6 Appendix

6.1 Experiments on CO3D dataset

To demonstrate the generalization ability across diverse categories, we conducted additional experiments on 10 categories from the CO3D dataset [63]: Toy Truck, Bench, Donut, Broccoli, Toy Train, Apple, Teddy Bear, Hydrant, Book, and Toaster. We compared our results with the baseline method vMap [45], as shown in Figure 5. Our approach consistently outperformed the baseline, benefiting from the incorporation of generative priors from a single pre-trained model.

6.2 More Qualitative Visualizations

Visualization on More Categories on ScanNet dataset. We further illustrate the effectiveness of our approach in generalizing to a variety of categories through qualitative visualizations, as shown in Figure 6. Unlike the DSP method [3], which is based on a single-category model, DeepSDF [27], our system is capable of supporting multiple object categories using a single pre-trained diffusion model, Shap-E [29].

Scene-level Visualization. Figure 7 showcases a 3D map visualization of a scene on the ScanNet dataset, containing multiple objects from various categories, including four chairs, a sofa, and a table. Each instance is independently reconstructed using 10 RGB-D views.

Visualization of Multi-view Inputs. We visualized 10 input images, 3D camera poses, and corresponding object meshes in Figure 8, which includes two objects from the ScanNet dataset and one object from the CO3D dataset. The ScanNet dataset presents a challenging camera trajectory, with all inputs gathered from a single direction relative to the objects. This scenario is typical in real-world applications such as robotics and augmented reality, where a robot or user cannot easily circle an object. It also underscores the importance of sparse view mapping, where prior information is crucial for completing and providing reasonable estimates for occluded parts. An extreme case is illustrated where all camera views have small baselines, resembling single-view mapping. In the CO3D dataset, the cameras are intentionally positioned to circle the objects; however, we randomly sampled only ten views from this trajectory, resulting in a sparse coverage of the objects. We demonstrate that our system can effectively handle all these situations.

6.3 More System Evaluations and Discussions

Performance on Single-view Inputs. Single-view mapping is particularly challenging due to its highly ill-posed nature. All systems experience significant performance losses when relying on only a single view. Our main challenges stem from the high-dimensional shape representation required for NeRF, which models multiple categories of objects, encompassing both texture and shape. In contrast, DSP employs DeepSDF, an SDF-based representation specifically trained on a single category, focusing solely on shapes. This approach has a much smaller parameter space and requires fewer constraints. However, this limited parameter space restricts its ability to generalize to other categories. Our method, on the other hand, can generalize across multiple categories using a single model and can continue to improve with additional observations. With the assistance of prior knowledge, we achieve effective enhancements in single-view scenarios compared to the baseline method, vMap. Additionally, we have a parameter that allows us to increase the weight of prior constraints during optimization, enabling us to place greater trust in the prior information for producing complete and reasonable mapping results.

Performance on Out-of-distribution Objects. Dealing with out-of-distribution objects is a critical challenge for developing a general mapping system. In this paper, we present a formulation that leverages prior knowledge from a pre-trained generative model. While we use Shap-E as a representative example at the time of writing, it's important to note that our method is not specifically tailored to Shap-E but applies to a broader class of diffusion models. The field of 3D generation

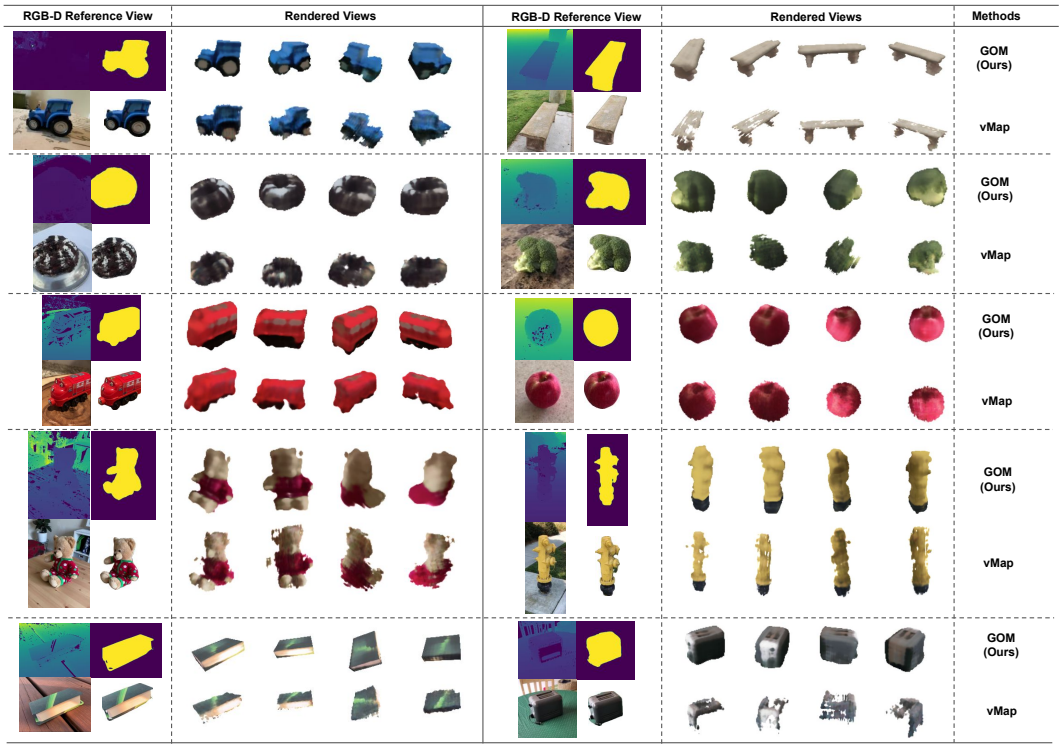

Figure 5: Effectiveness of Priors Across 10 Categories on the CO3D Dataset: Toy Truck, Bench, Donut, Broccoli, Toy Train, Apple, Teddy Bear, Hydrant, Book, Toaster, compared to vMap. The results are based on 10 RGB-D views.

models is rapidly evolving, with larger datasets (e.g., ObjaverseXL [39], OmniObject3D [64]) and more powerful generative models emerging.

Our proposed formulation allows for seamless integration of more advanced models without requiring any fine-tuning, enabling a plug-and-play approach. We demonstrate that the current Shap-E model is effective for a diverse range of objects, as shown by our results on both ScanNet and CO3D. While generative models may not provide detailed structural constraints for out-of-distribution objects, they still offer valuable general priors, such as completeness and smoothness. This is illustrated in the experiment in Figure 9, where a prompt like "an object" yields useful results.

Our system effectively combines observations and prior constraints to identify an optimal solution from both sources. When the prior information is less accurate, we can adaptively adjust the weights to rely more on observations for out-of-distribution objects, tailored to specific applications. Furthermore, it would be valuable future work to quantify the level of out-of-distribution characteristics, such as uncertainty [9, 65], and to self-adjust the weights to enhance mapping performance.

Text Prompts and Segmentation Methods. When selecting text prompts as prior conditions, our system demonstrates certain flexibility regarding the content of the text. In our experiments, we provide an example using the prompt "a category," where category can be easily obtained from off-the-shelf object detection or segmentation algorithms. We also conducted a new experiment to analyze the sensitivity to the quality of the text prompt, as shown in Figure 9. In this qualitative analysis, we focused on a chair that was only observed from the front, using three types of prompts: (1) Precise label: "a chair"; (2) General label: "an object"; and (3) Two incorrect labels: "a table" and "a ball." Our findings indicate that our system can robustly leverage prior knowledge from varying levels of label specificity by optimizing both the observations and the prior information to converge on a solution that integrates both aspects. Particularly for occluded areas with insufficient observational data, the prior knowledge effectively constrains the reconstruction to produce a reasonable

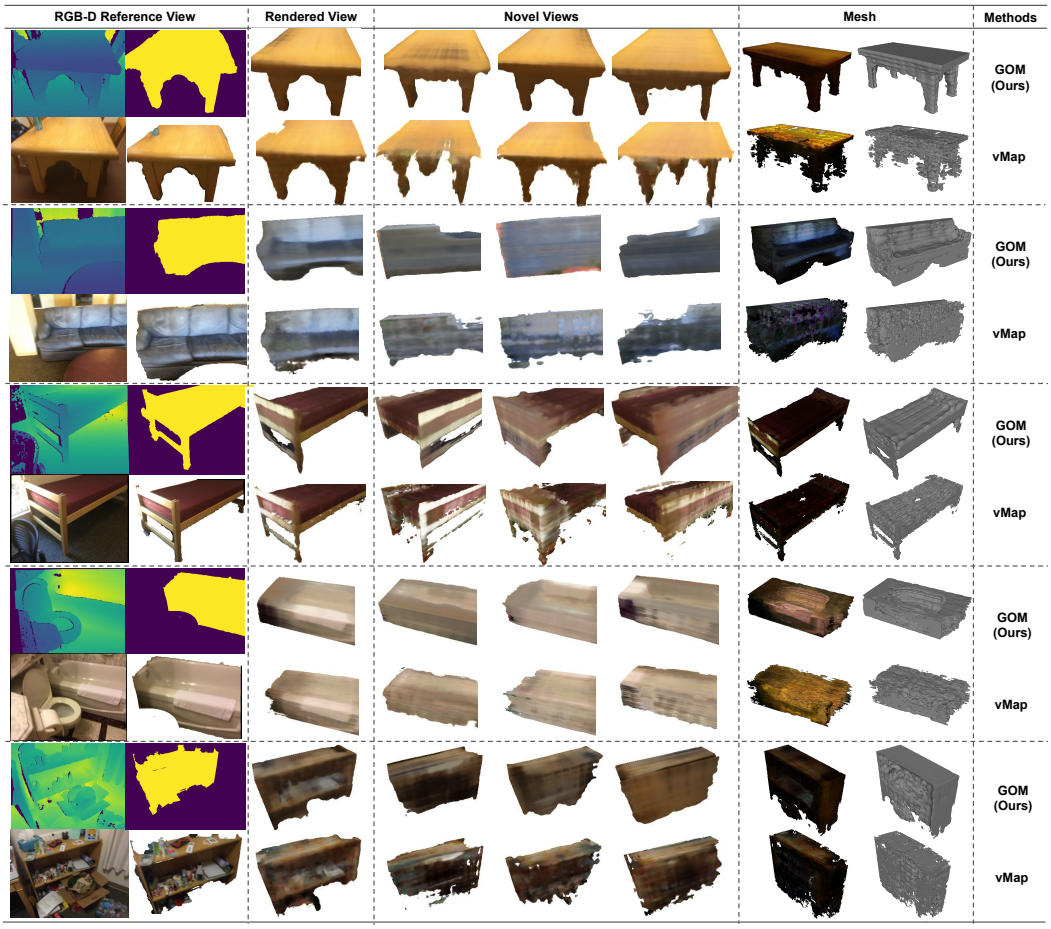

Figure 6: Effectiveness of Priors Across Multiple Categories: Leveraging priors, our method (GOM) can produce higher-quality, 3D-consistent views and generates 3D meshes with fewer artifacts compared to vMap. The results are based on 10 RGB-D views.

and smooth surface compared to vMap without any prior, even when using a general prompt like "an object."

In cases where incorrect labels were used, such as "a table" and "a ball," some artifacts were introduced into the structure (e.g., a ball-like back part of the chair). However, the results were still significantly more coherent than the blurry areas observed in the baseline vMap. For the segmented areas, we utilized the provided "ground truth" masks from the ScanNet dataset, which are not manually annotated but projected from segmented point clouds. These masks may still contain artifacts and missing parts, and our experiments on ScanNet highlight the system's performance under imperfect input conditions. Additionally, we have parameters in place to balance the weight between observation confidence and prior knowledge (as detailed in Eq. 14-16 of the main paper). For future work, it will be valuable to model the uncertainty of the segmentation algorithm to enable adaptive weight balancing for improved robustness.

Global Map Consistency. We provide a qualitative scene-level result in Figure 7, demonstrating that our reconstructed shapes from multiple objects are consistent with the overall scene point cloud. In our main experiments, we evaluate each object individually to focus on validating the effectiveness of the generative prior model, which influences each object independently. Our contribution in this paper is distinct from other research efforts aimed at constructing a globally consistent map.

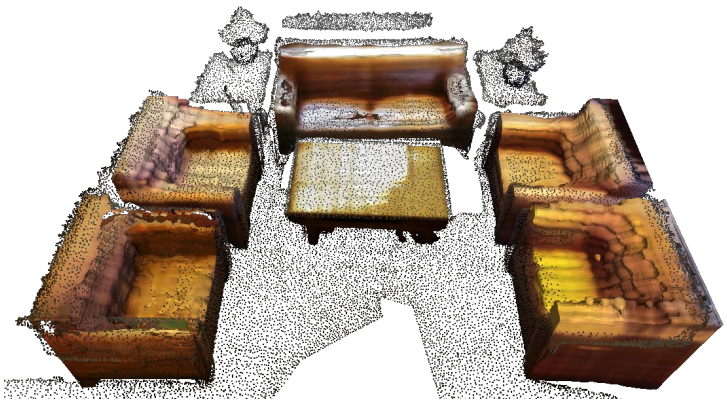

Figure 7: Scene-Level Visualization: An example of a reconstructed 3D map of a scene including four chairs, one sofa, and one table, all constrained with the same prior network.

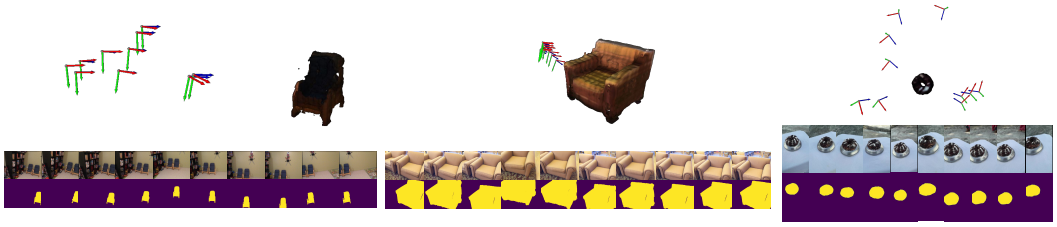

Figure 8: Input images from 10 RGB-D views. Sequences from the ScanNet dataset contain only observations from one side, which is more challenging for the occluded unseen parts. Sequences from CO3D contain 360-degree observations of objects.

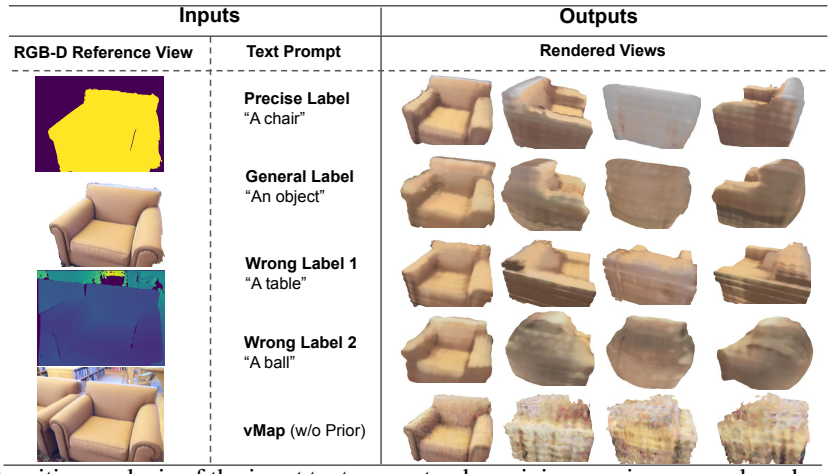

Figure 9: Sensitive analysis of the input text prompt, when giving precise, general, and wrong labels for the generative prior model. We find that our system can robustly leverage prior from different information-level labels, by optimizing both information from observations and prior knowledge to find a converged point that can match both parts.

Thanks to our optimization-based formulation, which combines both priors and observations, we can flexibly introduce additional constraints for global consistency. We can leverage well-established techniques from optimization-based SLAM and Structure from Motion (SfM), such as sliding window optimization and loop closure, to enhance global map consistency. We believe our approach can be extended to incorporate further constraints for global consistency, including cross-object semantic relationships and geometric supporting relationships, which can be represented as non-linear

constraints in an optimization framework. However, these enhancements require non-trivial engineering and are currently beyond the scope of this paper.

Deployment in real applications. We provide guidance for deploying our proposed method in real-world applications such as robotics to facilitate multi-object mapping. Data association—assigning observations of the same object from multiple frames—is crucial. There are various well-studied methods for data association, including the use of image features (such as manually defined ORB/SIFT or learned features like SuperPoint), ICP matching with point clouds, or probabilistic data association approaches [66, 67]. For camera pose estimation, we can employ offline calibration to determine intrinsic and extrinsic parameters, possibly using additional sensors such as IMUs, or implement online methods like SLAM [43]. We support two types of input for prior conditions: text prompts and segmented images. Both inputs can be easily derived from the input RGB-D images, for example, by utilizing off-the-shelf segmentation algorithms like groundedSAM [68]. Our text prompts are highly flexible; as demonstrated in the experiments in Figure 9, even a general prompt like "an object" can effectively constrain occluded parts. In the context of an autonomous robot, integrating extra text descriptions opens up new avenues for fusing information to construct a map. For instance, in human-robot interaction scenarios, a robot could ask the user, "What is the object in front of me?" and the user might respond, "It's a chair." This interaction provides valuable prior knowledge for the object mapping system. This capability enhances traditional SLAM and mapping systems by allowing them to gather information and constraints beyond mere sensor observations.

Relative Pose Definition and Evaluation. In our experiments, the Intersection over Union (IoU) is calculated between the bounding boxes of the estimated and ground truth shapes in world coordinates, rather than between voxelized meshes. This choice means that the metric is not sensitive to the detailed reconstruction quality. Evaluating independent pose errors—such as directly assessing translation, rotation, and scale relative to ground truth poses—does not provide a reasonable metric in our pose formulation. This limitation is shared by other object mapping systems like vMap, DSP, and Shap-E+ICP. In these systems, the pose is defined as a transformation from a canonical object coordinate system to the world coordinate system, which is closely linked to the shapes of objects in that canonical space. However, shape representations are not unique. For example, when using DeepSDF (for DSP) or NeRF representations (for vMap, Shap-E+ICP, and our method), shapes can exhibit different size ratios and orientations (e.g., a 90-degree rotation around the UP axis). Consequently, reconstructing a 3D object in the world involves multiple combinations of estimated poses and shapes in canonical space that can lead to a correct result. Therefore, there is no single, unique pose to evaluate against a "ground truth" pose. It is important to emphasize that, in applications such as robotics, our primary concern is the performance of the final 3D representation in the world, which encompasses both shape and pose, rather than evaluating a relative "pose" alone. Thus, we opt to directly evaluate the final outputs in world coordinates, reporting IoU and Chamfer Distance (CD) as quantitative metrics, accompanied by qualitative images to demonstrate performance relative to the baselines.

6.4 Ablation Study

Methods to Fuse Observations and Diffusion Priors. We compare three strategies to fuse observations and diffusion priors, as shown in Table 3. (1) *Optimize then diffuse*, which first optimizes shape and pose with geometric loss only for a given number of steps, and then uses the diffusion model to diffuse the shape. We notice that the information from observations is often lost during the post-diffusion process. Consequently, the ultimate shape diverges from the ground truth, resulting in a large metric error. (2) *Diffuse then optimize*, which first uses the diffusion model to generate a shape with a text condition, then uses the geometric loss to optimize both shape and pose. We observe that the unobserved segment of the shape is prone to corruption during the post-optimization process. Ultimately, this leads to a performance level that is similar to optimizing using only geometric observations without priors, which also remains more artifacts in the meshes and renderings. (3) *Jointly Optimize and Diffuse*, which fuses prior and observations iteratively during optimization with both diffusion prior and geometric loss so that both sources of information are active. This com-

Items	IoU ↑	CD ↓
Ours w/ *Optimize then Diffuse*	0.344	0.182
Ours w/ *Diffuse then Optimize*	0.416	0.160
Ours w/ *Jointly Optimize and Diffuse*	**0.429**	**0.157**
Gradients - *DirectDiffuse*	0.338	0.222
Gradients - *NoisePredict (Ours)*	**0.429**	**0.157**
Ours w/ Image Condition	**0.436**	0.160
Ours w/ Text Condition	0.429	**0.157**

Table 3: Ablation study on the strategies to fuse both observations and diffusion prior. Results are from 10 RGB-D views on Chairs of ScanNet dataset.

bined optimization can effectively merge constraints from both sources, thereby achieving superior performance compared to the other strategies.

Methods to Calculate Gradients from a Pre-trained Diffusion Model. The gradients derived from both the diffusion model and the observations are high-dimensional. Employing a method to effectively combine these gradients to guide the variable toward a convergence point is not a straightforward task. We compare our method with another to demonstrate the effectiveness of our approach, as shown in Table 3. (1) *NoisePredict (Ours)*. In Section 3.4 of the main content, we discuss our method to use the pre-trained diffusion model to predict the added noise and propagate back the error as gradients. This implicitly constrains the shape variable to lie inside the distribution modeled by the diffusion model, where it is trained to accurately predict the added noise. (2) *DirectDiffuse*, which directly uses the diffusion model to predict a less noisy version of the current shape for one step, as $\Theta_{t-1} = \epsilon_\beta(\Theta_t, C, t)$. Yang et al. [56] also use this method to leverage shape prior constraints from a single-category diffusion model for object reconstruction. Our task is more difficult with the extra unknown variable of pose. As shown in Table 3, *DirectDiffuse* underperforms in comparison to *NoisePredict (Ours)*. We attribute this to two primary factors. Firstly, making a pre-trained diffusion model to accurately predict a denoised variable is a challenging task. Secondly, each step of *DirectDiffuse* necessitates a precise timestamp t to denote the level of noise within the current variable, which becomes particularly complex when jointly optimized with gradients from observations. In contrast, our gradients can be derived from randomly sampled, uniformly distributed timestamps. This allows for flexible diffusion across arbitrary steps without the stringent requirement to adhere to the noise schedule from T to 0.

Input Conditions. We evaluate both input conditions supported by Shap-E model [29], image and text, as shown in Table 3. Each has its unique strengths and weaknesses, contingent on the specific applications. The image modality, which contains detailed prior information of a specific instance such as texture and shape, is nonetheless limited by the quality of the segmentation task. A corrupted or occluded mask can result in a corrupted 3D shape prior. On the other hand, a simple text prompt like "a chair" can provide a general distribution of complete shapes within the category, albeit without some instance-specific details. This approach allows the details to be constrained by the observations. Future work could explore the use of more complex text prompts and the fusion of multiple multi-modal priors to enhance the effectiveness and accuracy of prior constraints.

6.5 Computation Analysis

We conducted an evaluation of the system's computation using 10 RGB-D views on a 16GB V100 GPU. For each instance, GOM (Ours) requires 43.0 seconds for 200 optimization iterations, which includes 100 diffusion steps. In comparison, vMap [45] requires 38.1 seconds to complete 200 optimization iterations, utilizing only geometric constraints. Our method, leveraging diffusion prior, significantly enhances the quality with minimal computational overhead. The Shap-E model [29] requires 45.6 seconds to generate a single instance by diffusing from random noise via a computationally intensive sampling process. Our method, utilizing the prior information stored inside

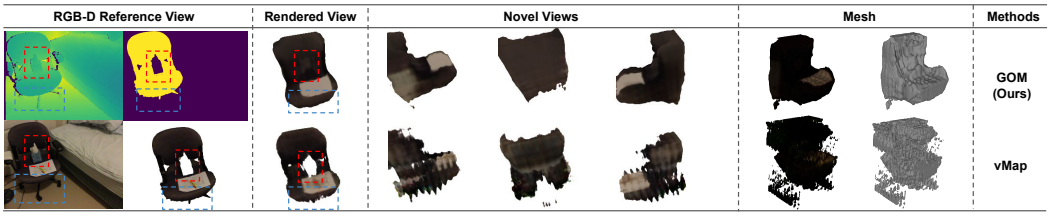

Figure 10: Failure Case: An instance where input observations are occluded and contain corrupted masks. While our method manages to complete part of the object compared to the baseline, it fails to fully complete thin elements such as legs and handles.

Shap-E, can achieve faster reconstruction than the original generation model. DSP [3] requires 32.3 seconds for 200 iterations, a speed benefiting from a smaller latent space provided by DeepSDF [27]. However, it is constrained to a single category and lacks texture information. We adopt a strategy of averaging the total number of rays sampled from multiple frames. Consequently, when more frames are available, the computation time remains nearly identical for 1, 3, and 10 views. BundleSDF [20] reconstructs an object's Signed Distance Function (SDF) and appearance field from scratch, without leveraging any prior information. It utilizes 17 keyframes from the same test scene to reconstruct the object and carries out 2500 iterations to optimize both the neural fields and the object pose. The cumulative running time is 188.6 seconds, partitioned into 118.2 seconds for pose graph optimization and 70.4 seconds for global optimization.

Depending on the specific applications, parameters such as the number of optimization steps, diffusion steps, and sampled rays can be adjusted to balance accuracy and computation. As a direction for future work, the implementation of an incremental mapping framework, as opposed to batch optimization from scratch, could further expedite online applications.

6.6 Failure Case and Limitation

The ScanNet dataset presents challenges due to occlusions and sensor noise. We illustrate a representative failure case in Figure 10. When the input observations are severely occluded or contain incomplete masks, our method can partially complete the object (for instance, the center occluded part by the tissue placed on the chair), but it fails to complete thin elements like legs and handles that the mask does not cover. The paper and pen placed on the chair are reconstructed as part of the texture. Despite these challenges, our method still generates a smoother surface with significantly fewer artifacts compared to the baselines. Future improvements could include the use of a more powerful segmentation model, such as SAM [69], and adaptively increasing the weights of the prior in areas with corrupted observations. Further, more flexible shape representation beyond a NeRF, such as Gaussian Splatting [70] can be explored to better model the details of objects.

We outline the limitations of our work to inform future research directions. Before deploying our method in real-world applications, practices such as data association and camera pose estimation are necessary for effective multi-object mapping. Our approach relies on text descriptions and segmented images from a segmentation algorithm as additional inputs. While these can be relatively easy to acquire using off-the-shelf models like Segment Anything, they do introduce extra input requirements. Additionally, the computation is not yet real-time, primarily due to the diffusion process and NeRF rendering. The challenges posed by the diffusion process could be mitigated by employing more lightweight generative models, such as those with smaller latent spaces. For the NeRF rendering, more efficient representations, like Gaussian splatting, could improve performance. Parameters such as the number of optimization steps can be adjusted to balance efficiency and effectiveness. Our current focus is on the priors of individual objects, leading us to evaluate them independently. However, more scene-level information and priors could enhance global consistency, such as cross-object relationships, structural knowledge about objects, and scene graphs. These aspects could provide valuable avenues for future work.

Figure 11: Latent Space Interpolation: visualizing the transition of Shap-E generated models from (1) a chair to another chair; (2) a chair to a table; (3) a chair to a car.

Text Prompt	Generated Shapes
A chair	
A green chair	

Figure 12: Text-Conditioned Generation: Shap-E can generate diverse shapes based on given text prompts. The application of more detailed text prompts presents an intriguing future direction for further constraining the shape and pose mapping process.

6.7 Analysis of the Generative Model Shap-E

Latent Space Interpolation. We illustrate a visualization of latent space interpolation from one chair to another, from a chair to a table, and from a chair to a plane in Figure 11. Unlike DeepSDF [27], which utilizes a 64-dimensional latent vector for an SDF-based shape, Shap-E employs a considerably larger latent space for a NeRF-based shape, with a dimension of 1024×1024. Despite its high dimensionality, linear interpolation still provides a meaningful transition for changes in both texture and geometry. A smooth latent space aids the optimization process when incorporating gradients from both observations and priors.

Generation from Text Prompt. The Shap-E model is capable of generating a variety of shapes based on a given text prompt, as demonstrated in Figure 12. The attributes specified in the text prompts, such as color, can influence the output shapes to a certain degree. The use of more complex text prompts, such as descriptions from large language models (LLMs) to assist in mapping object shapes and poses, presents an intriguing avenue for future research.

6.8 Derivation of Optimization with Prior

We provide the proof for Equation 5 in the main content. Given M observation frames $\{F_i\}_{i=1}^{M}$, and a condition C, we aim to estimate a Maximum Likelihood Estimation for the unknown variable pose $\mathbf{T}$ and shape Θ. We start from a joint distribution of $P(\mathbf{T}, \Theta | F_1, ..., F_M, C)$, and aim to get:

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} P(\mathbf{T}, \Theta | F_1, ..., F_M, C) \tag{17}$$

According to Bayes' rule:

$$P(\mathbf{T}, \Theta | F_1, ..., F_M, C) = \frac{P(F_1, ..., F_M | \mathbf{T}, \Theta, C) P(\mathbf{T}, \Theta | C)}{P(F_1, ..., F_M | C)} \tag{18}$$

Considering that any observation frames $F_1, ..., F_M$ are independent to the prior condition C, and we can assume the prior of the observation $P(F_i)$ is a constant, thus, $P(F_1, ..., F_M|C)$ is a constant. We can get:

$$P(\mathbf{T}, \Theta|F_1, ..., F_M, C) \propto P(F_1, ..., F_M|\mathbf{T}, \Theta, C)P(\mathbf{T}, \Theta|C) \tag{19}$$

Then, we consider the observation part $P(F_1, ..., F_M|\mathbf{T}, \Theta, C)$. Since the observations F_1, ..., F_M are conditionally independent among each other given $\mathbf{T}$ and Θ, and are independent to C, the likelihood can be factorized as:

$$P(F_1, ..., F_M|\mathbf{T}, \Theta, C) = \prod_i P(F_i|\mathbf{T}, \Theta) \tag{20}$$

Since we model the pose $\mathbf{T}$ and shape Θ separately, they are independent to each other. Further considering that the condition C only applies to the shape, we have:

$$P(\mathbf{T}, \Theta|C) = P(\mathbf{T}|C)P(\Theta|C) = P(\mathbf{T})P(\Theta|C) \tag{21}$$

We assume uniform distribution for the object pose $\mathbf{T}$, so that $P(\mathbf{T})$ is a constant. So we have:

$$P(\mathbf{T}, \Theta|C) \propto P(\Theta|C) \tag{22}$$

Inserting the observation part (Eq 20) and the prior part (Eq 22) into the joint distribution (Eq 19), we can estimate the unknown variables through:

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} \prod_i P(F_i|\mathbf{T}, \Theta)P(\Theta|C) \tag{23}$$

Finally, taking the logarithm, we can get a more convenient form for numerical optimization:

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} \sum logP(F_i|\mathbf{T}, \Theta) + logP(\Theta|C) \tag{24}$$