This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ActiveGrasp: Information-Guided Active Grasping with Calibrated Energy-based Model

Boshu Lei[1]* Wen Jiang[1]* Kostas Daniilidis[1,2]
[1] University of Pennsylvania [2] Archimedes, Athena RC
{leiboshu, wenjiang}@seas.upenn.edu, kostas@cis.upenn.edu

Abstract

Grasping in a densely cluttered environment is a challenging task for robots. Previous methods tried to solve this problem by actively gathering multiple views before grasp pose generation. However, they either overlooked the importance of the grasp distribution for information gain estimation or relied on the projection of the grasp distribution, which ignores the structure of grasp poses on the SE(3) manifold. To tackle these challenges, we propose a calibrated energy-based model for grasp pose generation and an active view selection method that estimates information gain from grasp distribution. Our energy-based model captures the multi-modality nature of grasp distribution on the SE(3) manifold. The energy level is calibrated to the success rate of grasps so that the predicted distribution aligns with the real distribution. The next best view is selected by estimating the information gain for grasp from the calibrated distribution conditioned on the reconstructed environment, which could efficiently drive the robot to explore affordable parts of the target object. Experiments on simulated environments and real robot setups demonstrate that our model could successfully grasp objects in a cluttered environment with limited view budgets compared to previous state-of-the-art models. Our simulated environment can serve as a reproducible platform for future research on active grasping. The source code of our paper will be made public when the paper is released to the public.

1. Introduction

Grasping in clutter is a fundamental challenging task in robot manipulation, since the initial given views could not provide useful information to predict successful grasps. Many attempts have been made to improve the performance of the model in a clutter environment by training better grasping models with larger amounts of data from real-world data and massive simulations [17, 23, 43, 46]. To the best of these methods, even large Vision-Language-Action (VLA) models like $\pi_{0.5}$ [28] struggle to grasp an object in a cluttered environment due to a lack of active information gathering. To solve this, a series of works [5, 45, 48, 71] actively take the next-best-view (NBV) by selecting the view with the highest information gain for the grasping task before prediction.

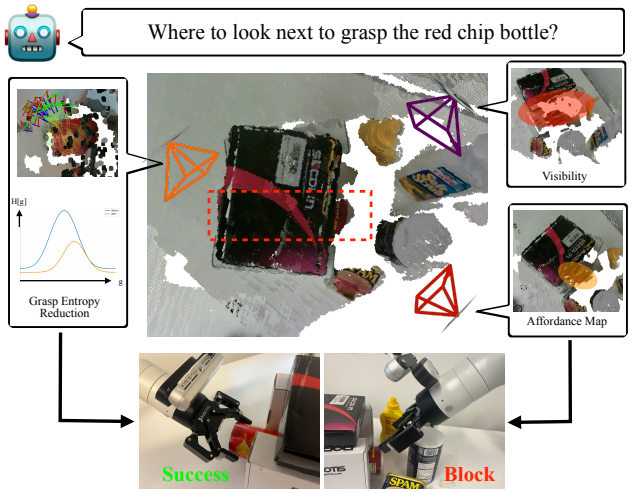


Figure 1. **ActiveGrasp Overview** The ActiveGrasp framework actively selects the next view with the highest information gain for the grasp task. Previous methods estimate the information gain using visibility or the affordance map, focusing on regions where visual features are rich but infeasible for grasp. In contrast, our method estimates information gain as entropy reduction of grasp pose distribution, selecting a region that observes highly uncertain grasps.

A key challenge for NBV selection is to estimate the information gain of candidate views for the grasping task. From information theory, we provide three necessary criteria for an unbiased estimate of information gain. First, the information gain should be computed from the grasp distribution instead of visibility [1, 5, 34], which will bias towards scene completion. Second, the grasp distribution should be formulated on the SE(3) manifold instead of 2D [45, 71] or 3D projection [48], which will bias towards the position of

*equal contribution

grasps. Third, the grasp distribution should be calibrated to the real grasp distribution such that the information gain computed on top of it is unbiased.

To meet all the criteria above, we propose ActiveGrasp, a novel approach to estimate the information gain for grasping tasks by using a calibrated energy-based model to estimate the grasp distribution on the SE(3) manifold. Rather than relying on other surrogates to estimate information gain, we define the information gain as the entropy reduction of the grasp distribution and estimate the entropy using the energy-based model, providing more task-specific and accurate information gain estimation.

To be specific, we define the entropy of grasp poses as the expectation for the entropy of a single grasp. The entropy of a single grasp can be computed using the success rate, as we model every single grasp as a Bernoulli distribution. The expectation is computed using the Gaussian Approximation of Posterior [37] (GAP) of the scene and our calibrated energy-based model, whose energy level aligns with the success rate. The next best view is selected by maximizing the entropy reduction given the candidate view without any other heuristics. Our experiments, both in simulation and the real world, show that the proposed ActiveGrasp framework achieves the highest success rate given a limited view budget. We summarize our contributions as follows:

- We give a clear definition of entropy of grasp distribution and derive the information gain of candidate views for the grasping task from an information-theoretic perspective.
- We estimate the entropy using the calibrated energy-based model and GAP such that the information gain estimation is accurate.
- We present an active grasping benchmark with YCB [7] objects and a physically informed simulator, enabling reproducible results for further research in the community.

2. Related Works

Grasp Learning on Manifolds Robotic Grasping is a vibrant research problem and has been studied for a long time. We refer the readers to literature reviews for broader interests on the robotic grasping [38, 52, 70]. Here, we mainly focus on grasp learning works on the SE(2) or SE(3) manifolds. To learn grasp poses on the SE(2) manifolds, early works used Sample and Test methods [41, 46, 51] to uniformly sample candidates on the manifold and train a classifier to test each sample. Inspired by object detection in the computer vision community, later works like [11, 33, 57] propose candidate regions and regress the parameters of the grasp pose on the manifold. Learning grasp poses on the SE(3) manifold differs from learning on the SE(2) manifold in terms of network architecture, such as 3D convolution [4, 6] and point models [66] To acquire a distribution of grasp poses, generative modeling of grasps on the manifold became popular recently. Weng *et al*. [67] develop CAPGrasp, an $R^3 \times SO(2)$-equivariant grasp pose generative model under the assumption of approach-constrained grasp. For the SE(3) manifold, Lim [44] learns an SE(3) equivariant flow field to generate grasp poses on the SE(3) manifold. Hu *et al*. [27] maps each point in the cloud to a continuous grasp quality function over the 2-sphere S^2 and generates samples along the perpendicular direction of the surface normal. The closest work to ours is Urain [64], which trains an energy-based model on the SE(3) manifold and runs the denoising steps to generate successful grasps. We leverage the energy function learned from [64] to compute the information gain for next-best-view selection.

Active Perception Active Perception [2, 3] has long been a significant area of research in robotics and computer vision. Various methods have been applied for uncertainty quantification for radiance field and 3DGS, including training neural networks [19, 24, 25, 32, 49, 55], reinforcement learning [8, 54], point cloud completion [14]. Second-order information, such as Fisher Information, is also employed for uncertainty estimation and next-best-view selection [20, 30, 31, 69].

Active perception on robotic grasping has mostly focused on other data representations. Gualtieri *et al*. [22] utilized viewpoint direction for view selection. Arruda *et al*. [1] select views that could refine the surface reconstruction quality near the model-predicted contact points. Breyer *et al*. [5] selects the next best view by counting the number of previously occluded voxels in the candidate view. Gregory *et al*. [34] parametrize the occluded region as a mixture of Gaussians and minimize uncertainty within the region. Chen *et al*. [10] uses visibility as one of the reward signals to train an RL policy to select viewpoints. Douglas *et al*. [48] parametrize the model prediction as a Bernoulli distribution and compute entropy within 3D grids. They select views that observe the largest entropy grid cells. Zhang *et al*. [71] builds a tri-plane feature and trains a network to predict the affordance score in unobserved views and perform view selection among them. Ma *et al*. [45] uses the inconsistency between the prediction from the graspness and the TSDF feature planes as information gain for view selection.

Model Calibration Model calibration aims to adjust a model so that its predicted probabilities better reflect the true likelihood of outcomes. Model calibration methods can be categorized into regularization, uncertainty estimation, and post-processing [18]. Regularization methods modify the training phase by label smoothing [62] or adding extra losses like weighted confidence-integrated loss [59] or entropy term [50]. Uncertainty estimation predicts an accurate estimation of the data uncertainty for calibration by using the Bayesian Network [29] or model ensembles [47]. Post-processing tunes the model on an extra calibration set

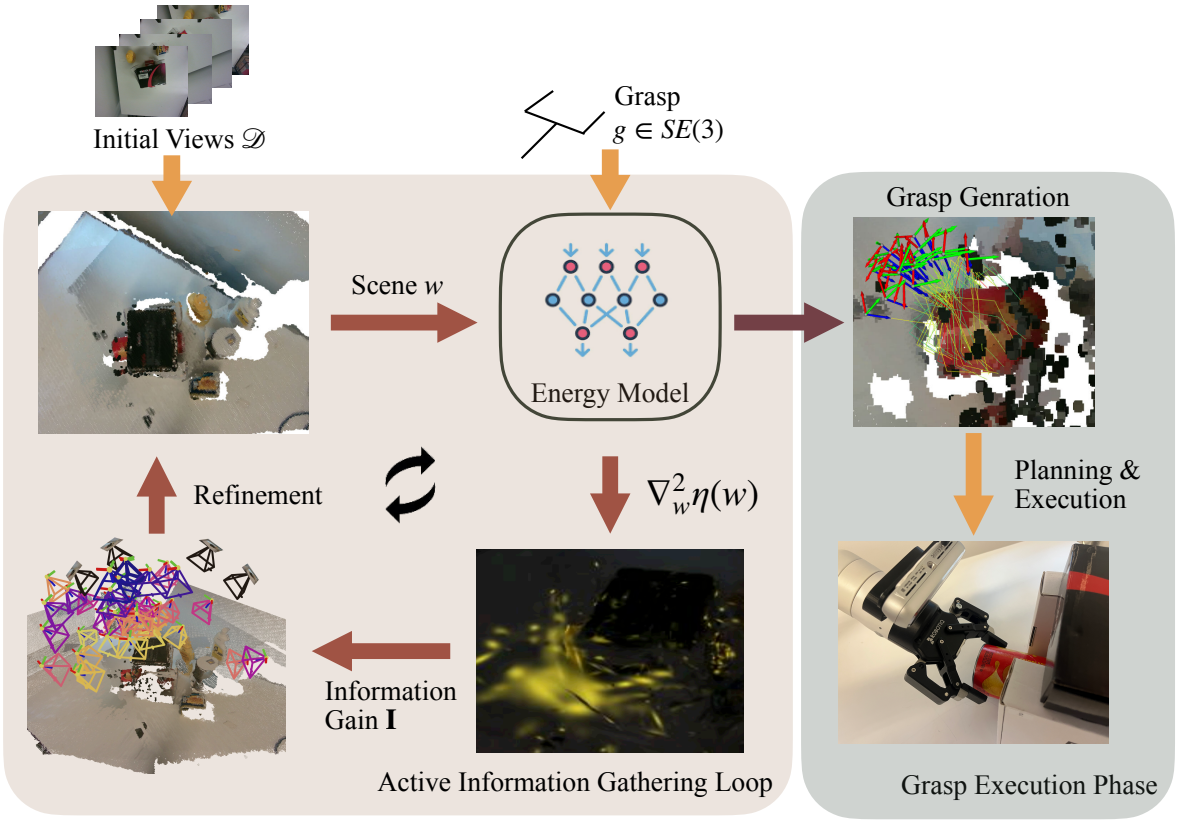


Figure 2. **Our Active Grasping Pipeline** Our method reconstructs 3D scene w from a set of initial views. The energy-based model estimates the grasp entropy $\eta(w)$ using w and the sampled grasp poses g. Then we compute the information gain $\mathbf{I}$ of candidate views from $\nabla^2_w \eta(w)$. We select the next best view with the highest information gain and repeat the procedure above until the view budget is met. Afterwards, we generate grasp poses using the refined scene representation. A planner generates paths to the grasps and executes them on a robot arm.

to align the prediction by temperature scaling [26], standard deviation scaling [42], or a Gaussian Process [68]. One work close to ours is [26], which tunes the temperature divided by logits on the calibration set. Kull *et al*. [39] extend [26] using the Dirichlet distribution to map uncalibrated scores to calibrated scores. Recently, Grathwohl *et al*. [21] find that a classifier trained using the energy-based model method can automatically calibrate itself. However, they do not explore how to calibrate the energy-based model on the regression task, like grasp pose detection. To the best of our knowledge, we are the first to calibrate the model on the grasp pose detection problem where the model's predicted success probability matches the real success rate.

3. Method

Overview Given a set of initial fixed views $\mathcal{D}$ and a target mask, ActiveGrasp first estimates the 3D scene representation w and then computes the information gain $\mathbf{I}$ for K candidate views using the energy-based model. We select the candidate view with the highest information gain, guide the robot to take one new observation at this view, and refine the scene w. This active information-gathering loop continues until the predetermined view budget is reached. Then we use the energy-based model to predict the grasp poses for the target. The whole pipeline is shown in Fig. 2. We use 3D Gaussian Splatting [36] (3DGS) as the scene representation. Each 3D Gaussian is parameterized with position, rotation, scale, RGB color, and a semantic channel indicating the target object. We introduce the Gaussian Approximation and the energy-based model in Sec 3.1. We derive the information gain of candidate views with respect to the grasping task in Sec 3.2. We show how to calibrate the energy level to compute the information gain in Sec 3.3.

3.1. Preliminary

Gaussian Approximation of Posterior For an estimate of scene representation $\mathbf{W}$ from the training set $\mathcal{D}$, we can use a Gaussian distribution to approximate the distribution near the Maximum A Posteriori (MAP) estimate w^* as [37]

$$p(w|\mathcal{D}) \sim \mathcal{N}(w^*, \mathbf{H}''[w|\mathcal{D}]^{-1}) \tag{1}$$

where $\mathbf{H}''(w|\mathcal{D})$ is the Hessian matrix for the Shannon entropy of the Posterior distribution $p(w|\mathcal{D})$. Using Bayes' Theorem and the additive property of logarithms, the Hessian of the posterior can be computed as

$$\mathbf{H}''[w|\mathcal{D}] = \mathbf{H}''[\mathcal{D}|w] + \mathbf{H}''[w] \tag{2}$$

The Hessian can be approximated by the Jacobian using Gauss-Newton and estimated approximately diagonal due to the sparsity of the 3DGS representation [30].

$$\mathbf{H}''[w|\mathcal{D}] \approx \sum_{x \in \mathcal{D}} \text{diag}(\nabla_w f(x,w) \nabla_w f(x,w)^T) + \lambda I \tag{3}$$

where $f(x,w)$ is the rendering equation at the camera pose x and λ is a regularizing constant.

Energy-based Model An energy-based model [40] for grasping is a neural network that maps a pair of scene representation w and grasp pose g to a scalar: $E_\theta : w \times g \to \mathbb{R}$, where w is the 3DGS model and $g \in SE(3)$. It models the conditional distribution of grasp poses given the scene as:

$$p(g|w) = \frac{e^{-E_\theta(g,w)}}{Z}; \quad Z = \int e^{-E_\theta(g,w)} dg \tag{4}$$

SE(3) Diffusion Field [64] extends the model to condition on noise scale σ_k as $E_\theta(g, \sigma_k, w)$ and trains the model on the SE(3) manifold using denoised score matching [60]. Noise is added to the ground-truth successful grasp poses as:

$$\hat{g} = g \exp(\sigma_k \hat{\epsilon}) \quad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

The model is trained with a combination of SDF Loss $\mathcal{L}_{\text{sdf}}$ and Denoised Score Matching Loss $\mathcal{L}_{\text{dsm}}$. The SDF Loss is the L1-Loss between the predicted Signed Distance Function (SDF) value on the reference points and the ground-truth SDF value on the same points. The ground truth SDF value is computed from depth images. The Denoised Score Matching Loss is:

$$\mathcal{L}_{\text{dsm}} = \frac{1}{L} \sum_{k=0}^{L} \mathbb{E}_{g,\hat{g}} \left\| \frac{\partial E_\theta(\gamma, \sigma_k, w)}{\partial \gamma}\bigg|_{\gamma=\hat{g}} + \frac{\partial \log \zeta(\gamma|g, \sigma_k \mathbf{I})}{\partial \gamma}\bigg|_{\gamma=\hat{g}} \right\|^2 \tag{6}$$

where the perturbed distribution ζ is:

$$\zeta(\gamma|g, \sigma_k I) \propto \exp\left(-\frac{1}{2} \|\text{Log}(g^{-1}\gamma)\|^2 / \sigma_k^2\right) \tag{7}$$

Annealed Langevin Dynamics [60] is used to sample on the $SE(3)$ manifold [64]:

$$g_{k-1} = g_k \exp\left(-\alpha_k^2 \frac{\partial E_\theta(\gamma, \sigma_k, w)}{\partial \gamma}\bigg|_{\gamma=g_k} + \alpha_k \epsilon\right) \tag{8}$$

where α_k is a step-dependent coefficient.

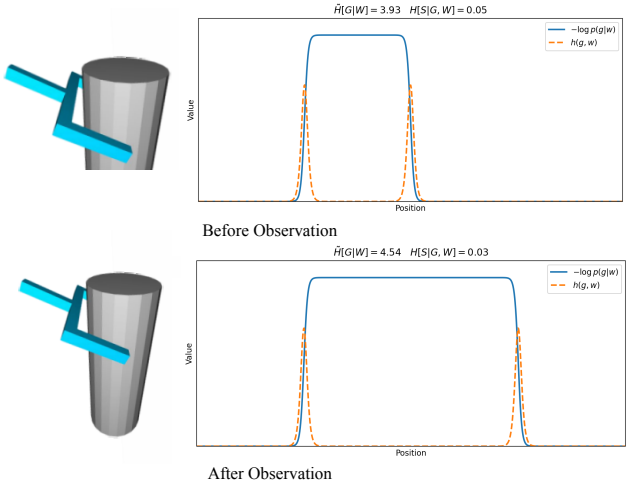


Figure 3. **Illustration of entropy on grasp poses** This is an example of the different definitions of entropy before the next observation (up) and after the next observation (down). We plot the success rate (blue) and entropy (dashed yellow) of a single grasp moving along the cup. $\tilde{\mathbf{H}}$ is the Shannon entropy of the grasp distribution and $\mathbf{H}$ is the conditional grasp entropy defined in our paper.

3.2. Expected Information Gain for Grasping

We define the entropy of grasps as the conditional entropy $\mathbf{H}[S|G,W]$ where G, W, and S are the random variables for the grasp pose, scene representation, and the final success event. We denote this as a function of w as $\eta(w)$:

$$\eta(w) = \mathbf{H}[S|G,W] = \mathbb{E}_{p(g|w)}[h(g,w)] \tag{9}$$

where $h(g,w)$ is the entropy of a single grasp pose g for the given scene w. The success of a single grasp pose follows a Bernoulli distribution parameterized by $s(g,w)$. The entropy for this single grasp is defined as

$$\begin{aligned} h(g,w) &= \text{h}(s(g,w)) \\ &= -s(g,w) \log s(g,w) - (1 - s(g,w)) \log(1 - s(g,w)) \end{aligned} \tag{10}$$

We differentiate Eq. 9 from the Shannon entropy of the grasp pose distribution

$$\tilde{\mathbf{H}}[G|W] = \mathbb{E}_{p(g|w)}[-\log p(g|w)] \tag{11}$$

As is shown in the Fig. 3, $\tilde{\mathbf{H}}$ increases after observation, which differs from the intuition that entropy should decrease after observation. On the other hand, $\mathbf{H}$ decreases due to more successful grasps being discovered. When the scene representation is estimated from the observations, the entropy of the grasp pose is

$$\mathbf{H}[S|G,W,\mathcal{D}] = \mathbb{E}_{p(w|\mathcal{D})}[\eta(w)] \tag{12}$$

where $\eta(w) = \int h(s,g) p(g|w) dg$. We apply Taylor expansion on the function $\eta(w)$ and only keep up to the second-

order term.

$$\eta(w) = \eta(w^*) + \nabla\eta^T(w^*)(w - w^*) + \frac{1}{2}(w - w^*)^T\nabla_w^2\eta(w^*)(w - w^*). \tag{13}$$

Using the Gaussian Approximation of the Posterior estimate from Eq. 1, the entropy can be approximated as:

$$\mathbf{H}[S|G, W, D] = \eta(w^*) + \frac{1}{2}\text{tr}\left(\nabla_w^2\eta(w^*)\mathbf{H}''[w^*|\mathcal{D}]^{-1}\right). \tag{14}$$

The information gain for the candidate views can be approximated as

$$\mathbf{I}[G, W; y_{\text{acq}}|x_{\text{acq}}, \mathcal{D}] \approx \frac{1}{2}\text{tr}\left\{\nabla_w^2\eta(w^*)\left[\mathbf{H}''[w^*|\mathcal{D}]^{-1} - \mathbf{H}''[w^*|x_{\text{acq}}, \mathcal{D}]^{-1}\right]\right\} \tag{15}$$

where x_{acq} is the candidate view pose and y_{acq} is the image at x_{acq}. We leave the proof of Eq. 14 and Eq. 15 to the appendix. The challenge to compute $\nabla_w^2\eta(w)$ is the lack of a relationship between the success rate $s(g, w)$ and the density $p(g|w)$. Even though EBM estimates the density $p(g|w)$, it does not provide more information about the success rate. Therefore, we need to calibrate the energy level to the success rate, and we introduce this in Sec 3.3.

To compute $\nabla_w^2\eta(w)$, we find that using the following approximation works well in practice.

$$\nabla_w^2\eta(w) = \nabla_w\eta(w)\nabla_w\eta(w)^T + \lambda I \tag{16}$$

where $\nabla_w\eta(w)$ can be computed as:

$$\nabla_w\eta(w) = \mathbb{E}[\nabla_w h] - \mathbb{E}[h\nabla_w E_\theta(g, w, \sigma_k)] - \mathbb{E}[h]\mathbb{E}[\nabla_w E_\theta(g, w, \sigma_k)] \tag{17}$$

The proof for Eq. 17 is left to the appendix.

3.3. Calibrated Grasp Generation

Prior work (JEM [21]) shows that EBMs can be calibrated for classification by combining Contrastive Divergence (CD) loss with classification losses, achieving a correspondence between energy and probability. However, applying CD loss directly to our setting is impractical because the conditioning variable w is continuous and high-dimensional, making the CD computation prohibitively expensive. We therefore develop a practical calibration method specifically designed for grasp generation.

We formulate grasp success/failure as a two-category classification problem and train the network to output logits (a_S, a_F) that parameterize success and failure probabilities:

$$(a_S, a_F) = f_\theta(g, w, \sigma_k) \tag{18}$$

The probabilities for the success and failure grasp poses are

$$p_S = \frac{e^{a_S}}{e^{a_S} + e^{a_F} + 2}; \quad p_F = \frac{e^{a_F}}{e^{a_S} + e^{a_F} + 2} \tag{19}$$

And the energy for success and failure is

$$E_S = -\log p_S; \quad E_F = -\log p_F \tag{20}$$

This formulation ensures that the energy is directly interpretable as a success probability, enabling reliable information gain computation.

A key insight is that the gradient of the energy function should point towards *successful* grasps while also pointing *away from* failure modes. To achieve this, we employ *dual score matching* losses: one that aligns the model's score with perturbations of successful grasps, and another that aligns with perturbations of failed grasps.

$$\mathcal{L}^+ = \frac{1}{L}\sum_{k=0}^{L}\mathbb{E}_{g^+,\hat{g}}\left\|\frac{\partial E_S(\mathbf{g}, w, \sigma_k)}{\partial \mathbf{g}} + \frac{\partial \log\zeta(\mathbf{g}|g^+, \sigma_k\mathbf{I})}{\partial \mathbf{g}}\right\|^2$$
$$\mathcal{L}^- = \frac{1}{L}\sum_{k=0}^{L}\mathbb{E}_{g^-,\hat{g}}\left\|\frac{\partial E_F(\mathbf{g}, w, \sigma_k)}{\partial \mathbf{g}} + \frac{\partial \log\zeta(\mathbf{g}|g^-, \sigma_k\mathbf{I})}{\partial \mathbf{g}}\right\|^2 \tag{21}$$

By learning from both successful and failed grasps, the model learns not only which grasps succeed, but also implicitly learns the failure modes, enabling the sampled grasps to avoid failure regions in pose space.

Unlike JEM [21], simply adding a standard cross-entropy loss to enforce $p_S > p_F$ can be too restrictive, potentially disrupting the delicate energy structure learned by DSM. Our key design choice is to use *Average Precision (AP) Loss* [58], which provides only as much supervision as necessary: once positive samples consistently rank higher than negatives, the loss stops applying gradients. This "lazy" supervisory signal preserves the score-matching structure while still enforcing calibration. To compute AP Loss, we first create a set of equally spaced bins in [0,1] for the success and failure categories. For each grasp with its predicted probability $\{g_i, \boldsymbol{p}^i\}$; $\boldsymbol{p} = (p_S^i, p_F^i)$, we assign them to each of the bins with weight as:

$$w_i^j = \max\left(0, 1 - \frac{|p^i - b_j|}{\Delta}\right) \tag{22}$$

where b_j is the center of j-th bin, Δ is the bin width, $p^i = p_S^i$ for success category and $p^i = p_F^i$ for failure category. Then we can compute the average precision for each category. Since all the operations are differentiable, we can directly optimize the average precision for each category.

During training, we find that the energy level is unstable with both losses. We attribute the phenomenon to the fixed temperature, which is the default setting of energy-based models. Comparing Eq.4 to a standard Boltzman definition,

Grasping Method + Active Learning Method	Success	Drop	Fail	Invalid	SR	ECE
ACE + ACE [71]	90	122	168	20	22.50%	0.35
Contact Graspnet [61] + ActiveNGF [45]	130	34	28	208	32.50%	0.32
GSNet [65] + ActiveNGF [45]	249	59	36	62	62.00%	0.30
Se3diff [64] + ActiveNGF [45]	233	26	72	69	58.25%	0.40
Se3diff Scene + ActiveGrasp	294	50	23	33	73.00%	0.28
Se3diff Calib + Random	297	25	33	45	74.25%	0.06
Se3diff Calib + Breyer [5]	295	26	27	52	73.75%	0.05
Se3diff Calib + FisherRF [30]	294	35	26	45	73.50%	0.05
Se3diff Calib + ActiveNGF [45]	296	28	33	43	74.00%	0.06
Se3diff Calib + ACE [71]	296	27	31	45	74.00%	0.07
Se3diff Calib + Random†	306	29	29	38	76.50%	0.05
Ours	316	19	24	41	79.00%	0.02

Table 1. **Evaluation results for active grasping in simulated environments.** The methods are listed as grasping pose generation methods plus active learning methods. *Se3diff Scene* is our scene augmented energy-based model without calibration. *Se3diff Calib* is our calibrated energy-based model. *ActiveGrasp* only selects the views with our proposed information gain. *Random*† takes **8 more views** than all other active methods. **Drop** is defined as the robot finding the target but failing to lift it. **Fail** is defined as the robot fails to get in contact with the target, but the grasp model predicts feasible grasps. **Invalid** is defined as the grasp model that cannot predict any feasible grasps. **SR** is the success rate. **ECE** is the Expected Calibration Error computed on all the executed grasps.

one finds that the EBM fixed the temperature as $T = 1$. By making the temperature learnable, the network can automatically adjust the energy scale to maintain numerical stability while still matching the score structure learned by DSM. This is a simple but crucial modification that allows the calibration losses to converge reliably.

In summary, the network is trained with the following loss:

$$\mathcal{L}_{\text{EBM}} = \frac{\mathcal{L}^+ + \mathcal{L}^-}{2} + \lambda_1 \mathcal{L}_{\text{ap}} + \lambda_2 \mathcal{L}_{\text{sdf}} \tag{23}$$

4. Experiment

4.1. Implementation Details

Energy-based Model Implementation The energy-based model takes scene w, grasp pose g, and noise level σ_k as input. We first preprocess the Gaussian splats by taking only the means for each splat. We augment the energy-based model by fusing means and one-hot semantic vectors for each splat. We use VNN [13] to extract features from the fused representation. The fused representation for VNN is $\mathcal{P} \in \mathbb{R}^{N\times 2\times 3}$ where $\mathcal{P}_i = \begin{bmatrix} p & l \end{bmatrix}^T \in \mathbb{R}^{2\times 3}$. $p \in \mathbb{R}^3$ is the position of the point, and $l \in \mathbb{R}^3$ is all ones for foreground points and all zeros for background points. We choose $N = 1024$. The gripper is represented as a set of fixed anchor points in the gripper coordinate and then converted to the world coordinate by g. We use MLP to extract features from these anchor points. Then, features from scene and anchor points are concatenated in spatial dimension and sent to PointNet [53]. The noise level σ_k is embedded and fused with features in the PointNet. The output of the PointNet is sent to an MLP, which predicts (a_S, a_F) in Eq. 18.

Model Training We train the energy-based model on the Acronym Dataset [15]. We randomly select object meshes and put them on the table in the simulation. Multiple views facing towards the target object are rendered on the sphere, centered on the target. We use Open3D [72] to compute the ground truth SDF value of points near the target object surface. We choose $\lambda_1 = 1,\ \lambda_2 = 0.1$ for Eq. 23. The model is trained for 200k steps with a batch size of 24. We use Adam optimizer with a learning rate of 1e-3. The noise schedule is chosen as $\sigma_k = (\sigma^{2k} - 1)/\log(\sigma),\ \sigma = 0.5,\ k \in [0, 1]$. The AP Loss is only computed for samples with noise level σ_k when $k < 0.01$.

Information Gain Computation We first segment the foreground object from the background using 2D segmentation [56] for each capture image. To initialize the 3DGS, we unproject the RGBD image to the 3D space and use it to initialize the 3DGS. We use the forward kinematics of the robot arm to get the camera pose for each captured view to train the 3DGS. Each Gaussian is set to isotropic. The Spherical Harmonics degree is set to 0. We train the 3DGS for 1k steps after we add one new view. The 3DGS semantic channel is trained with cross-entropy loss on rendered segmentation maps and segmentation labels. To filter out noise from 2D segmentation, we use DBSCAN [16] to cluster all foreground splats with a minimum of 50 and a distance of 0.01. For the largest cluster of foreground splats, we take

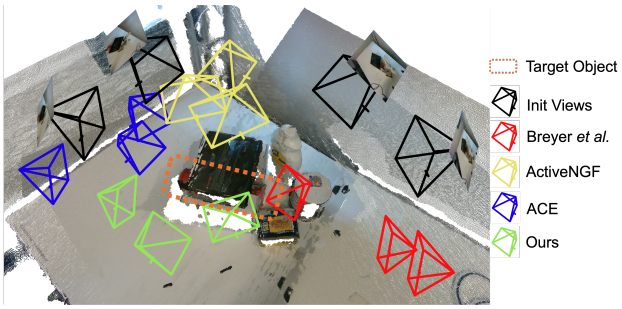


Figure 4. **Visualization for View Selections on Real-world Experiment** We showcase the top 3 view selections for different view selection methods at the same setup. Our method focuses more on perception for regions.

all the 3D Gaussians within a ball of radius r at the center of the foreground 3D Gaussians. Both the foreground and background 3D Gaussians are downsampled to 512 samples using farthest point sampling and concatenated as input for the energy-based model. For candidate view proposal, we use Spherical Fibonacci Sampling [35] to sample poses on a sphere centering the foreground 3D Gaussians, with a radius uniformly sampled between 0.3 and 0.5 m. The number of proposals is set to 128.

We compare our method against three different next-best-view planning methods in both simulation and the real world. For the grasp detection model, we compare against Contact GraspNet [61], ACE [71], $SE(3)$ Diffusion Fields [64] and GraspNet [65]. We use released checkpoints for Contact GraspNet and SE3Diffusion and train the ACE model from scratch using the official implementation. For the next-best-view selection, we also use official source code from Breyer [5], ACE [9] and ActiveNGF [45]

4.2. Simulation Setup

We use the PyBullet simulation environment to evaluate our method in simulation [12]. To evaluate our method in simulation, we build a cluttered scene from the meshes of the objects in the YCB dataset [7]. A 7 DoF Franka robot is placed on one end of the table. We randomly generate configurations of these objects and place them into the scene. The number of objects in each scene is 10. One object is dropped inside the rectangular region in front of the robot by another. An RGBD camera is mounted on the wrist of the robot, with a resolution of 800x800 for both color and depth cameras. Examples of the simulation scene are shown in appendix. We run all the experiments for 400 trials, including 20 different scenes, 5 different target objects in each scene, and 4 trials for each target. In each trial, the method takes 2 fixed views as initialization and actively selects 2 more views.

Scene	Method	Success	Drop	Fail	Invalid
	Ours	9/10	0/10	1/10	0/10
	Breyer [5]	6/10	2/10	2/10	0/10
	ActiveNGF [45]	7/10	0/10	3/10	0/10
	ACE [71]	8/10	0/10	2/10	0/10
	Ours	8/10	0/10	0/10	2/10
	Breyer [5]	6/10	2/10	0/10	2/10
	ActiveNGF [45]	6/10	0/10	0/10	4/10
	ACE [71]	4/10	2/10	2/10	2/10
	Ours	9/10	0/10	1/10	0/10
	Breyer [5]	8/10	2/10	0/10	0/10
	ActiveNGF [45]	8/10	0/10	2/10	0/10
	ACE [71]	8/10	0/10	2/10	0/10

Table 2. **Evaluation results for active grasping in real world environments** The targets for three setups are red cup, red chip bottle, and red tomato can.

4.3. Simulation Experiment

We summarize the results of each grasping with the following metrics: **Success**: The number of successful trials. **Drop**: The number of trials when the robot is in touch with the target object but fails to firmly grasp it. **Fail**: The number of trials when there are feasible grasp poses, but the robot fails to be in contact with the target object. **Invalid**: The number of trials when the grasp prediction model cannot predict any grasp candidates or all the grasp candidates are rejected by the planner, due to infeasible kinematics or collision. The results are summarized in Tab 1. Our calibrated energy-based model *Se3diff-Calib* has a higher success rate than all previous grasp detection models [9, 61, 64, 65]. We attribute this to the calibrated property of our energy-based model. Since the executed grasp is the grasp with the highest predicted success rate and the prediction is calibrated, it provides a guarantee for the success rate of execution. For next-based-view selection, we compare all previous methods using the same *Se3diff-Calib*. Our method achieves the highest success rate of 79%, higher than the best (74.25%) of the other methods with same number of selected views. Meanwhile, we also find that our proposed view selection brings the best of the calibrated energy-based model, reaching the lowest calibration error of 0.02. We also compare our method with a setting using the same view selection algorithm but an uncalibrated energy-based model, as *Se3diff Scene + ActiveGrasp*, to show the effectiveness of energy calibration. Finally, we compare with the setting, *Se3diff Calib + Random*†, to show the effectiveness of next-best-view selection for grasping task. We let Random† take **10 views** instead of 2 views. Even though taking more views could improve the success rate from 74.25% to 76.50%, it is still lower than our proposed active selection method with a higher execution time cost.

4.4. Real World Experiment

We run our real-world experiments on a Kinova 7 DoF Robot Arm. The robot is equipped with a parallel jaw end-

	AP ↑	ECE ↓	Ang (acc.) ↓	Trans (acc.) ↓	Ang (rec.) ↓	Trans (rec.) ↓	ECE (Bullet) ↓
Se3diff [64]	34.69	0.17	0.6244	0.0589	0.5233	0.0399	0.35
Se3diff+Scene	66.00	0.15	0.4369	0.0546	0.3235	0.0184	0.30
Se3diff+Scene+AP	83.31	0.14	0.3517	0.0201	0.3138	0.0200	0.17
Se3diff+Scene+FSM+AP	83.13	0.03	0.4119	0.0238	0.3217	0.0189	0.08
Se3diff+Scene+FSM+AP+T	87.84	0.03	0.3764	0.0212	0.3068	0.0178	0.05

Table 3. **Quantitative comparison for grasping pose average precision on the ACRONYM dataset [15]**. **AP**: average precision using the predicted energy. **ECE**: Expected Calibration Error on Acronym dataset. **Ang/Trans (acc.)**: Average closest angular and translational distance from the closest ground-truth successful grasps to the predicted grasps. **Ang/Trans(rec.)**: Average closest angular and translational distance from the predicted grasps to the closest ground-truth successful grasps. The unit of angular and translational distance is expressed as radians and meters. **ECE (Bullet)**: Expected Calibration Error after executing grasps in the Bullet simulator. **Scene**: Our implementation of model augmented with scene information. **FSM**: Model trained with both success and failure grasp. **AP** Model trained with AP Loss. **T**: Model trained with learnable temperature.

effector gripper and a RealSense camera at the end joint. A set of objects from the YCB dataset will be placed in front of the robot. During the experiments, the robot takes 4 fixed views at the beginning and uses each method to actively select 2 views to build a representation of the scene. Afterwards, the robot uses our calibrated energy-based model *Se3diff Calib* to predict grasps and attempts to grasp the target object. We compare ActiveGrasp with a series of other grasp detection methods, including [4], [64], and [63]. The results are summarized in Tab. 2. We visualize the results of the view selection in Fig. 4 for the second setup. As is shown in the visualization, Breyer et al. [5] (red) selects views on the right-hand side of the object, where there are many obstacles. This is because they only rely on the coverage, failing to focus on the potential grasp location. ACE [9] (blue) and ActiveNGF [45] (yellow) can focus on the left-hand side, but their selections are close to the captured views (black) compared to our method (green).

4.5. Ablation Study

4.5.1. Calibrated Energy-based Model

To show the effectiveness of our proposed calibrated energy-based model, we performed an ablation study on each module. We first evaluate our scene-aware grasp generation model *Se3diff scene* and *Se3diff* on the ACRONYM dataset [15]. We use the validation set of the augmented grasp dataset for evaluation. We compute four metrics: **AP**, **ECE**, **Ang/Trans (acc.)** **Ang/Trans (rec.)**. We also randomly create scenes in the bullet simulator using YCB [7] objects as in Sec 4.2 and place the gripper at the predicted grasp pose to grasp the object. The result is regarded as the ground truth to compute **ECE (Bullet)**. The result is summarized in Tab. 3. After adding scene information *Scene* and AP Loss *AP*, the model generates grasp poses closer to the ground truth distribution with lower distance measure **Ang/Trans (rec.)**, **Ang/Trans (acc.)**. This is because it learns to distinguish good grasp from bad grasp as it has a higher average precision and lower calibration error on the dataset. However, it still has high calibration error on its generation with large **ECE (Bullet)**. After including the failure grasps in the score matching loss, the calibration for generation decreases, but the quality decreases since both **Ang/Trans (rec.)** and **Ang/Trans (acc.)** increase. This issue is solved when we let the network learn the temperature, and the model performs better on all metrics.

4.5.2. Run Time Analysis

To demonstrate the efficiency of our method in terms of runtime, we provide a breakdown of the time spent on different stages of our active grasping system. Our ActiveGrasp takes 9.61 ± 0.09 seconds for view selection, which is significantly faster than ACE [71] (18.24 ± 0.15 s), ActiveNGF [45] (27.66 ± 0.32 s), and Breyer *et al*. [5] (32.97 ± 0.45 s). Although vanilla FisherRF [30] (7.43 ± 0.01 s) is slightly faster than our method, it does not consider the grasping task and has a worse successful rate in our Experiments, as shown in Tab. 1.

5. Conclusion and Limitation

In this paper, we addressed the challenging problem of active grasping, which improves robotic grasping quality by efficiently gathering information from the environment. We proposed a novel, information-theoretic framework that leverages a calibrated energy-based model to compute information gain for the next-best-view (NBV) selection process in the grasping task. We define the entropy of the grasp distribution and compute the entropy using the Gaussian Approximation of the Posterior of the scene and the energy-based model. Our method eliminates the need for separate grasp affordance models or predefined heuristic metrics, offering a more principled and task-aware active perception system. Our active grasping benchmark with physically informed simulators will also help follow-up work perform reproducible comparisons to boost the field's development further. The limitation of our work lies in the approximation of the second-order term $\nabla^2_w \eta(w)$ and the requirement of model calibration. Future work could explore better approximation and calibration techniques given the domain shift.

Acknowledgements The authors gratefully appreciate support through the following grants: NSF FRR 2220868, NSF IIS-RI 2212433, and ONR N00014-22-1-2677.

References

[1] Ermano Arruda, Jeremy Wyatt, and Marek Kopicki. Active vision for dexterous grasping of novel objects. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2881–2888, 2016. 1, 2

[2] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 2

[3] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42:177–196, 2018. 2

[4] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, pages 1602–1611. PMLR, 2021. 2, 8

[5] Michel Breyer, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Closed-loop next-best-view planning for target-driven grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, pages 1411–1416. IEEE, 2022. 1, 2, 6, 7, 8

[6] Junhao Cai, Jun Cen, Haokun Wang, and Michael Yu Wang. Real-time collision-free grasp pose detection with geometry-aware refinement using high-resolution volume. *IEEE Robotics Autom. Lett.*, 7(2):1888–1895, 2022. 2

[7] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22 (3):36–52, 2015. 2, 7, 8

[8] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[9] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 7, 8

[10] Xiangyu Chen, Zelin Ye, Jiankai Sun, Yuda Fan, Fang Hu, Chenxi Wang, and Cewu Lu. Transferable active grasping and real embodied dataset. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 3611–3618. IEEE, 2020. 2

[11] Fu-Jen Chu, Ruinian Xu, and Patricio A. Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics Autom. Lett.*, 3(4):3355–3362, 2018. 2

[12] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021. 7

[13] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12180–12189. IEEE, 2021. 6

[14] Harnaik Dhami, Vishnu D Sharma, and Pratap Tokekar. Pred-nbv: Prediction-guided next-best-view planning for 3d object reconstruction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7149–7154. IEEE, 2023. 2

[15] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 6222–6227. IEEE, 2021. 6, 8, 4

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231. AAAI Press, 1996. 6

[17] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023. 1

[18] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, 56(S1):1513–1589, 2023. 2

[19] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *ICRA*, 2022. 2

[20] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. *arXiv*, 2023. 2

[21] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3, 5

[22] Marcus Gualtieri and Robert Platt Jr. Viewpoint selection for grasp detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 258–264. IEEE, 2017. 2

[23] Marcus Gualtieri, Andreas ten Pas, Kate Saenko, and Robert Platt Jr. High precision grasp pose detection in dense clutter. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*, pages 598–605. IEEE, 2016. 1

[24] Antoine Guédon, Pascal Monasse, and Vincent Lepetit. Scone: Surface coverage optimization in unknown environments by volumetric integration. *Advances in Neural Information Processing Systems*, 35:20731–20743, 2022. 2

[25] Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. Macarons: Mapping and coverage anticipation with rgb online self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 940–951, 2023. 2

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1321–1330. PMLR, 2017. 3

[27] Boce Hu, Xupeng Zhu, Dian Wang, Zihao Dong, Haojie Huang, Chenghao Wang, Robin Walters, and Robert Platt. Orbitgrasp: Se(3)-equivariant grasp learning. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, pages 2456–2474. PMLR, 2024. 2

[28] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *CoRR*, abs/2504.16054, 2025. 1

[29] Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, pages 1169–1179. AUAI Press, 2019. 2

[30] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIII*, pages 422–440. Springer, 2024. 2, 4, 6, 8

[31] Wen Jiang, Boshu Lei, Katrina Ashton, and Kostas Daniilidis. Multimodal llm guided exploration and active mapping using fisher information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5392–5404, 2025. 2

[32] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11305–11312. IEEE, 2023. 2

[33] Edward Johns, Stefan Leutenegger, and Andrew J. Davison. Deep learning a grasp function for grasping under gripper pose uncertainty. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*, pages 4461–4468. IEEE, 2016. 2

[34] Gregory Kahn, Peter Sujan, Sachin Patil, Shaunak D. Bopardikar, Julian Ryde, Kenneth Y. Goldberg, and Pieter Abbeel. Active exploration using trajectory optimization for robotic grasping in the presence of occlusions. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 4783–4790. IEEE, 2015. 1, 2

[35] Benjamin Keinert, Matthias Innmann, Michael Sänger, and Marc Stamminger. Spherical fibonacci mapping. *ACM Trans. Graph.*, 34(6):193:1–193:7, 2015. 7

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[37] Andreas Kirsch and Yarin Gal. Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities. *Transactions on Machine Learning Research*, 2022. Expert Certification. 2, 3

[38] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1:239–249, 2020. 2

[39] Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *CoRR*, abs/1910.12656, 2019. 3

[40] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 4

[41] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *Int. J. Robotics Res.*, 34(4-5): 705–724, 2015. 2

[42] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022. 3

[43] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robotics Res.*, 37(4-5):421–436, 2018. 1

[44] Byeongdo Lim, Jongmin Kim, Jihwan Kim, Yonghyeon Lee, and Frank C. Park. Equigraspflow: Se(3)-equivariant 6-dof grasp pose generative flows. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, pages 5067–5086. PMLR, 2024. 2

[45] Haoxiang Ma, Modi Shi, Boyang Gao, and Di Huang. Active perception for grasp detection via neural graspness field. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 6, 7, 8, 3

[46] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *RSS*, 2017. 1, 2

[47] Alireza Mehrtash, William M. Wells III, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Medical Imaging*, 39(12): 3868–3878, 2020. 2

[48] Douglas Morrison, Peter Corke, and Jürgen Leitner. Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *International Conference on Robotics and*

Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019, pages 8762–8768. IEEE, 2019. 1, 2

[49] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *ECCV*, pages 230–246. Springer, 2022. 2

[50] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 2

[51] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 3406–3413. IEEE, 2016. 2

[52] Robert Platt. Grasp learning: Models, methods, and performance. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):363–389, 2023. 2

[53] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 6

[54] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 400–418. Springer, 2020. 2

[55] Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. *IEEE Robotics and Automation Letters*, 8(2):1125–1132, 2023. 2

[56] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[57] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 1316–1322. IEEE, 2015. 2

[58] Jérôme Revaud, Jon Almazán, Rafael S. Rezende, and César Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5106–5115. IEEE, 2019. 5

[59] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9030–9038. Computer Vision Foundation / IEEE, 2019. 2

[60] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. 4

[61] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444. IEEE, 2021. 6, 7

[62] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. 2

[63] Andreas Ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017. 8

[64] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 5923–5930. IEEE, 2023. 2, 4, 6, 7, 8

[65] Chenxi Wang, Haoshu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. *CoRR*, abs/2406.11142, 2024. 6, 7

[66] Wei Wei, Yongkang Luo, Fuyu Li, Guangyun Xu, Jun Zhong, Wanyi Li, and Peng Wang. GPR: grasp pose refinement network for cluttered scenes. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 4295–4302. IEEE, 2021. 2

[67] Zehang Weng, Haofei Lu, Jens Lundell, and Danica Kragic. Capgrasp: An $\mathbb {R}^{3}\times \text{SO(2)-Equivariant}$ continuous approach-constrained generative grasp sampler. *IEEE Robotics Autom. Lett.*, 9(4):3641–3647, 2024. 2

[68] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, pages 178–190. PMLR, 2020. 3

[69] Zike Yan, Haoxiang Yang, and Hongbin Zha. Active neural mapping. In *ICCV*, 2023. 2

[70] Hanbo Zhang, Jian Tang, Shiguang Sun, and Xuguang Lan. Robotic grasping from classical to modern: A survey. *arXiv preprint arXiv:2202.03631*, 2022. 2

[71] Xuechao Zhang, Dong Wang, Sun Han, Weichuang Li, Bin Zhao, Zhigang Wang, Xiaoming Duan, Chongrong Fang,

Xuelong Li, and Jianping He. Affordance-driven next-best-view planning for robotic grasping. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, pages 2849–2862. PMLR, 2023. 1, 2, 6, 7, 8

[72] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *CoRR*, abs/1801.09847, 2018. 6