1 Appendix

1.1 More Qualitative Visualization

Visualization on More Categories. We further illustrate the effectiveness of our approach in generalizing to a variety of categories through qualitative visualizations, as shown in Figure 1. Unlike the DSP method [1], which is based on a single-category model, DeepSDF [2], our system is capable of supporting multiple object categories using a single pre-trained diffusion model, Shap-E [3].

Scene-level Visualization. Figure 2 showcases a 3D map visualization of a scene containing multiple objects from various categories, including four chairs, a sofa, and a table. Each instance is independently reconstructed using 10 RGB-D views.

1.2 Ablation Study

Methods to Fuse Observations and Diffusion Priors. We compare three strategies to fuse observations and diffusion priors, as shown in Table 1. (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 simultaneously considers both diffusion prior and geometric loss during optimization steps so that both sources of information are active. This combined 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 1. (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. [4] 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 1, *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 [3], image and text, as shown in Table 1. 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,

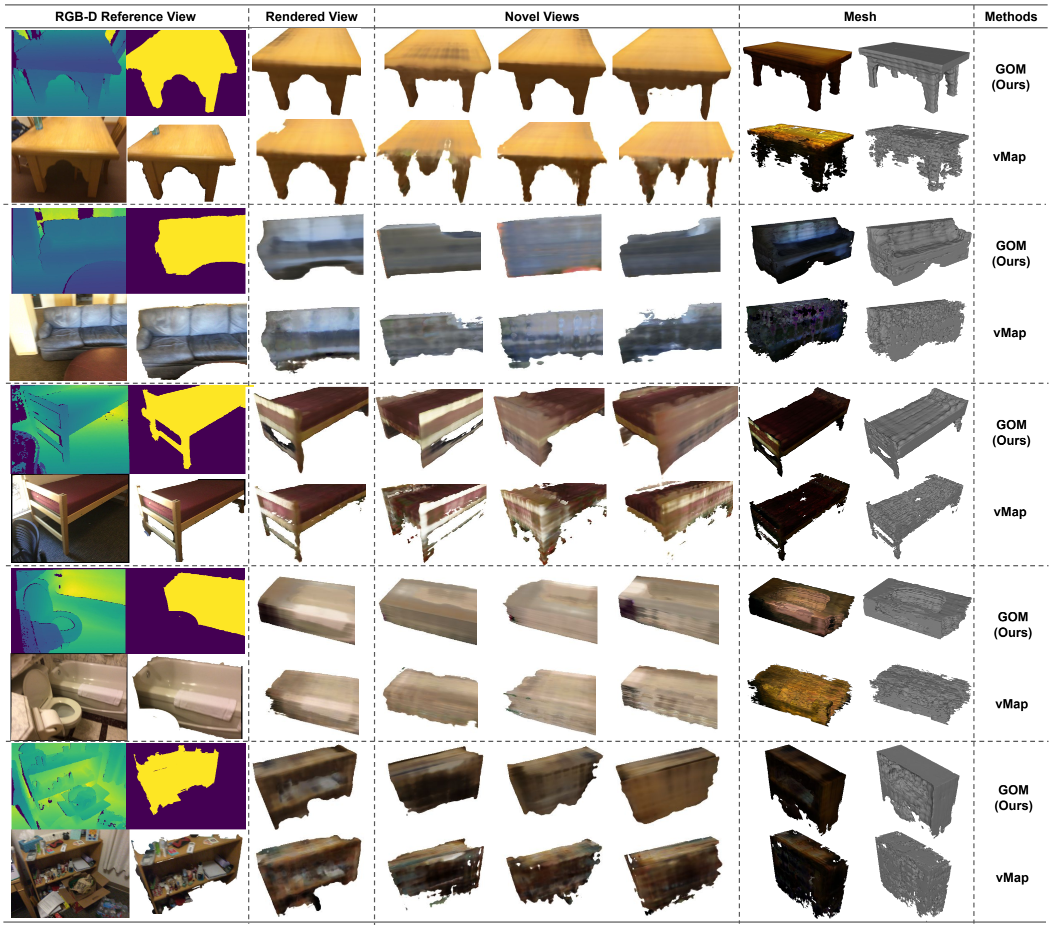


Figure 1: Effectiveness of Priors Across Multiple Categories: Leveraging priors, our method (GOM) can produce higher-quality, 3D-consistent views and generates 3D meshes with fewer artifacts compared to vMap. The results are based on 10 RGB-D views.

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

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.

1.3 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

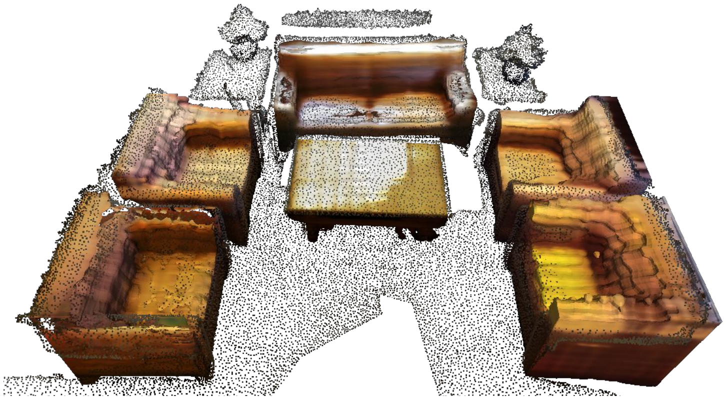


Figure 2: Scene-Level Visualization: An example of a reconstructed 3D map of a scene including four chairs, one sofa, and one table, all constrained from the same prior network.

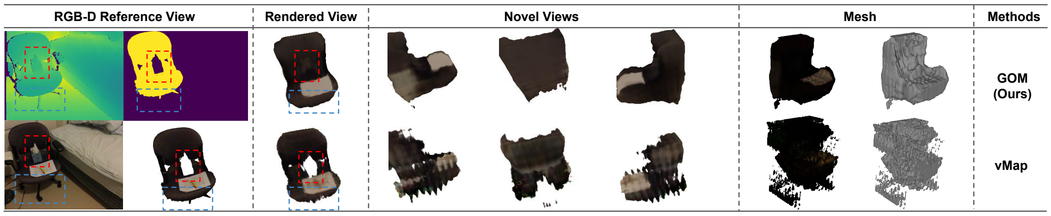


Figure 3: Failure Case: An instance where input observations are occluded and contain corrupted masks. While our method manages to complete part of the object compared to the baseline, it fails to fully complete thin elements such as legs and handles.

includes 100 diffusion steps. In comparison, vMap [5] 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 [3] 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 Shap-E, can achieve faster reconstruction than the original generation model. DSP [1] requires 32.3 seconds for 200 iterations, a speed benefiting from a smaller latent space provided by DeepSDF [2]. 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 [6] 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.

1.4 Failure Case and Discussion

The ScanNet dataset presents challenges due to occlusions and sensor noise. We illustrate a representative failure case in Figure 3. When the input observations are severely occluded or contain incomplete masks, our method can partially complete the object (for instance, the center occluded

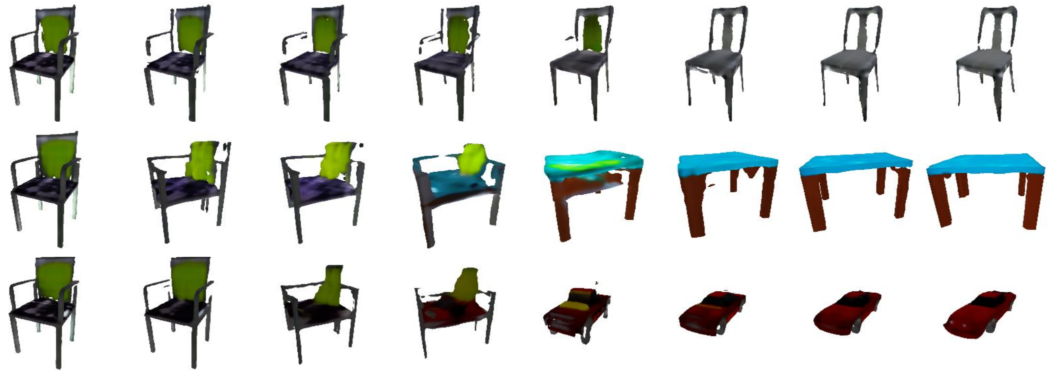


Figure 4: Latent Space Interpolation: visualizing the transition of Shap-E generated models from (1) a chair to another chair; (2) a chair to a table; (3) a chair to a car.

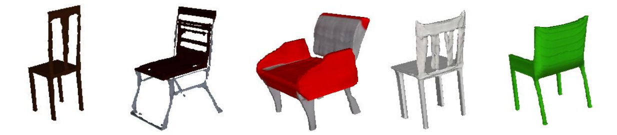
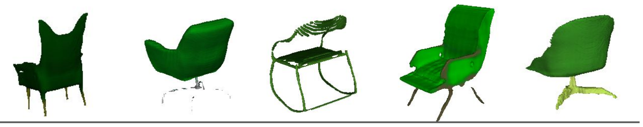
Text Prompt	Generated Shapes
A chair	
A green chair	

Figure 5: 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.

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 [7], 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 [8] can be explored to better model the details of objects.

1.5 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 4. Unlike DeepSDF [2], 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 5. 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.

1.6 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:

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} P(\mathbf{T}, \Theta | F_1, ..., F_M, C) \tag{1}$$

According to Bayes' rule:

$$P(\mathbf{T}, \Theta | F_1, ..., F_M, C) = \frac{P(F_1, ..., F_M | \mathbf{T}, \Theta, C) P(\mathbf{T}, \Theta | C)}{P(F_1, ..., F_M | C)} \tag{2}$$

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:

$$P(\mathbf{T}, \Theta | F_1, ..., F_M, C) \propto P(F_1, ..., F_M | \mathbf{T}, \Theta, C) P(\mathbf{T}, \Theta | C) \tag{3}$$

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:

$$P(F_1, ..., F_M | \mathbf{T}, \Theta, C) = \prod_i P(F_i | \mathbf{T}, \Theta) \tag{4}$$

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:

$$P(\mathbf{T}, \Theta | C) = P(\mathbf{T}|C) P(\Theta|C) = P(\mathbf{T}) P(\Theta|C) \tag{5}$$

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

$$P(\mathbf{T}, \Theta | C) \propto P(\Theta|C) \tag{6}$$

Inserting the observation part (Eq 4) and the prior part (Eq 6) into the joint distribution (Eq 3), we can estimate the unknown variables through:

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} \prod_i P(F_i | \mathbf{T}, \Theta) P(\Theta|C) \tag{7}$$

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

$$\hat{\mathbf{T}}, \hat{\Theta} = \arg\max_{\mathbf{T},\Theta} \sum log P(F_i | \mathbf{T}, \Theta) + log P(\Theta|C) \tag{8}$$

References

[1] J. Wang, M. Rünz, and L. Agapito. Dsp-slam: Object oriented slam with deep shape priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371. IEEE, 2021.

[2] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[3] H. Jun and A. Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[4] G. Yang, A. Kundu, L. J. Guibas, J. T. Barron, and B. Poole. Learning a diffusion prior for nerfs. *arXiv preprint arXiv:2304.14473*, 2023.

[5] X. Kong, S. Liu, M. Taher, and A. J. Davison. vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 952–961, 2023.

[6] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.