B^3-Seg: Camera-Free, Training-Free 3DGS Segmentation via Analytic EIG and Beta–Bernoulli Bayesian Updates

Supplementary Material

A. Beta Entropy and Notation

Let $X \sim \mathrm{Beta}(a, b)$ with $a, b > 0$. For simplicity, we define $B(\cdot, \cdot)$ as a Beta function, and $H(\cdot, \cdot)$ as the entropy of the Beta distribution. The entropy is

$$H(a,b) = \log B(a,b) - (a-1)\,\psi(a) - (b-1)\,\psi(b) + (a+b-2)\,\psi(a+b), \tag{14}$$

where ψ is the digamma function.

B. Lemma 1: Non-negativity (Adaptive Monotonicity)

Claim. For any candidate view v and any selected set S, the expected gain is nonnegative: $\mathbb{E}[\mathrm{EIG}(v \mid S)] \geq 0$.

Proof. For a single Gaussian i, define its mean m_i and concentration κ_i

$$m_i = \frac{a_i}{a_i + b_i}, \qquad \kappa_i = a_i + b_i.$$

From Eq. (11), the approximate posterior entropy of EIG is given by

$$\begin{aligned} &H(a_i + \tilde{e}_{i,1}(v), b_i + \tilde{e}_{i,0}(v)) \\ &= H(a_i + m_i\tau_i, b_i + (1-m_i)\tau_i) \\ &= H\Bigg(\Bigg(1 + \sum_{(j,k)\in I(v)} \alpha_i T_i \frac{1}{a_i + b_i}\Bigg) a_i, \\ &\qquad \Bigg(1 + \sum_{(j,k)\in I(v)} \alpha_i T_i \frac{1}{a_i + b_i}\Bigg) b_i\Bigg) \\ &:= H(c_i a_i, c_i b_i). \end{aligned}$$

This corresponds to multiplying the concentration κ_i by $c_i > 1$ while leaving the mean m_i unchanged. This can be interpreted as the observation of more trials of the same underlying Bernoulli probabilities.

In this situation, increasing the concentration κ_i makes the Beta distribution more peaked around m_i and therefore reduces its entropy. In particular, $\mathrm{Beta}(1, 1)$ is the uniform distribution, which has the maximum entropy among Beta distributions. In our setting of $a_i, b_i > 1$, the density becomes increasingly concentrated and the entropy decreases.

Equivalently, we obtain

$$H(c_i a_i, c_i b_i) \;\leq\; H(a_i, b_i)$$

for every Gaussian i. Thus the entropy drop

$$\Delta H_i(v \mid S) := H_i(S) - H_i\big(a_i(S) + e_{i,1}(v),\, b_i(S) + e_{i,0}(v)\big)$$

is nonnegative, and summing over i yields $\mathrm{EIG}(v \mid S) \geq 0$. Taking expectation over the randomness of the observation preserves the inequality, so we have

$$\mathbb{E}[\mathrm{EIG}(v \mid S)] \geq 0.$$

C. Lemma 2: Adaptive Submodularity

Claim. For any $S \subseteq S'$ and any candidate view v,

$$\mathbb{E}[\mathrm{EIG}(v \mid S)] \;\geq\; \mathbb{E}[\mathrm{EIG}(v \mid S')].$$

Proof. In our formulation, the uncertainty of each Gaussian i is represented by the entropy of its Beta posterior

$$H_i(S) = H(a_i(S),\, b_i(S)).$$

The per-view information gain is defined as the total entropy reduction:

$$\mathrm{EIG}(v \mid S) = \sum_i \Big[H_i(S) - H_i(a_i(S) + e_{i,1}(v),\; b_i(S) + e_{i,0}(v))\Big],$$

which is identical to Eq. (11) in the main paper.

The success/failure count increment

$$\tau_i(v) = e_{i,1}(v) + e_{i,0}(v)$$

depends only on the chosen view v and is independent of S. On the other hand, if $S \subseteq S'$, then

$$\kappa_i(S) = a_i(S) + b_i(S) \;\leq\; \kappa_i(S') = a_i(S') + b_i(S'),$$

meaning that the concentration monotonically increases as more views are observed.

By Lemma 1, the Beta entropy $H_i(\kappa_i)$ is a monotonically decreasing function of the concentration parameter κ_i (with the mean fixed). In other words, as the pseudo-counts accumulate and κ_i becomes larger, the posterior becomes more concentrated and its entropy becomes less sensitive to additional evidence. Therefore, adding the same pseudo-count increment $\tau_i(v)$ results in a smaller change in entropy when κ_i is large. When κ_i is small, the posterior distribution is still broad, so new evidence can significantly reduce uncertainty. In contrast, when κ_i is already large and the posterior is sharply concentrated, the same $\tau_i(v)$ leads

to only a minor decrease in uncertainty. Therefore, when $S \subseteq S'$,

$$\Delta H_i(\kappa_i(S); \tau_i(v)) \;\geq\; \Delta H_i(\kappa_i(S'); \tau_i(v)). \tag{15}$$

Summing Eq. (15) over all Gaussians i, and taking expectation over the random observation, yields

$$\mathbb{E}[\mathrm{EIG}(v \mid S)] \;\geq\; \mathbb{E}[\mathrm{EIG}(v \mid S')],$$

which proves the adaptive submodularity of the proposed EIG.

Intuitively, as more views are incorporated into S', the concentration κ_i of each Gaussian increases and the Beta posterior becomes more peaked. Additional observations therefore reduce uncertainty by a smaller amount, yielding a natural diminishing-return behavior characteristic of submodular set functions.

D. Greedy $(1-1/e)$ Guarantee

In the main paper, we use the notation $\mathrm{EIG}(v \mid S)$ for the one-step expected information gain of adding a candidate view v after a set of previously selected views S, and $\mathrm{EIG}(S_k)$ for the total information gain accumulated over k steps of our greedy policy:

$$\mathrm{EIG}(S_k^{\text{greedy}}) := \sum_{t=1}^{k} \mathrm{EIG}(v_t^{\text{greedy}} \mid S_{t-1}^{\text{greedy}}).$$

For clarity, we now make explicit how this notation relates to the adaptive submodularity framework of [7]. Let $F : 2^{\mathcal{V}} \to \mathbb{R}$ be the set-level objective defined by the expected total information gain after observing a set of views S_k:

$$F(S_k) \;:=\; \sum_{t=1}^{k} \mathrm{EIG}(v_t \mid S_{t-1}). \tag{16}$$

In the notation of [7], F corresponds to the utility function f, and the conditional marginal gain $\Delta(v \mid \psi)$ under a partial realization ψ is exactly our one-step $\mathrm{EIG}(v \mid S)$ evaluated after the views in ψ.

Lemmas 1 and 2 are stated in terms of $\mathrm{EIG}(v \mid S)$, but they can equivalently be viewed as establishing that the set-level objective $F(S)$ is adaptive monotone and adaptive submodular in the sense of [7].

Therefore, by Theorem 5 and 16 of [7], the greedy policy that selects at each step $v_t^{\text{greedy}} = \arg\max_v \mathrm{EIG}(v \mid S_{t-1})$ achieves a $(1-1/e)$ approximation to the optimal adaptive policy in terms of the expected total information gain $F(S_k)$:

$$\mathbb{E}\Big[F(S_k^{\text{greedy}})\Big] \;\geq\; (1-1/e)\, \max_{\pi} \mathbb{E}[F(S_k^{\pi})]\,, \tag{17}$$

where S_k^{π} denotes the random set of views selected by any adaptive policy π. Since $F(S_k)$ equals $\mathrm{EIG}(S_k)$ by definition (16), this matches the statement used in the main paper.

E. Relationship Between Posterior Entropy and Bayes Accuracy

In our Beta-Bernoulli setting, each Gaussian i has an unknown binary label $y_i \in \{0, 1\}$ with latent probability $p_i = \Pr(y_i = 1)$. Given observations, the posterior distribution of p_i is

$$p_i \mid \text{data} \sim \mathrm{Beta}(a_i, b_i),$$

and the corresponding predictive distribution of the label is Bernoulli with predictive mean

$$q_i := \mathbb{E}[p_i \mid \text{data}] = \frac{a_i}{a_i + b_i}.$$

The predictive uncertainty of the binary label is measured by the Bernoulli entropy

$$H_{\text{pred}}(q_i) = -q_i \log q_i - (1 - q_i) \log(1 - q_i).$$

This entropy is maximized at $q_i = 1/2$ and decreases monotonically as q_i moves away from $1/2$.

Under 0-1 loss for each Gaussian, the Bayes-optimal prediction is

$$\hat{y}_i = \begin{cases} 1, & q_i \geq 1/2, \\ 0, & q_i < 1/2, \end{cases}$$

yielding Bayes accuracy

$$A_i(q_i) = \max(q_i,\, 1 - q_i).$$

Because the Bernoulli entropy is symmetric, $H_{\text{pred}}(q) = H_{\text{pred}}(1 - q)$, it depends only on the smaller of $\{q, 1 - q\}$. Hence we can write

$$H_{\text{pred}}(q) = h\big(\min(q, 1 - q)\big),$$
$$\text{where} \quad h(u) = -u \log u - (1 - u) \log(1 - u),$$

with $\min(q, 1 - q) \in [0, 1/2]$. On this interval, the binary entropy satisfies the simple lower bound

$$h(u) \;\geq\; 2(\log 2)\, u, \qquad 0 \leq u \leq 1/2.$$

Therefore,

$$H_{\text{pred}}(q) \;=\; h\big(\min(q, 1-q)\big) \;\geq\; 2(\log 2)\, \min(q, 1-q),$$

which implies

$$\min(q, 1 - q) \;\leq\; \frac{H_{\text{pred}}(q)}{2 \log 2}.$$

The per-Gaussian Bayes accuracy is

$$A(q) = \max(q, 1 - q) = 1 - \min(q, 1 - q),$$

so we obtain the explicit entropy-based lower bound

$$A(q) \;\geq\; 1 - \frac{H_{\text{pred}}(q)}{2 \log 2}. \tag{18}$$

Thus, in our Beta–Bernoulli model, reducing the predictive entropy $H_{\text{pred}}(q)$ directly tightens a linear lower bound on the per-Gaussian Bayes-optimal accuracy.

Figure 8. **Example of assumed application of B^3-Seg.** A user specifies an object in an interactive 3DGS editor through text, clicking, or dragging. Our method then performs fast and stable 3D segmentation from the current viewpoint.

F. Approximation Gap Between IG and EIG

In the main paper, the information gain (IG) of a candidate view v is defined using the actual 2D mask $M(v)$ returned by the segmentation module:

$$\mathrm{IG}(v) = \sum_i \Big[H(\mathrm{Beta}(a_i, b_i)) - H(\mathrm{Beta}(a_i + e_{i,1}(v),\, b_i + e_{i,0}(v))) \Big].$$

This quantity is deterministic once the view is rendered and its mask has been inferred.

In contrast, our analytic EIG approximates the Beta update using the posterior mean $m_i = a_i/(a_i + b_i)$:

$$\tilde{e}_{i,1}(v) = m_i \tau_i(v), \quad \tilde{e}_{i,0}(v) = (1 - m_i)\tau_i(v),$$

where $\tau_i(v) = e_{i,1}(v) + e_{i,0}(v)$ is the total responsibility of Gaussian i. Thus,

$$\mathrm{EIG}(v) = \sum_i \Big[H(\mathrm{Beta}(a_i, b_i)) - H(\mathrm{Beta}(a_i + \tilde{e}_{i,1}(v),\, b_i + \tilde{e}_{i,0}(v))) \Big].$$

To analyze the difference, define for each Gaussian i the function

$$f_i(w) = H(\mathrm{Beta}(a_i + w,\ b_i + \tau_i - w)), \quad w \in [0, \tau_i].$$

Then

$$\mathrm{IG}_i(v) = H(\mathrm{Beta}(a_i, b_i)) - f_i(W_i),$$
$$\mathrm{EIG}_i(v) = H(\mathrm{Beta}(a_i, b_i)) - f_i(\mu_i),$$

where $W_i = e_{i,1}(v)$ is the true foreground responsibility and $\mu_i = m_i \tau_i$ is the posterior-mean prediction.

Hence the approximation error is

$$\Delta_i(v) := \mathrm{IG}_i(v) - \mathrm{EIG}_i(v) = f_i(\mu_i) - f_i(W_i).$$

Because $f_i(w)$ is continuously differentiable on the interval $[0, \tau_i]$, the mean value theorem applies directly. Therefore, for some $\xi_i \in [\min(W_i, \mu_i),\ \max(W_i, \mu_i)]$,

$$f_i(\mu_i) - f_i(W_i) = f_i'(\xi_i)\,(\mu_i - W_i).$$

Therefore,

$$|\mathrm{IG}_i(v) - \mathrm{EIG}_i(v)| \le \left(\sup_{u \in [0, \tau_i]} |f_i'(u)| \right) |\mu_i - W_i|.$$

The term $|\mu_i - W_i| = \tau_i\, |m_i - W_i/\tau_i|$ reflects the mismatch between the posterior belief m_i and the true proportion W_i/τ_i inferred by the mask. As the Beta posterior becomes aligned with the observed masks, this mismatch shrinks, and so does the approximation error.

Thus, once a few updates have been incorporated, the analytic EIG becomes a highly accurate surrogate of the true IG while requiring no mask inference for candidate views, explaining the strong correlation observed in Fig. 6.

G. Assumed Application

A primary motivation behind our method is the design of interactive 3DGS editing systems as web applications. In such an application, a user inspects a scene from an arbitrary camera viewpoint and selects a target object via text prompts (e.g., "bear statue"), clicks or drag gestures (Fig. 8). The system must immediately infer the corresponding 3D region, often within a fraction of a second, before further editing or manipulation (color changes, deletion, mesh extraction, etc.) can proceed.

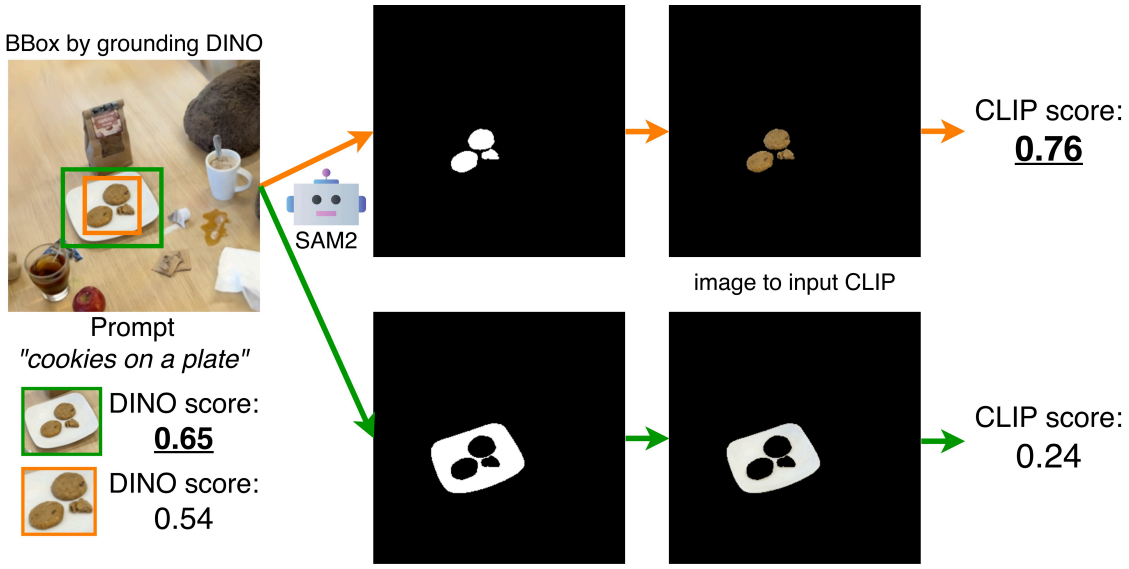


Figure 9. Effect of CLIP re-ranking in the LERF-Mask Teatime scene. Although GroundingDINO assigns a higher score to the wrong bounding box (green), CLIP correctly assigns a higher similarity score to the region corresponding to the true object described by the prompt "cookies on a plate" (orange).

This setting naturally provides a canonical initial view: the viewpoint from which the user initiates the interaction. Because the user always begins by examining the scene visually, an initial camera view for the initial object mask is available by design.

Thus, the assumption of an initial canonical view in our method is not an artificial restriction, but a realistic property of interactive 3DGS workflows. In fact, B^3-Seg is tailored for this usage scenario: its closed-form EIG, adaptively diminishing uncertainty, and $\sim$12s end-to-end inference make it particularly suitable for real-time editing environments.

H. Effectiveness of CLIP re-ranking

Text-driven region proposal by GroundingDINO is often coarse and sometimes selects a bounding box that does not match the user query with high confidence. To verify that CLIP re-ranking can correct such failure cases, we performed an ablation on the LERF-Mask Teatime scene using the prompt *"cookies on a plate"*. Figure 9 shows a representative example.

GroundingDINO produces two candidate bounding boxes for this prompt. Despite the lower DINO score, the orange box corresponds to the correct region, while the green box (incorrect) receives a higher DINO confidence (0.65 vs. 0.54). We apply SAM2 to each bounding box to obtain a candidate segmentation mask, crop the corresponding region, and compute the CLIP image–text similarity.

CLIP successfully assigns substantially higher similarity to the correct region, reversing the erroneous DINO ordering. This confirms that CLIP acts as a strong fine-grained verifier for text–region alignment, and that the two-stage DINO $\rightarrow$ SAM2 $\rightarrow$ CLIP re-ranking pipeline significantly improves the reliability of text-guided segmentation.

I. Future Directions

I.1. Multi-class Segmentation

In this work, we focused on binary foreground–background decisions with a Beta–Bernoulli model for each Gaussian. A natural extension is to support multi-class segmentation by replacing the Beta–Bernoulli pair with a Dirichlet–Categorical model.

Concretely, let each Gaussian label take values in a finite set of classes $\{1, \dots, K\}$. We introduce per-Gaussian class probabilities

$$\boldsymbol{\pi}_i = (\pi_{i,1}, \dots, \pi_{i,K}), \qquad \sum_{c=1}^{K} \pi_{i,c} = 1,$$

and place a Dirichlet prior/posterior on $\boldsymbol{\pi}_i$:

$$y_i \mid \boldsymbol{\pi}_i \sim \text{Categorical}(\boldsymbol{\pi}_i), \qquad \boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{a}_i),$$

where $\boldsymbol{a}_i = (a_{i,1}, \dots, a_{i,K})$ denotes the class-wise pseudo-counts.

Given a rendered view v and a multi-class mask, we aggregate class-specific evidence for each Gaussian in analogy to Eq. (5). Let $e_{i,c}(v)$ denote the responsibility of Gaussian g_i assigned to class c in view v (e.g., via per-pixel class masks or soft class scores):

$$e_{i,c}(v) = \sum_{(j,k)\in I(v)} \alpha_i T_i \, \mathbb{I}[M_{j,k}(v) = c], \quad c \in \{1, \dots, K\}.$$

The binary Beta update in Eq. (6) is then replaced by the Dirichlet update

$$\text{Dirichlet}(\boldsymbol{a}_i) \leftarrow \text{Dirichlet}\big(a_{i,1}+e_{i,1}(v),\ldots,a_{i,K}+e_{i,K}(v)\big),$$

and after multiple views we obtain

$$\boldsymbol{\pi}_i \sim \text{Dirichlet}\Big(a_{\text{init}}+\textstyle\sum_v e_{i,1}(v),\ldots,a_{\text{init}}+\sum_v e_{i,K}(v)\Big).$$

The posterior mean $\mathbb{E}[\pi_{i,c}] = \alpha_{i,c}/\sum_{c'}\alpha_{i,c'}$ then replaces the Beta mean used for binary decisions, enabling multi-class 3D labeling.

The information-theoretic view of Sec. 3.3 also extends naturally. The Beta entropy $H(\text{Beta}(a_i,b_i))$ in Eq. (9) and Eq. (11) is replaced by the Dirichlet entropy $H(\text{Dirichlet}(\boldsymbol{a}_i))$, and the per-view information gain becomes

$$\text{IG}(v) = \sum_i \Big[H\big(\text{Dirichlet}(\boldsymbol{a}_i)\big) - H\big(\text{Dirichlet}(\boldsymbol{a}_i+\mathbf{e}_i(v))\big)\Big],$$

where $\mathbf{e}_i(v) = (e_{i,1}(v),\ldots,e_{i,K}(v))$. The analytic EIG in Eq. (11) can analogously be defined by replacing the Beta mean m_i and entropy with their Dirichlet counterparts, using the posterior mean $m_{i,c} = \alpha_{i,c}/\sum_{c'}\alpha_{i,c'}$ to construct approximate evidence $\tilde{e}_{i,c}(v)$ for each class as follows.

$$\tilde{e}_{i,c}(v) = m_{i,c}\tau_i = \sum_{(j,k)\in I(v)} \alpha_i T_i \frac{a_{i,c}}{\sum_{c'} a_{i,c'}}$$

Thus EIG is

$$\text{EIG}(v) = \sum_i \Big[H\big(\text{Dirichlet}(\boldsymbol{a}_i)\big) - H\big(\text{Dirichlet}(\boldsymbol{a}_i+\tilde{\mathbf{e}}_i(v))\big)\Big].$$

In this Dirichlet-Categorical setting, the concentration parameter generalizes to $\kappa_i = \sum_{c=1}^{K}\alpha_{i,c}$, and the same intuition about monotonicity and diminishing returns carries over: increasing κ_i while keeping the class proportions fixed makes the posterior more concentrated and reduces the entropy. We therefore expect the adaptive monotonicity and adaptive submodularity arguments of Sec. 4 to extend to the multi-class regime, yielding a principled path toward fully multi-object, open-vocabulary 3DGS segmentation within the same EIG-driven framework.

I.2. Entropy-Based Early Stopping

The results in Appendix E imply that the predictive entropy of the Beta–Bernoulli posterior provides a quantitative lower bound on the achievable Bayes accuracy. For a Gaussian with posterior predictive probability q_i, the Bernoulli entropy $H_{\text{pred}}(q_i)$ satisfies

$$A_i(q_i) \;\geq\; 1 - \frac{H_{\text{pred}}(q_i)}{2\log 2},$$

where $A_i(q_i)$ denotes the Bayes-optimal per-Gaussian accuracy. Thus, decreasing the entropy directly tightens a guaranteed lower bound on the final segmentation accuracy.

This observation enables a principled *entropy-based early stopping* criterion for our sequential inference pipeline. Let

$$\bar{H}_t := \frac{1}{N}\sum_{i=1}^{N} H_{\text{pred}}(q_i^{(t)})$$

denote the average predictive entropy at iteration t. From the inequality above, the average Bayes accuracy after iteration t is lower bounded as

$$\bar{A}_t \;\geq\; 1 - \frac{\bar{H}_t}{2\log 2}.$$

Therefore, instead of fixing the number of iterations to a constant T, one can run the view-selection loop until $\bar{H}_t$ falls below a user-specified threshold $\bar{H}_{\text{target}}$, corresponding to a desired guaranteed accuracy level

$$\bar{A}_{\text{target}} \;=\; 1 - \frac{\bar{H}_{\text{target}}}{2\log 2}.$$

Such an adaptive stopping rule ensures that inference terminates once the posterior has become sufficiently concentrated, avoiding unnecessary queries while maintaining provable segmentation quality. Integrating this early-stopping mechanism with analytic EIG is a promising future direction for further improving the responsiveness and efficiency of B^3-Seg.

J. Additional Qualitative Results

To further illustrate the generality and robustness of B^3-Seg, Figures 10, 11, and 12 present additional qualitative results on each 3D dataset. Across a wide variety of object types—ranging from small items (e.g., fruit, plates, napkins) to deformable or texture-rich objects (e.g., stuffed animals, clothing, wooden materials), our method generally produces clean and spatially coherent 3D masks.

Notably, B^3-Seg successfully handles objects with thin structures, strong self-occlusion, low contrast against the background, and cases where multiple distractors appear in close proximity. These results highlight that (i) analytic EIG selects informative viewpoints even in cluttered scenes, and (ii) the sequential Beta–Bernoulli updates accumulate stable evidence despite imperfect 2D masks.

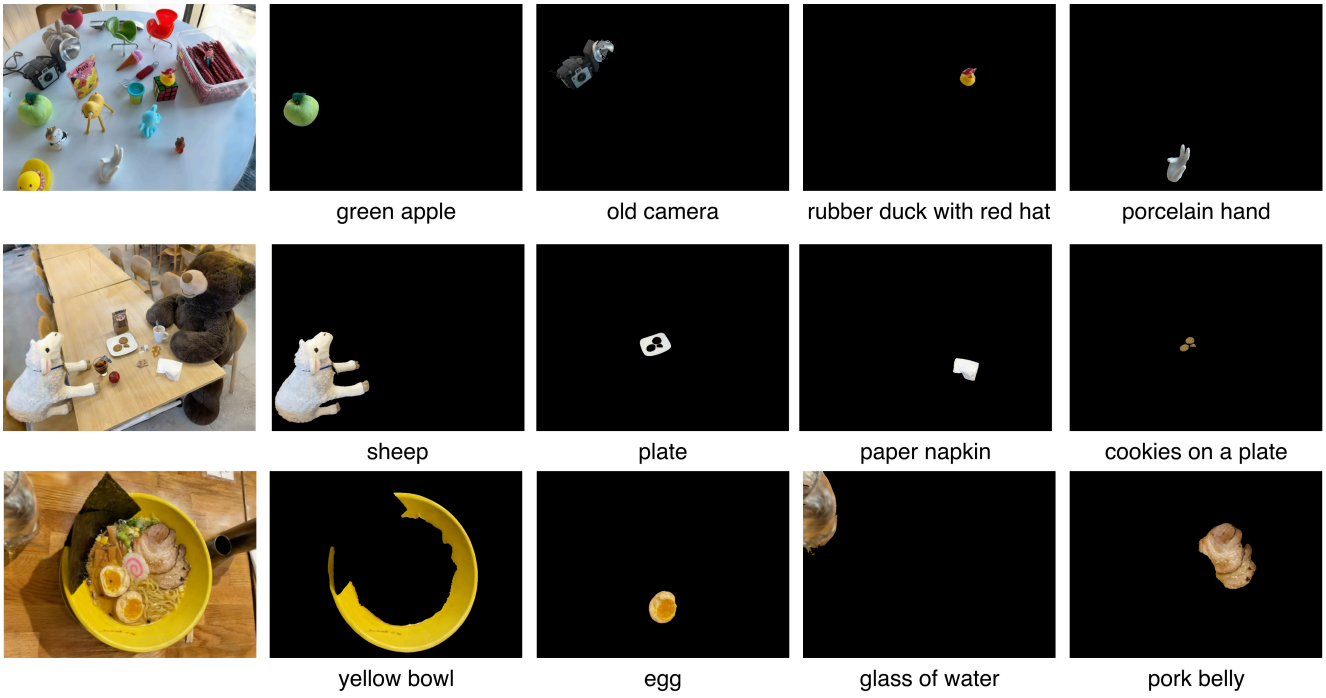


Figure 10. Qualitative results of B^3-Seg on LERF-Mask

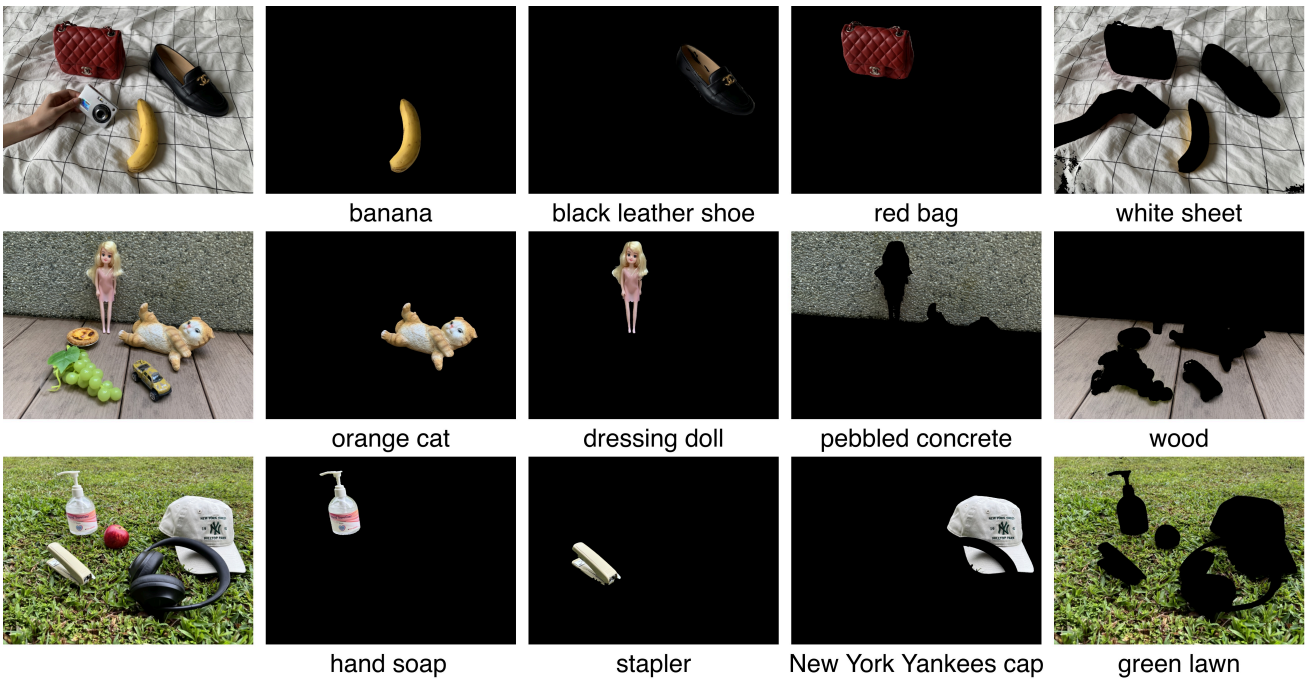
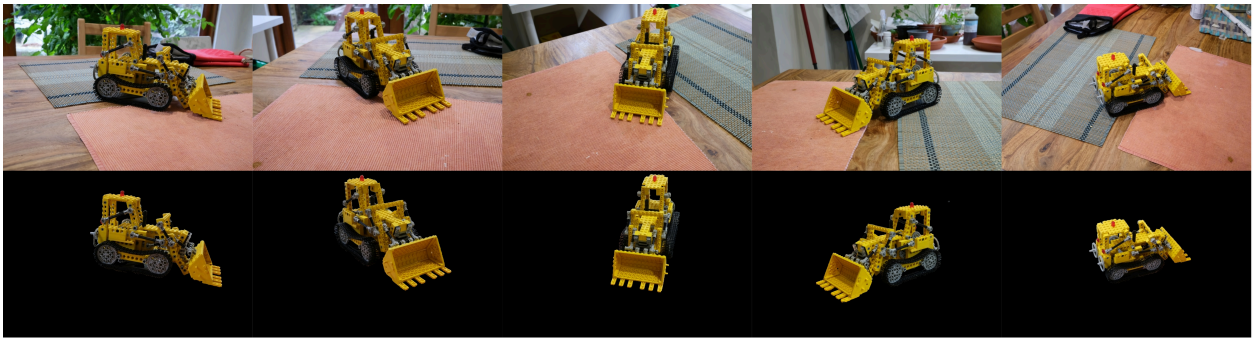


Figure 11. Qualitative results of B^3-Seg on 3D-OVS



Prompt: "Lego"

Prompt: "pinecone"

Prompt: "bear statue"

Prompt: "flowers"

Figure 12. Qualitative results of other 3D scenes