INFOFIELD3D: DIFFERENTIABLE INFORMATION FIELD FOR EMBODIED TASK-ORIENTED VIEW PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective task execution requires embodied agents to select observation positions that provide sufficient task-relevant information. Existing approaches typically predict viewpoints with learned models or score discrete candidates, but they either struggle to incorporate new constraints such as reachability or require extensive candidate evaluation to search. We introduce InfoField3D, which lifts geometry-bound information into a differentiable analytic field, whose value at each 3D observation position represents the task-required information accessible there. It uses low-order spherical harmonics to encode directional information responses and analytic line integrals of Gaussian occupancy functions to model occlusion along lines of sight. This formulation casts task-driven camera placement as a well-studied constrained continuous optimization problem, enabling both flexible incorporation of constraints through the feasible domain and efficient viewpoint search even in the presence of occlusion. Experiments show that InfoField3D compactly preserves the pointwise information landscape with high fidelity and generally recommends higher-quality, more robust viewpoints at lower search cost than the evaluated baselines. Controlled evaluations with a frozen vision-language-action (VLA) policy further suggest that the recommended views can improve downstream task performance under information-limited initial viewpoints.

1 INTRODUCTION

Effective task execution for embodied agents hinges on the acquisition of sufficient task-relevant information, whose availability is strongly conditioned by the agent's observation viewpoint (Cowan & Kovesi, 1988; Tarabanis et al., 1995). This necessity gives rise to the task-driven view planning task (Breyer et al., 2022; Shi et al., 2025). Existing approaches to this task commonly follow two main lines. The first line leverages learning-based models to directly predict the observation viewpoint (Bai et al., 2026), yet such implicit end-to-end approaches struggle to accommodate constraints in novel scenes, such as reachability. The second line adopts sampling-based models to evaluate abundant candidate viewpoints (Dai et al., 2025; Xie et al., 2025; Jiang et al., 2024), which can suffer from discontinuity over the continuous spatial domain. Moreover, both families of methods perform predicting or scoring operations that rely on geometry-bound surface features, for which effective occlusion handling remains nontrivial in this task and may affect the practical accuracy of the recommended viewpoints.

In this paper, we propose a new paradigm dubbed InfoField3D for embodied task-oriented view planning. Rather than directly operating on geometry-bound features, InfoField3D lifts the surface information to an analytic field, whose value at each 3D observation position represents the task-required information accessible from that location. Specifically, InfoField3D represents the inherent directional information distribution of dense surface samples with low-order spherical harmonics, enabling it to be computed in a differentiable and low-redundancy manner. Meanwhile, conventional discrete occlusion computation is reformulated as a continuous representation of object occupancy in space, with the degree of occlusion quantified by integrating occupancy along line-of-sight paths. In this way, information accessibility in the workspace is transformed into a fully differentiable analytic formulation.

InfoField3D offers three core advantages: (i) It casts viewpoint planning as a well-established constrained continuous optimization problem defined over the feasible domain, enabling embodied

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

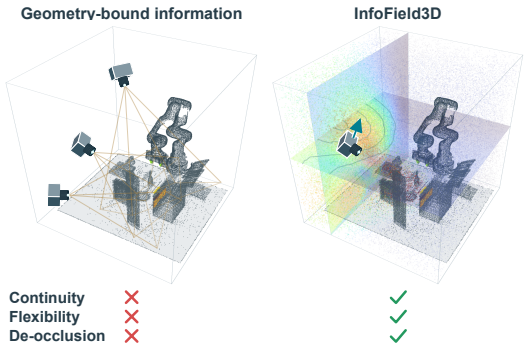


Figure 1: View planning based on (1) Geometry-bound information *vs.* (2) InfoField3D.

agents to efficiently search for high-quality viewpoints using analytic gradients. (ii) New constraints (e.g., reachability) can be flexibly incorporated into the field-based formulation, facilitating better adaptation to novel scenes and tasks. (iii) Occlusion is modeled in a differentiable form, providing gradients that guide viewpoints away from occlusion and toward positions with larger visibility margins.

Experiments on publicly available benchmarks show that InfoField3D closely approximates the spatial information distribution computed directly from the original point cloud using substantially fewer representation parameters. Compared with prior art, it generally recommends higher-quality observation positions with greater robustness to positional perturbations and lower online search cost. Controlled evaluations with a frozen VLA model suggest that the recommended views can provide modest improvements in downstream task performance.

2 Related Work

Existing methods for task-driven observation planning can be broadly grouped into the following two categories according to their primary mechanism for selecting or updating observation positions.

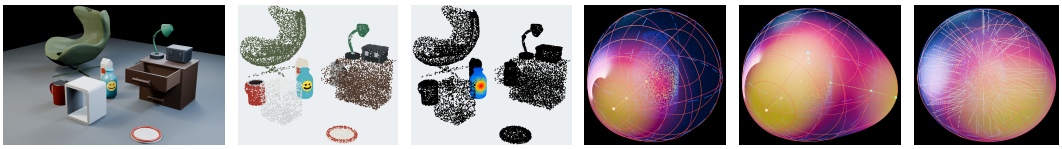
Sampling-based methods. These methods select observation positions by scoring candidate viewpoints or paths (Triggs & Laugier, 1995; Breyer et al., 2022; Lee et al., 2026; Jauhri et al., 2024), using geometric, semantic, or information-based criteria from 3D scenes (Liu et al., 2026b; Nagami et al., 2026; Dai et al., 2025; Ong et al., 2026; Burusa et al., 2024a; Song & Karydis, 2025) to rank candidates for tasks such as reconstruction, exploration, and grasping (Xie et al., 2025; Jiang et al., 2024; 2025; Lei et al., 2026; Strong et al., 2025; Wilson et al., 2025; Song et al., 2026). These methods can accommodate diverse scenes and tasks, but candidate scores themselves do not directly provide directions for continuous position updates, so selection quality depends on candidate coverage and the evaluation budget. Precomputed discrete fields move part of the evaluation to an offline stage and support online queries through interpolation, but introduce grid-dependent approximations (Zhang & Scaramuzza, 2020). Some methods differentiate utility surrogates fitted from samples (Wang et al., 2024) or discretized sensing estimators (Burusa et al., 2024b; Zhu et al., 2026), whose evaluation and gradients still depend on discrete sampling. Analytic geometric control is also used in some methods to directly generate motion directions based on target–occluder relationships or line-of-sight geometry (Shi et al., 2025; He et al., 2022); however, such guidance serves predefined geometric objectives and need not correspond to task-adjustable information utilities and their gradients.

Learning-based prediction methods. These methods use neural networks to directly predict camera poses or motions, conditioned on scene observations, task text, and other inputs (Bai et al., 2026; Li & Kasaei, 2025; Jayaraman & Grauman, 2018; Liu et al., 2025; Qin et al., 2026; Koo et al., 2025), or jointly learn camera policies together with manipulation policies (Liu et al., 2026a; Chuang et al., 2025; Xiong et al., 2025; Kerr et al., 2025; Wang et al., 2025). They support efficient inference, but their generalization to scenes, occlusion relationships, tasks, and embodiments outside the training distribution may be limited. In addition, such implicit end-to-end approaches typically lack explicit constraint interfaces, making it difficult to incorporate constraints in novel scenes, such as reachability.

3 InfoField3D: Formulation and Methodology

3.1 Problem Formulation and Overview

This paper considers task-oriented view planning in a reconstructed static scene represented by a point cloud enriched with geometry-bound semantic features (Kerr et al., 2023; Zhou et al., 2024; Qin et al., 2024). Our goal is to model the accessibility of geometry-bound scene information as a function of the observation position $\mathbf{x} \in \Omega \subseteq \mathbb{R}^3$, where Ω denotes the feasible domain satisfying the



(a) (b) (c) (d) (e) (f)

Figure 2: Task example and the construction from pointwise responses to a compact directional representation: (a) the task scene; (b) its point cloud; (c) the task-relevance map; (d–e) progressive superposition of pointwise angular responses on the reference sphere, weighted by task relevance; (f) the aggregated object-level distribution.

prescribed constraints. To remain independent of specific camera configurations, we assume ideal panoramic observation and optimize only the observation position, with practical camera constraints handled through orientation rules and Ω. To facilitate exposition, here we use an intuitive example to demonstrate the entire construction process of InfoField3D: *Consider an embodied agent tasked with writing a signature on a smiley-face sticker attached to a bottle (Fig. 2(a)). Which observation position best supports this task?*

To determine what needs to be observed for signing, a multimodal model interprets the task condition together with the current RGB images, drawing on learned priors to identify, for example, the smiley-face sticker as the primary observation target and formulate the query "the smiley-face sticker on the bottle." Applying this query to the semantic point cloud assigns each point a nonnegative task-relevance score s_i, forming a ***task-relevance map*** over the scene. The total relevance score of the points visible from a given position provides an intuitive measure of observation quality.

Let $J_{\mathcal{T}}(\mathbf{x})$ denote a scalar measure of the task-relevant information accessible at position $\mathbf{x}$ under task condition $\mathcal{T}$. The desired observation position maximizes this response over Ω:

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\Omega} J_{\mathcal{T}}(\mathbf{x}). \tag{1}$$

The problem is then reduced to determining $\mathbf{x}$ so that $J_{\mathcal{T}}(\mathbf{x})$ is maximized. InfoField3D constructs $J_{\mathcal{T}}(\mathbf{x})$ as a differentiable analytic function of the observation position. The following sections derive this construction in turn, covering directional information responses (Section 3.2), occlusion along the propagation path (Section 3.3), and the local geometric units that extend the construction to the full scene (Section 3.4). For a visual overview of the overall workflow, see Figure 3. Appendix A provides additional construction details and several practical extensions.

3.2 Differentiable Modeling of Information Propagation

The derivation first considers the bottle without occlusion from other objects. Throughout Sections 3.2–3.3, we fix $\mathcal{T}$ and suppress its dependence in the notation.

Pointwise response. For a surface point $\mathbf{p}_i$, its information exposure varies with the alignment between its outward surface normal and the direction toward the observation position $\mathbf{x}$. The cosine of the angle between these directions gives a continuous and differentiable measure of the information exposure of $\mathbf{p}_i$ at $\mathbf{x}$, and a smooth kernel $K(t)$ of the cosine value t (exponential in our implementation) can further adjust the response shape and angular concentration. For $\mathbf{p}_i$ with unit normal $\mathbf{n}_i$ and task-relevance score s_i, its information contribution at $\mathbf{x}$ ($\mathbf{x} \neq \mathbf{p}_i$) is $f_i(\mathbf{x}) = s_i K\left(\mathbf{n}_i \cdot \frac{\mathbf{x}-\mathbf{p}_i}{\|\mathbf{x}-\mathbf{p}_i\|}\right)$. For an approximately uniform N-point cloud, the individual point contributions sum to the total response at an observation position $\mathbf{x}$, $F(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x})$.

Spherical mapping. The response $F(\mathbf{x})$ is continuous and differentiable at valid observation positions, but its representation size and evaluation cost remain proportional to the point count. Dense geometric sampling, however, need not imply an equally complex information distribution. This motivates seeking an object-level representation whose complexity is no longer tied to the point count. To place the points' angular responses on a common domain, InfoField3D constructs a reference sphere S centered at the center $\mathbf{c}$ of the object's minimum enclosing sphere, with a radius R larger than the minimum enclosing radius. For each point, a ray along its outward normal intersects the

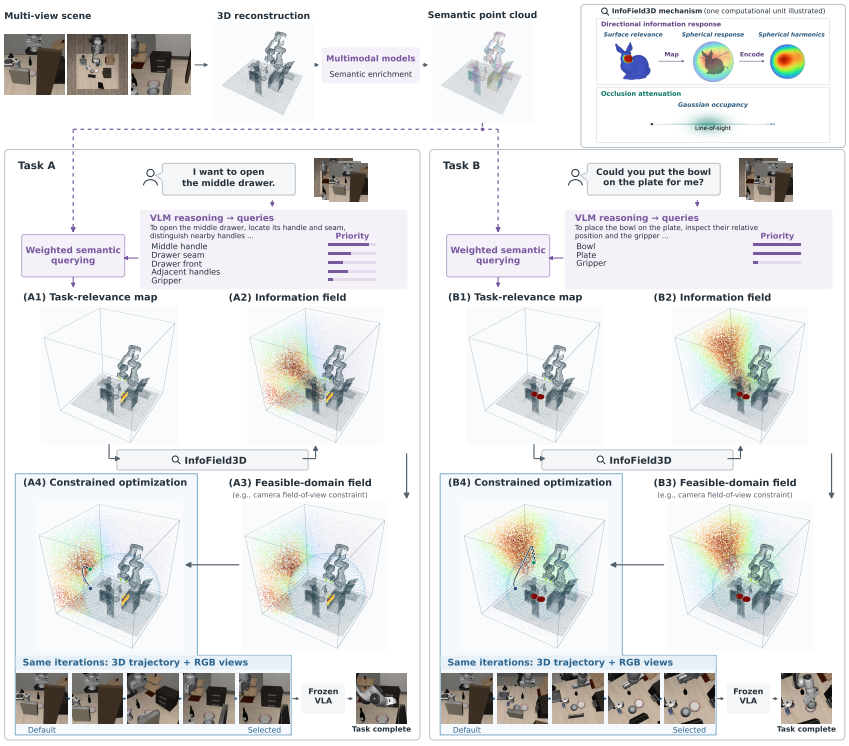


Figure 3: **Overview of InfoField3D.** Task-relevance maps (A1, B1) can be obtained from multimodal foundation models conditioned on the task and the reconstructed 3D scene; the figure shows one illustrative example. InfoField3D lifts geometry-bound information into a differentiable analytic field, whose value encodes task-relevant information accessibility at each observation position (A2, B2). The field is then restricted to the feasible domain (A3, B3), and the recommended observation position is obtained through constrained optimization over this field (A4, B4). Within the same scene, viewpoints recommended according to different task requirements are provided to a frozen VLA to execute the corresponding tasks.

sphere, defining the center of a continuous angular response. Weighted by the point's task-relevance score, this response forms a directional hill that decays with angular distance from the intersection. Superimposing the hills from all points yields an object-level directional response distribution whose high-response regions identify directions with greater task-relevant information (Figure 2(d-f)).

On the spherical domain, this mapping yields the approximate response $F_{\mathrm{sph}}(\mathbf{u}) = \sum_{i=1}^{N} s_i K(\mathbf{r}_i \cdot \mathbf{u})$. The above holds for all unit directions $\mathbf{u}$, where $\mathbf{r}_i$ denotes the unit direction from $\mathbf{c}$ to the intersection of the outward-normal ray from $\mathbf{p}_i$ with the reference sphere. This spherical construction provides a basis for seeking a compact representation of the object's directional response. Spherical harmonics provide an orthogonal basis adapted to the sphere's rotational symmetry and organized by angular frequency (Ramamoorthi & Hanrahan, 2001), motivating a finite-degree expansion to test whether a few low-frequency modes can preserve the distribution's main structure.

Harmonic encoding. The degree-L approximation of F_{sph} is $F_L(\mathbf{u}) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} c_{\ell m} Y_{\ell m}(\mathbf{u})$, where $Y_{\ell m}$ are the spherical harmonic basis functions and $c_{\ell m}$ are the coefficients. For a fixed degree L, the representation requires only $(L+1)^2$ coefficients, regardless of the point count. The evaluations in Section 4.1 show that this compact representation preserves the original pointwise spatial responses with high fidelity, supporting the hypothesis that a few low-frequency modes suffice. The following derivation therefore uses this compact representation for observation-position optimization. For $\mathbf{x} \neq \mathbf{c}$, the unit direction from the center to the observation position is $\mathbf{u}(\mathbf{x}) = (\mathbf{x} - \mathbf{c})/\|\mathbf{x} - \mathbf{c}\|$. The object-level response at $\mathbf{x}$ is then $F_L(\mathbf{u}(\mathbf{x}))$, serving as the objective function for observation-position search. The search therefore reduces to maximizing it over Ω:

$$\mathbf{x}^\star = \arg\max_{\mathbf{x} \in \Omega} F_L(\mathbf{u}(\mathbf{x})). \tag{2}$$

Since the spherical harmonic basis functions and direction mapping are continuously differentiable, the chain rule gives the objective's analytic spatial gradient: $\nabla_{\mathbf{x}} F_L(\mathbf{u}(\mathbf{x})) = \frac{1}{\|\mathbf{x}-\mathbf{c}\|}\left(\mathbf{I} - \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\top\right)\nabla_{\mathbf{u}} F_L(\mathbf{u}(\mathbf{x}))$. The problem of increasing the objective can now be formulated as an optimization problem.

3.3 Differentiable Modeling of Occlusion

How should this formulation account for occlusion along the propagation path? With a cube placed in front of the bottle, an intuitive score sums the task-relevance scores of the bottle's unoccluded points. However, classifying points as occluded or unoccluded provides no useful spatial gradients for position updates.

Path attenuation. To address this limitation, InfoField3D shifts the perspective on occlusion. Instead of classifying points as occluded or unoccluded, it quantifies the spatial occupancy accumulated along the propagation path. Once this occupancy is represented as a continuously differentiable function, the accumulated quantity can be given by a line integral, which remains differentiable with respect to the observation position. In this example, a Gaussian occupancy function ρ approximates the cube's occupied region. Using the object's reference center $\mathbf{c}$ as its representative source location, let $\tau(\mathbf{x})$ denote the occupancy accumulated along the finite segment from $\mathbf{c}$ to the observation position $\mathbf{x}$, given by $\tau(\mathbf{x}) = \int_{[\mathbf{c},\mathbf{x}]} \rho(\mathbf{z})\,\mathrm{d}s$, and let the corresponding transmittance be $T(\mathbf{x}) = \exp\left(-\tau(\mathbf{x})\right)$.

The transmittance $T(\mathbf{x})$ represents the fraction of the directional response retained after propagation, so multiplying it by the directional response $F_L(\mathbf{u}(\mathbf{x}))$ gives the task-relevant response available at $\mathbf{x}$ under occlusion. This product defines the single-object objective $J_{\mathrm{obj}}(\mathbf{x}) = T(\mathbf{x})F_L(\mathbf{u}(\mathbf{x}))$, and observation-position search becomes

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\Omega} T(\mathbf{x})F_L(\mathbf{u}(\mathbf{x})) = \arg\max_{\mathbf{x}\in\Omega} J_{\mathrm{obj}}(\mathbf{x}). \tag{3}$$

The objective $J_{\mathrm{obj}}(\mathbf{x})$ remains continuously differentiable and can be maximized in the same way as in Section 3.2. Its gradient guides the search toward positions with more task-relevant information available under occlusion.

Visibility margins. Notably, scores based solely on binary visibility saturate once the target is fully visible. The present objective does not stop at this point, because the Gaussian occupancy extends over the entire space; paths passing close to an occluder still accumulate attenuation, so the objective can favor positions with larger observation margins and lower occlusion risk. Consistent with this effect, the experiments in Section 4.3 show that InfoField3D recommends observation positions with greater robustness to positional perturbations.

3.4 Defining the Geometric Units of the Field

The preceding examples use whole objects as units. Representing an entire object's directional response around a shared center can introduce geometric mapping loss before spherical harmonic encoding, and likewise for a single-Gaussian approximation of an entire occluder. For example, a chair has spatially separated legs, each with its own directional information response and spatial occupancy, making the whole object difficult to represent around a shared center without loss. InfoField3D therefore does not adopt whole objects or predefined semantic parts as default computational units, since either may have complex geometry. Instead, it constructs local geometric units suited to its representations, each with its own reference center.

Geometric partitioning. Experiments in Appendix A.1.5 show that geometrically complex, less rounded units tend to incur greater loss under a single-center representation. Guided by this observation, we construct local units to be more geometrically compact using a simple recursive partitioning scheme without semantic part priors: the unit with the greatest spatial spread is repeatedly split into two balanced subsets along its principal axis until the desired granularity is reached. Each resulting unit serves as a basic computational unit. Appendix A.1.5 further shows that a small number of such splits substantially improves accuracy.

Field aggregation. In the resulting computation, each unit acts as both an information source and an occluder for other units, with the source unit's own spatial occupancy excluded from its propagation path. For example, once a cup's handle and body belong to different units, their mutual occlusion is included in the calculation. With M local units, the same information-response and occlusion rules define the objective over the full scene. For unit j, let $\mathbf{c}_j$ denote its reference center, ρ_j its Gaussian occupancy function, and $F_{j,L}^{(\mathcal{T})}$ its directional response under the task condition $\mathcal{T}$. The observation direction is $\mathbf{u}_j(\mathbf{x}) = (\mathbf{x} - \mathbf{c}_j)/\|\mathbf{x} - \mathbf{c}_j\|$. The accumulated path occupancy is $\tau_j(\mathbf{x}) = \int_0^1 \left[\sum_{k \neq j} \rho_k\big(\mathbf{c}_j + t(\mathbf{x} - \mathbf{c}_j)\big)\right] \|\mathbf{x} - \mathbf{c}_j\| \, \mathrm{d}t$, with the corresponding transmittance $T_j(\mathbf{x}) = \exp\big(-\tau_j(\mathbf{x})\big)$. Aggregating the attenuated responses over all units gives the full-scene objective under the task condition, $J_{\mathcal{T}}(\mathbf{x}) = \sum_{j=1}^{M} T_j(\mathbf{x}) F_{j,L}^{(\mathcal{T})}\big(\mathbf{u}_j(\mathbf{x})\big)$, which realizes the objective $J_{\mathcal{T}}(\mathbf{x})$ posed in Section 3.1. Observation-position search is then explicitly formulated as

$$\mathbf{x}^\star = \arg\max_{\mathbf{x} \in \Omega} J_{\mathcal{T}}(\mathbf{x}) = \arg\max_{\mathbf{x} \in \Omega} \sum_{j=1}^{M} T_j(\mathbf{x}) F_{j,L}^{(\mathcal{T})}\big(\mathbf{u}_j(\mathbf{x})\big). \tag{4}$$

As in the single-object examples, this objective remains continuously differentiable over Ω. This completes the construction of InfoField3D, enabling continuous optimization of task-conditioned observation positions with analytic gradients. Further extensions, including a formulation for frequent task queries, are presented in Appendix A.

4 Experimental Evaluation

The experiments evaluate representation fidelity, optimization quality and efficiency, viewpoint robustness, and downstream effects on a frozen VLA policy. Detailed experimental and implementation settings are provided in Appendices B, C, D, and E.

4.1 Representation Fidelity and Compression

Experimental setup. This section evaluates how faithfully InfoField3D's analytic field reproduces pointwise computation on the original point cloud. We use 1,655 static scenes from LIBERO and LIBERO-Plus (Liu et al., 2023; Fei et al., 2026), with approximately 800 uniformly sampled test positions per scene, under both uniform- and task-relevance maps. At each position, InfoField3D is compared with pointwise computation using exact occlusion. Table 1 reports the results.

Fidelity metrics. Optimal-value attainment measures how much of the best pointwise-reference value in the candidate set is attained at the field-selected position, emphasizing the preservation of high-value solutions most relevant to observation-position selection. Shape consistency is the Pearson correlation between response values, and ranking consistency is the correlation between their position rankings. All metrics are computed per scene and then averaged across scenes.

Results. As Table 1 shows, the representation retains high spatial fidelity under both relevance maps and, notably, uses only about 1/15 of the parameter budget of the original point cloud. To stress-test this compression, we densify the point clouds to 2×, 4×, and 16× while keeping the field budget fixed. Even at 16× density, with a point-cloud-to-field parameter ratio of about 239:1, all three metrics decrease by less than 0.4 pp, further supporting the low-complexity hypothesis in Section 3.2.

4.2 Quality and Efficiency of Observation-Position Optimization

Experimental setup. This section compares the quality and efficiency of observation-position optimization across methods. We evaluate all methods on 80 LIBERO-Plus object-layout scenes (20

Table 1: Mean fidelity of InfoField3D relative to pointwise spatial information responses.

Information setting	Optimal-value attainment ↑	Shape consistency ↑	Ranking consistency ↑
Uniform relevance map	97.93%	92.28%	90.51%
Task-relevance map	97.97%	93.86%	91.29%

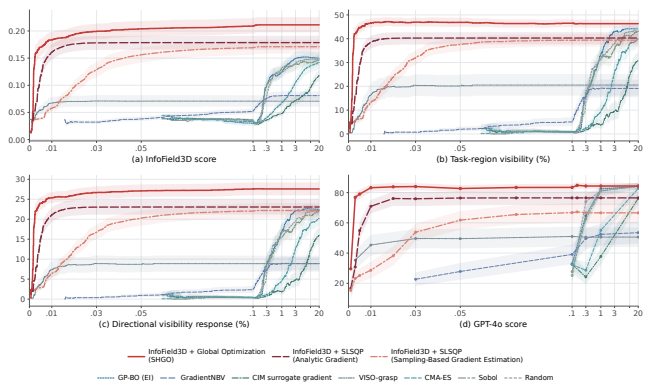


Figure 4: Viewpoint scores of different methods versus search time. Each panel reports the metric indicated by its title on the vertical axis, with search time (s) on the horizontal axis. The first three curves optimize the same InfoField3D objective; the two SLSQP variants differ only in gradient source.

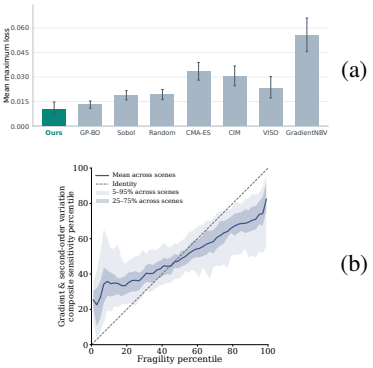


Figure 5: Fragility analysis: (a) perturbation loss of each method's recommended viewpoints; (b) Correlation analysis between InfoField3D's multi-order derivatives and fragility.

per suite), with 10 runs per scene under the same evaluation protocol. The result and time cost of each iteration are recorded. Results are first averaged within each scene and then averaged equally across scenes. All methods search over the same feasible domain within a cubic 3D workspace.

In the experiments, the task-relevant target regions and their corresponding weights are generated by a fully automated multimodal annotation pipeline (Kirillov et al., 2023; Radford et al., 2021; Cherti et al., 2023) and shared across all methods, thereby controlling for potential effects arising from differences in task interpretation. Viewpoint quality is evaluated from four complementary aspects. Task-region visibility $V(\mathbf{x})$ measures the relevance-weighted fraction of the task-relevant surface that is visible. Directional visibility response further weights visible task points by how directly their surfaces face the viewpoint. Although InfoField3D is the method proposed in this work, we believe its viewpoint score is still informative because it jointly considers viewing orientation, region exposure, and visibility margin. We therefore also report it as an auxiliary reference metric. In addition to these rule-based scoring metrics, we also use a multimodal model (GPT-4o), which scores each viewpoint based on the given task description and its rendered image, to provide an auxiliary evaluation perspective informed by the model's broad semantic and commonsense priors.

Optimization variants. For our method, once InfoField3D is constructed, view planning reduces to constrained optimization over the feasible domain, allowing standard optimization methods to be directly applied. We can therefore directly use an existing off-the-shelf optimizer to solve it. In our experiments, we simply use the SHGO solver (Endres et al., 2018) provided by SciPy (Virtanen et al., 2020), supplying our analytic gradients to its local optimizer. In addition, to isolate the benefit of analytic differentiation, we adopt a standard gradient-based constrained optimizer (SLSQP) (Kraft, 1994) on the same InfoField3D field, supplying either the analytic gradient or a sampling-based gradient estimate while keeping all other settings fixed.

Table 2: Viewpoint quality under different search budgets.

	InfoField score ↑		Task vis. (%) ↑		Dir. resp. (%) ↑		GPT-4o score ↑	
Method	0.1 s	20 s	0.1 s	20 s	0.1 s	20 s	0.1 s	20 s
InfoField3D (SHGO)	***0.2091***	***0.2113***	***46.51***	***46.32***	***27.59***	***27.58***	***83.42***	84.33
InfoField3D (Analytic Gradient)	0.1786	0.1786	40.30	40.30	23.06	23.06	76.50	76.50
InfoField3D (Sampling-Based Gradient)	0.1691	0.1710	39.29	39.37	22.02	22.20	66.77	66.60
GP-BO	0.0299	0.1497	0.71	44.12	0.34	22.46	32.89	84.10
CIM surrogate gradient	0.0360	0.1177	0.93	30.77	0.45	15.98	32.61	76.19
CMA-ES	0.0333	0.1419	0.87	40.01	0.42	20.24	32.36	82.73
GradientNBV	0.0526	0.0809	5.22	19.11	2.46	8.89	39.13	53.50
VISO-grasp	0.0708	0.0707	20.53	20.54	8.87	8.87	50.98	50.58
Sobol	0.0338	0.1470	0.78	43.37	0.38	22.21	27.74	***84.40***
Random	0.0332	0.1497	0.73	43.20	0.35	22.17	25.22	84.04

Baselines. We further compare Sobol-sampled global search (Sobol', 1967), uniform random search, CMA-ES (Hansen & Ostermeier, 2001), Bayesian optimization with expected improvement (GP-BO) (Jones et al., 1998), GradientNBV (Burusa et al., 2024b), a CIM-style surrogate-gradient method (Wang et al., 2024), and VISO-Grasp-style analytic control (Shi et al., 2025). Learned viewpoint predictors are omitted because reproducible alternatives use substantially different observation or action spaces, while methods under more comparable settings do not provide released models. Some methods, such as FIF (Zhang & Scaramuzza, 2020) and GO-PRE (Song et al., 2026), define information utility through localization observability and predictive rendering uncertainty, respectively, and can also be viewed, to some extent, as addressing camera placement. However, transferring them to the task considered in this experiment would require additional, unvalidated modeling choices. They are therefore excluded from the comparison to avoid conflating adaptation-induced errors with the capabilities of the original methods.

Results. As shown in Figure 4 and Table 2, InfoField3D optimized with SHGO shows a pronounced efficiency advantage, reaching high viewpoint quality within the first 0.1s of search time while remaining among the best-performing methods in final viewpoint quality across the evaluated metrics. Notably, its advantage extends beyond exposing more of the task-relevant region: it also achieves a markedly stronger directional visibility response, indicating that the visible task surfaces are observed more frontally. This behavior is consistent with InfoField3D's explicit modeling of directional information response. Separately, when optimizing the same InfoField3D field with SLSQP, the analytic gradient reaches high-quality solutions substantially earlier than the gradient estimated from six local samples, while the SHGO variant remains the strongest overall, showing that InfoField3D benefits both from an optimization-friendly field and from exact derivatives.

4.3 Robustness of Recommended Observation Positions

The preceding experiments evaluate how much task-relevant information a viewpoint can provide. However, practical deployment also requires robustness to positioning errors and environmental disturbances. If a viewpoint's observation ability collapses under slight perturbation, it is not robust—for example, when the line of sight threads a narrow gap between occluders or the target has only just emerged from behind an occluder's edge. This section evaluates whether the viewpoints recommended by each method remain reliable under such perturbations.

Perturbation tests. We define the *fragility* of a viewpoint, the opposite of robustness, as the worst-case information loss under small spatial perturbations, where information is quantified by the score $V(\mathbf{x})$ defined in Section 4.2. We perturb the final viewpoints recommended by each method and record their maximal loss in task-region visibility. As shown in Figure 5(a), InfoField3D demonstrates advantages in the robustness of its final recommended viewpoints.

Derivative-based analysis. Explicit perturbation tests can assess viewpoint robustness but incur additional computational overhead. Since fragile viewpoints often correspond to rapid information changes within their local neighborhoods, it is natural to seek local sensitivity signals from derivatives and their higher-order variations. As a concrete example, we sum first-order gradients with second-order information into a composite signal for analysis. As shown in Figure 5(b), this signal exhibits a certain degree of correlation with external fragility, suggesting that local derivative information of the field may provide useful auxiliary cues for fragility analysis.

4.4 Controlled Downstream Evaluation with a Frozen VLA

This section asks, under a frozen VLA policy ($\pi0.5$ (Black et al., 2025) in our experiments), whether InfoField3D's viewpoint optimization can improve task execution.

Experimental setup. Whether a VLA policy benefits from a viewpoint depends not only on observation quality but also on camera-pose distribution shift. A global comparison over widely separated viewpoints would therefore confound these effects. We instead use a local neighborhood-rescue protocol: from a third-person viewpoint with limited access to task-relevant information, InfoField3D proposes a nearby viewpoint with improved task information, requiring only a small displacement from the starting viewpoint. Because the two viewpoints remain very close in pose space, their proximity helps reduce camera-pose distribution shift, allowing the effect of improved information accessibility to be examined more directly. The displacement to the recommendation

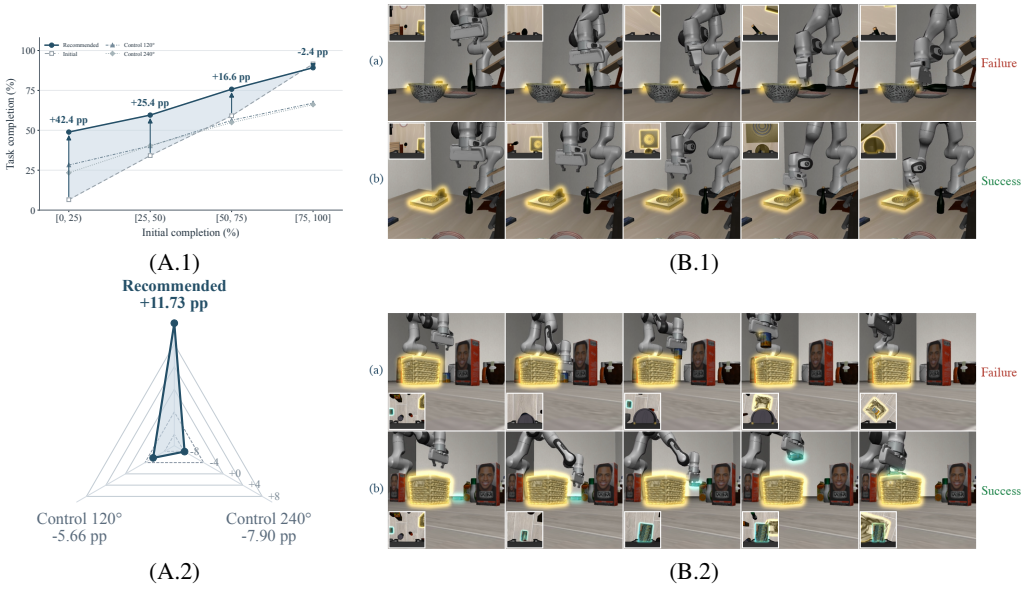


Figure 6: Trident-probe outcomes and representative VLA rescue episodes. (A.1) Mean task completion at the initial, recommended, and control viewpoints, stratified by initial-completion tier. (A.2) Mean completion change from the trapped viewpoint for the recommendation and two equal-distance controls. In (B.1–B.2), rows (a) and (b) show rollouts from trapped and recommended third-person viewpoints, respectively, with identical initial wrist-camera states; highlights mark task-relevant objects.

is rotated by $\pm 120^\circ$ to form two equal-length controls, yielding a trident probe. Each endpoint is evaluated with five frozen-VLA rollouts using different seeds. Across 80 LIBERO-Plus scenes, the above protocol yields 697 information-limited viewpoints, for which the corresponding probe groups are constructed. Quantitative outcomes and representative VLA episodes are shown in Figure 6.

Results. Figure 6(A.1) shows that the benefit is largest when initial completion is low ($+42.4$ pp for the 0–25% tier) and diminishes as initial completion increases. When completion is already high (75–100%), the recommendation produces a small decrease (-2.4 pp), suggesting that viewpoint optimization is most beneficial when observation is a limiting factor. Figure 6(A.2) shows that InfoField3D's recommendation improves task completion from the trapped viewpoint by $+11.73$ pp on average, substantially more than the two equal-distance controls. Figures 6(B.1–B.2) illustrate representative cases in which moving from a trapped to the recommended viewpoint changes VLA execution from failure to success.

5 Conclusion and Limitations

This work treats task-conditioned observation-position selection as optimization over an explicit information landscape in free space rather than repeated evaluation of isolated viewpoints. InfoField3D reorganizes geometry-bound scene information into a differentiable analytic field whose value measures task-relevant information accessibility and whose spatial gradient directly guides viewpoint improvement. Experiments show that the field preserves the underlying information landscape, enables efficient continuous optimization, favors less fragile observation positions, and can lead to downstream improvements for a frozen VLA policy under information-limited initial viewpoints. Together, these results suggest that InfoField3D, as an explicit and differentiable organization of scene information over free space, provides a useful interface between semantic scene understanding and task-oriented view planning.

The current formulation assumes a globally reconstructed, static scene. Future work may extend it to partially observed environments, dynamically changing scenes, and information beyond semantics. More broadly, this suggests a path toward treating diverse geometry-bound sensing objectives within a unified differentiable field formulation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and language polishing, including improving clarity, readability, and grammar. We also used GPT-4o for multimodal viewpoint scoring and task-relevance annotation in the experiments described in Section 4.2 and Appendix C. We have reviewed and revised all AI-assisted text to ensure its accuracy and consistency with our research. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yongjie Bai, Zhouxia Wang, Yang Liu, Kaijun Luo, Yifan Wen, Mingtong Dai, Weixing Chen, Ziliang Chen, Lingbo Liu, Guanbin Li, and Liang Lin. Learning to see and act: Task-aware virtual view exploration for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13386–13396, 2026.

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. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40, 2025.

Michel Breyer, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Closed-loop next-best-view planning for target-driven grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1411–1416, 2022. doi: 10.1109/IROS47612.2022.9981472.

Akshay K. Burusa, Joost Scholten, Xin Wang, David Rapado-Rincón, Eldert J. van Henten, and Gert Kootstra. Semantics-aware next-best-view planning for efficient search and detection of task-relevant plant parts. *Biosystems Engineering*, 248:1–14, 2024a. doi: 10.1016/j.biosystemseng.2024.09.018. URL https://doi.org/10.1016/j.biosystemseng.2024.09.018.

Akshay K. Burusa, Eldert J. van Henten, and Gert Kootstra. Gradient-based local next-best-view planning for improved perception of targeted plant nodes. In *2024 IEEE International Conference on Robotics and Automation*, pp. 15854–15860, 2024b. doi: 10.1109/ICRA57147.2024.10610397.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

Ian Chuang, Andrew Lee, Dechen Gao, M-Mahdi Naddaf-Sh, and Iman Soltani. Active vision might be all you need: Exploring active vision in bimanual robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7952–7959, 2025. doi: 10.1109/ICRA55743.2025.11128253. URL https://doi.org/10.1109/ICRA55743.2025.11128253.

Charles K. Cowan and Peter D. Kovesi. Automatic sensor placement from vision task requirements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(3):407–416, 1988. doi: 10.1109/34.3905.

Yixiang Dai, Siang Chen, Kaiqin Yang, Dingchang Hu, Pengwei Xie, Guosheng Li, Yuan Shen, and Guijin Wang. Active-perceptive language-oriented grasp policy for heavily cluttered scenes. *IEEE Robotics and Automation Letters*, 10(11):11094–11101, 2025. doi: 10.1109/LRA.2025.3604750.

Stefan C. Endres, Carl Sandrock, and Walter W. Focke. A simplicial homology algorithm for Lipschitz optimisation. *Journal of Global Optimization*, 72(2):181–217, 2018. doi: 10.1007/s10898-018-0645-y. URL https://doi.org/10.1007/s10898-018-0645-y.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, 2026.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398. URL https://doi.org/10.1162/106365601750190398.

Kerry He, Rhys Newbury, Tin Tran, Jesse Haviland, Ben Burgess-Limerick, Dana Kulić, Peter Corke, and Akansel Cosgun. Visibility maximization controller for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(3):8479–8486, 2022. doi: 10.1109/LRA.2022.3188430.

Snehal Jauhri, Sophie Lueth, and Georgia Chalvatzaki. Active-perceptive motion generation for mobile manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1413–1419, 2024. doi: 10.1109/ICRA57147.2024.10610714. URL https://doi.org/10.1109/ICRA57147.2024.10610714.

Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1238–1247, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jayaraman_Learning_to_Look_CVPR_2018_paper.html.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using Fisher information. In *Computer Vision – ECCV 2024*, pp. 422–440. Springer, 2024. doi: 10.1007/978-3-031-72624-8_24. URL https://doi.org/10.1007/978-3-031-72624-8_24.

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*, pp. 5392–5404, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_Multimodal_LLM_Guided_Exploration_and_Active_Mapping_using_Fisher_Information_ICCV_2025_paper.html.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: 10.1023/A:1008306431147. URL https://doi.org/10.1023/A:1008306431147.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19672–19682, 2023. doi: 10.1109/ICCV51070.2023.01807.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, tyler bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a BC-RL perception-action loop. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3647–3664. PMLR, 2025. URL https://proceedings.mlr.press/v305/kerr25a.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Juil Koo, Daehyeon Choi, Sangwoo Youn, Phillip Y. Lee, and Minhyuk Sung. Toward ambulatory vision: Learning visually-grounded active view selection. *arXiv preprint arXiv:2512.13250*, 2025. URL https://arxiv.org/abs/2512.13250.

Dieter Kraft. Algorithm 733: TOMP—Fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software*, 20(3):262–281, 1994. doi: 10.1145/192115.192124. URL https://doi.org/10.1145/192115.192124.

Taekbeom Lee, Dabin Kim, Youngseok Jang, and H. Jin Kim. HERE: Hierarchical active exploration of radiance field with epistemic uncertainty minimization. *IEEE Robotics and Automation Letters*, 11(3):3788–3795, 2026. doi: 10.1109/LRA.2026.3656771. URL https://doi.org/10.1109/LRA.2026.3656771.

Boshu Lei, Wen Jiang, and Kostas Daniilidis. ActiveGrasp: Information-guided active grasping with calibrated energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42418–42429, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lei_ActiveGrasp_Information-Guided_Active_Grasping_with_Calibrated_Energy-based_Model_CVPR_2026_paper.html.

Lun Li and Hamidreza Kasaei. Enhanced view planning for robotic harvesting: Tackling occlusions with imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13146–13152, 2025. doi: 10.1109/ICRA55743.2025.11127892. URL https://doi.org/10.1109/ICRA55743.2025.11127892.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023. URL https://arxiv.org/abs/2306.03310.

Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. SaPaVe: Towards active perception and manipulation in vision-language-action models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37164–37174, 2026a.

Yizhe Liu, Tong Jia, Haiyu Zhang, Guowei Yang, Hao Wang, and Dongyue Chen. RASP: Robot active scene perception with joint viewpoint planning and depth completion in cluttered environments. *IEEE Transactions on Automation Science and Engineering*, 22:20611–20623, 2025. doi: 10.1109/TASE.2025.3604455. URL https://doi.org/10.1109/TASE.2025.3604455.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. ActiveVLA: Injecting active perception into vision-language-action models for precise 3D robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8141–8151, 2026b.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 909–918, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Mo_PartNet_A_Large-Scale_Benchmark_for_Fine-Grained_and_Hierarchical_Part-Level_3D_CVPR_2019_paper.html.

Keiko Nagami, Timothy Chen, Javier Yu, Ola Shorinwa, Maximilian Adang, Carlyn Dougherty, Eric Cristofalo, and Mac Schwager. VISTA: Open-vocabulary, task-relevant robot exploration with online semantic gaussian splatting. *IEEE Robotics and Automation Letters*, 11(3):3150–3157, 2026. doi: 10.1109/LRA.2026.3653276.

Dexter Ong, Yuezhan Tao, Varun Murali, Igor Spasojevic, Vijay Kumar, and Pratik Chaudhari. ATLAS Navigator: Active task-driven LAnguage-embedded Gaussian splatting. *IEEE Transactions on Field Robotics*, 2026. doi: 10.1109/TFR.2026.3732100. URL https://doi.org/10.1109/TFR.2026.3732100.

Deyun Qin, Zezhi Liu, Hanqian Luo, Xiao Liang, and Yongchun Fang. A general one-shot multimodal active perception framework for robotic manipulation: Learning to predict optimal viewpoint. *arXiv preprint arXiv:2601.13639*, 2026. URL https://arxiv.org/abs/2601.13639.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024. doi: 10.1109/CVPR52733.2024.01895.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 497–500. ACM, 2001. doi: 10.1145/383259.383317. URL https://doi.org/10.1145/383259.383317.

Yitian Shi, Di Wen, Guanqi Chen, Edgar Welte, Sheng Liu, Kunyu Peng, Rainer Stiefelhagen, and Rania Rayyes. VISO-Grasp: Vision-language informed spatial object-centric 6-DoF active view planning and grasping in clutter and invisibility. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 14931–14938, 2025. doi: 10.1109/IROS60139.2025.11246329.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9. URL https://doi.org/10.1016/0041-5553(67)90144-9.

Xiao'Ao Song and Konstantinos Karydis. GS-NBV: a geometry-based, semantics-aware viewpoint planning algorithm for avocado harvesting under occlusions. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pp. 690–695, 2025. doi: 10.1109/CASE58245.2025.11164113. URL https://doi.org/10.1109/CASE58245.2025.11164113.

Yan Song, Zhihao Li, Chenglong Li, Li He, Yan Wang, and Wenqiang Zhang. GO-PRE: Goal-oriented next-best-view selection via predictive rendering entropy for active 3D reconstruction. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60821.

Matthew Strong, Boshu Lei, Aiden Swann, Wen Jiang, Kostas Daniilidis, and Monroe Kennedy. Next best sense: Guiding vision and touch with FisherRF for 3D Gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3204–3210, 2025. doi: 10.1109/ICRA55743.2025.11127233. URL https://doi.org/10.1109/ICRA55743.2025.11127233.

Konstantinos A. Tarabanis, Roger Y. Tsai, and Peter K. Allen. The MVP sensor planning system for robotic vision tasks. *IEEE Transactions on Robotics and Automation*, 11(1):72–85, 1995. doi: 10.1109/70.345939.

Bill Triggs and Christian Laugier. Automatic camera placement for robot vision tasks. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, volume 2, pp. 1732–1737, 1995. doi: 10.1109/ROBOT.1995.525522.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Guokang Wang, Hang Li, Shuyuan Zhang, Di Guo, Yanhong Liu, and Huaping Liu. Observe then act: Asynchronous active vision-action model for robotic manipulation. *IEEE Robotics and Automation Letters*, 10(4):3422–3429, 2025. doi: 10.1109/LRA.2025.3541334. URL https://doi.org/10.1109/LRA.2025.3541334.

Zhihao Wang, Haoyao Chen, Shiwu Zhang, and Yunjiang Lou. Active view planning for visual SLAM in outdoor environments based on continuous information modeling. *IEEE/ASME Transactions on Mechatronics*, 29(1):237–248, 2024. doi: 10.1109/TMECH.2023.3272910.

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. POp-GS: Next best view in 3D-Gaussian splatting with P-Optimality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3646–3655, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wilson_POp-GS_Next_Best_View_in_3D-Gaussian_Splatting_with_P-Optimality_CVPR_2025_paper.html.

Yuhan Xie, Yixi Cai, Yinqiang Zhang, Lei Yang, and Jia Pan. GauSS-MI: Gaussian splatting Shannon mutual information for active 3D reconstruction. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.030.

Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5450–5463. PMLR, 2025. URL https://proceedings.mlr.press/v305/xiong25a.html.

Zichao Zhang and Davide Scaramuzza. Fisher information field: An efficient and differentiable map for perception-aware planning. *arXiv preprint arXiv:2008.03324*, 2020. URL https://arxiv.org/abs/2008.03324.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024. doi: 10.1109/CVPR52733.2024.02048.

Jingsen Zhu, Silvia Sellán, and Alexander Terenin. A Bayesian approach for task-specific next-best-view selection with uncertain geometry. In *Proceedings of the ACM SIGGRAPH Conference Papers*, pp. 1–11, 2026. doi: 10.1145/3799902.3811119.

Appendix

CONTENTS

A. Method Details and Extensions

A.1 **Field Construction Details and Analysis**
Construction choices and the approximations introduced at each representation step.

A.1.1 **Area-Weighting Treatment for Nonuniform Point Clouds**
Area weights compensate for nonuniform sampling of the same surface.

A.1.2 **Coverage Normalization for Large, Low-Relevance Surfaces**
An optional readout measures task response relative to geometric exposure.

A.1.3 **Spherical Harmonic Degree and Approximation Error**
Randomized spherical patterns quantify approximation error and guide degree selection.

A.1.4 **Mapping Reference-Sphere Radius**
Reference-sphere radius controls mapping accuracy at different observation distances.

A.1.5 **Spherical-Mapping Loss and Obtaining Computational Units through Geometric Partitioning**
Shape-dependent mapping loss decreases as objects are partitioned into local units.

A.1.6 **Gaussian Occupancy**
Gaussian occupancy, analytic path integrals and gradients, and material transparency.

A.1.7 **Illustrations of the Effect of Gaussian Occlusion in a Differentiable Information Field**
Field values and gradient trajectories under single and multiple occluders.

A.2 **Vector-Valued Spherical Harmonics for Repeated Queries**
Reusable vector coefficients support repeated queries through linear readout.

A.3 **Construction Details When the Task Condition Contains Multiple Targets**
Separate target responses are combined to express joint observation requirements.

B. Representation Fidelity and Compression

B.1 **Representation-Fidelity Experiment Details**
Pointwise reference computation, fidelity metrics, and representation-size comparison.

B.2 **Stress Test under Different Point-Cloud Densities**
Fidelity under increased point-cloud density with a fixed computational-unit budget.

CONTENTS (CONTINUED)

C. Viewpoint Search

C.1 **Experimental Scene Selection and Feasible-Domain Setting**
Scene selection and the geometric constraints defining valid observation positions.

C.2 **Baselines and Task Adaptation**
Compared objectives, implementation sources, and adaptations to the shared search task.

C.3 **Evaluation Metrics and Numerical Results**
Definitions of evaluation metrics and detailed numerical results for the main experiments.

C.4 **GPT-4o Viewpoint Evaluation**
Reference-image design, task-specific scoring and aggregation, and repeated evaluations.

C.5 **Automated Multimodal Annotation Pipeline**
Task instructions and multiview images produce relevance weights on 3D surfaces.

C.6 **Preprocessing Cost**
Initial construction cost, task-update cost, and search efficiency including preprocessing.

D. Viewpoint Robustness

Perturbation protocol and task-region visibility losses at five displacement scales.

E. Controlled VLA Evaluation

E.1 **Controlled VLA Experimental-Design Details**
Local viewpoint rescue, equal-distance controls, frozen-policy rollouts, and completion scoring.

E.2 **Execution Examples**
Paired execution examples from initial and recommended third-person viewpoints.

APPENDIX A: METHOD DETAILS AND EXTENSIONS

A.1 FIELD CONSTRUCTION DETAILS AND ANALYSIS

A.1.1 AREA-WEIGHTING TREATMENT FOR NONUNIFORM POINT CLOUDS

The input point cloud consists of surface points $\mathbf{p}_i$, their associated unit outward normals $\mathbf{n}_i$, and semantic features $\mathbf{f}_i$. A task query produces point-level relevance s_i from the semantic features, representing the relevance of the surface at that location to the current observation requirement; the normal describes the directional response of that surface toward different observation positions.

For an approximately uniformly sampled point cloud, the surface area represented by each point is approximately the same, so the main text directly sums the contribution of every point. Under nonuniform sampling, surface regions of the same size may contain different numbers of points. If equal geometric weight is still assigned to each point, densely sampled regions receive a larger accumulated contribution even when their task relevance has not increased. Without resampling the point cloud, this sampling-density bias can be reduced by introducing, for each point, the surface area a_i that it represents and writing the weight used in response accumulation as

$$w_i = a_i s_i. \tag{5}$$

Here, a_i describes how much surface the sample point represents, and s_i describes the degree to which that surface is relevant to the task. When more sample points are added to the same surface, the area represented by each point should decrease correspondingly so that the total area weight of the region remains constant, preventing the response from being amplified solely by denser sampling. Under uniform sampling with $a_i = a$, area weighting and direct summation differ only by a common positive scale factor a, and therefore do not change the ranking of observation positions by the unnormalized response.

A.1.2 COVERAGE NORMALIZATION FOR LARGE, LOW-RELEVANCE SURFACES

The point-cloud sampling-density issue has been reasonably addressed in Section A.1.1, but even for uniform point clouds, the responses of large, low-relevance surfaces may still account for a substantial share of the sum, biasing the total response toward positions that expose more geometry. To reduce the influence of overall geometric coverage scale, coverage normalization may be used as an optional readout. It retains area weights and the other response-computation rules, but sets the task relevance of all surface points to 1 to obtain a geometric coverage response without task-relevance weights, and then uses this response as the normalization basis.

In this subsection, we fix the task condition $\mathcal{T}$ and write $J \equiv J_{\mathcal{T}}$ for brevity. For the multi-unit field constructed below, the numerator and denominator use the same occupancy attenuation:

$$C_L(\mathbf{x}) = \sum_j T_j(\mathbf{x}) F_{j,L}^{\text{coverage}}(\mathbf{u}_j(\mathbf{x})), \qquad \widetilde{J}(\mathbf{x}) = \frac{J(\mathbf{x})}{C_L(\mathbf{x}) + \varepsilon}. \tag{6}$$

Here, $F_{j,L}^{\text{coverage}}$ is constructed from area weights and does not include task-relevance weights, and $\varepsilon > 0$ is a numerical-stability term. The normalized readout measures the task response relative to overall geometric exposure. Whether to use this method depends on whether the task should emphasize the total response or the relative degree of relevance.

A.1.3 SPHERICAL HARMONIC DEGREE AND APPROXIMATION ERROR

To determine how many spherical-harmonic coefficients are sufficient, we isolate the approximation of F_{sph} by F_L defined in Section 3.2. We generate six families of relevance patterns directly on unit-sphere point clouds (Figure A1), avoiding the non-spherical mapping effects analyzed in Appendix A.1.5. Each family is evaluated over 100 random seeds, with 1,024 uniformly sampled surface points and newly randomized patterns per seed. The six families share each seed's point cloud. We use an exponential directional kernel, evaluate a range of spherical-harmonic degrees, and compare against the direct pointwise kernel sum over a dense set of fixed spherical test directions. The implementation applies the same normalization factor to both responses, which leaves all reported metrics unchanged.

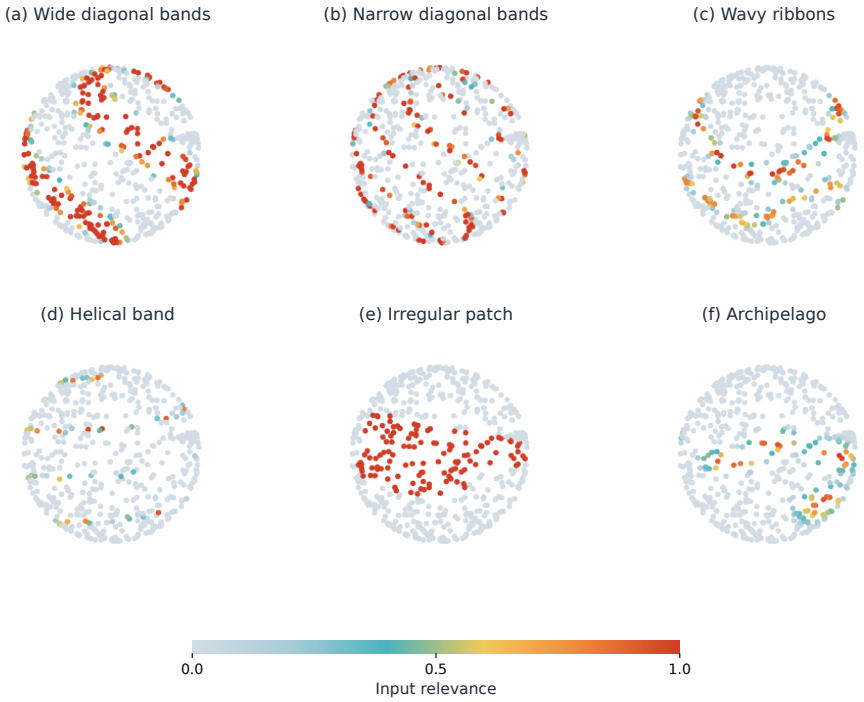


Figure A1. Six randomized pattern families: wide and narrow diagonal bands, wavy ribbons, a helical band, an irregular patch, and an archipelago.

Figure A2 reports relative L_2 error before and after separately removing each response's mean, peak-direction error, and top-10% direction-set overlap. The mean-removed error measures directional variation. Lines average 100 trials per family, and the two shaded bands are pointwise 90% and 95% bootstrap confidence intervals for the means.

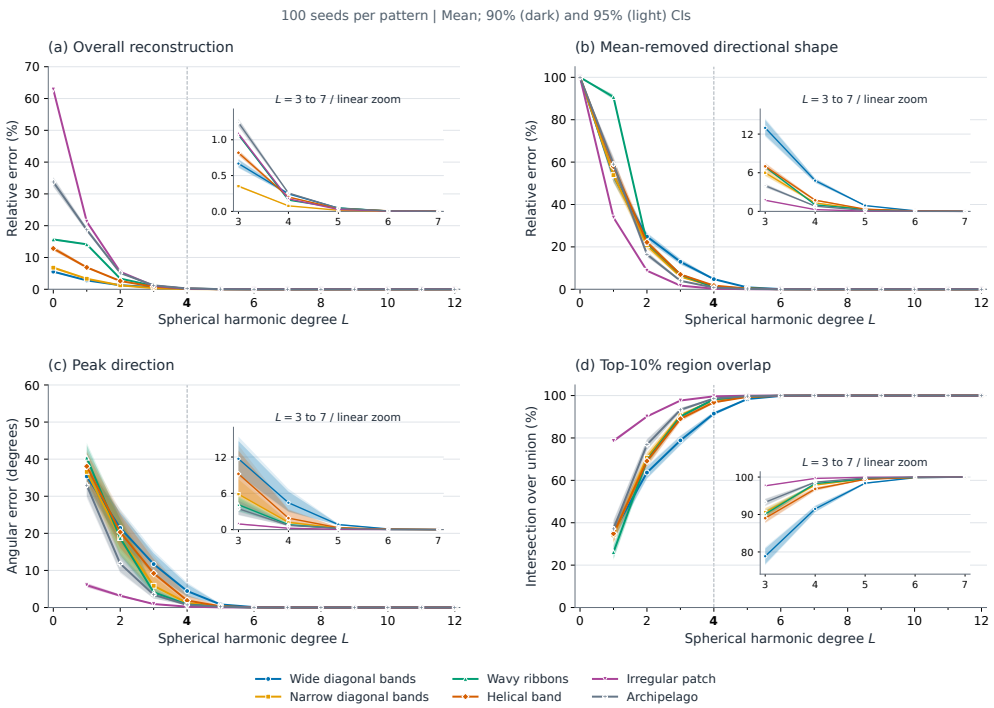


Figure A2. Approximation quality versus spherical-harmonic degree. Dark inner and light outer bands indicate 90% and 95% confidence intervals for the means. Insets use linear axes. The dashed line marks $L = 4$; peak and overlap metrics are undefined at $L = 0$ and omitted.

Most of the improvement occurs at low degrees. The largest family-mean overall error decreases from 5.51% at $L = 2$ to 1.24% at $L = 3$, 0.256% at $L = 4$, and 0.048% at $L = 5$. Peak-direction error can be less regular because distant peaks of similar height may exchange order. Figure A3 therefore also measures the relative loss of the direct response at the direction selected by F_L. At $L = 4$, the largest family-mean loss by this measure is only 0.0424%, indicating that a shift in peak location need not substantially reduce the attained response.

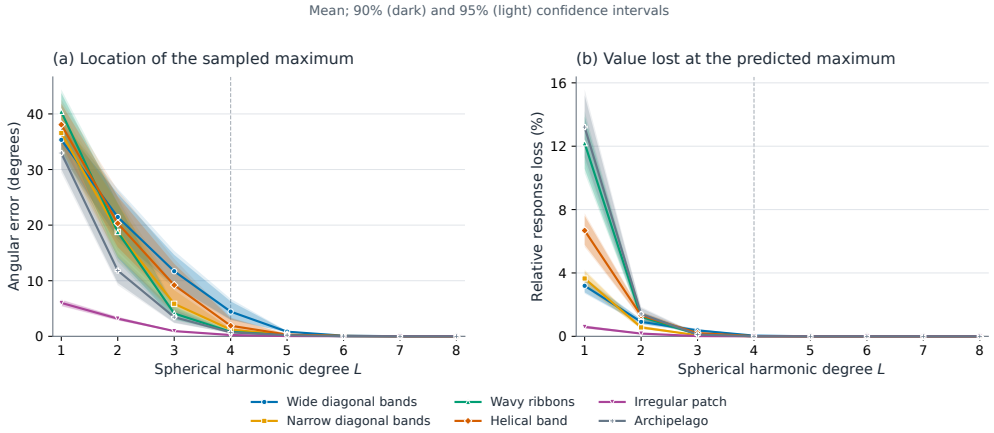


Figure A3. Peak-direction error and the relative direct-response loss at the predicted maximum, evaluated on the same test directions. Lines show means with 90% and 95% confidence bands. Similar response values can occur at widely separated directions.

These results suggest $L = 4$ as a compact default: 25 coefficients per scalar response give at most 0.256% overall error and 4.777% mean-removed error when comparing the six family means. For stricter accuracy requirements, $L = 5$ uses 36 coefficients and reduces these values to 0.048% and 0.898%; $L = 6$ uses 49 coefficients and reduces them further to 0.008% and 0.089%. Thus, $L = 5$ is a useful choice when sub-0.1% mean overall error and sub-1% mean directional-variation error are desired, while $L = 6$ provides a higher-precision option. Balancing representation size and the small response errors observed here, we use $L = 4$ in the paper's scene experiments.

In our implementation, we use orthonormal real spherical harmonics with the Condon–Shortley phase convention, with negative and positive orders corresponding to sine and cosine components, respectively. The basis is evaluated directly from unit Cartesian directions using polynomial recurrences, avoiding explicit angular coordinates and their coordinate singularities at the poles. Real-valued coefficients are obtained by area- and task-weighted projection of the mapped samples onto this basis, followed by degree-wise scaling with the angular kernel's Funk–Hecke multipliers.

A.1.4 Mapping Reference-Sphere Radius

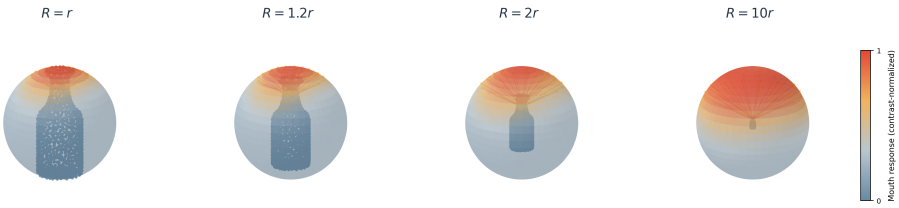


Figure A4. Differences in the directional responses formed under different reference-sphere radii R.

In Section 3.2, we unified the directional responses of the point cloud onto a reference sphere to form an object-level directional response. As shown in Figure A4, changing R changes the mapped directional response. A smaller reference sphere and a shorter observation distance can preserve more information about the spatial distribution of the point cloud; a larger reference sphere and a longer observation distance can more fully express the directional distribution of surface normals. In experimental observations, an extremely small mapping radius, such as the radius r of the minimum bounding sphere, may excessively truncate the directional distribution of surface normals and produce a large representation loss at ordinary observation distances. By contrast, when the mapping radius is between $4r$ and $10r$, a relatively small loss can usually be obtained, and the differences among values in this range are limited. This range can therefore serve as a recommended range for general scenes. As R tends to infinity, the mapping result can be regarded as retaining only the distribution of surface normals.

A.1.5 Spherical-Mapping Loss and Obtaining Computational Units through Geometric Partitioning

Section 3.4 of the main text briefly discusses why a complete object is not necessarily suitable as one computational unit and the corresponding solution strategy. A complete analysis and argument are provided here. For a curved faucet, for example, its front and rear ends are difficult to summarize accurately with the same center. When the directional responses of these surfaces are jointly mapped onto a reference sphere, the original relationship between position and normal changes. The choice of computational units must therefore consider geometric form rather than only object name or membership.

For a point set with reference center $\mathbf{c}$ and reference-sphere radius R, let $\mathbf{d}_i = \mathbf{p}_i - \mathbf{c}$. The positive intersection of a ray emitted from surface point $\mathbf{p}_i$ along the unit outward normal $\mathbf{n}_i$ with the reference sphere is

$$t_i = -\mathbf{n}_i^\top \mathbf{d}_i + \sqrt{(\mathbf{n}_i^\top \mathbf{d}_i)^2 + R^2 - \|\mathbf{d}_i\|^2}, \qquad \mathbf{r}_i = \frac{\mathbf{p}_i + t_i \mathbf{n}_i - \mathbf{c}}{R}. \tag{7}$$

Let $\mathbf{d}_i^\perp = (\mathbf{I} - \mathbf{n}_i \mathbf{n}_i^\top)\mathbf{d}_i$. The expression above can also be written as

$$\mathbf{r}_i = \frac{\mathbf{d}_i^{\perp}}{R} + \sqrt{1 - \frac{\|\mathbf{d}_i^{\perp}\|^2}{R^2}}\,\mathbf{n}_i. \tag{8}$$

When a sphere uses its center as the reference center and its normals are radial, $\mathbf{d}_i^{\perp} = 0$, and the mapped direction is identical to the original normal. For extended, bent, or separated structures, a shared center may not align the normals everywhere simultaneously. This geometric mapping occurs before spherical-harmonic encoding. Increasing the spherical-harmonic order can only improve the encoding of the mapped function and cannot eliminate the change introduced by the mapping itself.

To examine these geometric differences, Figure A5 compares spherical-response mapping loss across geometries ranging from a standard spherical point cloud to shapes that progressively depart from a sphere.

Directions are sampled uniformly and densely on the sphere. For each direction, the mean angular response over all points is computed using the original normals and the mapped directions, respectively. Each of the two response sets is divided by its own mean over these directions. Using the normalized original response as the reference, the relative L_2 error is computed and reported as a percentage. The results are shown in Figure A5.

As shown, spheres and near-spherical geometries have low loss, whereas elongated polyhedra and chairs have higher loss. Based on this observation, we conjecture that the applicability of a single-center representation is related to geometric shape and that using more rounded computational units may reduce the loss.

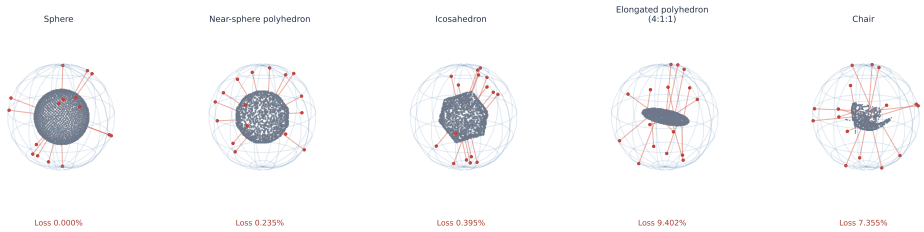


Figure A5. Spherical-response mapping loss for point clouds of different geometric shapes.

We therefore attempt to geometrically partition the point cloud so that the geometric structures of the computational units become more rounded. This paper uses simple recursive principal-axis bisection: at each step, the unit with the largest spatial extent is selected and split along its principal axis into two balanced subsets until the desired unit budget is reached. For the chair at the end of Figure A5, the effect of its geometric partitioning on representation loss is shown in Figures A6 and A7. This experiment focuses on the difference in directional response before and after spherical mapping and does not account for occlusion between units.

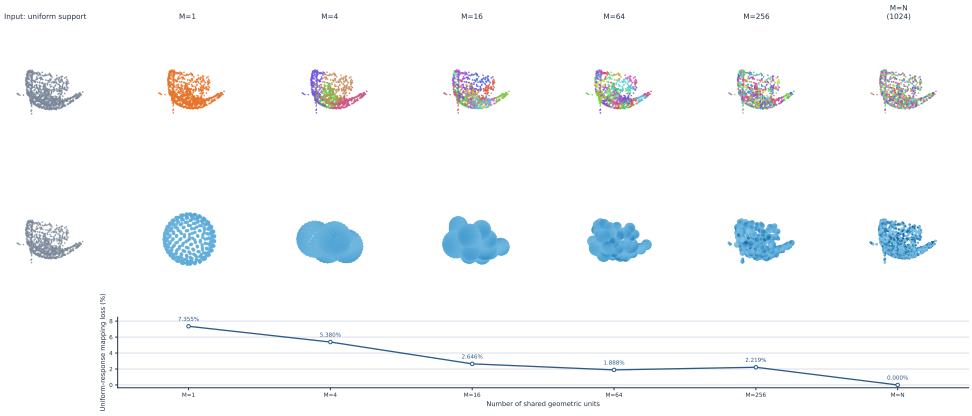


Figure A6. Geometric-partitioning loss for a uniformly weighted point cloud.

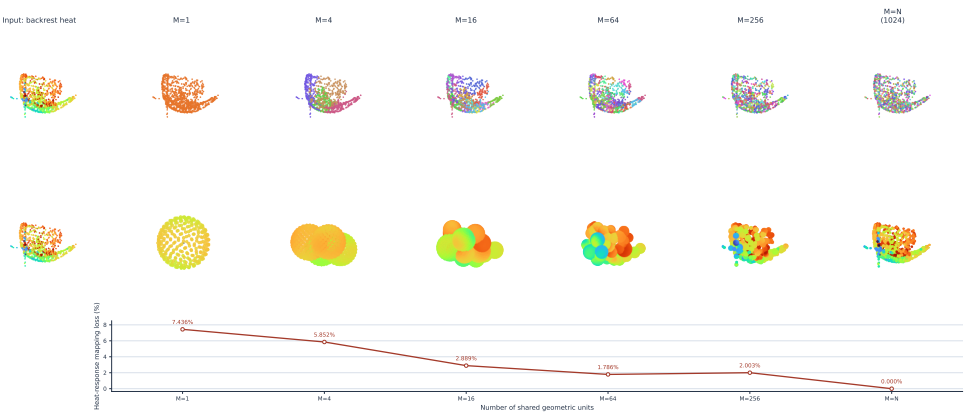


Figure A7. Geometric-partitioning loss for a point cloud whose query value is the backrest.

Figure A6 places greater emphasis on observing the loss caused by point-cloud geometry, whereas Figure A7 examines whether the weight distribution of a query-weighted point cloud can be retained. As shown, the accuracy loss decreases overall as the number of partitions increases. When the number of partitions reaches the limit, namely one point per computational unit, the loss falls to 0.

We further scale up this experiment and conduct the same experiment on 100 randomly selected objects from the PartNet dataset (Mo et al., 2019). The results are as follows.

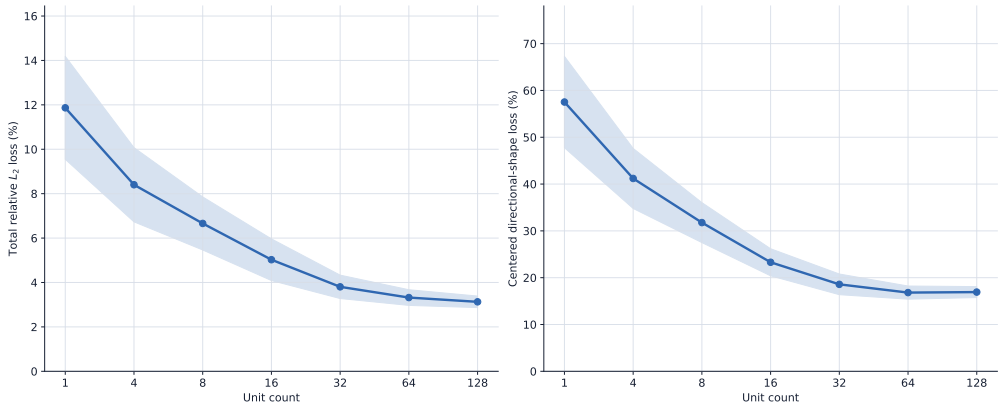


Figure A8. Trend analysis of the number of geometric partitions and representation loss.

Overall, the representation loss continues to decrease as the number of partitions increases. At the partitioning limit, the computation can be equivalent to pointwise computation, apart from the truncation loss of the spherical harmonics themselves. In practical use, however, such extreme partitioning is unnecessary. As shown, a small number of early partitions can recover most of the representation accuracy, while the benefit of continued partitioning diminishes later.

A.1.6 Gaussian Occupancy

The core of spatial occupancy modeling is to represent discrete geometry as a continuous, differentiable spatial occupancy function and to describe occlusion attenuation by integrating along observation paths, thereby making the occlusion computation differentiable as well. This construction is not restricted to any particular method. In principle, any functional form or construction method may be used, as long as the resulting spatial occupancy function remains differentiable and provides a reasonable description of occlusion. For example, more expressive continuous scene representations, such as 3D Gaussian Splatting, could in principle be used to construct a higher-fidelity occupancy model.

In this work, we use Gaussian functions to represent local occupancy, and this section presents only one example construction used in our experimental implementation.

Occupancy construction. For a geometric unit k with surface samples $\mathbf{p}_i$ and area weights $a_i > 0$, define its represented area and Gaussian mean as

$$\begin{aligned} A_k &= \sum_{i \in k} a_i, \\ \boldsymbol{\mu}_k &= \frac{1}{A_k} \sum_{i \in k} a_i \mathbf{p}_i. \end{aligned} \tag{9}$$

We recommend an area-weighted covariance with a positive minimum width:

$$\boldsymbol{\Sigma}_k = \frac{1}{A_k} \sum_{i \in k} a_i (\mathbf{p}_i - \boldsymbol{\mu}_k)(\mathbf{p}_i - \boldsymbol{\mu}_k)^\top + \sigma_0^2 \mathbf{I}. \tag{10}$$

Here a_i is the geometric area weight defined in Section A.1.1, without the task-relevance factor s_i; $\sigma_0 > 0$ ensures positive definiteness. The Gaussian mean $\boldsymbol{\mu}_k$ is distinguished from the directional-response reference center $\mathbf{c}_k^{\mathrm{ref}} \equiv \mathbf{c}_k$ defined in Section 3.4: the two may coincide, but the framework does not require this.

The resulting occupancy density is

$$\rho_k(\mathbf{z}) = \alpha_k \exp\left[-\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{z} - \boldsymbol{\mu}_k) \right], \tag{11}$$

with amplitude

$$\alpha_k = \frac{\kappa_k A_k}{(2\pi)^{3/2} |\boldsymbol{\Sigma}_k|^{1/2}}, \qquad \kappa_k \geq 0. \tag{12}$$

The covariance describes spatial shape, while κ_k controls attenuation strength. This normalization yields

$$\int_{\mathbb{R}^3} \rho_k(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \kappa_k A_k, \tag{13}$$

so increasing sample density while preserving area weights does not itself increase occupancy mass. Here ρ_k is an attenuation density.

Finite-segment attenuation. For a source unit with reference center $\mathbf{c}_j^{\mathrm{ref}}$ and an observation position $\mathbf{x}$, define

$$\begin{aligned} \ell &= \left\| \mathbf{x} - \mathbf{c}_j^{\mathrm{ref}} \right\|, \\ \mathbf{u} &= \frac{\mathbf{x} - \mathbf{c}_j^{\mathrm{ref}}}{\ell}. \end{aligned} \tag{14}$$

The contribution of an occluding unit k is integrated along this finite segment. Accumulating the contributions gives the optical depth and transmittance:

$$\begin{aligned} I_{jk}(\mathbf{x}) &= \int_0^\ell \rho_k\left(\mathbf{c}_j^{\mathrm{ref}} + s\mathbf{u}\right) \, \mathrm{d}s, \\ \tau_j(\mathbf{x}) &= \sum_{k \neq j} I_{jk}(\mathbf{x}), \\ T_j(\mathbf{x}) &= \exp[-\tau_j(\mathbf{x})]. \end{aligned} \tag{15}$$

This construction excludes the source unit itself; other units, including those belonging to the same object, remain potential occluders.

For one source–occluder pair, introduce

$$\begin{aligned} \mathbf{Q} &= \boldsymbol{\Sigma}_k^{-1}, \\ \mathbf{r} &= \mathbf{c}_j^{\mathrm{ref}} - \boldsymbol{\mu}_k, \end{aligned} \tag{16}$$

and the scalar coefficients

$$a = \mathbf{u}^\top \mathbf{Q}\mathbf{u},$$
$$b = \mathbf{u}^\top \mathbf{Q}\mathbf{r}, \tag{17}$$
$$h = \mathbf{r}^\top \mathbf{Q}\mathbf{r}.$$

Completing the square yields

$$I_{jk} = \alpha_k \exp\left[-\frac{1}{2}\left(h - \frac{b^2}{a}\right)\right]\sqrt{\frac{\pi}{2a}} \times \left[\operatorname{erf}\left(\frac{a\ell + b}{\sqrt{2a}}\right) - \operatorname{erf}\left(\frac{b}{\sqrt{2a}}\right)\right], \tag{18}$$

where erf is the error function.

Spatial derivatives. With scene parameters fixed, moving the observation position changes both path length and direction. Write the density along the segment as

$$g(s) = \alpha_k \exp\left[-\frac{1}{2}\left(as^2 + 2bs + h\right)\right]. \tag{19}$$

Its first two weighted moments have the closed forms

$$m_1 = \int_0^\ell s\, g(s)\, \mathrm{d}s = \frac{g(0) - g(\ell) - bI_{jk}}{a}, \tag{20}$$

$$m_2 = \int_0^\ell s^2 g(s)\, \mathrm{d}s = \frac{I_{jk} - \ell g(\ell) - bm_1}{a}. \tag{21}$$

The spatial derivative of the Gaussian contribution is therefore

$$\nabla_{\mathbf{x}} I_{jk} = g(\ell)\mathbf{u} - \frac{\mathbf{I} - \mathbf{u}\mathbf{u}^\top}{\ell}\left(m_2\mathbf{Q}\mathbf{u} + m_1\mathbf{Q}\mathbf{r}\right). \tag{22}$$

The first term accounts for path-length variation and the second for path-direction variation. The transmittance gradient follows directly:

$$\nabla_{\mathbf{x}} T_j = -T_j \sum_{k \neq j} \nabla_{\mathbf{x}} I_{jk}. \tag{23}$$

These expressions are valid for $\mathbf{x} \neq \mathbf{c}_j^{\mathrm{ref}}$. The integral and its derivative are exact for the chosen Gaussian density; agreement with physical scene occlusion remains a property of the occupancy approximation.

One possible direction for extension is that, when the information field is used solely to describe visual accessibility, the occupancy function could further account for the light-transmission properties of different materials. For example, transparent glass and an opaque brick with the same shape and size may occupy similar geometric extents but produce different occlusion effects. With the same Gaussian geometric parameters, this difference can be modeled by adjusting the occupancy strength, such that transparent units induce weaker attenuation along propagation paths.

A.1.7 Illustrations of the Effect of Gaussian Occlusion in a Differentiable Information Field

These two figures visualize the effect of Gaussian occlusion in a differentiable information field and are provided only as illustrations to aid understanding. They respectively show the field-value and gradient distributions formed by a single information source under one occluder and under multiple occluders, and illustrate how a given starting point moves along the gradient direction from an occluded region to a region with a higher field value.

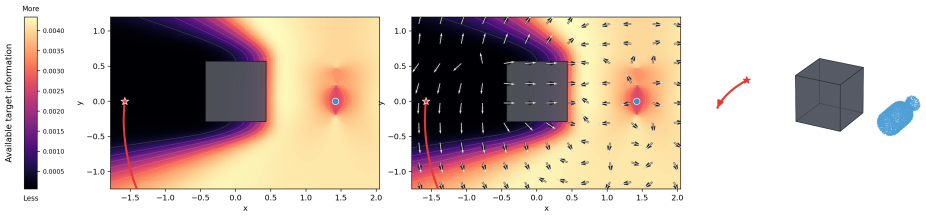


Figure A9. Illustration of InfoField3D under a single occluder.

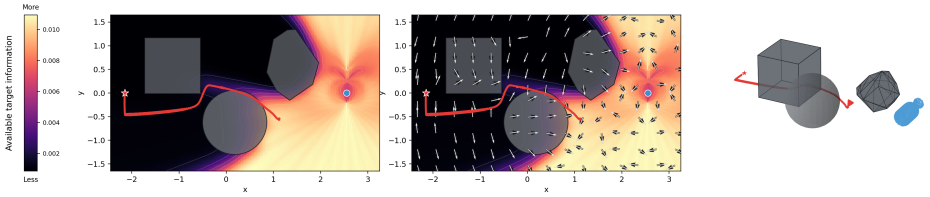


Figure A10. Illustration of InfoField3D under multiple occluders.

A.2 Vector-Valued Spherical Harmonics for Repeated Queries

The main text discusses first querying the semantic point cloud according to the task condition and then using spherical harmonics to encode the corresponding task response. In this process, the semantic point cloud is already reusable, but each new query still requires pointwise computation of semantic relevance, so the computational scale remains tied to the point cloud. To address this, the earlier idea of using spherical harmonics to compress and encode point-cloud responses can be extended by using each point's semantic vector as the vector weight of its directional response. This first yields an uncompressed vector directional field

$$\mathbf{F}_{j,\mathrm{sph}}(\mathbf{u}) = \sum_{i \in j} a_i \mathbf{f}_i K(\mathbf{r}_i^\top \mathbf{u}), \tag{24}$$

which is then encoded with spherical harmonics to obtain a finite-order vector-valued spherical-harmonic function

$$\mathbf{F}_{j,L}(\mathbf{u}) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \mathbf{c}_{j,\ell m} Y_{\ell m}(\mathbf{u}), \qquad \mathbf{c}_{j,\ell m} \in \mathbb{R}^D. \tag{25}$$

Here, D is the dimension of the semantic features. The original spherical-harmonic function produces a scalar response for a particular query; here, the point-cloud semantic field, a vector map, is directly encoded to obtain a vector-valued spherical-harmonic function. Its function value in each direction represents the semantic features aggregated along that direction and is independent of a particular query. When a query arrives, given a query vector $\mathbf{q}$, the vector coefficients or directional function of the vector-valued spherical-harmonic function can be inner-producted with it at the readout stage, thereby reading out the entire vector-valued spherical-harmonic function as a scalar spherical-harmonic function for that query semantics. Only after an observation direction is further specified is the scalar response in that direction obtained. In this way, once the scene geometry and

encoding settings are fixed, the same vector coefficients can serve multiple linear queries without traversing the original point cloud again for every query. When geometric partitioning, directional mapping, area weights, kernel functions, and fitting settings are fixed, and both encoding and querying are linear operations, querying before encoding and encoding before readout yield equivalent scalar spherical-harmonic functions. Formally,

$$F_{j,L}^{(\mathbf{q})}(\mathbf{u}) = \mathbf{q}^\top \mathbf{F}_{j,L}(\mathbf{u}) = \sum_{\ell,m} (\mathbf{q}^\top \mathbf{c}_{j,\ell m}) Y_{\ell m}(\mathbf{u}). \tag{26}$$

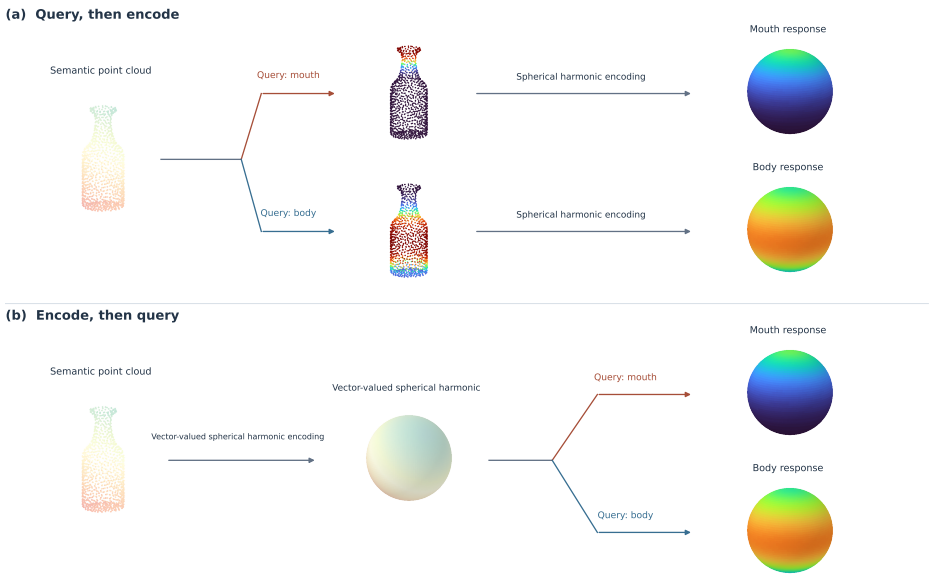


Figure A11. Two construction forms: querying before encoding and encoding before readout.

After fixing geometric partitioning, mapping, fitting weights, and occupancy functions, spherical-harmonic encoding can be written as a linear operator $\mathcal{E}$, so that

$$\mathcal{E}(\{\mathbf{q}^\top \mathbf{f}_i\}_i) = \mathbf{q}^\top \mathcal{E}(\{\mathbf{f}_i\}_i). \tag{27}$$

Likewise, the vector response after occlusion attenuation can be aggregated before readout:

$$\mathbf{G}(\mathbf{x}) = \sum_j T_j(\mathbf{x}) \mathbf{F}_{j,L}(\mathbf{u}_j(\mathbf{x})), \qquad J_{\mathbf{q}}(\mathbf{x}) = \mathbf{q}^\top \mathbf{G}(\mathbf{x}). \tag{28}$$

Therefore, under the linear conditions above, querying before encoding and encoding before readout produce the same response. With the latter, a new query no longer requires computation over the point cloud and only requires a readout over a small number of spherical-harmonic parameters, enabling faster task switching in scenarios that require frequent changes of task.

A.3 Construction Details When the Task Condition Contains Multiple Targets

When textual task descriptions are used to obtain the task-relevance map, one point worth noting is that, for tasks involving multiple targets, directly constructing a single objective from the entire task text may not always be the best choice.

For a task condition containing multiple targets, one direct approach is to query the complete task description as a whole. For example, “turn on the desk lamp and place the bottle in the drawer”

can be encoded as one query vector, producing a unified relevance map. This approach requires constructing and optimizing only one objective, and its computational form is no different from that of a single-target query. However, responses for different targets are merged into one function, and subsequent optimization considers only whether the total response increases, without explicitly controlling the contribution of each target. When the relevant region of a target is small or its response scale is relatively weak, it may be masked by the high response of other targets, so a position with a high overall score may still fail to observe that target sufficiently.

Another potentially better approach is to query multiple targets separately, for example, "desk lamp," "bottle," and "drawer," define an independent objective function for each target, and then aggregate these objective functions into a joint objective by a weighted product. Let $\mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_P\}$ be the set of queries corresponding to multiple targets, and let $\overline{J}_{\mathbf{q}_k}(\mathbf{x})$ denote the nonnegative response of target k at position $\mathbf{x}$ after accounting for occlusion and applying a common scale. Separate queries can be constructed through the pointwise scalar procedure described above or read directly from the vector-valued spherical-harmonic function in Appendix A.2. This requires multiple readouts but enables the joint objective to explicitly express the relative priority of each target and the requirement that they be jointly satisfied.

One feasible joint form is the weighted product

$$J_{\mathcal{Q}}^{\mathrm{prod}}(\mathbf{x}) = \prod_{k=1}^{P} \left[\overline{J}_{\mathbf{q}_k}(\mathbf{x})\right]^{\lambda_k}, \qquad \lambda_k > 0, \quad \sum_{k=1}^{P} \lambda_k = 1. \tag{29}$$

Here, λ_k denotes the relative priority of target k. In this form, a low response from any target directly affects the joint response: when the response of one target approaches zero, the product also approaches zero; when one target has a low response, the gain from further improving other targets while ignoring it is suppressed. The optimization therefore tends, within the feasible range, to find positions that satisfy all targets simultaneously without requiring an additional visibility threshold. Comparable channel scales aid interpretation of the joint score; when the responses are strictly positive, the product remains smooth. Depending on task preferences, a weighted sum, harmonic mean, or another aggregation form can also be used.

APPENDIX B: REPRESENTATION FIDELITY AND COMPRESSION

B.1 REPRESENTATION-FIDELITY EXPERIMENT DETAILS

The experiment uses 130 LIBERO and 1,525 LIBERO-Plus static scenes, for a total of 1,655 scenes. Each simulated scene is converted into a point cloud containing 47,500 points. Approximately 800 points are uniformly sampled in the feasible domain as the test set. At each test point, the information response is computed using the original surface point cloud, including point weights, and occlusion is determined using the simulation mesh; the analytic field is evaluated at the same position using the same weights and aggregation rules. Two settings are used for the weights on the scene point-cloud surface: uniform relevance, which places greater emphasis on geometry itself, and relevance weights obtained from the objects designated by the task in the scene, which introduce more variation. In the experiment, our method still obtains computational units through recursive principal-axis partitioning, using approximately 600 units per scene on average and a spherical-harmonic function with $L = 4$ for each unit, requiring 25 coefficients.

Among the metrics, the optimal-value attainment rate measures whether the position at which InfoField3D takes its maximum in the test set remains a high-value position under computation with the original point cloud. It is measured by comparing the value of that position under the original point-cloud computation with the maximum selected by the original point-cloud computation itself. Shape consistency uses the Pearson correlation of response values; ranking consistency uses the correlation of position-ranking indices, with equal response values ordered according to the original candidate order.

The compression ratio described in the main text compares geometric and scalar-response representations. Each pointwise sample stores 7 scalars (position, normal, and area weight), whereas each InfoField3D unit stores 34 scalars (center, covariance, and 25 spherical-harmonic coefficients). Aggregated over the dataset, this corresponds to approximately a $15\times$ reduction in representation size.

B.2 STRESS TEST UNDER DIFFERENT POINT-CLOUD DENSITIES

With the number-of-units budget of InfoField3D held constant, only the sampling density of the scene point cloud is increased to stress-test the representation capacity of InfoField3D. The results are shown in Table B1.

Table B1. Representation fidelity under the stress test.

Relevance	Reference density	Optimal-value attainment ↑	Shape consistency ↑	Ranking consistency ↑
Uniform	$1\times$	97.93%	92.28%	90.51%
Uniform	$2\times$	97.90%	92.17%	90.37%
Uniform	$4\times$	97.87%	92.16%	90.32%
Uniform	$16\times$	97.86%	92.11%	90.25%
Task-related	$1\times$	97.97%	93.86%	91.29%
Task-related	$2\times$	97.89%	93.74%	91.08%
Task-related	$4\times$	97.84%	93.73%	91.02%
Task-related	$16\times$	97.82%	93.68%	90.93%

Under both relevance settings, even when the reference density is increased by a factor of 16, the count ratio is approximately 239:1, and the change in every metric relative to its initial value remains below 0.4 percentage points.

APPENDIX C: VIEWPOINT SEARCH

C.1 EXPERIMENTAL SCENE SELECTION AND FEASIBLE-DOMAIN SETTING

The experiment uses 20 scenes randomly selected from each of the four task suites of the LIBERO-Plus object-layout-variation scenes, for a total of 80 scenes. The scene list is as follows.

Table C1. Experimental scene list.

Task suite	Scene IDs
LIBERO-Spatial	1904, 2062, 1991, 1791, 2098, 1733, 1772, 2018, 1976, 1830, 2033, 1989, 1732, 1859, 1901, 2022, 1799, 1889, 2051, 2010
LIBERO-Object	1888, 2071, 1962, 1878, 2013, 1858, 1940, 2129, 2169, 2182, 1958, 1989, 1864, 2015, 2051, 1924, 2087, 2156, 1982, 2009
LIBERO-Goal	2258, 2305, 2055, 2290, 2298, 1930, 2234, 2306, 2119, 1925, 2227, 2269, 2144, 2100, 2033, 2074, 2009, 2252, 2312, 2160
LIBERO-10	2157, 1981, 1972, 2162, 1982, 2141, 1970, 2160, 1980, 2130, 1971, 2156, 1976, 2080, 2051, 2112, 1946, 1967, 2190, 2115

For each scene, a cube above the tabletop/floor where the task is located is defined as the feasible domain. At the same time, to ensure that the camera frustum can at least cover the task region, a minimum bounding sphere is constructed for the set of task-related objects. A sphere is then formed using the radius at which the camera frustum, here the default LIBERO camera frustum, can cover the minimum bounding sphere and using the center of the minimum bounding sphere as its center. The region inside this sphere is defined as infeasible. The final feasible domain is the space inside the cube with the inner-sphere region removed; points on the boundaries of the cube and inner sphere are allowed. The side length of the cube is set to 2.2 times the radius of the inner sphere.

C.2 BASELINES AND TASK ADAPTATION

Details of the compared methods are provided in Table C2.

Adaptation note. For baselines requiring adaptation to the shared task and camera-position setting, we retain the core algorithmic mechanisms of the original methods and make only the minimal changes needed to provide the common inputs, objective interface, and feasible-domain constraints.

Table C2. Compared methods, objectives, implementation sources, and experimental adaptations.

Method	Objective and implementation
Gaussian-process Bayesian optimization (GP-BO)	Task pixel score and expected improvement. Uses an in-house implementation of Gaussian-process regression and expected improvement. Camera positions are normalized to the feasible domain, evaluated using the shared task pixel scorer, and restricted to feasible candidates.
CIM surrogate gradient	Surrogate gradients fitted to sampled task pixel scores. Builds on the per-line sampling, polynomial fitting, and gradient-update mechanisms of the official IGLOV code. The task pixel score is used as the task-specific information measure, with the corresponding threshold adjustment; sampling and movement are adapted to three-dimensional camera positions.
Covariance matrix adaptation evolution strategy (CMA-ES)	Task-weighted hard-ray pixel score. Uses the public pycma library while retaining its optimization core. The optimization variables are three-dimensional camera positions, the initial search scale is adjusted to scene size, and infeasible candidates are resampled.
GradientNBV	Task-adapted differentiable ray gain. Builds on the official voxel-grid and ray-sampling components while retaining their differentiable sampling and occlusion mechanisms. Frozen geometry provides the input, task relevance is used in place of semantic uncertainty, and multiple task roles are aggregated. The method is interfaced with the shared camera model and three-dimensional position-update formulation.
VISO-grasp	Analytic geometric velocity field defined by targets and occluders. Builds on the geometric velocity-field mechanism from the official vMF-Contact repository, retaining its anchor construction, occluder filtering, and velocity aggregation. Frozen geometry is used in place of online perception, task objects jointly define the target, and position updates are expressed within the shared feasible-domain constraints.

Table C2 (continued). Compared methods, objectives, implementation sources, and experimental adaptations.

Method	Objective and implementation
Sobol	Task-weighted hard-ray pixel score. Uses SciPy's Sobol sampler, maps three-dimensional samples to the feasible domain, and evaluates them with the shared task pixel scorer.
Random	Task-weighted hard-ray pixel score. Uses NumPy to sample camera positions uniformly within the feasible domain and evaluates them with the shared task pixel scorer.

For InfoField3D, we use SciPy's SHGO, with InfoField3D's analytic gradient supplied to its local optimizer, and SLSQP implementations with default solver options, keeping all SLSQP settings fixed across the two gradient variants.

Specifically, in this experiment, we sum the occlusion-attenuated responses of the local units separately for each task target and normalize each target's response by its total area-weighted relevance, which is independent of viewpoint. This preserves the objective's emphasis on the total amount of available task-relevant information. We do not apply viewpoint-dependent coverage normalization, so the score cannot increase merely because less irrelevant geometry is observed. For multi-role tasks, we adopt the product aggregation described in Appendix A.3, specifically the equally weighted geometric mean of the normalized target scores, to balance the observation requirements of all targets. Low-scoring targets thus constrain the overall score, reducing the risk of a well-observed target masking another that is barely visible. SHGO, SLSQP with analytic gradients, and SLSQP with sampling-based gradient estimates all optimize this same objective. The comparison metrics are evaluated consistently across methods under this experimental setting.

Several learning-based viewpoint-prediction methods were also considered, including TVVE (repository name: TAVP), MVPNet, and VG-AVS. However, their currently available implementations either do not provide all components required for direct reproduction under our setting or adopt viewpoint-prediction formulations that differ from the free-space observation-position search considered in this paper. We therefore do not include them in the experimental comparison.

C.3 Evaluation Metrics and Numerical Results

The evaluation metrics are calculated as shown in Table C3, and the numerical results in the figures of the main text are shown in Tables C4 through C7.

Additional details on the figures in the main text: In Figure 4, curves show mean performance across scenes, and shaded bands indicate 95% confidence intervals for the mean, obtained by bootstrap resampling of scenes with replacement. Error bars in Figure 5(a) indicate 95% confidence intervals for the mean. In Figure 5(b), the solid curve shows the cross-scene mean, while the dark and light bands indicate the 25th–75th and 5th–95th percentile ranges across scenes, respectively.

Table C3. Evaluation metrics and scoring mechanisms.

Metric	Definition or configuration
Task-region visibility (geometric mean across task roles)	Measures the visible fraction of task-relevant surfaces, weighted by surface area and task relevance. A surface sample counts as visible only when it lies within the camera's field of view, faces the camera, and is not occluded. Visibility is computed separately for each task role and combined using the geometric mean, so severe occlusion of one role lowers the overall score.
Directional visibility response	Extends task-region visibility by weighting visible surfaces according to their viewing angle. The weight is highest for a frontal view and decreases exponentially as the view becomes more oblique. Occluded samples contribute nothing. Scores are combined using the geometric mean across task roles, reflecting both visibility and viewing direction.
InfoField3D score	Although this is the objective value of our own method, we believe it is still informative for viewpoint evaluation because it jointly considers favorable viewing orientation, maximal exposure of the observation region, and reduced occlusion risk. We therefore include it as an additional reference metric.
GPT-4o semantic score	Presents a high-quality reference image and the evaluated image from the same scene to GPT-4o. The reference calibrates the upper end of the scoring scale. Based on visible evidence, the model uses two items for single-role tasks and three for multi-role tasks, as defined in Section C.4. Each item receives an integer score from 0 to 10 and a brief justification. Item scores are averaged equally and converted to a 0–100 scale.

Note. Higher scores indicate better performance for all metrics.

"—" indicates that the method had not produced an output within the corresponding time budget.

Table C4. InfoField3D Score.

Method / search time (s)	0.03	0.1	0.3	1	20
InfoField3D (SHGO)	***0.1995***	***0.2091***	***0.2113***	***0.2113***	***0.2113***
InfoField3D (Analytic Gradient)	0.1781	0.1786	0.1786	0.1786	0.1786
InfoField3D (Sampling-Based Gradient)	0.1282	0.1691	0.1710	0.1710	0.1710
GP-BO	—	0.0299	0.0993	0.1243	0.1497
CIM surrogate gradient	—	0.0360	0.0363	0.0482	0.1177
CMA-ES	—	0.0333	0.0405	0.0676	0.1419
GradientNBV	0.0321	0.0526	0.0760	0.0810	0.0809
VISO-grasp	0.0710	0.0708	0.0707	0.0707	0.0707
Sobol	—	0.0338	0.1020	0.1308	0.1470
Random	—	0.0332	0.0814	0.1244	0.1497

Table C5. Task-Region Visibility (%).

Method / search time (s)	0.03	0.1	0.3	1	20
InfoField3D (SHGO)	***46.96***	***46.51***	***46.32***	***46.32***	***46.32***
InfoField3D (Analytic Gradient)	40.27	40.30	40.30	40.30	40.30
InfoField3D (Sampling-Based Gradient)	31.92	39.29	39.37	39.37	39.37
GP-BO	—	0.71	16.45	33.67	44.12
CIM surrogate gradient	—	0.93	1.82	3.73	30.77
CMA-ES	—	0.87	2.39	9.53	40.01
GradientNBV	0.72	5.22	15.80	19.28	19.11
VISO-grasp	20.37	20.53	20.54	20.54	20.54
Sobol	—	0.78	17.66	32.50	43.37
Random	—	0.73	13.97	31.25	43.20

Table C6. Directional visibility response (%).

Method / search time (s)	0.03	0.1	0.3	1	20
InfoField3D (SHGO)	***26.70***	***27.59***	***27.58***	***27.58***	***27.58***
InfoField3D (Analytic Gradient)	23.03	23.06	23.06	23.06	23.06
InfoField3D (Sampling-Based Gradient)	16.71	22.02	22.20	22.20	22.20
GP-BO	—	0.34	8.80	17.02	22.46
CIM surrogate gradient	—	0.45	0.87	1.82	15.98
CMA-ES	—	0.42	1.15	4.61	20.24
GradientNBV	0.36	2.46	7.38	8.96	8.89
VISO-grasp	8.88	8.87	8.87	8.87	8.87
Sobol	—	0.38	9.30	16.66	22.21
Random	—	0.35	6.76	15.47	22.17

Table C7. GPT-4o Score.

Method / search time (s)	0.03	0.1	0.3	1	20
InfoField3D (SHGO)	***83.98***	***83.42***	***84.33***	***84.33***	84.33
InfoField3D (Analytic Gradient)	75.90	76.50	76.50	76.50	76.50
InfoField3D (Sampling-Based Gradient)	53.77	66.77	66.60	66.60	66.60
GP-BO	—	32.89	64.67	82.35	84.10
CIM surrogate gradient	—	32.61	24.27	37.77	76.19
CMA-ES	—	32.36	28.69	55.13	82.73
GradientNBV	22.69	39.13	49.63	52.38	53.50
VISO-grasp	49.65	50.98	50.52	50.58	50.58
Sobol	—	27.74	62.54	81.00	***84.40***
Random	—	25.22	74.00	81.17	84.04

C.4 GPT-4o Viewpoint Evaluation

In our viewpoint-quality evaluation, to provide a more diverse set of evaluation metrics, we use a multimodal model in addition to scores based purely on geometric rules. This model-based score serves mainly as a relatively subjective external reference. We make every effort to ensure fairness in the scoring procedure and conduct additional tests to verify that the scores are largely stable for the same viewpoint. The detailed evaluation procedure is described below.

GPT-4o receives the task description, a reference image, and the image to be scored. The reference is selected by the highest task-region visibility among candidate viewpoints in the same scene and calibrates the upper end of the scale. All methods share this reference. Images are rendered at 224×224 pixels with consistent identity annotations to reduce object ambiguity; method names and search scores are withheld. The model evaluates task-related visibility from visible evidence, with temperature set to 0.

Scoring items follow the scene's fixed task roles and relevant-part lists (Table C8). These are shared across methods. Source and placement roles are predefined, except for three-role LIBERO-Spatial tasks, where the placement target is resolved from the supplied task statement.

Table C8. Task-specific GPT-4o scoring items.

Task roles	Scoring items
Single role	Target identification; visibility of task-relevant parts and their spatial relationships.
Multiple roles	Source-object identification; grasp-region clarity; visibility of the placement target's receiving region.

For each item, GPT-4o gives a brief observation and an integer score from 0 to 10. The fixed rubric anchors 0 at an invisible or unrecognizable target or region, 5 at approximately half of the key region being visible with ambiguity or occlusion, and 10 at fully clear and complete task-relevant content. For placement, the criterion concerns the receiving region itself. The two scores for single-role tasks or three for multi-role tasks are averaged equally and multiplied by 10, giving a 0–100 score.

Fixed-seed subsets comprising approximately 20% of unique evaluated views are scored again under the same task, reference, and model settings. The mean absolute difference between the original and repeated total scores ranges from 3.22% to 6.42% of the full scoring scale across methods.

C.5 Automated Multimodal Annotation Pipeline

The task-relevance map indicates which regions of the 3D scene are important to observe for task execution. Its construction is not restricted to any particular method: it may be provided by multimodal models based on task understanding, derived from predefined rules, or even specified manually by humans. In our experiments, we use GPT-4o for task understanding and open-source models to generate the corresponding annotations.

The figure below shows an example from the LIBERO task "Pick the ketchup and place it in the basket." GPT-4o identifies task-relevant basket regions and produces weighted positive and negative text queries from the task and nine initial RGB views (one shown). Modified-SAM (Kirillov et al., 2023) proposes image regions, and OpenCLIP (Radford et al., 2021; Cherti et al., 2023) scores them against the queries to produce image-space relevance. Depth and simulator instance correspondence bind these scores to the basket's 3D surface, where the rim and interior contributions are fused into a task-relevance map.

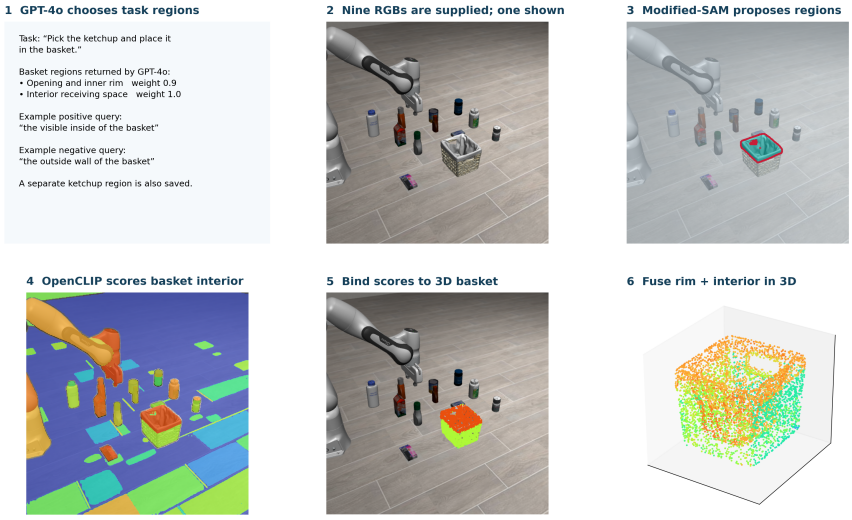


Figure C1. From task instructions to a task-conditioned 3D surface relevance map.

C.6 Preprocessing Cost

We measure construction from preloaded geometry and task-relevance weights, excluding input generation, file I/O, query-engine compilation and warm-up, and search. All point-based methods use the same 43,955-point inputs; our implementation uses 600 units. VISO retains its native collision-mesh input. Measurements use the same RTX 4090-equipped server, with one PyTorch CPU thread and GPU synchronization at completion. The five discrete baselines share the same semantic spatial-index builder and therefore receive the same measured cost.

We additionally evaluate the vector-valued spherical harmonics described in Appendix A.2 as a comparison variant. It compresses both the original point cloud and the semantic features attached to the points into the representation, enabling fast responses to subsequent queries. Here, we evaluate 128- and 512-dimensional semantic features.

The results are shown in Table C9. When the initial construction cost is included, the search efficiency of the compared methods is shown in Figure C2.

Table C9. Representation construction and task-update costs with preloaded geometry and task relevance.

Method	Representation	Construction (ms)	Update (ms)	Time saved (%)
Ours	Continuous field	66.81	2.35	96.48
Ours	Continuous field (128D)	117.38	0.222	99.81
Ours	Continuous field (512D)	153.56	0.278	99.82
GradientNBV	Task-weighted voxel grid	23.00	11.78	48.78
VISO	Bounding-box-based scoring function	23.39	0.095	99.59
GP-BO	Semantic spatial index	3.59	0.198	94.48
CIM	Semantic spatial index	3.59	0.198	94.48
CMA-ES	Semantic spatial index	3.59	0.198	94.48
Sobol	Semantic spatial index	3.59	0.198	94.48
Random	Semantic spatial index	3.59	0.198	94.48

Construction times are averaged over 80 scenes with five runs per scene in a warmed runtime. Updates use three task changes per scene with five runs per change. Furthermore, we fix the scene and vary only the task-relevance weights by randomly selecting a new focus center and generating a new Gaussian relevance map, in order to measure the cost of task switching within the same scene. For each scene, we test three such task changes, each repeated five times, and measure the update time when reusing the cached geometry information. For the vector-valued SH variants, scene geometry and point features remain fixed; an update changes the query vector and reads out the corresponding field.

It can be seen that, although our method incurs a relatively higher upfront construction cost, it achieves a high degree of reuse across different tasks. With a larger upfront investment, the representation can be further extended to vector-valued spherical harmonics, enabling even faster switching between task queries. Unlike conventional discrete precomputation, however, the resulting representation is a general semantic 3D vector field that can directly accommodate different task queries.

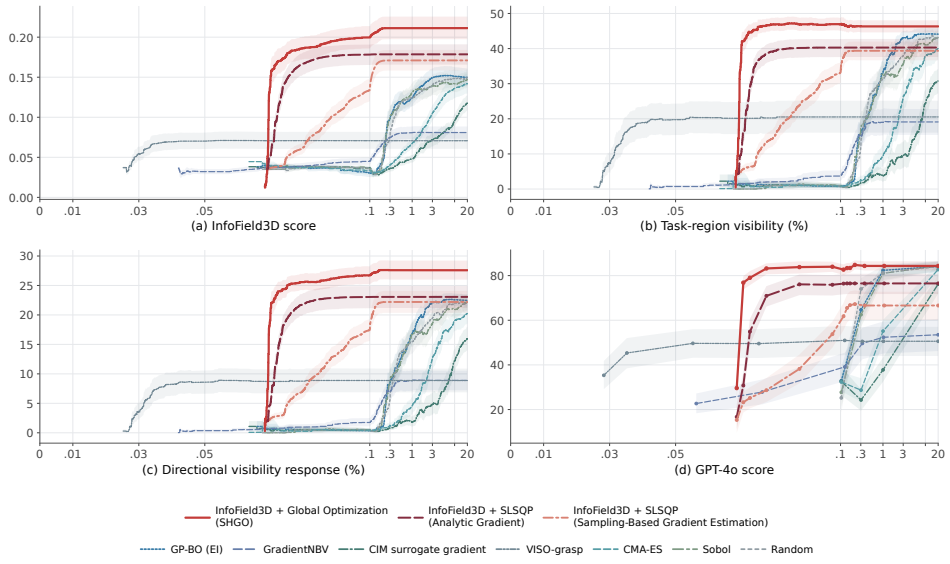


Figure C2. Efficiency curve accounting for preprocessing cost.

APPENDIX D: VIEWPOINT ROBUSTNESS

We evaluate the robustness of the viewpoints recommended by eight methods to positional perturbations in the 80 scenes of Table C1. For each recommended viewpoint, six positional perturbations are applied along the positive and negative directions of a random orthonormal basis shared within the scene. The perturbation magnitudes are $\varepsilon \in \{0.005, 0.01, 0.02, 0.04, 0.08\}l$, where l is the side length of the feasible-domain cube. After perturbation, the camera continues to face the same scene center. Evaluation uses the unified external visibility metric V. Perturbation loss is defined as the maximum nonnegative visibility decrease over valid test directions. The losses of the points recommended by each method under different perturbation strengths are shown in Table D1.

Table D1. Mean sampled perturbation loss at five displacement scales (percentage points; lower is better).

Method	Loss $(0.005l)\downarrow$	Loss $(0.01l)\downarrow$	Loss $(0.02l)\downarrow$	Loss $(0.04l)\downarrow$	Loss $(0.08l)\downarrow$
InfoField3D (Ours)	**0.122**	**0.224**	**0.513**	**0.998**	**2.522**
GP-BO	0.174	0.308	0.595	1.301	2.841
Sobol	0.266	0.509	0.939	1.870	3.689
Random	0.281	0.507	0.935	1.920	3.969
CIM	0.450	0.796	1.627	3.039	5.931
CMA-ES	0.308	0.629	1.534	3.324	7.079
GradientNBV	0.501	1.001	2.537	5.556	9.838
VISO-grasp	0.464	0.825	1.431	2.316	2.833

APPENDIX E: CONTROLLED VLA EVALUATION

E.1 CONTROLLED VLA EXPERIMENTAL-DESIGN DETAILS

This experiment uses the same scene set as Table C1.

In this experiment, we seek to minimize interference from the VLA model's own training distribution in the result analysis, thereby focusing more directly on the relationship between the information sufficiency of an observation viewpoint and downstream VLA performance, and on whether the movement direction recommended by InfoField3D is better than other equidistant directions. To this end, we use a controlled-variable experimental design and constrain additional changes beyond viewpoint as much as possible.

First, we constrain the feasible viewpoint region. A camera that is too far from or too close to the task region, or that departs substantially from the third-person observation distribution commonly seen by a VLA, may cause large changes in policy performance, and these changes may not arise from whether task information is sufficient. To reduce this factor, the experiment fixes the camera-to-task-center distance. The task center is defined as the center of the minimum bounding sphere of the task-related object set, and the camera radius is the distance from the default third-person camera to this center. Candidate positions are therefore first constrained to a sphere of fixed radius. Further considering that the default third-person camera is usually on the side opposite the robot arm, we exclude a spherical region of approximately 50° on the robot-arm side and also remove physically invalid positions below the floor or tabletop. The result is an upper-hemisphere candidate region on the side opposite the robot arm with an opening angle of approximately 130°.

On this basis, we further construct a local viewpoint-rescue experiment. Initial viewpoints with a clear information bottleneck are first selected according to their hard-visibility scores. This is because, when an initial viewpoint can already observe the task region sufficiently, changes in performance caused by a further small viewpoint shift may reflect random variation in the policy itself more than improved information accessibility. For these information-limited initial viewpoints, InfoField3D recommends an optimized viewpoint in the local neighborhood and limits its angular displacement from the starting point to within 20° to prevent a large viewpoint change from substantially altering the camera distribution.

Under these rules, we obtain 697 information-limited initial viewpoints, referred to below as trapped viewpoints. For each trapped viewpoint, we evaluate both the starting viewpoint and the viewpoint recommended by InfoField3D. At the same time, to test whether the recommended movement direction is better than other local directions, we construct a trident probe: while keeping the distance from the starting point unchanged, we rotate the recommended movement direction by $\pm 120^\circ$ to obtain two equidistant control viewpoints, as shown in Figure E1. Because the three endpoints have the same movement magnitude and all lie in a local region near the starting point, this design can control changes in camera distribution to a substantial extent, allowing comparisons among directions to reflect differences in task-relevant information accessibility more directly.

The experiment uses a frozen $\pi 0.5$ as the VLA policy and replaces only its third-person camera viewpoint. For the starting viewpoint, recommended viewpoint, and two control viewpoints in each trident probe, VLA rollouts are conducted with 5 different random seeds. In total, 2,788 viewpoint locations are evaluated, producing 13,940 VLA execution results.

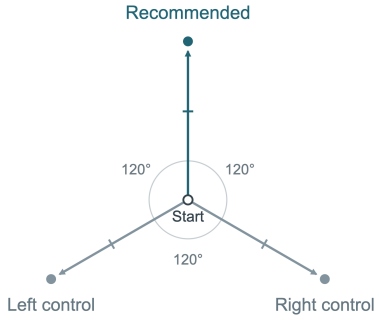


Figure E1. Local trident-probe schematic. Three equal-length directions extend from the starting viewpoint: the InfoField3D recommendation and two controls rotated by $+120^\circ$ and -120°. The drawing illustrates their local directional relationship; the observation-sphere constraint is described in the text.

To analyze changes in the downstream performance of the VLA more finely, this experiment specifically examines the completion degree of each VLA execution. Completion-degree scores are determined by the specific task type, using the criteria shown in Table E1.

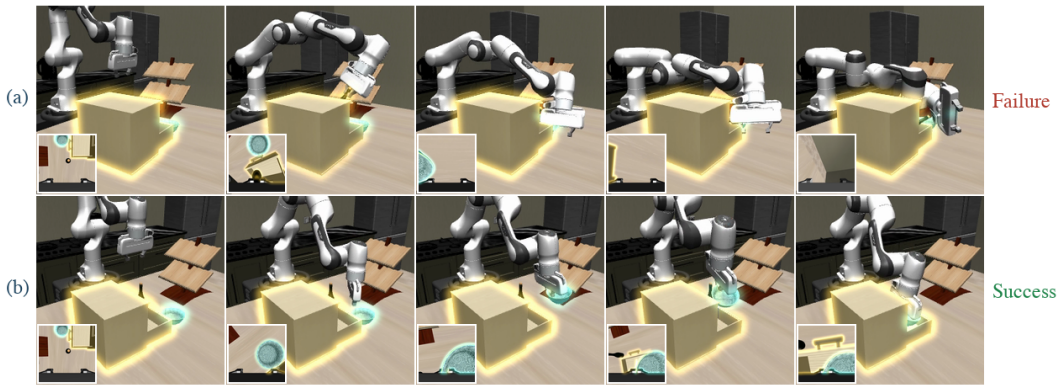
Table E1. Quantitative scoring criteria for task completion degree.

Task type	25 points	50 points	75 points
Single-object grasp and placement	The left or right finger group makes real contact with the correct source object, with contact distance no greater than 0.	Both finger groups are in contact, and the height of the source-object root is at least 2 cm above its initial height for 3 consecutive control steps.	The 50-point stage has been reached, and the source object is currently within the valid placement range of the target region, but the full official success criterion has not yet been met.
Placement into a target region	The correct fingers contact the source object.	Two-finger contact is established and the object is lifted by at least 2 cm for 3 consecutive control steps.	The object has previously been lifted, and it is currently within the horizontal extent of the official region and above the effective target surface.
Multi-object placement	Any not-yet-completed object is contacted by the correct fingers.	If only one object remains incomplete, that object is supported by both fingers and lifted for 3 consecutive control steps; if multiple objects remain, at least one native placement subgoal becomes true.	Targets satisfied initially remain satisfied; at least one active target is currently true; and the other objects must previously have been lifted correctly and are currently in deliverable positions.
Articulation operation	The correct movable part is contacted; contact with the cabinet body, base, or another drawer does not count.	Progress from the initial joint value to the native success boundary reaches $1/3$ for 3 consecutive control steps.	Progress reaches $2/3$ for 3 consecutive control steps.
Pushing into a target region	The correct palm or fingers contact the target object.	Planar distance to the official target region decreases by at least $1/3$, while planar displacement of the source object is at least $5\,\mathrm{mm}$ and its height change is no more than $1\,\mathrm{cm}$, for 3 consecutive control steps.	The same conditions hold, but the planar distance decreases by at least $2/3$, for 3 consecutive control steps.
Switch-and-placement combination	The correct source object or movable part is contacted.	The native switch subgoal is completed, or the prescribed sustained lifting or articulation progress is achieved.	The placement, lifting, delivery geometry, and maintenance of initially satisfied targets are all satisfied simultaneously.

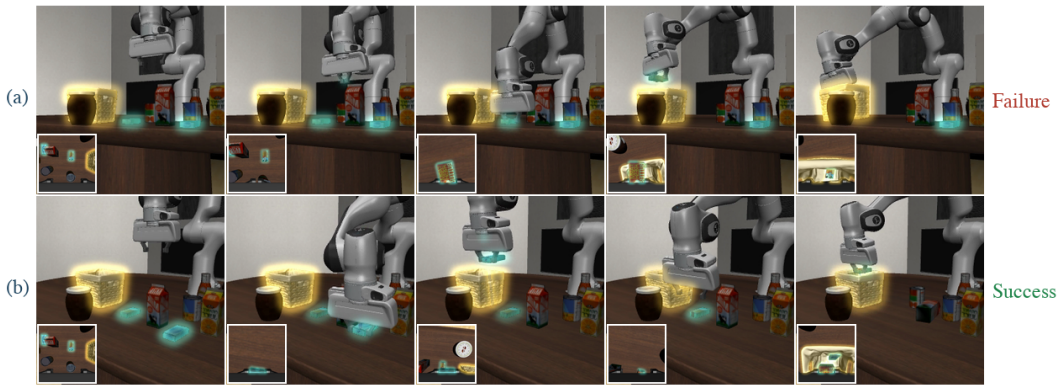
The stages in this table represent task-specific discrete execution progress rather than the physical proportion of the task completed; 100 points means that the official task-success criterion is met.

E.2 Execution Examples

Figure E2, including its continuations, collects eight execution examples. In each image strip, row (a) uses the initial third-person viewpoint, and row (b) uses the recommended viewpoint from the same scene; the inset shows the first-person camera observation, and the yellow overlay marks the task-relevant region.

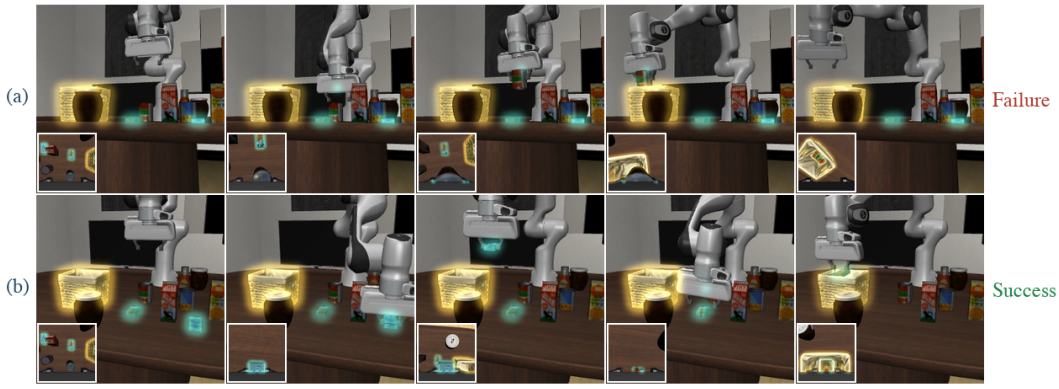


(a) LIBERO-10, scene 1976.

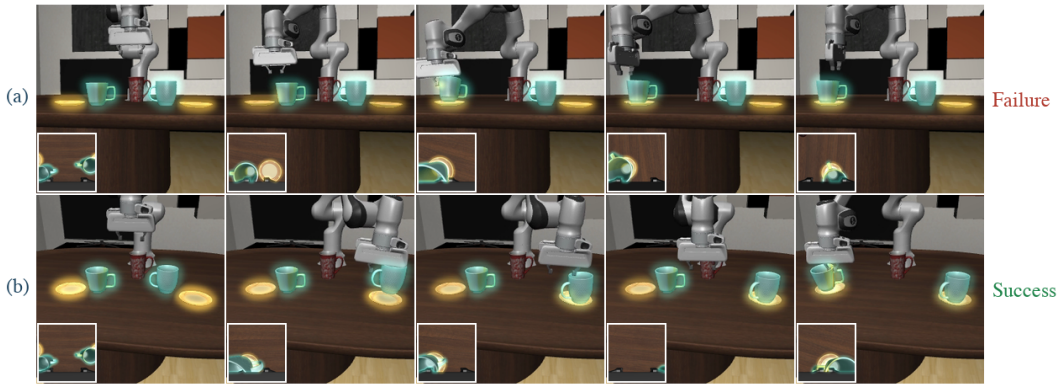


(b) LIBERO-10, scene 2112, starting point g05.

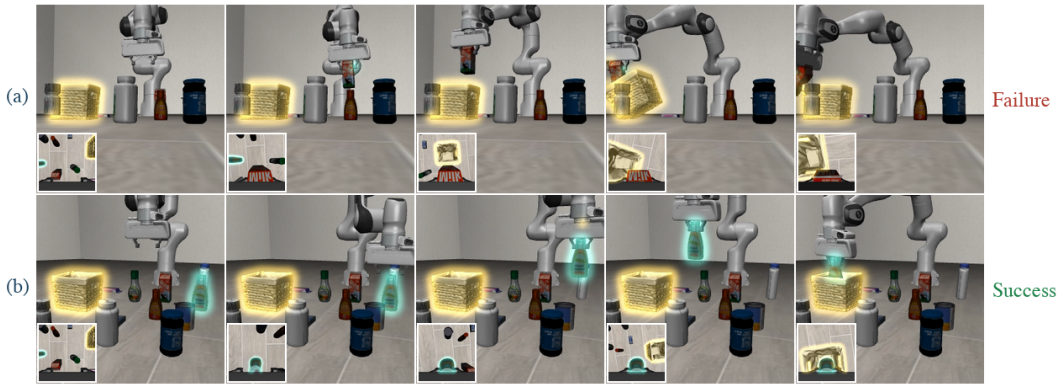
Figure E2. Paired execution examples. Within each case, image row (a) uses the initial third-person viewpoint and row (b) uses the recommended viewpoint. Insets show wrist-camera observations; yellow overlays mark task-relevant regions. Case labels (a)–(h) run across this figure and its continuations.



(c) LIBERO-10, scene 2112, starting point g07.



(d) LIBERO-10, scene 2156.

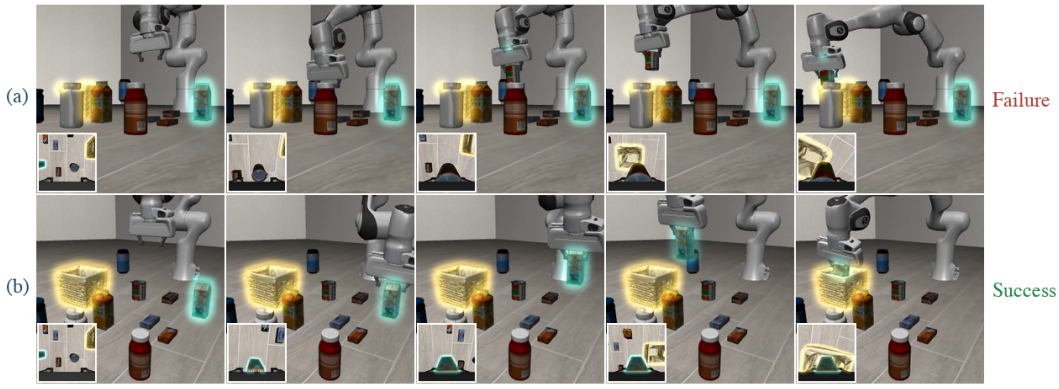


(e) LIBERO-Object, scene 2015.

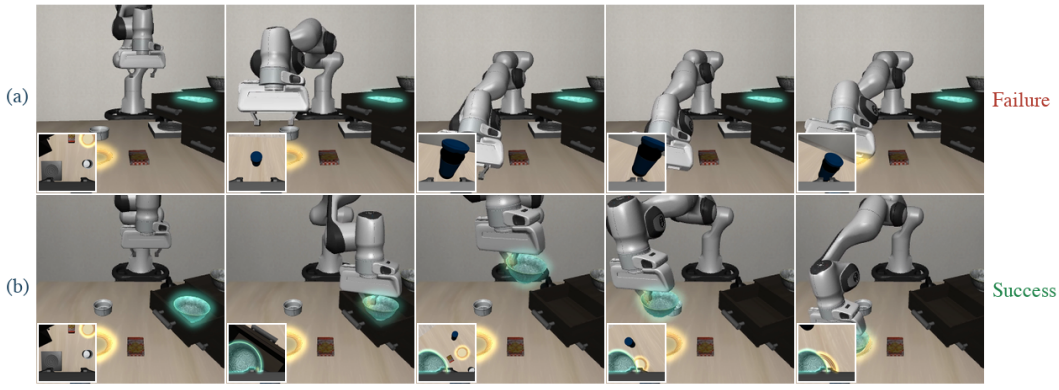
Figure E2 (continued). Paired execution examples, cases (c)–(e).



(f) LIBERO-Object, scene 2051, starting point g02.



(g) LIBERO-Object, scene 2051, starting point g07.



(h) LIBERO-Spatial, scene 1830.

Figure E2 (continued). Paired execution examples, cases (f)–(h).