Believing is Seeing: Unobserved Object Detection using Generative Models

Subhransu S. Bhattacharjee, Dylan Campbell, Rahul Shome
School of Computing, The Australian National University
{subhransu.bhattacharjee, dylan.campbell, rahul.shome}@anu.edu.au

Abstract

Can objects that are not visible in an image—but are in the vicinity of the camera—be detected? This study introduces the novel tasks of 2D, 2.5D and 3D unobserved object detection for predicting the location of nearby objects that are occluded or lie outside the image frame. We adapt several state-of-the-art pre-trained generative models to address this task, including 2D and 3D diffusion models and vision–language models, and show that they can be used to infer the presence of objects that are not directly observed. To benchmark this task, we propose a suite of metrics that capture different aspects of performance. Our empirical evaluation on indoor scenes from the RealEstate10k and NYU Depth V2 datasets demonstrate results that motivate the use of generative models for the unobserved object detection task.

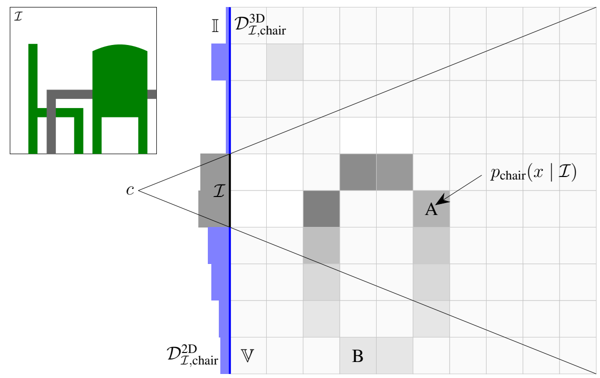


Figure 1. *Unobserved object detection* aims to infer the location of objects that were not directly observed in an image. Consider this toy example of a dining table and chairs. Here we visualize a top-down view (slice) of the predicted discrete distributions $\mathcal{D}^{2D}_{\mathcal{I}o}$ and $\mathcal{D}^{3D}_{\mathcal{I}o}$ for the object label o of "chair," conditioned on the image $\mathcal{I}$, where darker is more probable. The presence of an occluded chair (A) is predicted as relatively likely, as is the presence of an out-of-frame chair (B). Crucially, the domain $\mathbb{V}$ of the predicted 3D distribution exceeds the camera frustum (drawn in black), and the domain $\mathbb{I}$ of the 2D distribution extends beyond the image plane $\mathcal{I}$.

1. Introduction

Object detection [90] is a fundamental building block of autonomous systems that are capable of making sense of their surroundings. However, it is limited to visible surfaces within the camera's field-of-view. While amodal object detection and segmentation [35, 89] relaxes this requirement to permit partially visible objects, we propose to go further and detect those objects that are near the camera but were not observed at all.

In this paper, we propose the novel task of unobserved object detection in 2D, 2.5D and 3D for inferring the presence of objects not seen in an image within a domain that extends beyond the camera frustum and behind the visible surfaces. We are particularly motivated by perceptual and planning tasks such as visual search [84, 88], where a robot might be expected to use its understanding of a scene *beyond* what it can directly observe. For example, consider the setup in Fig. 1. A robot tasked with locating someone who is seated might find it useful to look to the right, where it is likely that additional chairs would be found, or to move to view a potentially occluded chair. Predicted spatial likelihoods of objects in a scene that was only partially observed can inform probabilistic models [28, 31, 45] for planning under uncertainty. In this work we recover such spatio-semantic distributions over observed and unobserved parts of the scene. We expect progress on this task to benefit downstream applications in which decisions are made in the presence of uncertainty arising from partial observations, such as active vision, navigation, visual search, adaptive planning and scene exploration [1, 5, 18, 24, 27, 63, 84].

While explicitly learning spatio-semantic distributions of scenes and objects is challenging [78], they can be variationally inferred using generative models trained on 2D image datasets [2, 30]. Novel view synthesis and semantic scene completion [7, 11, 17, 33, 55, 60, 73, 77] can offer useful spatio-semantic priors that can be used to render full 3D volumes given partial observations. However, occlusions and incomplete or sparse data still pose significant challenges to 3D scene generation methods. Image diffusion models [22, 70] have been shown to be useful at expanding beyond the visible frame, known as outpainting [51, 57], and some 3D diffusion models [73] can generate 3D representations consistent with an input image. These approaches typically aim for photorealism, of the completed scene, which may not be necessary or desirable for unobserved object detection, where we want to infer the spatio-semantic distribution of an object situated within possible scenes. Despite this weakness, the generative capability of such models is still highly useful.

In this work, we investigate the ability of three categories of generative models (2D and 3D diffusion models and vision–language models) to predict spatio-semantic distributions via a sampling procedure. We develop the associated detection pipelines, metrics, and evaluation protocol for the unobserved object detection task. Our contributions are

1. defining the novel task of unobserved object detection;
2. extending three major classes of generative models (2D and 3D diffusion models and vision–language models) to address unobserved object detection; and
3. benchmarking these approaches under a standardized evaluation protocol with respect to the accuracy and diversity of the predictions.

We evaluate performance under 2D, 2.5D, and 3D settings, on the RealEstate10k [87] and NYU Depth V2 [65] datasets using MS-COCO object categories [34]. The results from these pre-trained models is promising, but also reveals several challenges, motivating further research on the unobserved object detection task.

2. Related Work

Object detection [90] focuses on identifying and localizing objects within images, with recent deep learning models such as Faster R-CNN [54], YOLO [53], and transformer-based approaches like DETR [9] significantly enhancing both accuracy and efficiency. However, occlusions in real-world scenarios, such as autonomous driving and robotics, remains a critical challenge for traditional models. Amodal object detection and segmentation [35, 89] extends traditional approaches by predicting the full projected geometry of partially occluded objects. Li and Malik [32] introduced amodal bounding box regression including unseen parts. Recent improvements [6, 71] leverage Bayesian or hierarchical occlusion modeling. However, these methods cannot reason about fully occluded or out-of-frame objects.

2D diffusion models [22, 68, 70] generate high-quality, plausible images via an iterative denoising process, which can be guided by conditioning inputs such as text or images [29, 58]. Classifer-free guidance [21], capable of learning the conditional and unconditional models simultaneously, has been applied to scene completion and conditional synthesis [59, 72]. Stable Diffusion [51, 57] is used for extending high-resolution image boundaries, conditioned on text, in image outpainting, generating high-resolution images, conditioned on text [21]. We explore spatio-semantic relationships learned by such models in the context of our task.

Novel View Synthesis (NVS) [20, 67] is the task of generating new views of a scene from a set of posed input images. Neural Radiance Fields (NeRF) [41] parameterize scenes as continuous volumetric radiance fields, enabling high-fidelity rendering. However, NeRF typically requires many input images and known camera poses. Several related methods [56, 79, 82] leverage pre-trained scene priors to improve generalization in the sparse input setting, but the reconstruction is typically blurry in the unobserved regions. More recent approaches [2, 60, 83] integrate a diffusion model into the NVS pipeline to overcome these averaging issues. Diffusion with Forward Models (DFM) [73] combines a transformer encoder with PixelNeRF [49, 82] and uses a Denoising Diffusion Implicit Model (DDIM) [69] to generate diverse 3D representations conditioned on a single RGB image. GeNVS [11] follows a similar approach to sample novel views instead of full 3D scenes. Newer models like MultiDiff [44], using depth-warped priors, show improvements in quality and speed, but code has not yet been publicly released. Despite its computational demands and challenges with out-of-distribution data, DFM remains a strong benchmark for generating 3D-consistent views of a scene from a single image. We investigate whether DFM-generated scenes capture plausible spatio-semantic relationships for our task.

Vision–Language Models (VLMs) [85] are effective for spatio-semantic reasoning, excelling at one-shot and zero-shot tasks, but struggle with context-based reasoning. Models like CLIP [52] or ALIGN [25] which align visual and textual data for object recognition and segmentation, encounter difficulties in contexts with incomplete or sparse information [12]. These limitations are evident in tasks requiring spatial and compositional reasoning [86] and benchmarks [40, 74] which highlight difficulties in ambiguous scenes. While models like ChatGPT and LLaVA [12, 37, 38, 46] have improved multi-modal reasoning, they have limitations in understanding 3D spatial relationships [8]. Cirik et al. [14] leveraged contextual semantic co-occurrences, learned by language models, from the unobserved parts of a scene and relied on semantic affinities for active prediction. In contrast, our work evaluates whether modern VLMs [3, 4, 38, 46] can perform this passive inference, implicitly from a single observation.

3. Unobserved Object Detection

The task of unobserved object detection, as shown in Fig. 1, is to detect objects that are in the scene, but are not captured within the camera frustum. The approaches considered in this paper address this task by predicting a conditional distribution $\mathcal{D}_{\mathcal{I}}$ over a bounded space and set of semantic labels given a single RGB image $\mathcal{I}$, referred to as a spatio-

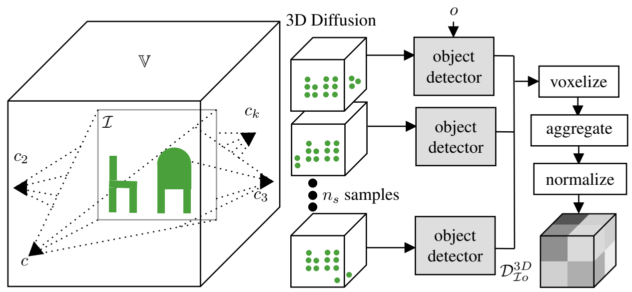


Figure 2. The 3D diffusion-based pipeline.

semantic distribution (SSD). This facilitates the inference of the probable locations of unobserved objects outside the visible region captured by the image and its associated camera frustum, using contextual cues from the image. The spatial domain $\mathcal{X}$ defines the support of this distribution. The SSD is an estimate of the ground-truth conditional distribution $\Gamma_{\mathcal{I}}$. While this distribution is not observable, we have access to additional posed images from a realization of this conditional distribution for the purpose of evaluation. For each object label o, we denote the associated SSD as $\mathcal{D}_{\mathcal{I}o}$. This maps each $x \in \mathcal{X}$ to a likelihood in $[0, 1]$. That is, the likelihood of detecting object o at location x. For example, it assigns a likelihood to the presence of a chair one meter in front of and two meters to the right of the camera, which may not be visible in the input image. Note that we consider all distributions to be normalized. We consider distributions over two discretized spatial domains $\mathcal{X}$: (1) a 2D domain $\mathbb{I}$, discretized as a pixel grid, corresponding to an expansion of the input image; and (2) a 3D domain $\mathbb{V}$, discretized as a voxel grid, extending beyond the input camera's frustum. The 2D domain $\mathbb{I}$ symmetrically extends the $H \times W$ pixel grid of the input image to $H \times W'$, where $W' > W$. The 3D domain $\mathbb{V}$ expands the frustum to an expanded voxel grid aligned with the input camera pose.

4. Generative Model-based Detection Pipelines

This section details how three pipelines based on generative models—(1) a 3D diffusion model (Fig. 2), a 2D diffusion model (Fig. 3), and (3) vision–language models (Fig. 4), which can be used to estimate the 2D or 3D SSD $\mathcal{D}_{\mathcal{I}o}$ from an RGB image $\mathcal{I}$, for object o.

3D Diffusion-based Model. We utilize Diffusion with Forward Models (DFM) [73] to estimate both the 2D and 3D $\mathcal{D}_{\mathcal{I}o}$. This approach generates a 3D representation conditioned on a single RGB image, from which an RGBD image can be rendered from any camera pose. We select a set of $k = 4$ target camera poses, including the input image's camera pose, as visualized in Fig. 2, render the corresponding RGBD images, and combine the results into a single point cloud. An off-the-shelf object detector, YOLOv8x [26, 53], is run on these frames. For 3D detection, the 2D bounding boxes for each object o are back-projected onto the point

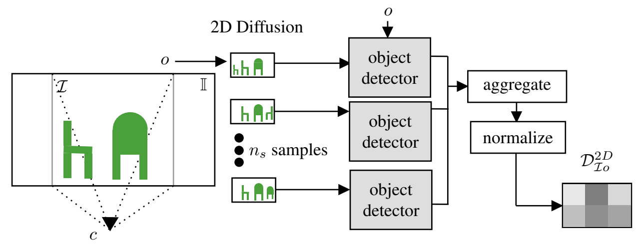


Figure 3. The 2D diffusion-based pipeline.

cloud alongside detection confidences. We take multiple samples from the conditional distribution by repeating the generation process a total of n_s times over additional camera poses, conditioned on the same input image $\mathcal{I}$, and the results are aggregated into the estimated distribution. We combine all the 3D samples into a common frame of reference, voxelize the resulting 3D confidence map into grid cells $\mathbb{V}$, average the non-zero confidences of the points across each voxel, and then apply softmax normalization to obtain the 3D distribution $\mathcal{D}^{3D}_{\mathcal{I}o}$. The DFM pipeline can expand the detection volume beyond the initial visible frame and recover occlusions in 3D using the novel camera views. More details on the sampling process is in Appendix C.

2D Diffusion-based Model. To evaluate the capacity of a 2D diffusion model at estimating the conditional distribution, we utilize a 2D diffusion model, SDXL [51, 75], which expands the image canvas of $\mathcal{I}$ using outpainting to perform conditional generation to fill the unobserved regions to the left and right of the input image (360 $\times$ 360 input to a 360 $\times$ 640 output). The process is repeated for n_s generated samples (shown in Fig. 3) We then run an object detector, YOLOv8x [26, 53] to assign confidence scores for object o to each pixel, followed by softmax normalization, resulting in a 2D SSD $\mathcal{D}^{2D}_{\mathcal{I}o}$. To obtain the 3D distribution $\mathcal{D}^{3D}_{\mathcal{I}}$, we re-projected 2D bounding boxes into 3D using metric depth values from DepthAnythingv2 [80, 81] and camera parameters from the RealEstate10k dataset, with additional parameters from COLMAP reconstructions [61, 62] for the NYU Depth V2 dataset, then voxelized the 3D point map, averaging confidence scores within each voxel.

VLM-based Models. We adapt several VLMs to the task of predicting the (coarsely-discretized) 2D conditional distribution: ChatGPT-4o [46], Claude-3.5 Sonnet [4], Gemini 1.5 Ultra [3] and LLaVa-34B-v1.6 [37, 38]. To do so, each model was provided with the 360 $\times$ 360 input image,

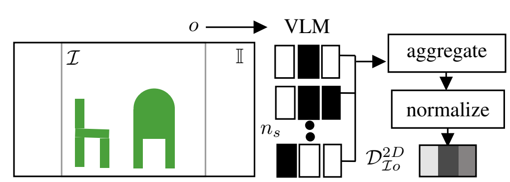


Figure 4. The VLM-based pipeline.

padded with 140 black pixels to the left and right, forming a 640×360 image. For each object o, the model was queried for a constrained binary "Yes"/"No" response whether the object was within the original frame or the extended region to the left or to the right. A fraction of "Yes" responses out of n_s queries was produces a prediction confidence for the left/center/right regions, further normalized to generate $\mathcal{D}^{\text{2D}}_{\mathcal{I}o}$ (Fig. 4) for region-wise estimates. Prompt and experiment design details are provided in Sec. 5.2 and Appendix D.

5. Experiments

We evaluate model performance on three distinct tasks—2D, 2.5D, and 3D—for indoor scenes [65, 87] with MS-COCO objects [34]. Each experiment generates SSDs, aligned with the input image's camera pose. The 2D task involves estimating $\mathcal{D}^{\text{2D}}_{\mathcal{I}o}$ within the expanded image domain $\mathbb{I}$ from the viewpoint of the input camera. The 2.5D task focuses on estimating the portion of $\mathcal{D}^{\text{3D}}_{\mathcal{I}o}$ visible in the expanded image $\mathbb{I}$, centered at the input camera location. Finally, the 3D task extends this estimation for $\mathcal{D}^{\text{3D}}_{\mathcal{I}o}$ over an expanded 3D region that accounts for out-of-frame detections and occlusions.

5.1. Metrics for Unobserved Object Detection

To evaluate the unobserved object detection task, we propose metrics to quantify how well the predicted conditional distribution $\mathcal{D}_{\mathcal{I}o}$ for an object o over a discrete domain $\mathcal{X}$ aligns with the true conditional distribution. However, we only have access to a single realization of this distribution, denoted by $\mathcal{G}_{\mathcal{I}o}$. Significantly, the true conditional distribution $\Gamma_{\mathcal{I}o}$ is likely to have more modes than the realization $\mathcal{G}_{\mathcal{I}o}$. For example, given an image $\mathcal{I}$ of a dining table, the ground-truth scene (the realization) happened to have several chairs clustered to the right of the camera. However, the exact same image could equally well have corresponded to a scene with a chair to the left of the camera (another, unrealized, scene conditioned on the same image). We expect the models to predict all plausible modes, and our metrics are informed by this asymmetry between the ground-truth and the prediction.

Normalized Entropy ($\mathcal{H}$) and Cross-Entropy ($\mathcal{H}^{\times}$). The normalized entropy $\mathcal{H}$ [64] of the predicted distribution $\mathcal{D}_{\mathcal{I}o}$ and the normalized cross-entropy $\mathcal{H}^{\times}$ [15] between the ground-truth distribution $\mathcal{G}_{\mathcal{I}o}$ and $\mathcal{D}_{\mathcal{I}o}$ are given by

$$\mathcal{H}_{\mathcal{I}o} := \frac{-1}{\log|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathcal{D}_{\mathcal{I}o}(x)\log\mathcal{D}_{\mathcal{I}o}(x), \tag{1}$$

$$\mathcal{H}^{\times}_{\mathcal{I}o} := \frac{-1}{\log|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathcal{G}_{\mathcal{I}o}(x)\log\mathcal{D}_{\mathcal{I}o}(x), \tag{2}$$

for input image $\mathcal{I}$ and object o. The normalization factor $\log|\mathcal{X}|$ makes the metric independent of the grid resolution and ensures that a uniform distribution evaluates to unit entropy. This asymmetric measure penalizes the prediction of low object likelihood where the ground-truth indicates the presence of an object. However, as required, it does not penalize the prediction of high object likelihoods where the ground-truth has not seen an object. For the remaining metrics, we consider high-likelihood detections above a threshold τ. The thresholded ground truth is denoted by $\mathcal{G}^{\tau}_{\mathcal{I}o}$ for all $x \in \mathcal{X}$ where $\mathcal{G}_{\mathcal{I}o}(x) > \tau$. Let the detection label $y_{\mathcal{I}o}$ be 1 if $\mathcal{G}^{\tau}_{\mathcal{I}o}$ is non-empty and 0 otherwise. The thresholded prediction $\mathcal{D}^{\tau}_{\mathcal{I}o}$ and label $\hat{y}_{\mathcal{I}o}$ are defined over $\mathcal{D}_{\mathcal{I}o}$.

Normalized Nearest Neighbor Distance (Δ). This asymmetric metric quantifies the spatial distance between the peaks (local maxima) in the thresholded ground truth $\mathcal{P}_{\mathcal{G}^{\tau}_{\mathcal{I}o}}$ to the nearest peaks in the thresholded predictions $\mathcal{P}_{\mathcal{D}^{\tau}_{\mathcal{I}o}}$. For a single input image $\mathcal{I}$ and object o, this is given by

$$\Delta_{\mathcal{I}o} := \frac{1}{|\mathcal{P}_{\mathcal{G}^{\tau}_{\mathcal{I}o}}|}\sum_{x\in\mathcal{P}_{\mathcal{G}^{\tau}_{\mathcal{I}o}}} \min_{x'\in\mathcal{P}_{\mathcal{D}^{\tau}_{\mathcal{I}o}}} \frac{\|x-x'\|}{\mathtt{diam}(\mathcal{X})}, \tag{3}$$

where $\mathtt{diam}(\mathcal{X})$ is the maximum distance within $\mathcal{X}$ (i.e., the diagonal in the 2D or 3D grid) that normalizes the measure with respect to the domain size. If predicted peaks $\mathcal{P}_{\mathcal{D}^{\tau}_{\mathcal{I}o}}$ is empty, $\Delta_{\mathcal{I}o} := \infty$. Large distances indicate that the object was not predicted near where it is located in the ground truth.

False Negative Rate (FNR). Measured over the entire set of evaluations (scenes $\mathcal{S}$ and objects $\mathcal{O}$), the thresholded ground-truth and predicted labels are used in this metric to compute the false negative rate of classification, given as.

$$\text{FNR} := \frac{\sum_{\mathcal{I}\in\mathcal{S}}\sum_{o\in\mathcal{O}} y_{\mathcal{I}o}(1-\hat{y}_{\mathcal{I}o})}{\sum_{\mathcal{I}\in\mathcal{S}}\sum_{o\in\mathcal{O}} y_{\mathcal{I}o}}. \tag{4}$$

2D Region-wise Accuracy ($\mathcal{A}$). This metric evaluates the coarse-grained classification accuracy of the predictor in 2D. Specifically, we decompose the domain $\mathcal{X}$ into ℓ subdomains $\mathcal{X}^i$ and define corresponding ground-truth and prediction labels for thresholded detections over the regions as $y^i_{\mathcal{I}o}$ and $\hat{y}^i_{\mathcal{I}o}$ respectively. For image $\mathcal{I}$ and object o, this is given by

$$\mathcal{A}_{\mathcal{I}o} := \frac{1}{\ell}\sum_{i=1}^{\ell} y^i_{\mathcal{I}o}\hat{y}^i_{\mathcal{I}o} + (1-y^i_{\mathcal{I}o})(1-\hat{y}^i_{\mathcal{I}o}). \tag{5}$$

5.2. Experiment Setup

In this section we provide details pertaining to the datasets, baselines, implementation, and evaluation protocol.

Datasets. All models were evaluated on 10 indoor scenes from the RealEstate10k test set [87] and the NYU Depth V2 dataset [65], automatically selected based on object detection frequency. After filtering, each scene has an average of 190 images in RealEstate10k (720×1280, resized to 360×640 for $\mathbb{I}$, and center-cropped to 360×360 for $\mathcal{I}$) and

1359 images in NYU Depth V2 (480×640, center-cropped to 360×640 for $\mathbb{I}$ and center-cropped to 360×360 for $\mathcal{I}$). One randomly selected frame per scene was used as the input image $\mathcal{I}$ for evaluation. COLMAP [61] and GLOMAP [48] were used for camera pose estimation (for NYU Depth V2) and for reconstruction to obtain depth maps. Object annotations for each expanded 360×640 image $\mathbb{I}$ were generated using YOLOv8x [26, 53], pre-trained on the MS-COCO dataset [34], retaining those with confidence greater than τ_{conf}. The 2D ground-truth spatial distributions $\mathcal{G}^{\text{2D}}_{\mathcal{I}o}$ were computed by a softmax normalization of the detected object confidence logits. The 3D ground-truth spatial distributions $\mathcal{G}^{\text{3D}}_{\mathcal{I}o}$ were computed by back-projecting the object confidence logits using the 3D reconstructions, voxelizing with average pooling, then applying softmax normalization. The voxel grid has resolution $20 \times 20 \times 20$ with grid size 1.0. The 2.5D ground truth spatial distributions $\mathcal{G}^{\text{2.5D}}_{\mathcal{I}o}$ were computed similarly, by back-projecting the confidence logits within the input camera frustum and using a voxel grid of resolution $10 \times 10 \times 10$, with a grid size of 1.0. Note that we only retain non-empty ground-truth distributions in the dataset: if a scene s does not contain an object o, (s, o) is not part of the dataset. See Appendices A and B for details on automatic scene pre-processing, filtering, selection, and reconstruction.

Baselines. To provide context for the benchmark, we report results for a *uniform* baseline, whereby equal probabilities are assigned across the grid. We also report results for an *oracle*, which has access to the ground truth. For the 2D task, the oracle uses the 2D ground-truth distribution corresponding to the expanded image $\mathbb{I}$. For the 2.5D task, the oracle uses the unprojected 2D ground-truth distribution. For the 3D task, the oracle uses the 3D ground-truth distribution.

Implementation Details. The hyperparameters are set as follows: the classification threshold $\tau = 1.4/|\mathcal{X}|$, the minimum detection confidence $\tau_{\text{conf}} = 0.1$, and the number of subdomains $\mathcal{X}^i$ for region-wise accuracy $\ell = 3$, corresponding to the 360×360 input image $\mathcal{I}$, the 360×140 domain to the left of the image and the 360×140 domain to the right of the image. For DFM [73] inference, we use autoregressive sampling with 3 timesteps per frame, generating 50 intermediate frames with default settings: a temperature of 0.85 and a guidance scale of 2.0. Generating each sample took 25 minutes per target pose per sample on average; that is, 48 days to generate all samples for each dataset.

For Stable Diffusion sampling [51], the Diffusers library SDXL inpainting model was employed. Input images were padded by 140 pixels on the left and right sides for masked inpainting. The temperature was set to 0.25, the prompt guidance scales were drawn from $\{0.25, 0.5, 0.75, 1.0\}$, the seeds from $\{0, 42, 123, 999, 2021\}$, and the prompts from a set of 10, yielding $4 \times 5 \times 10 = 200$ samples per scene. For the VLMs, we used ChatGPT-4o [46], Claude-Sonnet-3.5 [4], and Gemini 1.5 Ultra [3] via their official APIs, and LLaVa-v1.6-34b [37, 38] via the Replicate API. The temperature was set to 0.25, the top-p value was set to 1.0 where applicable, and the order of the image–question pairs was randomly shuffled. The VLMs were queried 100 times per object, region, and scene, resulting in 6000 calls each (20 images $\times$ 30 questions $\times$ 10 shuffles). Data acquisition and post-processing through these APIs took $\sim$ 8h. Details are in Appendices C and D. Future updates or access restrictions to APIs used in this study could affect reproducibility. All experiments were run on one NVIDIA RTX A6000 GPU. Our code and evaluation protocols are on the project page[1].

5.3. Results

In Tabs. 1 and 2, we report the performance of the generative models at the 2D, 2.5D and 3D unobserved object detection tasks. The metrics are reported as averages and standard deviations over scene–object combinations. The distance average excludes $\Delta = \infty$ entries; these are captured by the false negative rate. Methods were only evaluated on tasks for which they are applicable: we lift the VLM outputs from their binary labels to a 2D distribution, and lift the 2D diffusion outputs to a 2.5D distribution, but cannot obtain the 3D distribution from either. For the 2D task (Tab. 1), the 2D diffusion model shows the best performance, while the 3D diffusion model remains competitive. The coarse-grained VLMs struggle in most metrics with large spatial errors, Δ, as expected. For the 2.5D task (Tab. 2, left), the 3D diffusion model consistently outperforms the 2D diffusion model (SDXL) [51, 57], augmented with monocular depth estimates [81], with the exception of cross-entropy and the false negative rate when the outpainting model is prompted with the target object. Finally, for the 3D task (Tab. 2, right), we see that the 3D diffusion model remains competitive. Lower values of $\mathcal{H}^\times$ and Δ indicate better agreement with $\mathcal{G}_{\mathcal{I}o}$. For example, a low Δ indicates that the predicted modes that overlap with those in $\mathcal{G}_{\mathcal{I}o}$.

Qualitative results are shown in Fig. 5, with the 360×360 center-cropped input image $\mathcal{I}$ shown within the dashed lines in the first column, a top-down view of the DFM-predicted 3D distribution in the second column, the SDXL-predicted 2D distribution in the third column, and the ChatGPT-predicted 2D distribution in the last column. In particular, the DFM-based model usually predicts high likelihoods near the ground-truth object's 3D location, including when the object is occluded or out-of-frame, which is a significantly harder task than the 2D prediction task; see, for example, the final row. More qualitative results can be found in Appendix E. Failure cases are shown in Fig. 6, where all models perform poorly and an object category (door) that elicits many false negatives in the object detector. We also observe that the diffusion-based outpainting approach has its own modes of

[1] https://1ssb.github.io/UOD/

Table 1. Performance of all models at the 2D unobserved object detection task on the RealEstate10k [87] and NYU Depth V2 [65] datasets. We report the cross-entropy $\mathcal{H}^{\times}$, the prediction entropy $\mathcal{H}$, the nearest neighbor distance Δ, the false negative rate (FNR), and the region-wise classification accuracy $\mathcal{A}$. The best non-oracle result is indicated in **bold**.

	RealEstate10k [87]					**NYU Depth V2 [65]**				
Method	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta\%\downarrow$	FNR $\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta\%\downarrow$	FNR $\downarrow$	$\mathcal{A}\uparrow$
Uniform	1.000 ±0.000	1.000 ±0.000	∞	1.000	0.553 ±0.220	1.000 ±0.000	1.000 ±0.000	∞	1.000	0.558 ±0.214
Oracle	0.690 ±0.123	0.690 ±0.123	0.000 ±0.000	0.000	1.000 ±0.000	0.714 ±0.143	0.714 ±0.143	0.000 ±0.000	0.000	1.000 ±0.000
DFM	1.555 ±1.460	**0.774** ±0.119	5.990 ±39.470	0.032	0.747 ±0.139	1.661 ±1.336	**0.696** ±0.128	6.217 ±36.112	0.055	0.765 ±0.236
SDXL	1.257 ±2.033	0.848 ±0.171	**0.446** ±25.490	**0.000**	**0.918** ±0.147	**1.223** ±2.107	0.778 ±0.212	**0.510** ±26.929	**0.000**	**0.944** ±0.142
ChatGPT-4o	1.332 ±0.265	0.968 ±0.056	25.448 ±13.504	0.024	0.816 ±0.251	1.415 ±0.935	0.981 ±0.043	18.118 ±26.556	0.000	0.810 ±0.427
Claude-3.5-Sonnet	1.880 ±1.240	0.935 ±0.053	22.107 ±23.443	0.076	0.772 ±0.148	1.575 ±1.553	0.976 ±0.040	19.091 ±22.609	0.024	0.750 ±0.280
Gemini-1.5-Ultra	1.968 ±0.541	0.968 ±0.044	24.389 ±12.981	0.124	0.733 ±0.286	1.896 ±0.909	0.957 ±0.040	16.911 ±15.458	0.039	0.705 ±0.277
LLaVa-v1.6-34b	**1.232** ±0.602	0.945 ±0.050	25.321 ±15.532	0.262	0.770 ±0.276	1.455 ±0.594	0.970 ±0.043	24.369 ±17.803	0.142	0.743 ±0.283

Table 2. Performance of all models on the 2.5D and 3D unobserved object detection tasks. We report the cross-entropy $\mathcal{H}^{\times}$, prediction entropy $\mathcal{H}$, nearest neighbor distance Δ, and false negative rate FNR. The best non-oracle result is in **bold**.

	2.5D Task								**3D Task**							
	RealEstate10k [87]				**NYU Depth V2 [65]**				**RealEstate10k [87]**				**NYU Depth V2 [65]**			
Method	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$
Uniform	1.000 ±0.000	1.000 ±0.000	∞	1.000	1.000 ±0.000	1.000 ±0.000	0.000 ±0.000	1.000	1.000 ±0.000	1.000 ±0.000	∞	1.000	1.000 ±0.000	1.000 ±0.000	∞	1.000
Oracle	0.238 ±0.014	0.238 ±0.014	0.000 ±0.000	0.000	0.297 ±0.006	0.297 ±0.006	0.000 ±0.000	0.000	0.288 ±0.136	0.288 ±0.136	0.000 ±0.000	0.000	0.168 ±0.176	0.168 ±0.176	0.000 ±0.000	0.000
DFM	1.955 ±2.033	**0.530** ±0.441	**5.128** ±3.042	0.032	1.785 ±2.101	**0.610** ±2.077	**4.214** ±2.803	0.055	**2.471** ±4.415	**0.303** ±0.202	**9.190** ±6.223	**0.017**	**2.062** ±3.915	**0.541** ±0.374	**3.948** ±5.801	**0.000**
SDXL	**1.752** ±2.043	0.617 ±0.305	6.277 ±2.229	**0.000**	**1.533** ±2.916	0.655 ±0.323	5.245 ±2.351	**0.000**	–	–	–	–	–	–	–	–

Table 3. Ablation and analysis on the combined RealEstate10k and NYU Depth V2 dataset. The best variant of each method is in **bold**.

	2D Task					**2.5D Task**				**3D Task**			
Method	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$	$\mathcal{H}^{\times}\downarrow$	$\mathcal{H}\downarrow$	$\Delta(\%)\downarrow$	FNR $\downarrow$
DFM (25 samples, 4 poses)	**1.595**	**0.744**	**6.076**	**0.04**	**0.754**	**1.890**	**0.560**	**4.781**	**0.04**	**2.303**	**0.401**	**7.041**	**0.01**
DFM (25 samples, 3 poses)	1.944	0.860	19.019	0.14	0.612	2.037	0.808	12.339	0.13	3.226	0.558	8.797	0.08
DFM (25 samples, 2 poses)	2.466	0.931	14.734	0.88	0.567	2.562	0.994	13.205	0.86	4.931	0.703	11.105	0.38
DFM (15 samples, 4 poses)	1.804	0.925	16.059	0.58	0.592	2.140	0.873	8.759	0.67	4.047	0.858	9.980	0.07
DFM (10 samples, 4 poses)	2.342	0.952	14.787	0.72	0.574	2.480	0.996	17.678	0.75	4.695	0.965	12.481	0.22
SDXL (w/ object cues)	**1.244**	**0.870**	**0.470**	**0.00**	**0.928**	**1.668**	**0.631**	**5.884**	**0.00**	–	–	–	–
SDXL (w/o object cues)	1.977	0.885	6.037	0.49	0.716	2.129	0.860	16.051	0.48	–	–	–	–
SDXL (w/o prompts)	2.963	0.926	6.482	0.51	0.767	2.229	0.952	17.949	0.51	–	–	–	–

failure, also shown in Fig. 6, where the results are not informative in a scene if the model is not prompted with the label of the object to be detected.

5.4. Ablation Studies & Analysis

In Tab. 3, we report the ablation study and analysis for the 2D and 3D diffusion-based detectors. First, we analyze the effect of reducing the number of camera poses k at which the DFM model is queried, from 4 to 2, which affects the cov-

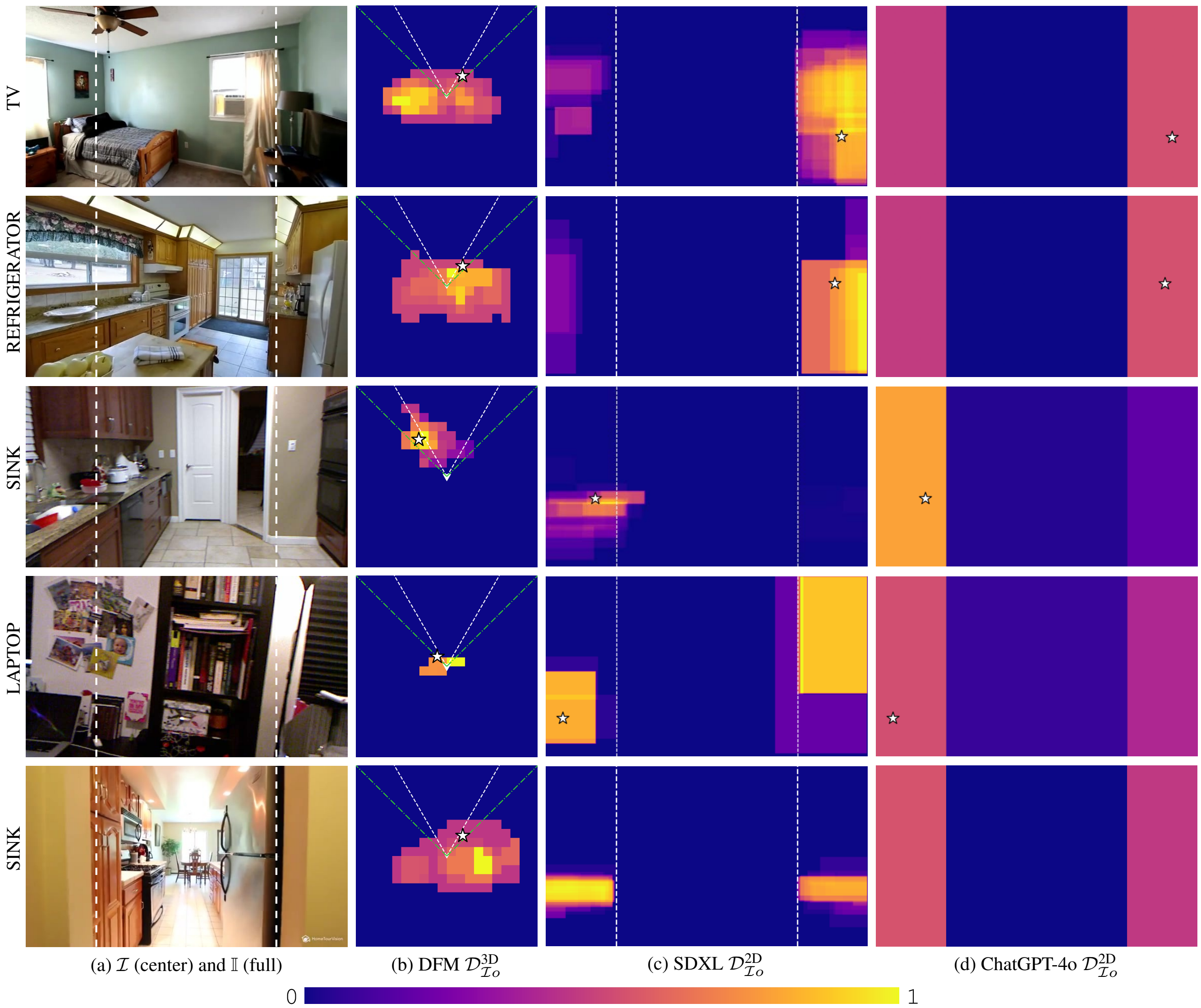
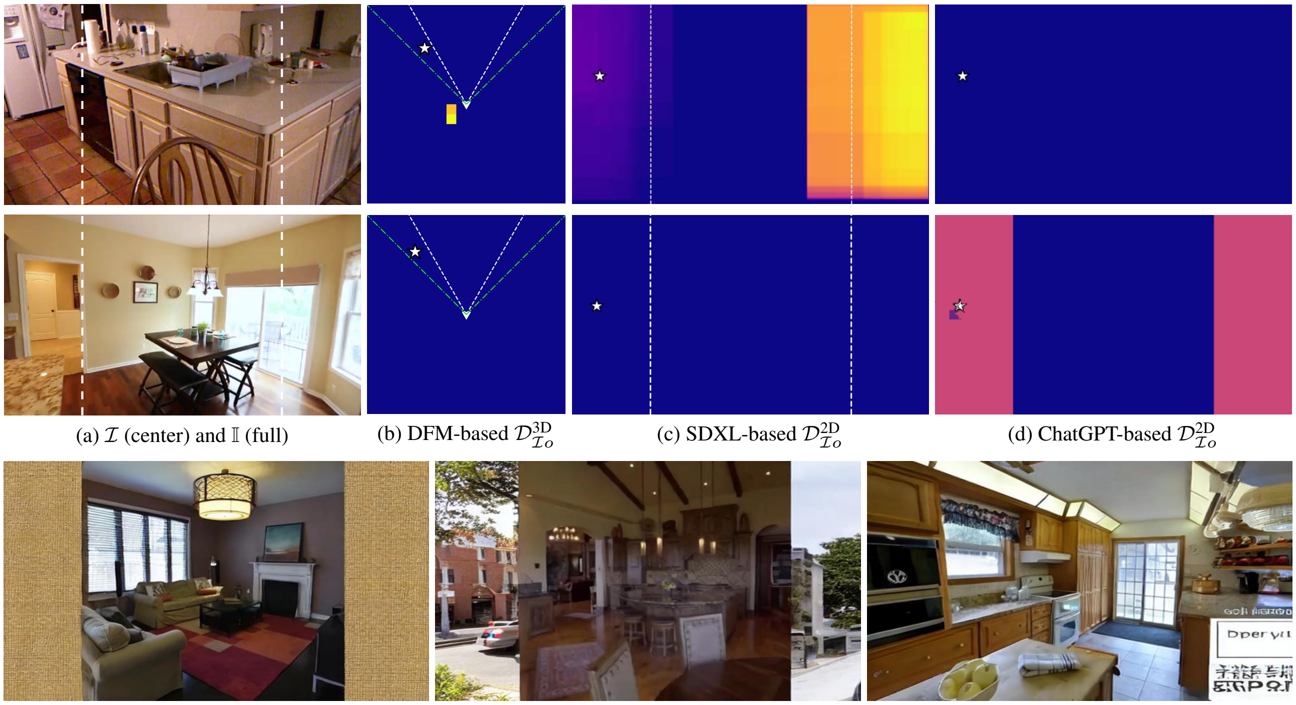


Figure 5. **Qualitative results.** Each row shows the predicted 2D and top-down 3D spatial distributions generated by each method for various object categories: TV (first row), refrigerator (second row), sink (third row), laptop (fourth row), and sink (fifth row). Notably, in the bottom row, the DFM-based model infers the likely presence of a sink, occluded by the refrigerator, albeit not with a high likelihood. A white triangle marks the camera position, while dashed and dot-dashed lines depict the camera frustums for $\mathcal{I}$ and $\mathbb{I}$. The white star indicates the ground-truth position of the object, when visible in 2D. Heatmap colors indicate object likelihood, with warmer tones representing higher probabilities. Since these are spatially-normalized distributions, we use a log-scale for visualization.

erage of the scene. To do so, we randomly drop one or two camera poses and report the performance. We find that the performance decreases considerably as the number of camera poses decreases, across all tasks. Second, we analyze the effect of reducing the number of conditionally-independent samples per additional camera pose from 25 to 10, which reduces the diversity of the predicted conditional distribution which negatively affects performance. We also ablate the 2D diffusion-based approach, investigating the effect of object cues in the text prompts on the results, evaluating the pipeline (a) without prompts; (b) without object cues, where the model is provided with text guidance but without referring to objects (*e.g.*, "extend this indoor scene naturally to the left and right of the given frame"); and (c) with object cues in the prompts (*e.g.*, "extend this scene naturally on both sides, with objects like [OBJECT]"). Prompts with object cues appear to significantly impact performance of the SDXL-based method in the task, indicating that these methods make heavy use of textual guidance in the generative process, as demonstrated in the 2D and 2.5D ablation study.



(a) $\mathcal{I}$ (center) and $\mathbb{I}$ (full) (b) DFM-based $\mathcal{D}^{3D}_{\mathcal{I}o}$ (c) SDXL-based $\mathcal{D}^{2D}_{\mathcal{I}o}$ (d) ChatGPT-based $\mathcal{D}^{2D}_{\mathcal{I}o}$

(e) SDXL w/o object cue prompt failure (f) SDXL w/o object cue prompt failure (g) Failure with object cues

Figure 6. **Failure cases.** In the first row, the DFM-based model misses plausible locations where a refrigerator might be and the ChatGPT-4o-based model fails to report any detections. In the second row, the presence of the door on the left is not inferred by the DFM or SDXL-based models, primarily due to the object detector's false negatives on synthetic images for that class. The last row shows outpainting failure cases. The SDXL outpainting model, when prompted without object cues, often extends images with textures (e) or unrelated imagery (f). Even when object cues (here, "refrigerator") are provided (g), results may ignore the prompt and generate unnaturally extended images.

Details of the complete analysis and ablation study can be found in the Appendix E. We used default hyperparameters without any dataset-specific changes for all our experiments in the generation process (details in Appendices A, C and D).

6. Conclusion

This work has introduced the unobserved object detection task in 2D, 2.5D and 3D, has presented three detection pipelines using state-of-the-art generative models, and has evaluated their performance on a proposed set of metrics in a dataset of indoor scenes and common objects. We find that unobserved object detection presents different characteristics in 2D and 3D. Expanding the camera frame can be used in both 2D and 3D, with detection pipelines showing the ability to predict objects that may be fully outside the visible frustum. In contrast, predicting objects within 3D regions beyond the camera frustum suggests that 3D pipelines have the capability to detect fully occluded objects. While the 2D SDXL-based pipeline and VLMs cannot detect occluded objects, the 3D pipeline leveraging DFM achieves this.

This work has several significant limitations. In particular, the inference time of 3D diffusion models, is currently very large and are costly to re-train. Since our approaches require many samples to approximate the conditional distribution, this is a significant bottleneck and stymies their use in real-time applications like in robotics. Another limitation is the need for object prompts in outpainting-based methods, since this is likely to bias the obtained empirical conditional distribution. Also, VLM-based evaluations may not be fully reliable, since it is possible these models could have seen and memorized the test images during training. Since the task requires that the model not over-fit to the ground truth realization—which is an instance of the true conditional distribution—there are inherent challenges in measuring performance that reflects accurate modeling of the underlying true distribution that is unavailable in real-world settings.

Future applications will apply the proposed pipeline to outdoor and large-scale scenes [36]. Synthetic datasets [13] can let us access known conditional distributions in simulation [16]. Beyond object detection, different spatio-semantic features like occupancy or affordances [23] are promising next steps. Improvements in solving this task could pave the way for robotics applications in reactive control [47], visual search [5], planning [27], navigation [18], and exploration [24]. This motivates the current work as a significant step toward improving machine vision capabilities for unobserved object detection.

Acknowledgments

Subhransu is supported by the University Research Scholarship at the Australian National University. This research was partially funded by the U.S. Government under DARPA TIAMAT HR00112490421. The views and conclusions expressed in this document are solely those of the authors and do not represent the official policies or endorsements, either expressed or implied, of the U.S. Government. This research was also funded by the Australian Research Council under ITRH IH210100030.

References

[1] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2), 2022. 1

[2] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J. Mitra, and Paul Guerrero. RenderDiffusion: Image diffusion for 3D reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[3] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. 2, 3, 5, 18

[4] Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. 2, 3, 5, 18

[5] Alper Aydemir, Andrzej Pronobis, Moritz Göbelbecker, and Patric Jensfelt. Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*, 29(4), 2013. 1, 8

[6] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *Proceedings of the International Conference on Robotics and Automation*, 2022. 2

[7] Miguel Ángel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander T Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Joshua M. Susskind. GAUDI: A neural architect for immersive 3D scene generation. In *Proceedings of the Conference on Neural Information Processing Systems*, 2022. 1

[8] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. PUG: Photorealistic and semantically controllable synthetic data for representation learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023. 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[10] Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert. Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2015. 13

[11] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3D-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 16

[13] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023. 8

[14] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. HOLM: Hallucinating objects with language models for referring expression recognition in partially-observed scenes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022. 2

[15] Thomas M. Cover and Joy A. Thomas. Differential entropy. In *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. 4

[16] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[17] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360 (6394), 2018. 1

[18] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79), 2023. 1, 8

[19] Rafael C. Gonzalez and Richard E. Woods. Intensity transformations and spatial filtering. In *Digital Image Processing*, chapter 3. Pearson Education, 3rd edition, 2008. 15

[20] Richard Hartley and Andrew Zisserman. N-view computational methods. In *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2, 15

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Conference on Neural Information Processing Systems*, 2020. 2

[23] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do As I Can, Not As I Say: Grounding language in robotic affordances. In *Proceedings of the Annual Conference on Robot Learning*, 2022. 8

[24] Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 8

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 2

[26] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO. https://ultralytics.com, 2023. 3, 5, 13, 14

[27] Pascal Jutras-Dubé, Ruqi Zhang, and Aniket Bera. Adaptive planning with generative models under uncertainty. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2024. 1, 8

[28] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D. Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *Proceedings of the International Conference on Robotics and Automation*, 2019. 1

[29] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[30] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Proceedings of the Conference on Neural Information Processing Systems*, 2021. 1

[31] Hanna Kurniawati. Partially observable markov decision processes and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1), 2022. 1

[32] Ke Li and Jitendra Malik. Amodal instance segmentation. In *Proceedings of the European Conference on Computer Vision*, 2016. 2

[33] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. InfiniteNature-Zero: Learning perpetual view generation of natural scenes from single images. In *Proceedings of the European Conference on Computer Vision*. Springer, 2022. 1

[34] 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 *Proceedings of the European Conference on Computer Vision*, 2014. 2, 4, 5, 13

[35] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. In *Proceedings of the Conference on Neural Information Processing Systems*, 2020. 1, 2

[36] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3DV-10k: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 8

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023. 2, 3, 5, 18

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 5, 18

[39] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1999. 14

[40] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[41] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[42] Raúl Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. In *IEEE Transactions on Robotics*. IEEE, 2017. 13

[43] Raúl Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5), 2015. 13

[44] Norman Müller, Katja Schwarz, Barbara Roessle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. MultiDiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[45] Keiko Nagami and Mac Schwager. State estimation and belief space planning under epistemic uncertainty for learning-based

perception systems. *IEEE Robotics and Automation Letters*, 9(6), 2024. 1

[46] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2, 3, 5, 13, 18

[47] Kei Ota, Devesh Jha, Tadashi Onishi, Asako Kanezaki, Yusuke Yoshiyasu, Yoko Sasaki, Toshisada Mariyama, and Daniel Nikovski. Deep reactive planning in dynamic environments. In *Proceedings of the Conference on Robot Learning*, 2021. 8

[48] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision*, 2024. 5, 15

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[50] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013. 15

[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*, 2024. 2, 3, 5

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3, 5, 13

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 2016. 2

[55] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3D scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[56] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3D-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5

[58] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM Transactions on Graphics*, 2022. 2

[59] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the Conference on Neural Information Processing Systems*, 2022. 2

[60] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[61] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2016. 3, 5, 15

[62] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 3, 15

[63] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *Proceedings of the Annual Conference on Robot Learning*, 2023. 1

[64] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 1948. 4

[65] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. In *Proceedings of the European Conference on Computer Vision*. Springer, 2012. 2, 4, 6, 13, 14, 18

[66] Josef Sivic, Michael Isard, Andrew Zisserman, James Philbin, and Ondrej Chum. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2007. 15

[67] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. In *ACM Transactions on Graphics*, 2006. 2

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2015. 2

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[70] Yang Song and Stefan Ermon. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[71] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a Bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[72] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Niessner. DiffuScene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[73] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Fredo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 5, 15, 16, 18

[74] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[75] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3, 17

[76] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004. 15

[77] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2023. 1

[78] Chuan Wen, Dinesh Jayaraman, and Yang Gao. Can transformers capture spatial relations between objects? In *Proceedings of the International Conference on Learning Representations*, 2023. 1

[79] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. SinNeRF: Training neural radiance fields on complex scenes from a single image. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[80] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 16

[81] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Proceedings of the Conference on Neural Information Processing Systems*, 2024. 3, 5, 16

[82] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[83] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[84] Zhen Zeng, Adrian Röfer, and Odest Chadwicke Jenkins. Semantic linking maps for active visual object search. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021. 1

[85] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[86] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. In *Proceedings of the International Conference on Machine Learning*, 2024. 2

[87] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *Proceedings of ACM SIGGRAPH*, 2018. 2, 4, 6, 13, 14, 15, 18

[88] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017. 1

[89] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[90] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 1, 2