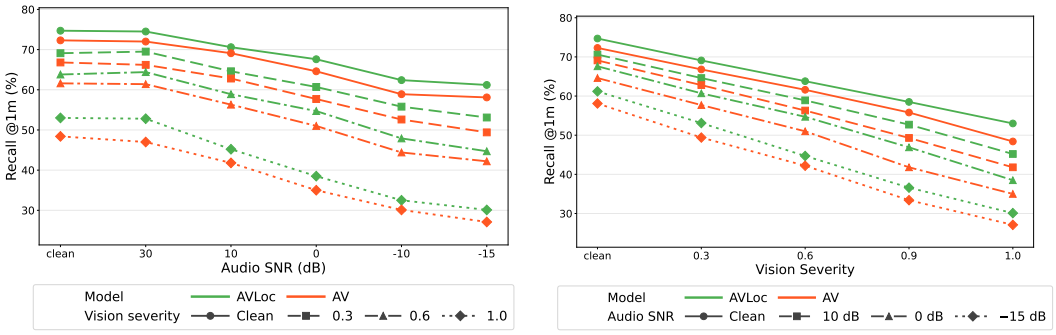
Table 6: Performance on HM3D-Multi.

I Contribution of the E^{AV} Expert

Fig. 8 examines what the E^{AV} expert contribute, by comparing AVLoc against E^{AV} alone over the same degradation grid used in the main paper. E^{AV} shows good robustness, as it learns to dynamically adapt to modality degradation by encountering diverse degradation during training. Nevertheless, AVLoc outperforms it consistently across all audio SNRs and visual corruption severities. This gap emphasizes that the unimodal experts capture evidence that the joint expert does not fully exploit, since their predictions remain unaffected when the other stream is corrupted. Routing to the unimodal experts thus provides a consistent gain on top of a strong joint expert, rather than compensating for a weak one.

J Qualitative results

Adaptive Expert Router Analysis Fig. 9 extends the main-paper analysis with a finer-grained grid of visual severity and audio SNR. Weights averaged over the same 500 samples reveal how the router progressively adjusts each expert's contribution as modality quality degrades.



(a) Varying audio degradation with fixed vision

(b) Varying vision degradation with fixed audio

Figure 8: **Robustness to modality degradation on Gibson(f).** Recall@1m as one modality is progressively degraded, while the other is fixed at different noise levels. (a) Audio SNR is varied under fixed vision severity. (b) Vision severity is varied under fixed audio SNR. Colors denote models, and line styles and markers denote the fixed modality conditions.

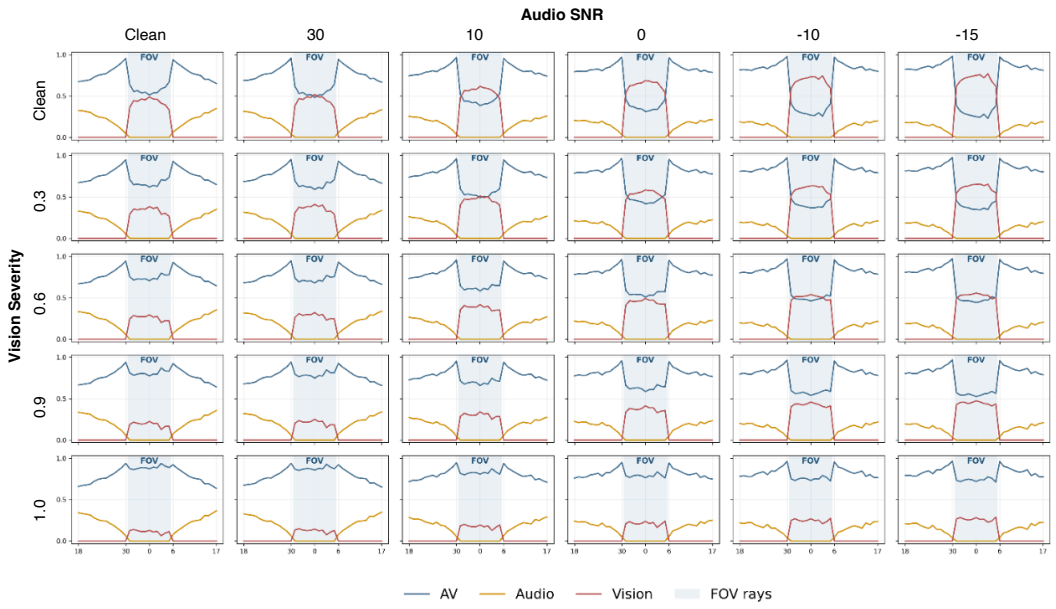


Figure 9: Ray-wise Modality weights

Beyond this aggregate analysis, Fig. 10 provides a sample-level view of how the router responds to input quality and ray direction. The vision expert receives higher weights around distinctive structures such as corners, but its contribution vanishes under severe visual degradation, while the audio weight decreases with increasing acoustic noise. Outside the FOV, the audio-visual expert is favored when visible context supports extrapolation of nearby, unseen geometry, such as a wall extending beyond the FOV boundary. Thus, the router adapts each expert's contribution to both the reliability of the modality and directional geometric evidence.

Comparison with state-of-the-art model In this section, we present further qualitative results. Fig. 11 provides a qualitative comparison of AVLoc and UnLoc on Gibson and realworld data examples. Overall, AVLoc tends to produce more deterministic predictions, concentrating higher-probability volume on a smaller set of candidate locations than UnLoc does. Moreover, even when AVLoc does not predict correctly due to floorplan ambiguity, it still assigns a substantial probability to the Ground Truth location, as illustrated in the last row of the figure.

K NOISE VISUALIZATION

In this section, we visualize the types of noise across varying severities to help clarify which noise augmentations are used during training and analysis.

Fig. 12 shows different types of vision degradation at varying severities. While vision contains most information at mild vision noise, it loses most information at the most extreme vision degradation.

Fig. 13 shows different types of audio noises across different SNRs. Similar to vision degradation, audio also retains most information at low SNRs but loses much signal at high degradation. However, unlike vision degradation, audio degradation often doesn't remove all of the input signal, depending on the frequency coverage of the ambient noise source.

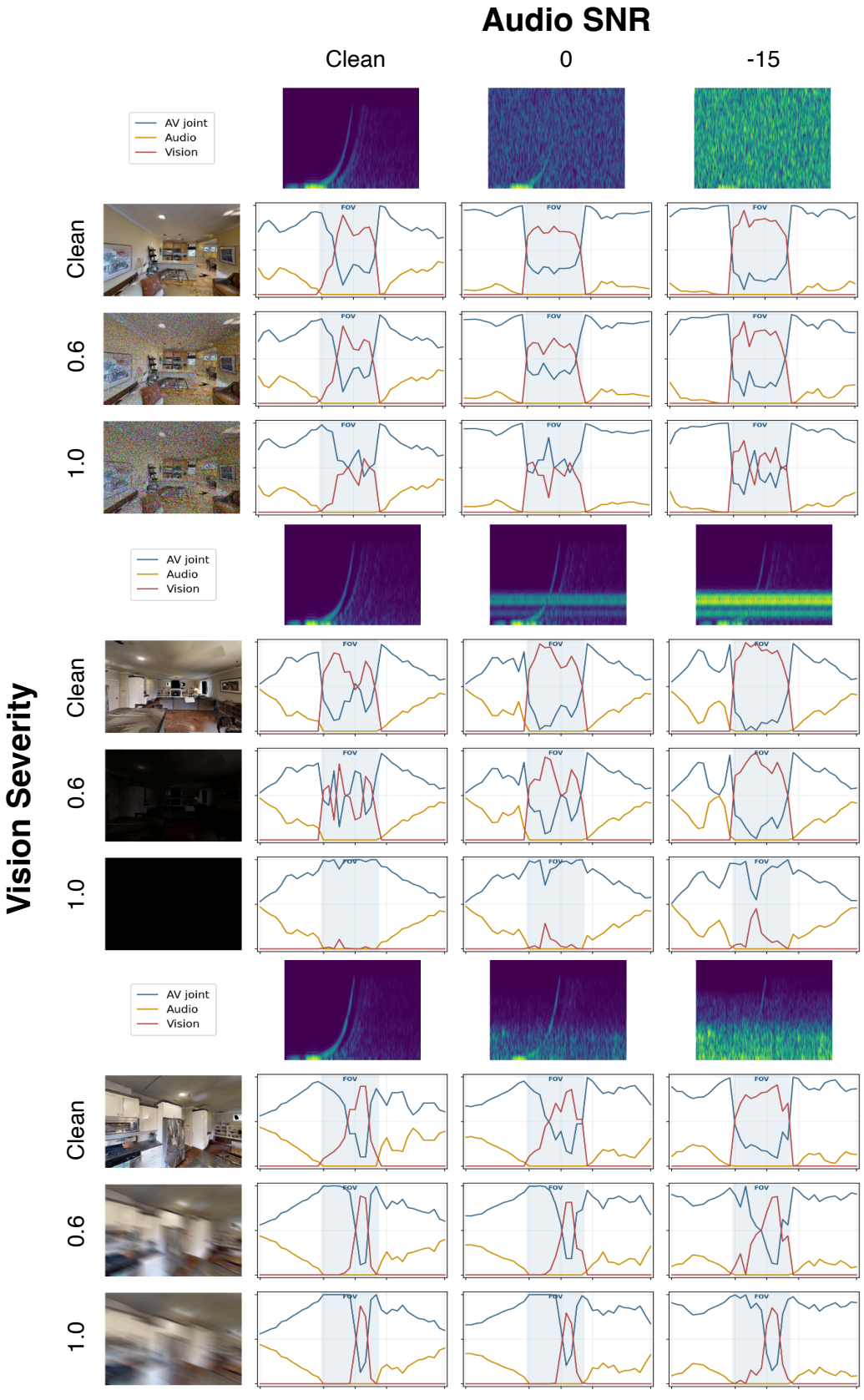


Figure 10: Ray-wise Modality weights

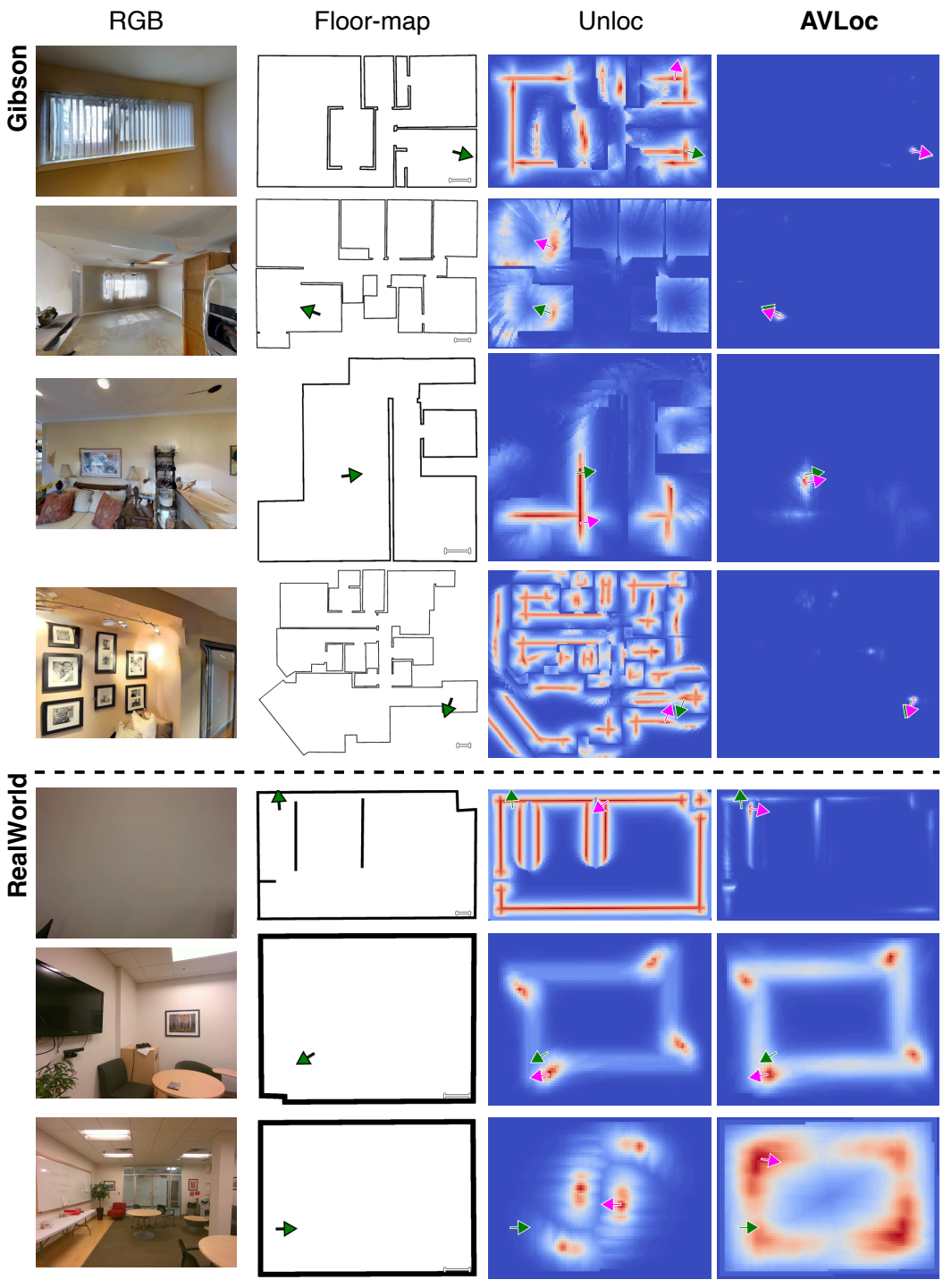


Figure 11: Qualitative results in real-world and Gibson observations

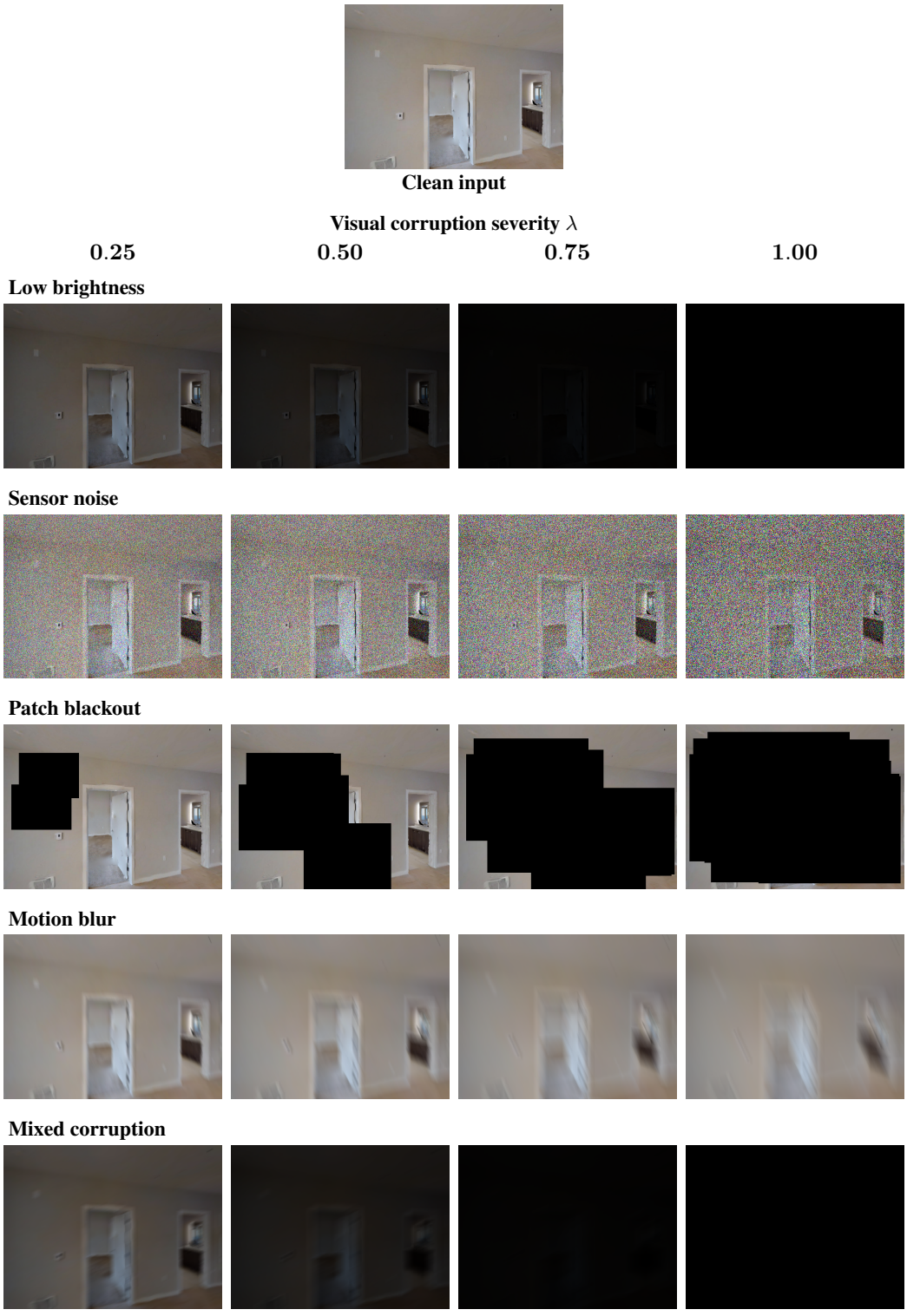


Figure 12: Examples of visual corruptions used in our robustness evaluation. Each row shows one corruption type at increasing severity λ.

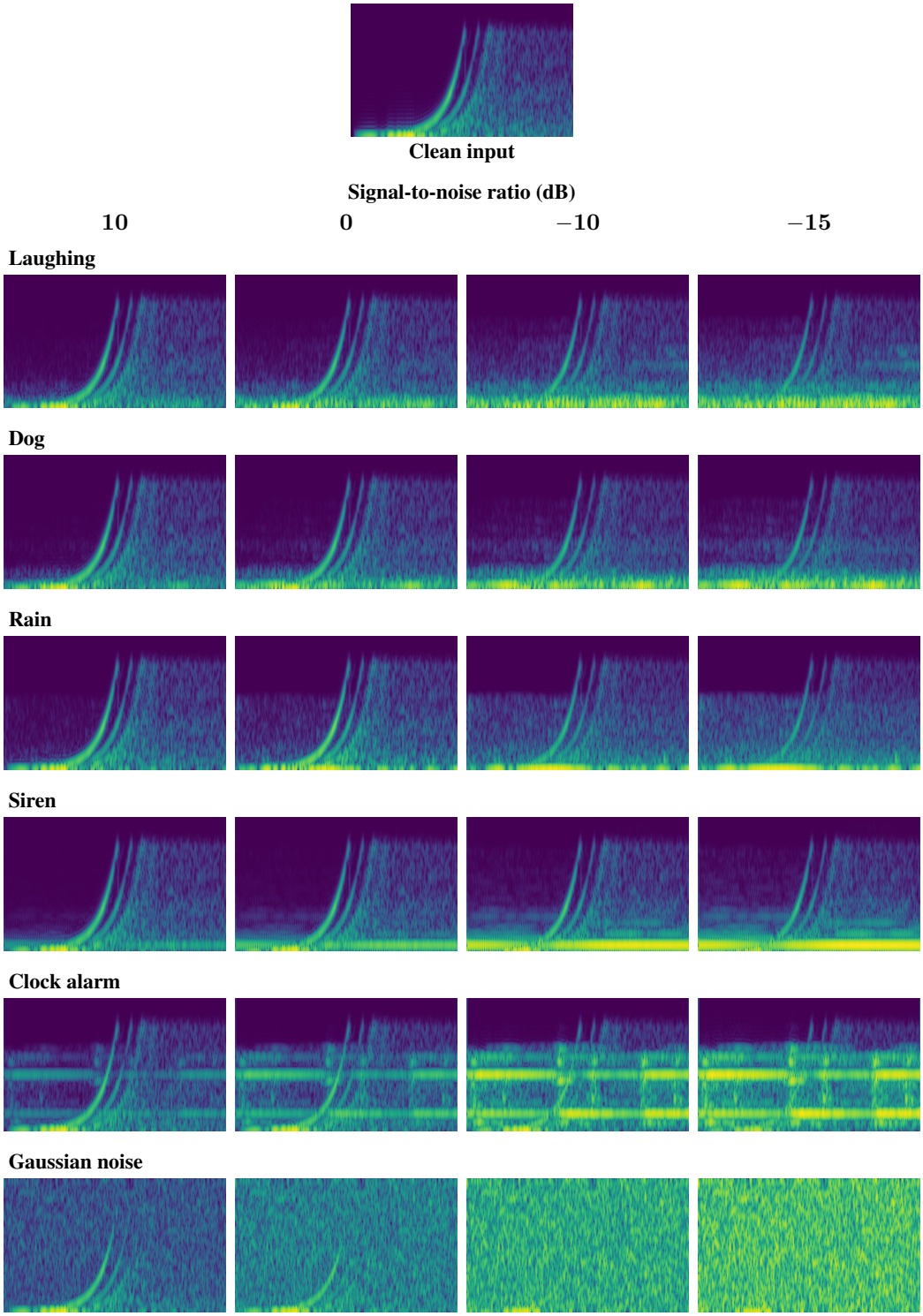


Figure 13: Examples of audio corruptions used in our robustness evaluation. The clean input is shown at the top. Each row presents one noise type at decreasing signal-to-noise ratios, where lower SNR indicates stronger corruption.