VISUAL GENERATIVE FOUNDATION MODELS ARE REUSABLE PRIORS FOR NON-VISUAL SENSORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual foundation models have acquired transferable knowledge of scene geometry, spatial structure, and semantics from internet-scale imagery, yet sensors that remain informative where cameras fail, such as radio-frequency (RF) sensors, lack a comparable data ecosystem. Existing approaches either pretrain a separate foundation model for each sensor or distill one task at a time from a visual teacher, requiring substantial new effort for every new sensor or task. We ask whether a sensor can instead access the knowledge of a pretrained visual foundation model through a lightweight interface. In this paper, we present RFception, which maps RF measurements into the latent space of a pretrained visual generative perception model. Because the pretrained model selects tasks through text prompts, a single RF representation supports multiple geometric and semantic perception tasks without task-specific RF backbones or prediction heads. Across six held-out buildings, RFception outperforms task-specific RF models in depth estimation and, without task-specific RF supervision, performs surface-normal estimation, human pose estimation, and category-conditioned instance segmentation zero-shot. Its visual outputs can further be consumed by off-the-shelf vision-language models without fine-tuning, extending RF sensing to language-level reasoning. These results suggest that the knowledge captured by visual foundation models is not inherently tied to images, and that new sensing modalities can inherit this knowledge by learning an interface to the model rather than building a foundation model of their own.

1 INTRODUCTION

Physical intelligence requires machines to perceive and reason about the world through sensors, and real systems rarely rely on cameras alone. LiDAR and radio-frequency (RF) sensors such as radar remain informative in darkness, smoke, glare, and under occlusions that degrade or defeat vision. Yet the recent foundation-model revolution in perception has been overwhelmingly vision-centric. Internet-scale image collections have enabled visual models to acquire remarkably transferable knowledge of object semantics, scene geometry, spatial structure, and visual correspondence (Radford et al., 2021; Kirillov et al., 2023; Oquab et al., 2023; Yang et al., 2024a). Much of this knowledge concerns the physical world itself rather than the visual modality through which it was acquired. Non-visual sensors lack a comparable data ecosystem. Their measurements are sensor-specific, expensive to collect and annotate at scale, and vary substantially across hardware and deployment settings. Perception models for these sensors are therefore still largely trained for one modality and one task at a time, leaving a widening capability gap between visual and non-visual perception.

Two strategies are commonly used to close this gap, yet both face a structural limitation. The first is to build a foundation model directly for each modality, and recent radar work has begun to scale modality-specific pretraining to increasingly large datasets (Hao et al., 2024). Reproducing the data and computational scale of modern visual foundation models for every sensor is nevertheless prohibitively expensive, and every new sensor generation restarts the process. The second is to distill knowledge from a stronger visual or LiDAR model into the target sensor (Zhao et al., 2018a; Bang et al., 2024). Such transfer can be effective, but in most existing formulations the unit of transfer remains a single task, with task-specific supervision, architecture, or prediction heads. Supporting an additional task therefore typically requires new supervision, a new prediction head, and another training run. In both cases, the visual foundation model is treated primarily as a source of answers.

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.

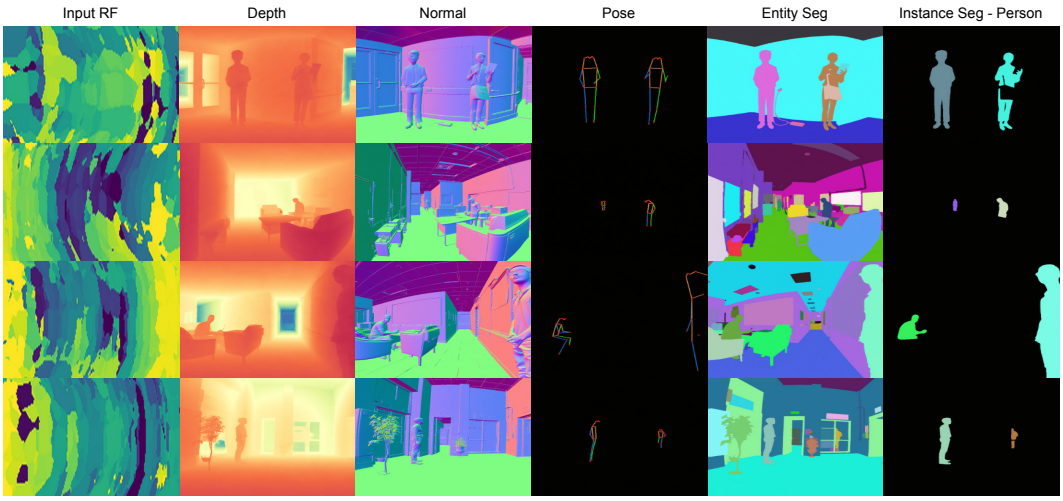


Figure 1: **One RF input, many visual perception outputs.** By changing only the text task prompt to a pretrained visual generative foundation model, RFception turns a single RF heatmap into depth, surface normals, human pose, entity segmentation, and category-conditioned instance segmentation. More qualitative results are provided in Figures 7, 8, 9, and 10.

We instead ask whether a sensor can be given access to the model itself, allowing its capabilities to be reused through a shared interface rather than transferred one task at a time.

We study this question through RF sensing. An RF heatmap records reflected electromagnetic energy shaped by specular reflection, multipath propagation, and limited angular resolution. It shares neither the appearance statistics nor the projection geometry of natural images. Our approach, RFception, learns a lightweight sensor-to-foundation interface, an RF adapter that maps RF measurements into the latent space of a pretrained visual generative perception model (Rombach et al., 2022; Ke et al., 2024; Zhao et al., 2026). The pretrained weights remain frozen while only lightweight low-rank updates are introduced, so training modifies less than 1% of the model parameters.

Simply anchoring the adapter to the visual latent space through numerical regression is not sufficient. RF and light do not resolve the same information about a scene, so a residual persists even when the adapter is well optimized, and what matters is not only its magnitude but where it lies in the model's latent space. We therefore also optimize the interface through the generative model, shaping this residual according to its effect on the generated output. Without this objective, the outputs collapse to flat, empty maps. Because the pretrained model selects tasks through text instructions, the same adapter supports depth, surface normals, human pose, entity segmentation, and category-conditioned instance segmentation without any task-specific RF backbone or prediction head (Figure 1).

We evaluate RFception on indoor RF sensing using a building-level split. On depth estimation, RFception outperforms the task-specific PanoRadar (Lai et al., 2024), reducing the mean absolute error across six held-out buildings from 0.194 m to 0.170 m. Surface-normal estimation, human pose estimation, and category-conditioned instance segmentation receive no RF supervision at any stage and are performed zero-shot by changing only the task prompt. Despite this, RFception reaches 51.2 AR on human pose and 55.2 AP on person instance segmentation, compared with 63.8 and 60.2 for PanoRadar trained with 15K labeled samples per task. Four off-the-shelf VLMs can also consume RFception outputs without fine-tuning, retaining 87.9% of the compositional-reasoning accuracy achieved when the same VLMs are given LiDAR reference depth. Together, these results establish an end-to-end path from raw electromagnetic measurements to language-level understanding.

2 Related Work

2.1 RF Imaging and Sensing

RF signals remain informative in darkness, smoke, and fog, and lower-frequency signals can even penetrate walls, motivating a long line of work on sensing the physical world with WiFi (Pallaprolu

et al., 2022; Song et al., 2024) and radar (Adib et al., 2014; 2015; Prabhakara et al., 2023; Lai et al., 2024). Classical RF imaging recovers scene structure through signal processing and synthesizes a larger effective aperture through antenna motion (Moreira et al., 2013; Yanik & Torlak, 2019). Its resolution, however, remains fundamentally limited by aperture size and bandwidth, leaving the resulting heatmaps coarse, noisy, and strongly affected by specular reflection and multipath.

Learning-based methods have consequently become a primary approach for mapping RF heatmaps to perception outputs, but most are designed around a single downstream task. Human-centric systems estimate pose (Zhao et al., 2018a;b), recognize actions (Li et al., 2019), and reconstruct body meshes (Xue et al., 2021). Scene-level systems recover LiDAR-like geometry for indoor mapping (Lu et al., 2020), dense point clouds (Qian et al., 2020; Prabhakara et al., 2023; Song et al., 2026), and panoramic range images that further support surface normals, segmentation, detection, and SLAM (Lai et al., 2024; 2025). Efforts beyond this per-task paradigm largely follow two directions. One scales RF-specific pretraining (Hao et al., 2024; Huang et al., 2025), but still operates at a much smaller data scale than visual foundation models and must be adapted to new sensor distributions. The other transfers knowledge from richer supervision through camera teachers (Zhao et al., 2018a), LiDAR features (Bang et al., 2024), or paired text (Lai et al., 2026; Mishra et al., 2025), but typically optimizes the RF representation for a particular alignment target or downstream behavior. In contrast, RFception trains no task-specific RF model. It learns a lightweight interface to a pretrained generalist visual perception model, allowing a single RF representation to support depth, surface normals, human pose, and segmentation, including tasks that receive no RF supervision during training.

2.2 Visual Foundation Models and Generative Perception

Visual foundation models acquire broad knowledge of object semantics, scene geometry, and spatial structure, whether trained for specific perception tasks such as monocular depth estimation (Yang et al., 2024a;b; Bochkovskiy et al., 2025) and segmentation (Kirillov et al., 2023), or as general-purpose representations through language supervision (Radford et al., 2021) and self-supervision (Caron et al., 2021; He et al., 2022; Oquab et al., 2023). Diffusion and flow-based generative models (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2022; Liu et al., 2022; Esser et al., 2024) encode similarly transferable priors. Fine-tuning pretrained latent generative models has produced depth and surface-normal estimators with strong zero-shot generalization (Zhao et al., 2023; Ke et al., 2024; Fu et al., 2024; Gui et al., 2025; Ye et al., 2024), while later work simplifies this process to a single deterministic step (Xu et al., 2025; He et al., 2025; Garcia et al., 2025), typically for one task at a time. Multi-task generalists instead support multiple perception tasks within a single model through visual in-context learning (Wang et al., 2023a;b), sequence modeling with modality-specific tokenizers (Lu et al., 2022; 2024; Mizrahi et al., 2023; Bachmann et al., 2024), or instruction-conditioned generation (Gan et al., 2024; Geng et al., 2024; Wang et al., 2025b; Le et al., 2025). We build on DICEPTION (Zhao et al., 2026), which unifies depth, surface normals, human pose, and segmentation within a single diffusion through token-wise conditional generation.

These generalist models, however, assume visual inputs. A separate line of work connects new modalities to pretrained models through lightweight interfaces, including vision-language adapters (Li et al., 2023; Liu et al., 2023), shared embedding spaces (Girdhar et al., 2023), and fMRI decoding into diffusion latent spaces (Takagi & Nishimoto, 2023; Scotti et al., 2023). Related approaches have also adapted visual foundation models to other sensing modalities, such as thermal imagery (Fan et al., 2024), but these interfaces are typically designed around a particular representation or downstream behavior. RFception instead replaces only the visual input pathway while preserving the pretrained model's text-driven task interface, allowing a sensor with neither the appearance statistics nor the projection geometry of natural images to access the same generalist perception priors.

3 Method

3.1 Problem Formulation and Overview

We study whether a visual generative foundation model can serve as a reusable prior for sensing modalities whose measurements differ fundamentally from natural images. Our central hypothesis is that large-scale visual generative pretraining captures transferable regularities of the physical world,

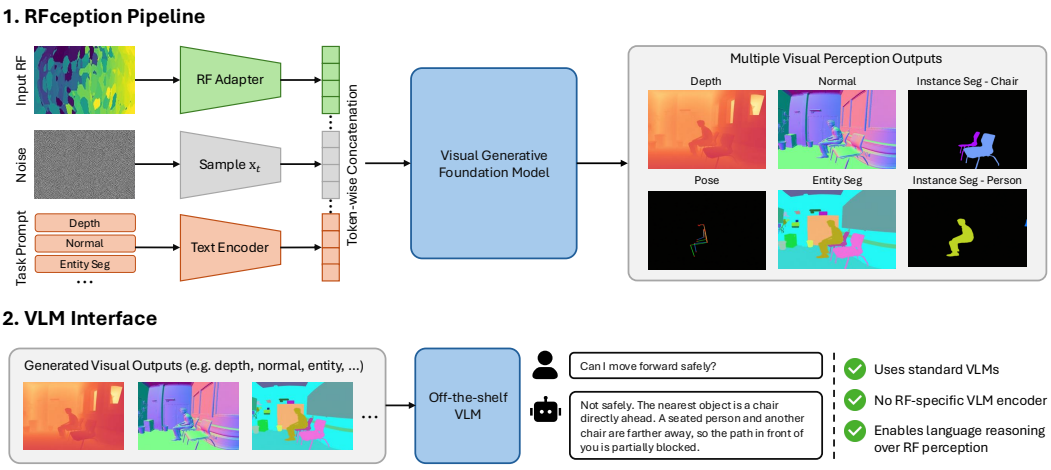


Figure 2: **Overview of RFception.** *Top*: The RF adapter A_ϕ maps an RF heatmap into the latent space of a pretrained visual generative foundation model. The resulting RF condition tokens are concatenated with noisy output tokens and task-prompt embeddings, with the prompt alone specifying the perception task. *Bottom*: The generated outputs are standard visual perception representations and can therefore be consumed by an off-the-shelf VLM through its native visual interface, enabling language-level reasoning over RF sensing without an RF-specific VLM encoder or VLM fine-tuning.

including geometry, structure, and semantics, and that non-visual sensors can access these priors through a lightweight interface rather than requiring a separately pretrained foundation model.

We instantiate this problem with RF sensing. Let $\mathbf{x}^r \in \mathbb{R}^{C_r \times H_r \times W_r}$ denote an RF observation. Rather than learning a task-specific predictor directly from $\mathbf{x}^r$, we seek a mapping from RF measurements into the latent representation space of a pretrained visual generative model. Once this mapping is established, the pretrained model can be reused across a range of perception tasks through task instructions, including tasks for which no RF supervision is provided.

3.2 RFCEPTION: ALIGNING RF WITH VISUAL GENERATIVE PRIORS

Architecture. RFception builds upon a pretrained generalist visual perception model using the token-wise conditional generation formulation of DICEPTION (Zhao et al., 2026), as illustrated in Figure 2. Our design principle is to preserve the pretrained visual model as much as possible. We therefore keep its VAE decoder, text encoder, and generation procedure unchanged, and replace only the input VAE encoder with a lightweight RF adapter A_ϕ. Given an RF observation $\mathbf{x}^r$, the adapter directly produces a latent representation

$$\mathbf{z}^r = A_\phi(\mathbf{x}^r) : \mathbb{R}^{C_r \times H_r \times W_r} \to \mathbb{R}^{C_z \times h \times w}, \tag{1}$$

whose dimensionality matches the latent representation expected by the pretrained visual model. After applying the same patchification as in the original model, the resulting RF condition tokens are concatenated with noisy tokens and task-prompt embeddings along the token dimension. The generative backbone then predicts the corresponding perception output. During iterative generation, the RF condition and task tokens remain fixed, while only the noisy tokens are iteratively updated.

Since RF measurements follow a distribution different from that of visual observations, matching the latent dimensionality alone does not ensure that the RF representation is interpreted correctly by the pretrained backbone. We therefore additionally adapt the backbone to RF inputs in a parameter-efficient manner. Let $\mathcal{F}_\theta$ denote the pretrained generative backbone with weights θ. We keep the pretrained weights fixed and apply LoRA (Hu et al., 2021) to the query, key, and value projections of every attention layer. For each projection with pretrained weight $\mathbf{W}_0 \in \{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$, $\mathbf{W}_0 \in \mathbb{R}^{d \times d'}$, the adapted weight is

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\mathbf{A}, \quad \mathbf{B} \in \mathbb{R}^{d \times \rho}, \ \mathbf{A} \in \mathbb{R}^{\rho \times d'}, \ \rho \ll \min(d, d'), \tag{2}$$

where only $\mathbf{A}$ and $\mathbf{B}$ are trainable. We denote the collection of low-rank parameters by $\Delta\theta$, which constitutes a small fraction of the backbone parameters, and denote the adapted backbone by $\mathcal{F}_{\theta+\Delta\theta}$.

Finally, RFception selects perception tasks through simple textual instructions, such as "image to depth", "image to normal", or "image to segmentation". For category-conditioned segmentation, the target category is incorporated directly into the instruction, such as "image to instance – chair". We retain the exact instruction templates used to train the pretrained model, since replacing "image" with the RF modality would shift the prompts away from the text distribution seen during training. Consequently, all RF perception tasks share the same RF adapter and adapted generative backbone, with the task specified only through the text prompt.

RF-to-visual representation alignment. During training, we assume that each RF observation $\mathbf{x}^r$ is paired with a spatially aligned visual observation $\mathbf{x}^v$ of the same scene. The visual observation is used only to establish cross-modal latent correspondence and is not required at inference. Let E_{vis} denote the frozen VAE encoder of the pretrained visual model. We obtain the target visual latent as

$$\mathbf{z}^v = \mathcal{T}\left(E_{\text{vis}}(\mathbf{x}^v)\right) \in \mathbb{R}^{C_z \times h \times w}, \tag{3}$$

where $\mathcal{T}$ denotes the latent normalization, scaling, and shifting operations used by the pretrained model. We first align $\mathbf{z}^r$ with $\mathbf{z}^v$ through element-wise regression, which encourages the RF latent to match the visual latent at each spatial location. We further adopt a dense contrastive objective (Wang et al., 2021) to preserve local cross-modal correspondence by distinguishing the matched visual feature from features at other spatial locations. Let $\mathbf{q}_u$ and $\mathbf{k}_u$ denote the ℓ_2-normalized embeddings of $g_\psi(\mathbf{z}^r)$ and $g_\psi(\mathbf{z}^v)$ at location u, respectively, where g_ψ is a lightweight per-location projection head. The representation alignment loss is

$$\mathcal{L}_{\text{rep}} = \lambda_{\text{mse}} \frac{1}{C_z hw} \left\|\mathbf{z}^r - \mathbf{z}^v\right\|_2^2 - \lambda_{\text{dense}} \frac{1}{|\Omega|} \sum_{u \in \Omega} \log \frac{\exp\left(\mathbf{q}_u^\top \mathbf{k}_u / \tau\right)}{\sum_{u' \in \Omega} \exp(\mathbf{q}_u^\top \mathbf{k}_{u'} / \tau)}, \tag{4}$$

where Ω denotes the set of spatial locations considered for contrastive matching, τ is the temperature, and λ_{mse} and λ_{dense} control the two loss terms.

Functional alignment through the generative model. Numerical similarity in the latent space alone does not guarantee that the RF representation induces the intended behavior in the pretrained model. Because RF and visual sensors do not resolve exactly the same information about a scene, a residual between $\mathbf{z}^r$ and $\mathbf{z}^v$ can remain even after representation alignment. What matters is not latent distance alone, but whether this residual preserves the behavior of the pretrained model. We address this by introducing a functional alignment objective that optimizes the RF interface through the generative model, shaping the RF representation according to its effect on the generated output.

For task $k \in \mathcal{K}$, where $\mathcal{K}$ is a small set of perception tasks used for functional alignment, let $\mathbf{y}^k$ denote the target output and $\mathbf{e}_k$ its corresponding task embedding. Following the flow-matching formulation (Lipman et al., 2022; Liu et al., 2022), we encode the target into the latent space and interpolate it with Gaussian noise:

$$\mathbf{z}_t^k = (1-t)\mathbf{z}_0^k + t\boldsymbol{\epsilon}, \quad \mathbf{z}_0^k = \mathcal{T}\big(E_{\text{vis}}(\mathbf{y}^k)\big), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

where $t \in [0,1]$ is sampled from the timestep distribution of the pretrained model. Conditioned on the RF latent $\mathbf{z}^r$ and task embedding $\mathbf{e}_k$, the adapted backbone is trained to predict the target velocity

$$\mathcal{L}_{\text{func}} = \lambda_{\text{func}} \mathbb{E}_{k \sim \mathcal{K},\, t,\, \boldsymbol{\epsilon}} \left[\left\| \mathcal{F}_{\theta+\Delta\theta}\left(\mathbf{z}_t^k, \mathbf{z}^r, \mathbf{e}_k, t\right) - \left(\boldsymbol{\epsilon} - \mathbf{z}_0^k\right) \right\|_2^2 \right], \tag{6}$$

where λ_{func} controls the strength of functional alignment. Gradients from $\mathcal{L}_{\text{func}}$ update both the low-rank parameters $\Delta\theta$ and the RF adapter A_ϕ. The pretrained backbone weights θ remain frozen.

3.3 A Visual Interface for Multimodal Foundation Models

A further benefit of aligning non-visual sensing with a visual generative foundation model is compatibility with the broader ecosystem of pretrained multimodal models. Raw RF measurements differ substantially from the natural-image distribution on which modern vision-language models (VLMs) are trained, making direct integration difficult without additional modality-specific encoders. RFception instead converts RF observations into standard visual perception representations through the visual generative model. These outputs can be consumed directly by an off-the-shelf VLM through its native visual interface, without modifying or fine-tuning the VLM itself.

This setting probes a broader consequence of our hypothesis. If visual generative foundation models encode reusable priors about the physical world, then aligning non-visual sensing with both their internal latent space and generated visual outputs provides a natural bridge from specialized physical sensors to existing multimodal foundation models.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We evaluate on the indoor RF dataset released with PanoRadar (Lai et al., 2024), which pairs 3D RF heatmaps with co-captured LiDAR range images collected across 15 buildings. To evaluate generalization to unseen environments, we split the dataset by building. Buildings 1–6 are held out for testing, while buildings 7–15 are used for training. Because the pretrained model is trained on perspective images, we extract a $135^\circ \times 90^\circ$ window from each 360° panorama, yielding an RF input of size $256 \times 192 \times 64$ with 256 range bins. The window is sampled at a random azimuth during training and centered at the 180° azimuth at test time. Per-building statistics, the cropping procedure, and the construction of task labels are detailed in Appendix A.

Tasks. We consider five tasks: depth estimation, surface-normal estimation, entity segmentation, human pose estimation, and category-conditioned instance segmentation of people. Only depth and entity segmentation are used for the functional-alignment objective, with $\mathcal{K} = \{\text{depth}, \text{entity}\}$. Surface normals, human pose, and instance segmentation receive no task-specific RF supervision and are evaluated zero-shot by changing only the text prompt. We motivate this choice of $\mathcal{K}$ in Appendix B. Depth is predicted in metric units using the metric decoder described in Appendix C.

Implementation details. We build RFception on the released DICEPTION model (Zhao et al., 2026). The RF adapter A_ϕ is a ResNet with nine residual blocks that treats the 256 range bins as input channels and maps an RF heatmap of size $256 \times 192 \times 64$ to a $16 \times 96 \times 96$ latent, matching the latent representation produced by the pretrained VAE for a 768×768 image. The projection head g_ψ used for the dense contrastive loss consists of two strided convolutions for downsampling, a 4×4 average-pooling layer, and a two-layer MLP. We adapt the backbone using rank-128 LoRA ($\rho = 128$), applied to the query, key, and value projections of every attention layer. In total, RFception has approximately 70M trainable parameters, accounting for less than 1% of the 7B parameters in the pretrained model. We use a batch size of 64 and set $\lambda_{\text{mse}} = 1.0$, $\lambda_{\text{dense}} = 0.01$, and $\lambda_{\text{func}} = 1.0$. We optimize the model with AdamW using a learning rate of 1.0×10^{-5}. At inference, we follow the default DICEPTION setting and generate each output using 28 sampling steps with a classifier-free guidance scale of 2.0. All experiments are conducted on four NVIDIA RTX PRO 6000 GPUs.

4.2 MAIN RESULTS

We compare against PanoRadar (Lai et al., 2024), a state-of-the-art RF imaging system that trains task-specific networks, and report two variants. PanoRadar† follows the original panoramic setting, where the model is trained and evaluated on uncropped 360° panoramas at a resolution of 512×64. Because our model operates on a 135° window, we additionally train PanoRadar under the same cropped setting, using random 192×64 crops during training and a fixed crop centered at the 180° azimuth for evaluation. Tables 1, 4, and 5 report quantitative results on the six held-out buildings, while Figures 3, 7, and 8 provide qualitative comparisons of depth and surface-normal estimation.

RFception achieves the highest δ_1 in all six held-out buildings, averaging 0.974 compared with 0.953 for the cropped PanoRadar baseline. It also achieves the lowest average MAE, 0.170 m versus 0.194 m. The largest gains occur in the most challenging environments. In Buildings 2, 3, and 4, where PanoRadar incurs its highest MAE, RFception reduces MAE by 26%, 30%, and 21%, respectively. The two methods nevertheless exhibit complementary error profiles. PanoRadar achieves lower AbsRel in all six buildings, indicating tighter fits on nearby, smooth surfaces. Qualitatively, however, its predictions tend to over-smooth surfaces and miss thin structures and people. RFception instead recovers fine scene structure, including furniture, doorways, and human bodies, with substantially sharper boundaries, consistent with the structural prior inherited from the pretrained visual model.

Table 1: **Quantitative comparison on the six held-out test buildings.** Columns 1–6 index the test buildings of Table 3. "Res." is the output resolution and "Training Samples" is the number of RF-supervised samples. 0 denotes zero-shot. PanoRadar† denotes the original setting with full 360° panoramas, and the others use 135° crops. Building 3 has no object annotations (N/A, Appendix A).

Task	Method	Res.	Training Samples	1		2		3		4		5		6	
				MAE↓	$\delta_1^\uparrow$	MAE↓	$\delta_1^\uparrow$	MAE↓	$\delta_1^\uparrow$	MAE↓	$\delta_1^\uparrow$	MAE↓	$\delta_1^\uparrow$	MAE↓	$\delta_1^\uparrow$
Depth	PanoRadar†	512×64	15K	0.133	0.959	0.260	0.943	0.331	0.906	0.198	0.949	0.126	0.969	0.125	0.978
	PanoRadar	192×64	15K	0.145	0.958	0.249	0.946	0.282	0.922	0.218	0.944	0.139	0.968	0.132	0.977
	Ours	768×768	15K	0.161	0.963	0.185	0.987	0.198	0.950	0.172	0.974	0.136	0.986	0.169	0.987
				$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$	$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$	$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$	$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$	$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$	$22.5^{\circ\uparrow}$	$30^{\circ\uparrow}$
Normal	PanoRadar†	512×64	15K	87.14	89.46	85.43	87.67	79.73	83.44	84.39	87.02	86.32	88.78	90.70	92.36
	PanoRadar	192×64	15K	86.78	89.13	85.36	87.73	81.16	84.70	83.15	85.91	84.96	87.52	90.32	92.07
	Ours	768×768	0	62.94	71.02	68.16	74.98	53.65	62.46	64.78	71.97	63.89	71.37	69.51	76.70
				AR↑		AR↑		AR↑		AR↑		AR↑		AR↑	
Pose	PanoRadar	192×64	15K	63.62		58.09		N/A		63.56		65.11		68.72	
	Ours	768×768	0	60.85		41.00		N/A		47.07		51.34		55.53	
				AP↑		AP↑		AP↑		AP↑		AP↑		AP↑	
Instance	PanoRadar	192×64	15K	61.88		54.89		N/A		57.23		70.16		56.60	
	Ours	768×768	0	57.53		46.60		N/A		53.53		66.28		51.92	

4.3 Zero-Shot Transfer to Unsupervised Tasks

We next ask whether the learned interface transfers directly to tasks that receive no RF supervision during training. We evaluate this capability on surface-normal estimation, human pose estimation, and person instance segmentation, none of which is included in the RF training objectives.

Surface normals. Without any normal supervision, RFception places 63.82% of pixels within 22.5° of the ground truth on average and 71.42% within 30°. PanoRadar is explicitly supervised on surface normals and performs better on these pixel-averaged metrics. Part of this gap reflects how the metrics weight the scene. Large planar regions such as walls, floors, and ceilings dominate indoor images and can be fit accurately by a task-specific supervised model. In contrast, the zero-shot predictions of RFception preserve richer local structure, recovering the geometry of furniture, ceiling fixtures, and people that is often smoothed away by the supervised baseline.

Human pose and instance segmentation. RFception achieves an average AR of 51.16 for human pose estimation and an average AP of 55.17 for person instance segmentation. The supervised PanoRadar baseline remains higher at 63.82 AR and 60.15 AP, respectively. Both tasks require capabilities beyond the supervision used to train the RF interface, including localizing people, separating individual instances, and recovering body structure. Neither the depth nor the entity alignment target directly provides this supervision. The ability to perform these tasks reflects knowledge carried by the pretrained visual model and accessed through the shared RF interface.

What enables transfer. Zero-shot transfer depends critically on how the RF interface is aligned with the pretrained model (Figure 5). When the functional alignment objective is removed, the adapter is trained only to match $\mathbf{z}^r$ to $\mathbf{z}^v$, and the outputs collapse to nearly flat maps. Numerical similarity in latent space is therefore insufficient to establish a functional interface. The remaining mismatch must also be shaped according to its effect on the generative model's outputs. Functional alignment with depth alone is sufficient to recover meaningful depth, surface normals, and human pose, whereas segmentation-type tasks still fail. We hypothesize that this is because RF provides no color information, leaving appearance-related components of the visual latent weakly constrained. Adding the class-agnostic entity task provides complementary semantic supervision, after which category-conditioned instance segmentation emerges zero-shot. Appendix B provides the full ablation.

4.4 Language-Level Reasoning with Off-the-Shelf VLMs

The outputs of RFception are standard visual perception representations that can be consumed directly by existing vision-language models. We evaluate this compatibility end to end, from a raw RF heatmap to a natural-language answer, without modifying or fine-tuning the VLM.

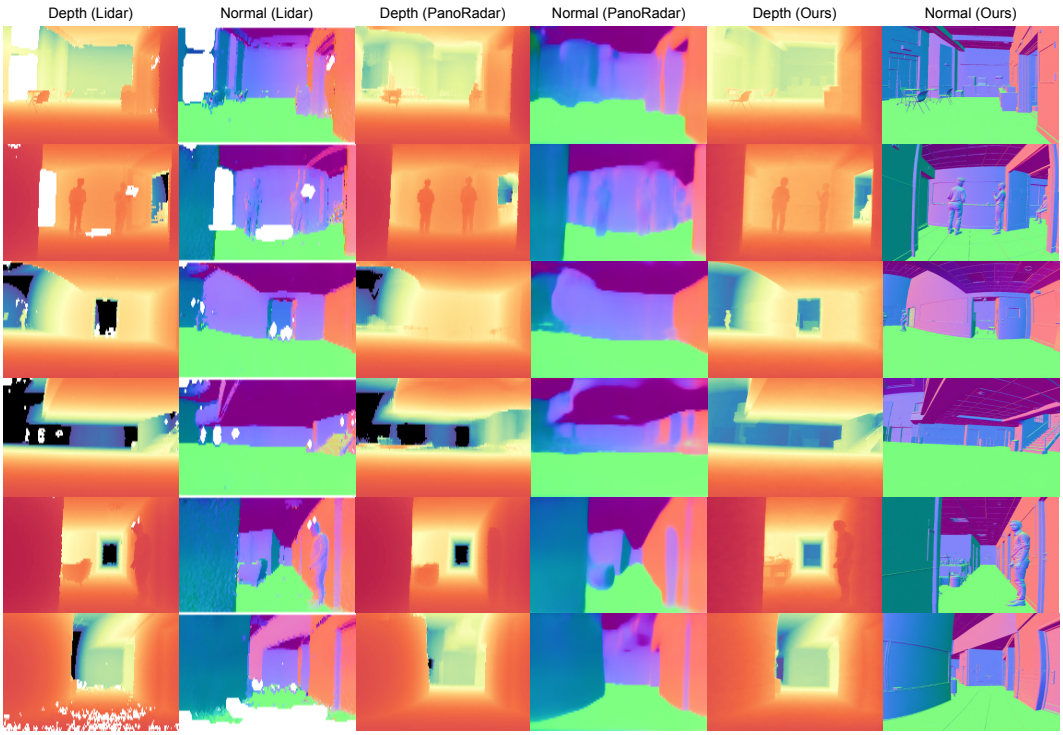


Figure 3: **Qualitative comparison of depth and surface-normal estimation.** Each scene shows three depth–normal pairs: LiDAR, PanoRadar, and RFception. PanoRadar recovers the coarse layout but over-smooths surfaces and loses thin structures, whereas RFception reproduces furniture, doorways, and human bodies with sharp boundaries. More results are provided in Figures 7, 8, 9, 10.

Benchmark. We sample 100 scenes from the six held-out test buildings and annotate each scene with three types of questions, yielding 300 questions in total. The first type provides a set of candidate categories and asks which are present in the current view, evaluating object understanding. The second provides the categories known to be present and asks the model to order them, for example from nearest to farthest relative to the sensor or from left to right, evaluating spatial relations. The third combines the two, requiring the model to first identify the present categories from a candidate set and then order them, evaluating compositional reasoning. Reference answers are annotated using the LiDAR depth map of the same scene. Appendix E provides examples of the benchmark.

Baselines and base VLMs. The base VLM remains frozen throughout the evaluation, and the methods differ only in the representation provided to it. RFception and PanoRadar (Lai et al., 2024) provide the depth map predicted from RF through the VLM's native visual interface. We additionally feed the corresponding LiDAR reference depth as a reference for the performance attainable from high-quality geometric input. We further compare against two radar-language interfaces that connect RF measurements to the VLM through its embedding space. RSLM (Pushkareva et al., 2024) trains a radar encoder to align with the VLM's image embedding, whereas RLM (Mishra et al., 2025) aligns a radar encoder with its text embedding. Because these interfaces depend on the embedding space of the target VLM, we retrain each of them on our RF training split for every base VLM. To ensure that the results are not specific to a single VLM, we repeat the full evaluation with four open-weight models of similar scale: Qwen3-VL-8B (Yang et al., 2025), InternVL3.5-8B (Wang et al., 2025a), LLaVA-OneVision2-8B (An et al., 2026), and Gemma3-12B (Team et al., 2025).

Results. Averaged over the four base VLMs, RFception achieves the best performance among the RF interfaces across all three reasoning categories. On compositional reasoning, it reaches an exact-match accuracy of 43.50, compared with 39.75 for PanoRadar. Both methods provide the VLM with an RF-derived depth map, so their performance difference reflects the scene information preserved by that representation (Figure 3). PanoRadar captures the coarse scene layout but tends to over-smooth surfaces, whereas RFception preserves finer structures with sharper boundaries.

Table 2: **Language-level reasoning with off-the-shelf VLMs.** Frozen VLMs answer the same 300 questions using the representation provided by each method (Appendix E). "Training Free" indicates whether the interface requires no training against the target VLM, while "Alignment" specifies which VLM embedding space is used by methods that require such training.

Method	Base VLM	Training Free	Alignment	Object Understanding		Spatial Relations		Comp. Reasoning
				F1	Exact Match	Pairwise Acc.	Exact Match	Exact Match
LiDAR	Qwen3-VL 8B	✓	–	83.17	56.00	84.36	62.00	47.00
RSLM		✗	Image enc.	40.22	9.00	55.33	22.00	1.00
RLM		✗	Text enc.	50.31	14.00	47.92	19.00	0.00
PanoRadar		✓	–	77.46	48.00	80.71	54.00	37.00
Ours		✓	–	81.92	51.00	83.04	57.00	41.00
LiDAR	InternVL3.5 8B	✓	–	79.66	49.00	86.24	63.00	52.00
RSLM		✗	Image enc.	44.96	3.00	54.67	19.00	0.00
RLM		✗	Text enc.	47.95	6.00	63.79	26.00	8.00
PanoRadar		✓	–	75.14	52.00	79.97	58.00	44.00
Ours		✓	–	80.43	46.00	83.51	67.00	43.00
LiDAR	LLaVA OneVision2 8B	✓	–	81.31	57.00	80.76	51.00	46.00
RSLM		✗	Image enc.	47.84	5.00	56.09	26.00	0.00
RLM		✗	Text enc.	54.10	4.00	50.34	23.00	6.00
PanoRadar		✓	–	78.09	45.00	78.72	56.00	33.00
Ours		✓	–	79.42	54.00	83.18	49.00	39.00
LiDAR	Gemma3 12B	✓	–	85.27	60.00	87.12	68.00	53.00
RSLM		✗	Image enc.	56.53	11.00	48.33	14.00	6.00
RLM		✗	Text enc.	65.81	17.00	62.87	31.00	11.00
PanoRadar		✓	–	79.83	56.00	83.47	59.00	45.00
Ours		✓	–	84.06	63.00	85.91	64.00	51.00

The two embedding-level interfaces perform substantially worse. On compositional reasoning, RSLM and RLM achieve average exact-match accuracies of only 1.75 and 6.25, respectively. Two factors may help explain this gap. First, both methods must learn a radar encoder from the 15K RF samples in our training split, far fewer than the data used to establish the pretrained VLM embedding spaces. Second, the supervision differs substantially in spatial granularity. Reconstructing a visual representation provides dense supervision across the output and directly encourages the RF adapter to preserve local scene structure, whereas matching a global embedding constrains the representation at the sample level and does not explicitly preserve spatial layout. The resulting degradation is consistent with this distinction. RSLM and RLM retain 58% and 66% of the LiDAR F1 score for object understanding, yet achieve only 53% and 56% pairwise accuracy on spatial relations, close to the 50% of a random ordering. This suggests that coarse semantic information can survive embedding-level alignment, while detailed spatial organization is much more difficult to preserve.

5 Conclusion

This paper asked whether the knowledge acquired by a visual foundation model from internet-scale images can be transferred to a sensor with appearance statistics different from those of natural images. RFception addresses this by learning a lightweight interface into a pretrained visual generative foundation model. Because tasks are selected through text prompts, the resulting RF representation is not tied to any single perception task. Across six held-out buildings, RFception improves over task-specific RF models in depth estimation, while surface-normal estimation, human pose estimation, and category-conditioned instance segmentation transfer zero-shot. Its visual outputs can also be consumed by four off-the-shelf vision-language models without fine-tuning, with the RF-derived depth retaining 87.9% of the compositional-reasoning accuracy of LiDAR reference depth.

These results also come with limitations. RFception inherits the capabilities and the limitations of the pretrained visual model, so transfer is restricted to tasks that the model already supports. Moreover, when dense RF supervision is available, a specialized task-specific model can still perform better on pixel-averaged metrics, while the zero-shot predictions of RFception preserve richer local structure. The underlying approach, however, is not specific to RF. In principle, other sensing modalities that can be paired with corresponding visual observations during training could access the same pretrained visual priors through a learned interface, including thermal, ultrasonic, and tactile sensing. We therefore view these findings as evidence that useful knowledge captured by visual foundation models need not remain tied to the image modality in which it was learned, and that advances in visual foundation models may be transferable to sensing modalities that lack comparable data ecosystems.

AI USE STATEMENT

In this work, we did not use generative AI tools for research ideation, methodology or experiment design, implementation, data processing or analysis, or the interpretation of results. We used generative AI tools only for language editing and polishing of the manuscript. Specifically, these tools were used to improve grammar, clarity, and phrasing. All AI-assisted text was reviewed and revised by the authors, and we verified that these edits did not alter the technical content, experimental results, claims, or conclusions of the work. We take full responsibility for the final content of this work, including text, claims, and results.

REFERENCES

Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 3d tracking via body radio reflections. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, pp. 317–329, USA, 2014. USENIX Association. ISBN 9781931971096.

Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Trans. Graph.*, 34(6), November 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818072. URL https://doi.org/10.1145/2816795.2818072.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Roman Bachmann, Oğuzhan F Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *Advances in Neural Information Processing Systems*, 37:61872–61911, 2024.

Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15491–15500. IEEE, 2024.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ruoyu Fan, Wang Zhao, Matthieu Lin, Qi Wang, Yong-Jin Liu, and Wenping Wang. Generalizable thermal-based depth estimation via pre-trained visual foundation model. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14614–14621, 2024. doi: 10.1109/ICRA57147.2024.10610394.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Yulu Gan, Sung Woo Park, Alexander Schubert, Anthony Philippakis, and Ahmed Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. In *International Conference on Learning Representations*, volume 2024, pp. 54090–54106, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 753–762, 2025.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12709–12720. IEEE, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Yiduo Hao, Sohrab Madani, Junfeng Guan, Mohammed Alloulah, Saurabh Gupta, and Haitham Hassanieh. Bootstrapping autonomous driving radars with self-supervised learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15012–15023. IEEE, 2024.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89454–89467, 2025.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tianshu Huang, Akarsh Prabhakara, Chuhan Chen, Jay Karhade, Deva Ramanan, Matthew O'toole, and Anthony Rowe. Towards foundational models for single-chip radar. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24655–24665. IEEE, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502. IEEE, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Haowen Lai, Gaoxiang Luo, Yifei Liu, and Mingmin Zhao. Enabling visual recognition at radio frequency. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 388–403, 2024.

Haowen Lai, Zhiwei Zheng, and Mingmin Zhao. Rf-based 3d slam rivaling vision approaches. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*, ACM MOBICOM '25, pp. 170–185, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711299. doi: 10.1145/3680207.3723467. URL https://doi.org/10.1145/3680207.3723467.

Zengyuan Lai, Jiarui Yang, Songpengcheng Xia, Lizhou Lin, Lan Sun, Renwen Wang, Jianran Liu, Qi Wu, and Ling Pei. Radarllm: Empowering large language models to understand human motion from millimeter-wave point cloud sequence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5791–5799, 2026.

Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2671–2682. IEEE, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 872–881. IEEE, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pp. 14–27, 2020.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26429–26445. IEEE, 2024.

Pushkal Mishra, Kshitiz Bansal, and Dinesh Bharadia. Rlm: A vision-language model approach for radar scene understanding. *arXiv preprint arXiv:2511.21105*, 2025.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, 2013. doi: 10.1109/MGRS.2013.2248301.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Anurag Pallaprolu, Belal Korany, and Yasamin Mostofi. Wiffract: a new foundation for rf imaging via edge tracing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom '22, pp. 255–267, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391818. doi: 10.1145/3495243.3514261. URL https://doi.org/10.1145/3495243.3514261.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4135–4142. IEEE, 2023.

Mariia Pushkareva, Yuri Feldman, Csaba Domokos, Kilian Rambach, and Dotan Di Castro. Radar spectra-language model for automotive scene parsing. In *2024 International Radar Conference (RADAR)*, pp. 1–6. IEEE, 2024.

Kun Qian, Zhaoyuan He, and Xinyu Zhang. 3d point cloud generation with millimeter-wave radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4), December 2020. doi: 10.1145/3432221. URL https://doi.org/10.1145/3432221.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Kunzhe Song, Qijun Wang, Shichen Zhang, and Huacheng Zeng. Siwis: Fine-grained human detection using single wifi device. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, pp. 1439–1454, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704895. doi: 10.1145/3636534.3690703. URL https://doi.org/10.1145/3636534.3690703.

Kunzhe Song, Geo Jie Zhou, Xiaoming Liu, and Huacheng Zeng. Rascene: High-fidelity 3d scene imaging with mmwave communication signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36454–36463, June 2026.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3023–3032. IEEE, 2021.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839. IEEE, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023b.

Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, and Tongliang Liu. Lavin-dit: Large vision diffusion transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20060–20070. IEEE, 2025b.

Guangkai Xu, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, Chunhua Shen, et al. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations*, volume 2025, pp. 6786–6799, 2025.

Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, pp. 269–282, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384438. doi: 10.1145/3458864.3467679. URL https://doi.org/10.1145/3458864.3467679.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024a.

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, 2024b.

Muhammet Emin Yanik and Murat Torlak. Near-field mimo-sar millimeter-wave imaging with sparsely sampled aperture data. *IEEE Access*, 7:31801–31819, 2019. doi: 10.1109/ACCESS.2019.2902859.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (ToG)*, 43(6):1–18, 2024.

Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. *Advances in Neural Information Processing Systems*, 38:40108–40155, 2026.

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7356–7365, 2018a. doi: 10.1109/CVPR.2018.00768.

Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '18, pp. 267–281, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450355674. doi: 10.1145/3230543.3230579. URL https://doi.org/10.1145/3230543.3230579.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5706–5716. IEEE, 2023.

A Dataset Details

Data source. All experiments use the indoor dataset released with PanoRadar (Lai et al., 2024), which contains spatially and temporally aligned RF and LiDAR measurements collected across 15 buildings. We index the buildings from 1 to 15 in alphabetical order by name. Table 3 reports the number of samples in each building and the corresponding data split. Buildings 1–6 are held out entirely for evaluation, while buildings 7–15 are used for training, ensuring that all reported results are evaluated on environments unseen during training.

Table 3: **Per-building sample counts and dataset split.** Buildings 1–6 are held out for evaluation and are the columns reported in Table 1, and buildings 7–15 are used for training.

Building	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Samples	631	698	601	854	200	1,765	1,449	670	1,985	2,366	1,759	1,685	1,082	1,664	2,079
Split	test	test	test	test	test	test	train	train	train	train	train	train	train	train	train

Sensor format. Each sample consists of a paired RF heatmap and LiDAR range image. The RF heatmap has a resolution of $256 \times 512 \times 64$, corresponding to 256 range bins over a 512×64 azimuth–elevation grid. The LiDAR range image is defined on the same 512×64 angular grid. The horizontal dimension spans a 360° azimuth field of view, while the vertical dimension spans a 90° elevation field of view. The LiDAR range image provides dense metric depth and serves both as the paired visual observation $\mathbf{x}^v$ in Equation 3 and as the ground truth for depth estimation.

Annotations. Apart from depth, the released dataset provides semantic segmentation masks and object bounding boxes, but no annotations for surface normals, entity segmentation, instance segmentation, or human pose. We construct the labels from the available annotations as follows.

- **Surface normal.** Surface normals are computed from the LiDAR depth using the same procedure as the official PanoRadar (Lai et al., 2024) implementation.
- **Entity segmentation.** The semantic mask assigns a category to each pixel but does not distinguish different objects of the same category. We separate each semantic region into individual entities using the provided object bounding boxes and treat each resulting region as a distinct entity.
- **Instance segmentation - person.** We extract pixels labeled as person from the semantic mask and assign them to individual instances using the corresponding person bounding boxes.
- **Human pose.** We detect human keypoints using an off-the-shelf Keypoint R-CNN with a ResNet-50-FPN backbone (He et al., 2017) and match each detected skeleton to a person bounding box provided by the dataset. Only matched skeletons are retained as pose labels.

For building 3, the released annotations contain neither valid semantic labels nor object bounding boxes. The segmentation and pose labels above therefore cannot be constructed for this building, and the corresponding entries in Table 1 are marked as N/A. These derived labels are used to supervise the baseline models and to evaluate the corresponding tasks, but RFception is not trained on any of them. For the entity segmentation task in $\mathcal{K}$, the functional alignment target $\mathbf{y}^k$ is instead generated by the pretrained DICEPTION model, thereby preserving the task representation learned during visual pretraining. Surface normals, human pose, and category-conditioned instance segmentation receive no label supervision and are evaluated zero-shot.

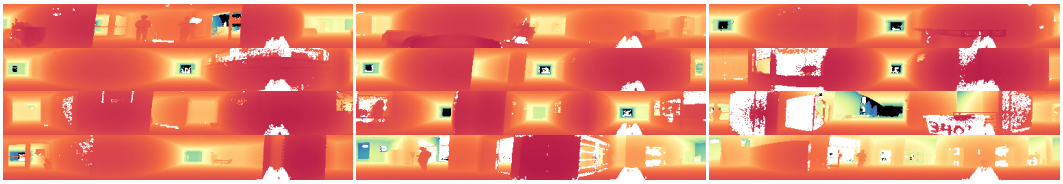


Figure 4: **Example LiDAR range images.** Each strip shows a full 360° LiDAR panorama at 512×64 resolution with an 8:1 aspect ratio. This panoramic geometry differs substantially from the perspective-image inputs used to pretrain DICEPTION, motivating our field-of-view cropping.

Field-of-view cropping. DICEPTION (Zhao et al., 2026) is pretrained on perspective images with conventional camera geometry. In contrast, the original 512×64 LiDAR range image represents a full $360°$ panorama with an aspect ratio of $8{:}1$ (Figure 4). Directly resizing this panorama to the square input resolution of the pretrained model would introduce a large geometric mismatch from its pretraining distribution. We therefore crop both the RF heatmap and the LiDAR range image to a 192×64 angular window, corresponding to a $135°$ horizontal and $90°$ vertical field of view. During training, the azimuth center of the window is sampled randomly for each sample. At test time, we use a fixed window centered at the $180°$ azimuth. The cropped LiDAR depth maps and derived labels are then resized to 768×768, the input resolution used by the pretrained model.

B Choice of Tasks for Functional Alignment

The functional alignment objective in Equation 6 is computed over a small task set $\mathcal{K}$. Our goal is to keep $\mathcal{K}$ as small as possible and rely on the pretrained generative model to transfer to the remaining tasks. This section explains how we select the task set used in the main experiments.

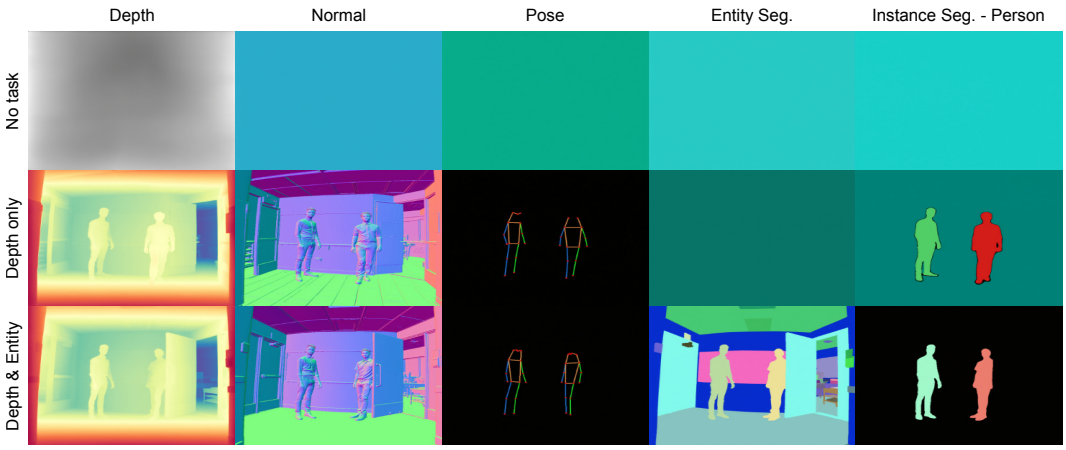


Figure 5: **Effect of the alignment task set $\mathcal{K}$.** *Top*: no functional alignment task. *Middle*: alignment with depth. *Bottom*: joint alignment with depth and entity segmentation. Without functional alignment every task collapses to a flat output. Depth alone already transfers geometry and pose, but entity and instance segmentation still collapse. Adding entity segmentation restores both.

No functional alignment. We first remove $\mathcal{L}_{\text{func}}$ and train the adapter using only the representation alignment objective in Equation 4, corresponding to $\mathcal{K} = \emptyset$. As shown in the top row of Figure 5, all task outputs collapse to nearly flat maps with no meaningful scene content. Numerically matching the RF latent to the visual latent is therefore insufficient to form a usable interface. Although the optimized RF latent remains close to $\mathbf{z}^v$, the residual can still lie along directions to which the pretrained backbone is sensitive, causing the RF latent to fall outside the region that the generative model interprets as a valid scene representation. Functional alignment is therefore necessary to shape the interface according to its effect on the generative model, rather than latent-space distance alone.

Depth-only alignment. We next use depth as the only aligned task, $\mathcal{K} = \{\text{depth}\}$, since it is directly coupled to the geometric information captured by RF. As shown in the middle row of Figure 5, a single functionally aligned task is already sufficient to make the interface usable, and both surface normals and human pose emerge zero-shot with plausible structure. This suggests that the adapter now produces an RF latent that the pretrained model can interpret as a meaningful scene representation.

Segmentation tasks, however, fail to transfer. Entity segmentation collapses into a single undifferentiated region, while instance segmentation loses the scene structure needed to separate individual objects. We attribute this failure to the modality gap between RF measurements and the RGB images used to pretrain DICEPTION. In natural images, color and texture provide strong cues for grouping pixels into coherent objects. RF contains no such appearance information, so depth-only alignment constrains the geometric content of the latent while leaving object-grouping cues underdetermined.

Joint depth and entity alignment. To constrain this missing structure, we add entity segmentation to the aligned task set, $\mathcal{K} = \{\text{depth}, \text{entity}\}$. Entity segmentation is class-agnostic and asks the model to partition the scene into coherent object regions, directly supervising the grouping behavior that depth alone does not determine. As shown in the bottom row of Figure 5, adding this single task restores entity segmentation and also enables zero-shot category-conditioned instance segmentation. We therefore use $\mathcal{K} = \{\text{depth}, \text{entity}\}$ for all experiments in the main paper.

C Metric Depth Decoding

Relative depth in DICEPTION. DICEPTION (Zhao et al., 2026) does not predict metric depth. Instead, its depth target is a relative map normalized in inverse-depth (disparity) space using per-image statistics. Given a metric depth map $\mathbf{d}$ with per-image minimum $d_{\min}$ and maximum $d_{\max}$, each pixel is normalized as

$$\tilde{d}_i = 2\,\frac{1/d_i - 1/d_{\max}}{1/d_{\min} - 1/d_{\max}} - 1 \in [-1, 1], \tag{7}$$

and the resulting single-channel map is replicated across three channels before being encoded by the VAE. Near surfaces therefore take values close to 1, while far surfaces take values close to -1.

At inference, the pretrained VAE decoder produces a prediction $\hat{\tilde{\mathbf{d}}}$ in the same normalized space. Recovering metric depth requires inverting Equation 7,

$$\hat{d}_i = \left(\frac{\hat{\tilde{d}}_i + 1}{2}\left(\frac{1}{d_{\min}} - \frac{1}{d_{\max}}\right) + \frac{1}{d_{\max}}\right)^{-1}, \tag{8}$$

which requires the per-image extrema $d_{\min}$ and $d_{\max}$. These quantities are discarded during normalization and are unavailable at test time. The prediction is therefore defined only up to an unknown affine transformation in inverse-depth space. Consequently, the decoded map preserves relative scene geometry but does not uniquely determine metric depth, as illustrated in the top row of Figure 6.

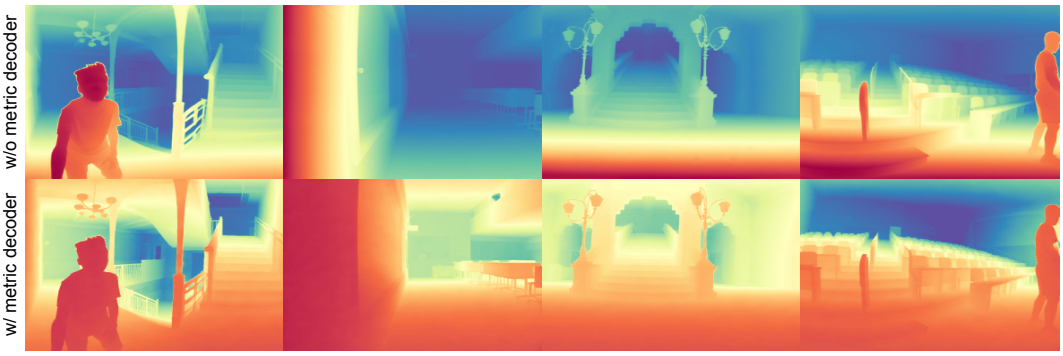


Figure 6: **Relative versus metric depth.** *Top:* Relative depth decoded by the pretrained VAE, represented in per-image normalized disparity space. *Bottom:* Metric depth produced by our lightweight decoder, which recovers absolute range while preserving the scene structure of the relative prediction.

Metric decoder. To recover metric depth without modifying the pretrained components, we attach a lightweight metric decoder M_ω after the frozen VAE decoder. RF sensing preserves absolute range through time-of-flight measurements, and the 256 range bins of the heatmap $\mathbf{x}^r$ correspond to known metric distances. We therefore introduce a direct RF branch that bypasses the generative backbone and provides absolute range cues to the final prediction. Let $\hat{\mathbf{y}} \in \mathbb{R}^{3\times H\times W}$ denote the relative depth output decoded by the VAE. The RF heatmap is first projected by a 1×1 convolution P_ω, upsampled to the output resolution using nearest-neighbor interpolation, and concatenated with $\hat{\mathbf{y}}$ along the channel dimension. The metric depth is then predicted as

$$\hat{\mathbf{d}} = M_\omega\left([\hat{\mathbf{y}},\, \mathrm{Up}(P_\omega(\mathbf{x}^r))]\right) \in \mathbb{R}^{1\times H\times W}, \tag{9}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation and $\mathrm{Up}(\cdot)$ denotes upsampling. M_ω is a ResNet with eight residual blocks. In this design, $\hat{\mathbf{y}}$ provides the fine scene structure inherited from the generative

prior, while $\mathbf{x}^r$ supplies absolute range cues that anchor this structure to metric scale. As shown in the bottom row of Figure 6, the metric decoder recovers absolute range while preserving the structural detail of the relative prediction.

The metric decoder is trained in a separate stage. The RF adapter, LoRA parameters, and all pretrained components are frozen, and only P_ω and M_ω are optimized using a mean squared error loss. Because the remainder of the model is unchanged, this stage does not affect the behavior of any other task. All RFception depth results reported in Table 1 and Table 4 are produced by this metric decoder.

D Additional Results

D.1 Evaluation Protocol

Setup. RFception generates predictions at 768×768, whereas the LiDAR ground truth for each test crop is defined on a 192×64 angular grid. We therefore resize the ground truth to 768×768 and compute all metrics at this resolution. Metrics are computed only over pixels with valid LiDAR returns. All results are reported separately for each test building. PanoRadar† follows its original protocol and is evaluated on the full 360° panorama at its native 512×64 resolution, whereas PanoRadar is evaluated on the same 135° test crop as RFception.

Depth. Let d_i and $\hat{d}_i$ denote the ground-truth and predicted metric depth at pixel i, respectively, and let $\mathcal{V}$ denote the set of valid pixels. We report the mean absolute error $\mathrm{MAE} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} |\hat{d}_i - d_i|$, the absolute relative error $\mathrm{AbsRel} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} |\hat{d}_i - d_i|/d_i$, the root mean squared error $\mathrm{RMSE} = \left(\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} (\hat{d}_i - d_i)^2\right)^{1/2}$, and the threshold accuracies δ_j, defined as the fraction of valid pixels satisfying $\max(\hat{d}_i/d_i,\, d_i/\hat{d}_i) < 1.25^j$ for $j \in \{1, 2, 3\}$. MAE and RMSE are reported in meters.

Surface normal. Let θ_i denote the angular error between the predicted and ground-truth unit normals at pixel i. We report the mean and median angular error in degrees, together with the percentage of valid pixels for which θ_i is below 11.25°, 22.5°, and 30°.

Human pose and instance segmentation. Both tasks follow the standard COCO evaluation protocol (Lin et al., 2014), using the labels constructed in Appendix A as ground truth. For human pose, we report average recall (AR) over object keypoint similarity (OKS) thresholds from 0.50 to 0.95 in increments of 0.05. For category-conditioned instance segmentation of people, we report average precision (AP) over mask intersection-over-union (IoU) thresholds from 0.50 to 0.95 with the same increment. Both metrics are reported as percentages.

D.2 Depth

Table 4 complements Table 1 with AbsRel, RMSE, and the δ_2 and δ_3 accuracies. RFception achieves the lowest RMSE and the highest δ_2 on all six test buildings. Its RMSE is 43%–60% lower than that of PanoRadar. Because RMSE places greater weight on large errors, this gap is consistent with fewer severe failures, such as missing objects or misplaced far-range surfaces. AbsRel shows the opposite trend, with PanoRadar achieving lower error on all six buildings. Since AbsRel normalizes each error by ground-truth depth, it places greater relative weight on errors at nearby surfaces, which in indoor scenes are often large, smooth walls and floors. A pixel-wise regression network can fit these surfaces tightly, whereas RFception decodes depth through the generative prior and trades some local smoothness for greater structural fidelity elsewhere. The two methods therefore exhibit complementary error profiles. Figure 7 illustrates this contrast qualitatively.

D.3 Surface Normals

Table 5 reports the mean and median angular errors and the 11.25° accuracy. PanoRadar is directly supervised for this task, whereas RFception receives no normal supervision at any stage and predicts normals entirely through the pretrained generative prior. The two methods also exhibit markedly different error distributions. For PanoRadar, the mean angular error is roughly four times the median,

Table 4: **Additional depth metrics on the six test buildings.** AbsRel, RMSE, and the δ_2/δ_3 threshold accuracies, complementing the MAE and δ_1 results of Table 1. Methods, training data, and evaluation settings are identical to Table 1, and RMSE is reported in meters.

Method	Res.	Training Samples	Metric	1	2	3	4	5	6
PanoRadar†	512×64	15K	AbsRel↓	0.056	0.072	0.099	0.060	0.058	0.043
			RMSE↓	0.412	0.721	0.909	0.617	0.370	0.376
			$\delta_2^\uparrow$	0.984	0.975	0.952	0.981	0.991	0.990
			$\delta_3^\uparrow$	0.993	0.986	0.970	0.990	0.996	0.995
PanoRadar	192×64	15K	AbsRel↓	0.056	0.062	0.082	0.069	0.054	0.043
			RMSE↓	0.445	0.708	0.791	0.684	0.397	0.382
			$\delta_2^\uparrow$	0.984	0.977	0.962	0.977	0.991	0.991
			$\delta_3^\uparrow$	0.992	0.987	0.978	0.987	0.996	0.995
Ours	768×768	15K	AbsRel↓	0.093	0.076	0.092	0.077	0.067	0.081
			RMSE↓	0.226	0.296	0.390	0.271	0.195	0.219
			$\delta_2^\uparrow$	0.989	0.986	0.979	0.982	0.995	0.997
			$\delta_3^\uparrow$	0.997	0.994	0.993	0.990	0.998	0.998

Table 5: **Additional surface-normal metrics on the six test buildings.** Mean and median angular errors in degrees and the percentage of valid pixels whose predicted normal falls within $11.25°$ of the ground truth, complementing the $22.5°$ and $30°$ accuracies of Table 1. Methods and evaluation settings are the same as in Table 1. RFception is evaluated zero-shot.

Method	Res.	Training Samples	Metric	1	2	3	4	5	6
PanoRadar†	512×64	15K	Mean↓	9.94	10.65	14.15	11.06	10.25	7.39
			Med↓	2.77	2.13	4.40	2.48	2.78	1.87
			$11.25°$↑	80.09	79.68	69.71	77.60	78.72	86.01
PanoRadar	192×64	15K	Mean↓	10.10	10.71	13.47	11.84	10.71	7.53
			Med↓	2.77	2.16	4.42	2.71	2.64	1.81
			$11.25°$↑	79.50	79.42	70.97	76.12	78.07	85.68
Ours	768×768	0	Mean↓	23.99	21.47	27.71	23.22	24.15	20.79
			Med↓	15.78	12.89	19.72	14.72	16.33	13.88
			$11.25°$↑	39.56	47.31	36.33	42.55	38.38	45.88

indicating a strongly heavy-tailed distribution in which most pixels have small errors while a smaller fraction incurs much larger ones. For RFception, this ratio is about 1.5, indicating a less heavy-tailed distribution with errors spread more evenly across pixels. The key difference is where these errors occur. Large planar surfaces such as walls, floors, and ceilings dominate the pixels in indoor scenes, and a supervised network can fit them to within a few degrees. RFception is less precise on these planes but better preserves the geometry of furniture and people, which occupy fewer pixels and therefore contribute less to pixel-averaged metrics. Figure 8 illustrates this contrast directly.

E RF Question-Answering Benchmark

Figures 11 and 12 show example scenes from the benchmark described in Section 4.4, together with the three questions and reference answers associated with each scene. The image shown for each example is the LiDAR reference depth used to annotate the reference answers. At test time, the same questions are posed using the representation provided to the VLM by each method. For RFception, this is the depth map predicted from RF alone.

Querying protocol. We query each model once per question and limit generation to 512 new tokens. In preliminary runs, Qwen3-VL-8B occasionally revises its reasoning repeatedly and reaches the token limit before producing a final answer. The prompt therefore instructs the model to reason briefly, avoid repeating or reconsidering alternative answers, and provide its best final answer without unnecessary revision. The same prompt, generation budget, and answer parser are used for every method and base VLM, so the only variable across rows of Table 2 is the input presented to the VLM.

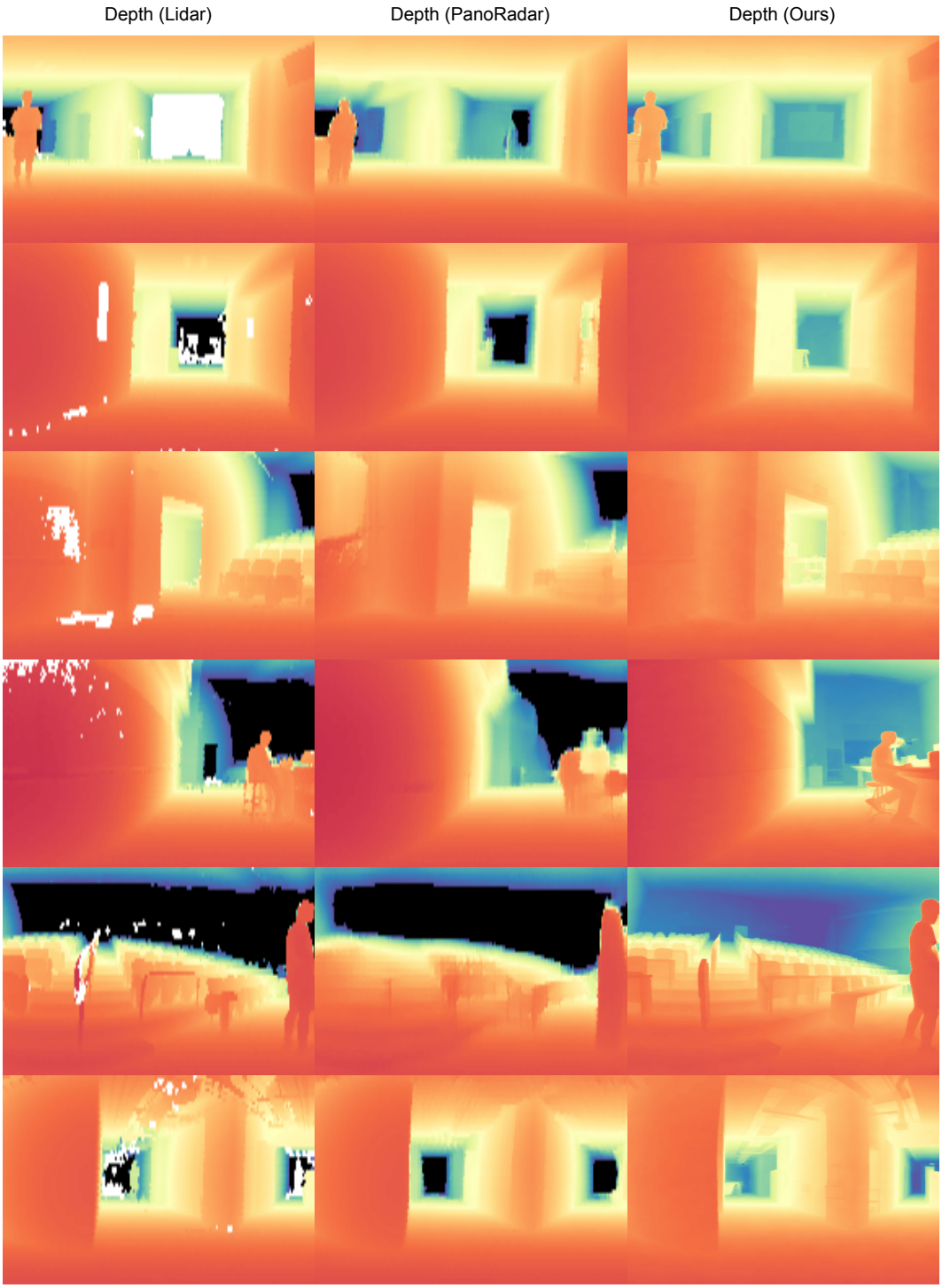


Figure 7: **Additional depth comparisons.** From left to right, the panels show the LiDAR ground truth, PanoRadar predictions, and RFception predictions. RFception preserves windows, doorways, and seated or standing people that PanoRadar tends to blur into the surrounding surfaces.

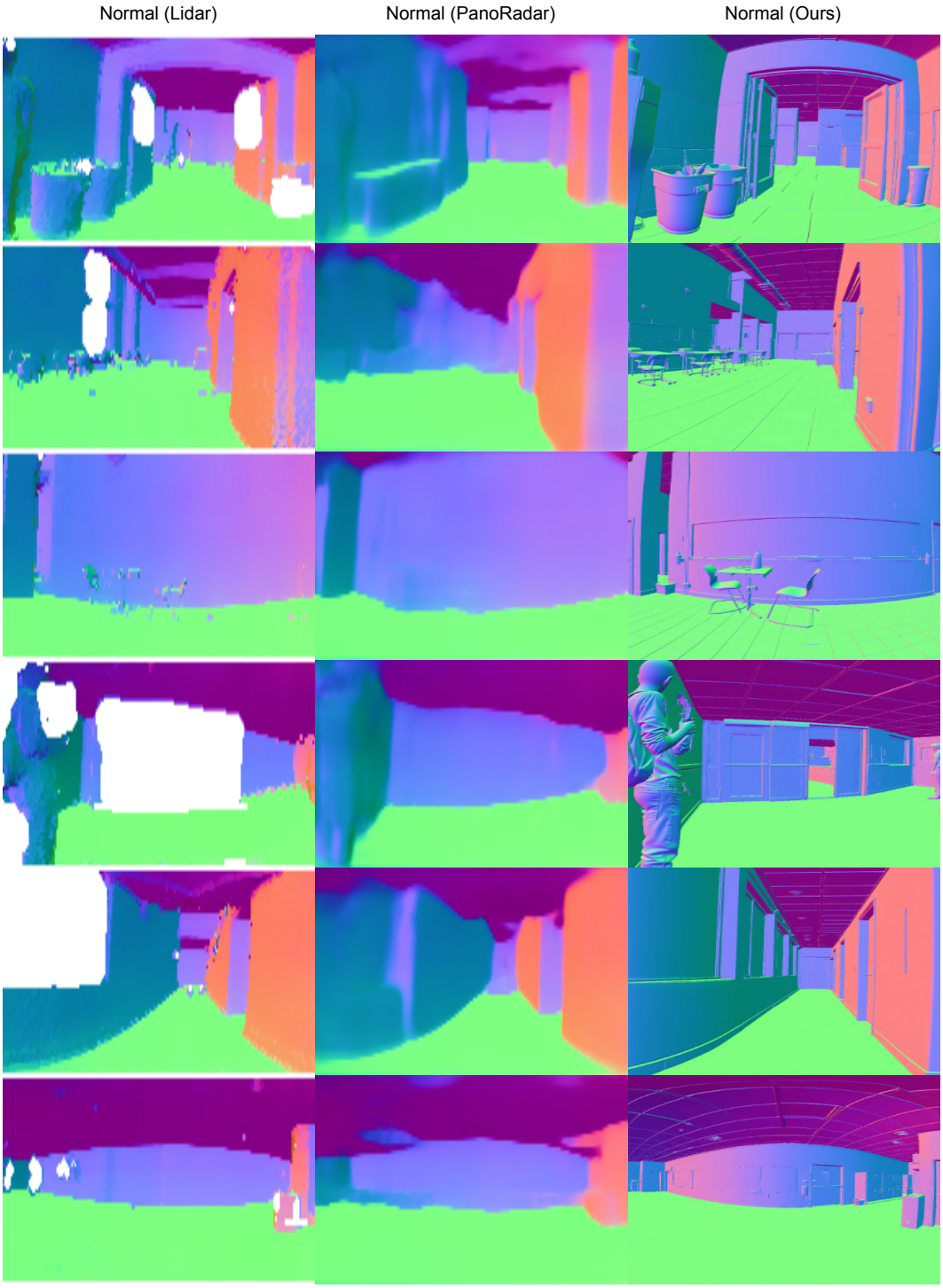


Figure 8: **Additional surface-normal comparisons.** From left to right, the panels show the LiDAR ground truth, PanoRadar predictions, and RFception predictions. Although RFception receives no normal supervision, it preserves the geometry of furniture and people that PanoRadar tends to smooth into the surrounding surfaces.

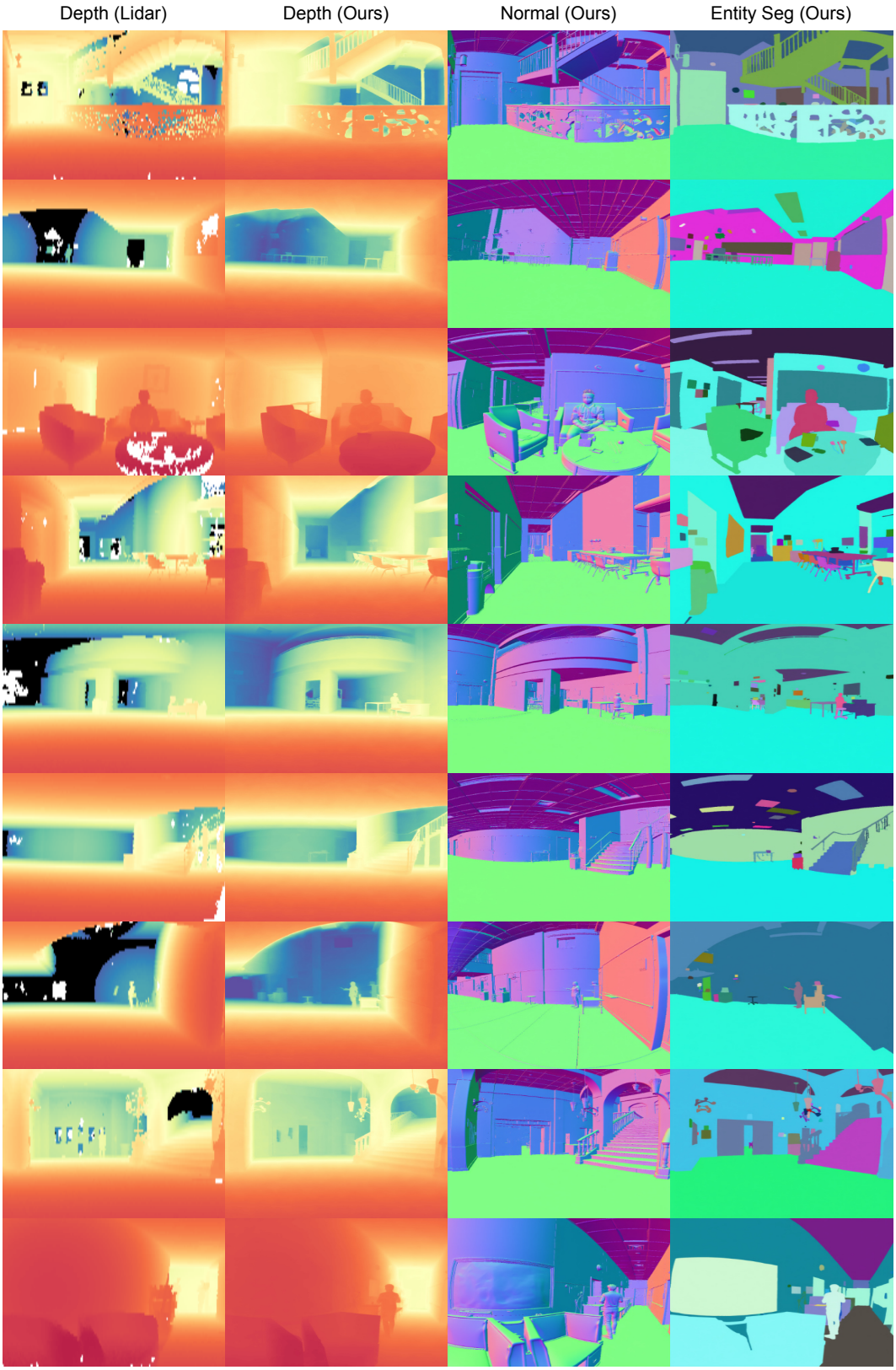


Figure 9: **Additional qualitative results of RFception (part 1).** Each row shows, from left to right, the LiDAR ground-truth depth and the depth, surface-normal, and entity-segmentation predictions produced by RFception from the same RF input. All three predictions share the same adapter and generative backbone and are selected solely by the text prompt.

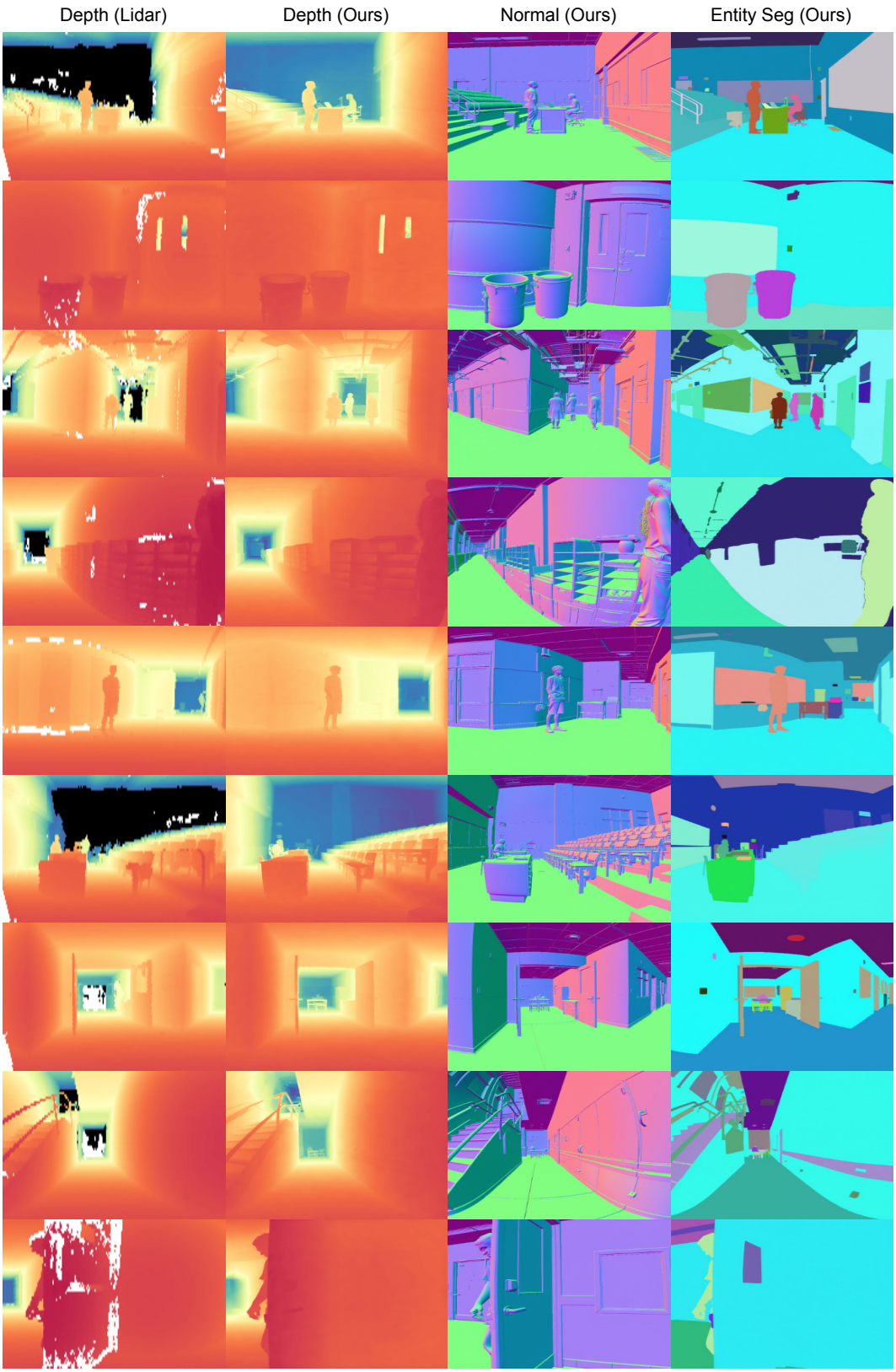
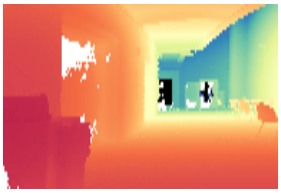


Figure 10: **Additional qualitative results of RFception (part 2).** Each row follows the same layout as Figure 9, showing the LiDAR ground-truth depth followed by the depth, surface-normal, and entity-segmentation predictions produced by RFception from the same RF input.

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, railing, door
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: person, railing, door.	person → door → railing
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: person, railing, trash can, stairs, door.	present nearest-to-farthest: person → door → railing; absent: trash can, stairs

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, chair/table, railing, stairs
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: person, stairs, floor.	floor → stairs → person
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: person, stairs, window, ceiling, floor.	present nearest-to-farthest: floor → stairs → person; absent: window, ceiling

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	chair/table, railing, trash can, window
	Spatial relations	Order these categories from left to right according to their predominant positions: chair/table, railing, trash can.	chair/table → trash can → railing
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: person, chair/table, trash can, stairs, window.	present nearest-to-farthest: trash can → chair/table → window; absent: person, stairs

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	chair/table, trash can, door
	Spatial relations	Order these categories from left to right according to their predominant positions: chair/table, trash can, door.	trash can → door → chair/table
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: railing, trash can, stairs, door, ceiling.	present nearest-to-farthest: trash can → door → ceiling; absent: railing, stairs

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: ceiling, wall, floor.	wall → floor → ceiling
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: chair/table, railing, ceiling, wall, floor.	present nearest-to-farthest: wall → floor → ceiling; absent: chair/table, railing

Figure 11: **Examples from the RF question-answering benchmark (part 1).** Each block shows one scene together with its three questions and reference answers, covering object understanding, spatial relations, and compositional reasoning. The displayed image is the LiDAR reference depth used to construct the reference answers.

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, door, window
	Spatial relations	Order these categories from left to right according to their predominant positions: person, door, window.	person → window → door
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: chair/table, railing, door, window, floor.	present nearest-to-farthest: floor → door → window; absent: chair/table, railing

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, railing, trash can, door
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: person, railing, wall.	wall → railing → person
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: person, chair/table, railing, stairs, wall.	present nearest-to-farthest: wall → railing → person; absent: chair/table, stairs

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, trash can
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: ceiling, wall, floor.	floor → wall → ceiling
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: stairs, door, ceiling, wall, floor.	present nearest-to-farthest: floor → wall → ceiling; absent: stairs, door

LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	person, chair/table, window
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: window, ceiling, wall.	wall → ceiling → window
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: trash can, stairs, window, ceiling, wall.	present nearest-to-farthest: wall → ceiling → window; absent: trash can, stairs

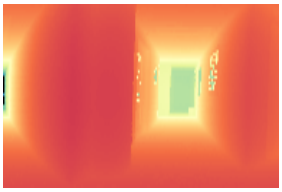
LiDAR Reference	Category	Question	Answer
	Object understanding	From the following candidate categories, list every category present in the current scene view: person, chair/table, railing, trash can, stairs, door, window.	door, window
	Spatial relations	Order these categories from nearest to farthest relative to the sensor: window, ceiling, wall.	wall → ceiling → window
	Compositional reasoning	For the following candidate categories, identify every category that is present in the current scene view and then order the present ones from nearest to farthest relative to the sensor: trash can, stairs, window, ceiling, wall.	present nearest-to-farthest: wall → ceiling → window; absent: trash can, stairs

Figure 12: **Examples from the RF question-answering benchmark (part 2).** Additional benchmark examples following the same layout as Figure 11. Each block shows the LiDAR reference depth together with the three questions and reference answers for object understanding, spatial relations, and compositional reasoning.