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

Supplementary Material

6. Derivation for Entropy Definition

For the conditional entropy $\mathbf{H}[S|G,W]$, we can factorize it as:

$$\begin{aligned}\mathbf{H}[S|G,W] &= -\int \sum_s p(s,g|w)\log\frac{p(g,s|w)}{p(g|w)}dg \\ &= -\int \sum_s p(g|w)p(s|g,w)\log p(s|g,w)dg \\ &= -\int p(g|w)\sum_s \left(p(s|g,w)\log p(s|g,w)\right)dg \\ &= \int p(g|w)\sum_s \left(-p(s|g,w)\log p(s|g,w)\right)dg \\ &= \mathbb{E}_{p(g|w)}[h(g,w)]\end{aligned} \tag{24}$$

Here we prove Eq. 9. For conditional entropy given the observations $\mathcal{D}$,

$$\begin{aligned}&\mathbf{H}[S|G,W,\mathcal{D}] \\ &= -\int\int \sum_s p(s,g,w|\mathcal{D})\log\frac{p(s,g,w|\mathcal{D})}{p(g,w|\mathcal{D})}dgdw \\ &= -\int\int \sum_s p(w|\mathcal{D})p(g|w)p(s|g,w)\log p(s|g,w,\mathcal{D})dwdg \\ &= -\int p(w|\mathcal{D})\int p(g|w)\sum_s \left(p(s|g,w)\log p(s|g,w)\right)dgdw \\ &= \mathbb{E}_{p(w|\mathcal{D})}[\eta(w)]\end{aligned} \tag{25}$$

Here we assume the grasp pose G and observations G are independent given the scene representation W.

7. Derivation for Information Gain

We define the entropy of grasps $\mathbf{H}[g|w]$ as the expectation of the entropy for a single grasp $h(g,w)$, and denote it as a function $\eta(w)$:

$$\eta(w) = \mathbf{H}[S|G,W] = \mathbb{E}_{p(g|w)}[h(g,w)] \tag{26}$$

When the scene w is estimated from a set of collected views $\mathcal{D}$, the entropy of the grasp distribution is

$$\begin{aligned}\eta(w) &= \mathbb{E}_{p(g,w|\mathcal{D})}[h(g,w)] \\ &= \int p(g,w|\mathcal{D})h(g,w)dgdw \\ &= \int p(g|w)p(w|\mathcal{D})h(g,w)dgdw \\ &= \int p(w|\mathcal{D})\left(\int p(g|w)h(g,w)\right)dw \\ &= \int p(w|\mathcal{D})\eta(w)dw \\ &= \mathbb{E}_{p(w|\mathcal{D})}[\eta(w)]\end{aligned} \tag{27}$$

From line 2 to line 3, we assume the grasp pose g and the dataset $\mathcal{D}$ are independent when the scene w is given. Using the Taylor expansion of $\eta(w)$ and plugging Eq. 13 into Eq. 27, we get

$$\begin{aligned}&\mathbf{E}_{p(w|\mathcal{D})}[\eta(w)] \\ &\approx \mathbf{E}_{p(w|\mathcal{D})}[\eta(w^*) + \nabla_w\eta^T(w^*)(w-w^*) \\ &\qquad + \frac{1}{2}(w-w^*)^T\nabla_w^2\eta(w^*)(w-w^*)]\end{aligned} \tag{28}$$

The expectation of the first term is just itself. The expectation of the second term is 0.

$$\begin{aligned}&\mathbf{E}_{p(w|\mathcal{D})}[\nabla_w\eta^T(w^*)(w-w^*)] \\ &= \nabla_w\eta^T(w^*)\mathbf{E}_{p(w|\mathcal{D})}[(w-w^*)] \\ &= \nabla_w\eta^T\mathbf{0} = 0\end{aligned} \tag{29}$$

Using $\text{tr}(ABC) = \text{tr}(BCA)$ for any matrices A, B, C, the third term is

$$\begin{aligned}&\frac{1}{2}\mathbb{E}_{p(w|\mathcal{D})}\left[(w-w^*)^T\nabla_w^2\eta(w^*)(w-w^*)\right] \\ &= \frac{1}{2}\mathbb{E}_{p(w|\mathcal{D})}\left[\text{tr}\left((w-w^*)^T\nabla_w^2\eta(w^*)(w-w^*)\right)\right] \\ &= \frac{1}{2}\mathbb{E}_{p(w|\mathcal{D})}\left[\text{tr}\left(\nabla_w^2\eta(w^*)(w-w^*)(w-w^*)^T\right)\right] \\ &= \frac{1}{2}\text{tr}\left(\nabla_w^2\eta(w^*)\mathbb{E}_{p(w|\mathcal{D})}\left[(w-w^*)(w-w^*)^T\right]\right) \\ &= \frac{1}{2}\text{tr}\left(\nabla_w^2\eta(w^*)\mathbf{H}''[w^*|\mathcal{D}]^{-1}\right)\end{aligned} \tag{30}$$

We get the results in Eq. 14. The mutual information for the candidate view is defined as the difference of the entropy

Method	Success	Drop	Fail	Invalid	SR	ECE
ActiveGrasp ($\lambda = 10^{-1}$)	309	20	32	39	77.25%	0.04
ActiveGrasp ($\lambda = 10^{-2}$)	312	20	29	39	78.00%	0.04
ActiveGrasp ($\lambda = 10^{-3}$)	302	29	32	37	75.50%	0.06
ActiveGrasp ($\lambda = 10^{-4}$)	316	19	24	41	79.00%	0.02

Table 4. **Ablation study in simulated environments.** We test 4 different settings of λ from 10^{-1} to 10^{-4}. **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.

given the candidate $\{x_{\text{acq}}, y_{\text{acq}}\}$:

$$\begin{aligned}&\mathbf{I}[g, w; y_{\text{acq}}|x_{\text{acq}}, \mathcal{D}] \\ &= \mathbf{H}[g, w|D] - \mathbf{H}[g, w|\{x_{\text{acq}}, y_{\text{acq}}\}, \mathcal{D}] \\ &= \mathbf{E}_{p(w|\mathcal{D})}[\eta(w)] - \mathbf{E}_{p(w|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D})}[\eta(w)]\end{aligned} \tag{31}$$

Using the Gaussian approximation, $p(w|\mathcal{D}) \sim \mathcal{N}(w^*, \mathbf{H}''[w^*|\mathcal{D}]^{-1})$ and $p(w|x_{\text{acq}}, y_{\text{acq}}\mathcal{D}) \sim \mathcal{N}(\tilde{w}^*, \mathbf{H}''[\tilde{w}^*|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D}]^{-1})$, where $\tilde{w}^*$ is the new MAP estimate. We expand $\eta(w)$ near the previous MAP estimate w^* instead of the new MAP estimate $\tilde{w}^*$,

$$\begin{aligned}&\mathbf{E}_{p(w|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D})}[\eta(w)] \\ &\approx \mathbf{E}_{p(w|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D})}[\eta(w^*) + \\ &\quad \nabla_w \eta^T(w^*)(w - w^*) + \frac{1}{2}(w - w^*)^T \nabla_w^2 \eta(w^*)(w - w^*)]\end{aligned} \tag{32}$$

The expectation is:

$$\begin{aligned}&\mathbf{E}_{p(w|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D})}[\eta(w)] \\ &\approx \eta(w^*) + \nabla\eta(w^*)^T(\tilde{w}^* - w^*) \\ &\quad + \frac{1}{2}(\tilde{w}^* - w^*)^T \nabla^2 \eta(w^*)(\tilde{w}^* - w^*) \\ &\quad + \frac{1}{2}\text{tr}\left(\nabla^2\eta(w^*)\mathbf{H}''(w^*|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D})^{-1}\right)\end{aligned} \tag{33}$$

Plugging back into the mutual information gain formula, we get:

$$\begin{aligned}&\mathbf{I}[G, W; y_{\text{acq}}|x_{\text{acq}}, \mathcal{D}] \\ &\approx \frac{1}{2}\text{tr}\left(\nabla_w^2 \eta(w^*)\left[\mathbf{H}''[w^*|\mathcal{D}]^{-1} - \mathbf{H}''[\tilde{w}^*|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D}]^{-1}\right]\right) \\ &\quad - \nabla\eta(w^*)^T(\tilde{w}^* - w^*) - \frac{1}{2}(\tilde{w}^* - w^*)^T \nabla^2 \eta(w^*)(\tilde{w}^* - w^*)\end{aligned} \tag{34}$$

If we assume the shift from w^* to the new estimate $\tilde{w}^*$ is small, then the mutual information is approximated as

$$\begin{aligned}&\mathbf{I}[g, w; y_{\text{acq}}|x_{\text{acq}}, \mathcal{D}] \\ &\approx \frac{1}{2}\text{tr}\left(\nabla_w^2 \eta(w^*)\left[\mathbf{H}''[w^*|\mathcal{D}]^{-1} - \mathbf{H}''[w^*|x_{\text{acq}}, y_{\text{acq}}, \mathcal{D}]^{-1}\right]\right)\end{aligned} \tag{35}$$

Since the computation of $\mathbf{H}''$ (Eq. 2) only involves x, we can omit y_{acq} in the equation and get

$$\begin{aligned}&\mathbf{I}[G, W; y_{\text{acq}}|x_{\text{acq}}, \mathcal{D}] \\ &\approx \frac{1}{2}\text{tr}\left(\nabla_w^2 \eta(w^*)\left[\mathbf{H}''[w^*|\mathcal{D}]^{-1} - \mathbf{H}''[w^*|x_{\text{acq}}, \mathcal{D}]^{-1}\right]\right)\end{aligned} \tag{36}$$

Here we get the results in Eq. 15.

8. Derivation for $\nabla_w \eta(w)$

$\eta(w)$ is defined as:

$$\eta(w) = \mathbb{E}_{p(g|w)}[h(g, w)] = \int p(g|w)h(g, w)dg \tag{37}$$

Taking the first-order derivative w.r.t w, we have

$$\nabla_w \eta(w) = \int (\nabla_w h(g, w))p(g, w) + h(g, w)\nabla_w p(g, w)dg \tag{38}$$

Using the equation

$$\nabla_w p(g|w) = p(g|w)\frac{\nabla_w p(g|w)}{p(g|w)} = p(g|w)\nabla_w \log p(g|w) \tag{39}$$

Plugging in, we get the following.

$$\begin{aligned}&\nabla_w \eta(w) \\ &= \int p(g|w)\left(\nabla_w h(g, w) + h(g, w)\nabla_w \log p(g|w)\right) dg \\ &= \mathbb{E}_{p(g|w)}[\nabla_w h(g, w) + h(g, w)\nabla_w \log p(g|w)]\end{aligned} \tag{40}$$

Since we model the grasp distribution using an energy-based model, we have

$$\log p(g|w) = -E_\theta(g, w, \sigma_k) + \log Z;\ Z = \int e^{-E_\theta(g, w, \sigma_k)}dg \tag{41}$$

Take the first-order derivative,

$$\nabla_w \log p(g|w) = -\nabla_w E_\theta(g, w, \sigma_k) + \frac{1}{Z}\nabla_w Z \tag{42}$$

The first-order derivative of Z is

$$\begin{aligned}\frac{1}{Z}\nabla_w Z &= \frac{1}{Z}\nabla_w \int e^{-E_\theta(g,w,\sigma_k)}dg \\ &= \frac{1}{Z}\int e^{-E_\theta(g,w,\sigma_k)}(-\nabla_w E_\theta(g,w,\sigma_k))dg \\ &= \int p(g|w)(-\nabla_w E_\theta(g,w,\sigma_k))dg \\ &= -\mathbb{E}_{p(g|w)}[\nabla_w E_\theta(g,w,\sigma_k)]\end{aligned} \tag{43}$$

From line 2 to line 3, we put Z inside the integral and use the definition in Eq. 4. Putting this back in Eq. 38, we get

$$\begin{aligned}\nabla_w \eta(w) &= \int p(g|w)\left(\nabla_w h(g,w) + h(g,w)\nabla_w \log p(g|w)\right) dg \\ &= \mathbb{E}_{p(g|w)}[\nabla_w h(g,w) + \\ &\quad h(g,w)(-\nabla_w E_\theta(g,w,\sigma_k) - \\ &\quad\quad \mathbb{E}_{p(g|w)}[\nabla_w E_\theta(g,w,\sigma_k)])] \\ &= \mathbb{E}_{p(g|w)}[\nabla_w h(g,w)] - \\ &\quad \mathbb{E}_{p(g|w)}[h(g,w)\nabla_w E_\theta(g,w,\sigma_k] - \\ &\quad\quad \mathbb{E}_{p(g|w)}[h(g,w)]\mathbb{E}_{p(g|w)}[\nabla_w E_\theta(g,w,\sigma_k)])]\end{aligned} \tag{44}$$

Here we prove Eq. 17. We sample from the energy-based model to compute the expectation.

Figure 5. **Visualization for Real World and Simulation Experiment Setup** We show the objects used in the real world experiment (left) and one randomly generated simulation scene (right).

9. Experiment Details

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 pruning and duplicating process runs every 500 steps. To compute the expectation of information gain, we sample 512 grasp poses from the energy-based model. We also show the experiment setup for both simulation and real-world experiments in Fig. 5.

10. Regularization Parameter λ

We do an ablation study on the regularization constant λ. We follow the same setting as in the simulation experiment in Sec 4.3. Each setting starts with 2 fixed initial views and actively takes 2 more views before grasping. The result is summarized in Tab. 4. $\lambda = 10^{-4}$ achieves the best result, a success rate of 79.00% as we report in the paper in Tab. 1. Even the worst case ($\lambda = 10^{-3}$) has a higher success rate than other active view selection methods (74.25%), when the number of selected views is the same.

11. Ablation on Information Gain Estimation

We do an ablation study on the number of grasp poses used to estimate the information gain. We follow the same setting as in the simulation experiment in Sec 4.3. The result is summarized in Tab. 6. As the number of grasp poses increases, the estimation becomes better, leading to better performance. We use 512 grasp poses in the main paper. As shown in Fig. 6, even though we use 512 grasp poses, the view selection process only accounts for 6.5% of the entire running time.

12. Network Parameterization

We formulate the predicted probability as unnormalized in Eq. 19 because we find it is better than a normalized formulation as:

$$p_S = \frac{e^{a_S}+1}{e^{a_S}+e^{a_F}+2}; \quad p_F = \frac{e^{a_F}+1}{e^{a_S}+e^{a_F}+2} \tag{45}$$

We carry out an ablation study, and the result is summarized in Tab. 5. As shown in the table, the unnormalized version has a lower Ang/Trans(acc.) and Ang/Trans(rec.) error than the normalized version, showing that the generation quality is better. We think the reason is that the unnormalized formulation allows the network to predict a grasp pose as neither success nor failure. Therefore, it gives the network more freedom to adjust the energy landscape.

13. Time Analysis

We provide a runtime analysis of our active grasping system in Fig 6. The results show that the view selection module and 3D-GS data representation do not create significant overhead, and the overall runtime is efficient for real-world applications, as most of the time is spent on the motion of the robot arm.

14. Comparison on GraspNet-1billion

We evaluated our method using the GraspNet-1Billion dataset following the same setup in ActiveNGF [45]. We use the same parameter setting and select 10 views from

	AP ↑	ECE ↓	Ang (acc.) ↓	Trans (acc.) ↓	Ang (rec.) ↓	Trans (rec.) ↓	ECE (Bullet) ↓
Se3diff+Scene+FSM(normalized)+AP+T	81.52	0.03	0.3900	0.0242	0.3294	0.0203	0.06
Se3diff+Scene+FSM(unnormalized)+AP+T	87.84	0.03	0.3764	0.0212	0.3068	0.0178	0.05

Table 5. **Ablation study of energy-based model 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 the model augmented with scene information. **FSM(normalized)**: Model trained with both success and failure grasp, and the probability is normalized. **FSM(unnormalized)**: Model trained with both success and failure grasp, and the probability is unnormalized. **AP** Model trained with AP Loss. **T**: Model trained with learnable temperature.

Method	Success Rate
ActiveGrasp (N=128)	74.50%
ActiveGrasp (N=256)	75.25%
ActiveGrasp (N=512)	79.00%

Table 6. **Number of grasp poses for information gain estimation** N is the number of grasp poses used to compute in Eq. 17.

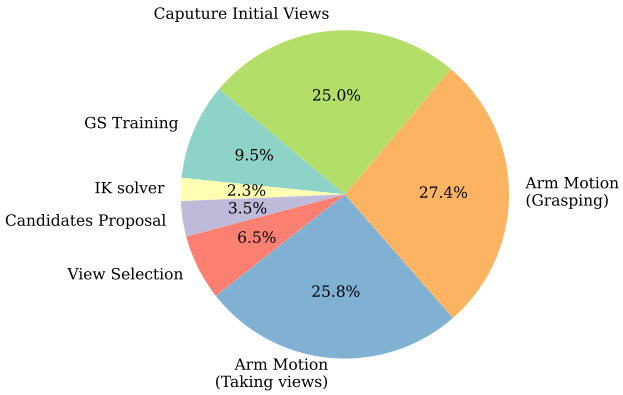

Figure 6. **Runtime Analysis of our Active Grasping System.** As we can see, the view selection module and 3D-GS data representation do not create significant overhead.

each scene. Then we use the same grasp predictor from ActiveNGF to predict the poses and compute the metric. The result is summarized in Table. 7.

Methods	Seen			Similar			Novel		
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$
ActiveNGF	55.12	65.07	48.88	52.85	62.63	46.49	24.74	30.21	12.00
Ours	58.44	66.33	51.67	56.22	65.67	48.33	26.44	35.33	14.00

Table 7. **Graspnet-1Billion Experiment**