Object Reconstruction under Occlusion with Generative Priors and Contact-induced Constraints

Minghan Zhu*[1,2], Zhiyi Wang*[2], Qihang Sun[2], Maani Ghaffari[2], and Michael Posa[1]

***Abstract*— Object geometry is key information for robot manipulation. Yet, object reconstruction is a challenging task because camera observations are partial due to occlusions. The scene may not offer the flexibility for a robot to alter its viewpoint to obtain a full observation of the object of interest. In this paper, we leverage two extra sources of information to reduce the ambiguity of vision signals under occlusion. First, generative models learn priors of the shapes of commonly seen objects, allowing us to make reasonable guesses of the unseen part of geometry. Second, contact information, which can be obtained from videos and physical interactions, provides sparse constraints on the boundary of the geometry. We combine the two sources of information through contact-guided 3D generation. The guidance formulation is inspired by drag-based generative image editing. We explore different guidance strategies and highlight the importance of short gradient paths for guided generation. Experiments on synthetic and real-world data show that our approach improves the object reconstruction compared to pure 3D generation and contact-based optimization methods. Project page: https://contactgen3d.github.io/**

I. INTRODUCTION

While vision is a primary perception modality, it fundamentally provides only partial scene information, making environment state estimation a classic ill-posed problem for robotics. In object manipulation, for instance, recovering essential geometry and pose is often hindered by self-occlusion and occlusions by other objects and the robot itself. Furthermore, changing viewpoints to resolve these occlusions is frequently infeasible in dense environments where workspace and visibility are strictly limited. This long-standing challenge steered some robotic research efforts to end-to-end approaches [1, 2], where policies and actions are generated directly from raw observations, but they still suffer from limited generalizability and interpretability.

The problem of object reconstruction from partial observations may be better solved if we go beyond the raw pixels and leverage other sources of information. We can obtain knowledge about the geometry of an object from its contacts with the environment. Contacts can only happen on the boundary of the geometry, thus imposing constraints on the object shape. Consider, for example, a robot exploring objects in a cluttered kitchen cabinet. Items in the back are heavily occluded from all viewpoints, but as the robot pushes and nudges within the cabinet, contacts provide critical information. On the other hand, by learning priors of object geometry from large datasets, generative models can also build a full object model from a partially occluded image. Interestingly, the information from contact-induced constraints and data-driven shape priors is usually complementary, as the "inpainting" of unobserved geometry by the generative model often has multiple solutions, and the sparse contact constraints reduce the ambiguity.

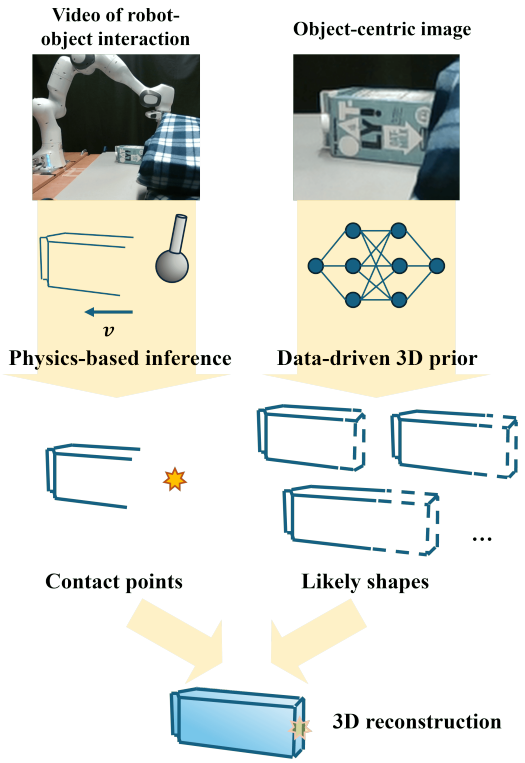

Fig. 1. This work develops a novel framework for 3D object reconstruction under occlusion by integrating data-driven 3D priors and physics-based contact information through guided generation. The two perspectives bring complementary insights that lead to high-quality and accurate 3D reconstruction.

In this paper, we tackle the problem of visual object reconstruction under occlusion by combining data-driven 3D priors and contact-induced constraints. A high-level overview is depicted in Figure 1. We use the paradigm of image-conditioned object generation to harness the powerful generative prior of object shapes. Specifically, we use a flow-matching 3D generative model dedicatedly designed to work with occluded 2D images, Amodal3R [3], as our base model for 3D generation. On the other hand, we model contact information as a sparse set of contact points. In the real-world experiments, we employ Vysics [4] to estimate contact

*These authors contributed equally to this work.

[1]University of Pennsylvania, Philadelphia, PA 19104 {minghz, posa}@seas.upenn.edu

[2]University of Michigan, Ann Arbor, MI 48109 {minghanz, jerrywzy, qhsun, maanigj}@umich.edu

points from very brief (about 10 seconds long) videos and physical interactions.

We formulate the fusion as 3D generation with training-free contact guidance, where a contact-based energy function is formulated based on the contact points and used to guide the inference of the flow-matching generative model. Specifically, we are inspired by drag-based editing in 2D image generation, and propose a drag-based 3D energy function for contact guidance. In this way, we can build the object geometry model from heavily occluded visual observations with high quality and good metric accuracy.

We summarize the core contributions of this work as follows:

- We propose an object reconstruction framework that improve the accuracy of generative 3D priors with sparse physical cues when vision is partially occluded.
- We develop a guidance method for native 3D generation using a drag-based energy function.
- Our method achieves superior object reconstruction in synthetic and real-world data, highlighting the importance of synthesizing data-driven and physics-driven insights.

II. RELATED WORK

A. 3D Object Reconstruction

There are different paradigms that can be applied on the problem of object reconstruction. Object SLAM [5, 6] approaches rely on low-level tracking and shape-pose optimization, achieving good efficiency and generalizability, but lack robustness against visual signal degradation. In this paper, we mainly discuss object reconstruction by 3D generation, in which data-driven priors are used to build the complete geometry (and sometimes appearance) model of an object from one or a few images.

The development of 3D generative methods is largely inspired and supported by the progress in 2D image generative models. The denoising steps in 2D diffusion [7] can be interleaved with the regression [8, 9] or optimization [10] of 3D representations to reconstruct the 3D target. Moreover, multi-view image generation from a single image input [11, 12] provides strong 3D priors. Many works connect multi-view image generation with multi-view 3D reconstruction for 3D generation from a single image [13–15], or repurpose multi-view image generation capability for pixel-aligned 3D representation generation [16]. On the other hand, 3D generative priors can also be learned natively in 3D without relying on image generative priors. Earlier attempts [17, 18] are limited by the relatively small size of 3D datasets [19]. With larger datasets of 3D objects [20] being released, 3D generation models become much more powerful. Some [21–23] simply use implicit learnable tokens as the latent features for 3D generation, while others [24–26] use explicit 3D primitives in the latent space for better interpretability and controllability. We used a model based on explicit latent to formulate the contact-induced energy function more easily.

B. Occlusion-aware 3D Reconstruction

Despite the significant progress in 3D generation, most such models are trained to take an unoccluded image of the target as input, and their performance degrades significantly when occlusion is present. To overcome this limitation, recent research has explored three main directions.

1) Completing occluded areas in 2D: In this paradigm, the occluded single-view image is first processed by a 2D amodal segmentation and completion model [27, 28] or an image inpainting model [29, 30] to recover the missing regions in the image space. The completed image is then provided to a 3D generative model for reconstruction. A representative example is SceneComplete [31], which integrates BrushNet [32] with InstantMesh [33]. CAST [34] employs a Masked Auto Encoder (MAE) [35] to complete the occluded area in the image latent space. However, the effectiveness of this pipeline is constrained by the quality of the 2D amodal completion or inpainting. It is also less straightforward to inject 3D guidance into the 2D generation process, which is important for our work.

2) Shape completion from partial point clouds: It is possible to use partial point clouds as the input observation to 3D generative models for shape completion [17, 18, 22, 36] . However, a lot of them require the input points to reside in the canonical object frame, which is usually unavailable in real-world use cases. They also typically require extra image, text, or category condition, as a partial point cloud alone may not carry enough information to recover the whole object. Another paradigm is point cloud completion [37–39], some of which do not require canonical-frame input [40, 41], but they are typically trained on a smaller dataset [19] than the 3D generative models, limiting their performance and generalizability. Therefore, we use images as the input for our 3D reconstruction task.

3) Fine-tuning 3D generative models for occlusion: Amodal3R [3] fine-tunes a native 3D generation framework, TRELLIS [24], with occlusion and visibility masks as additional input modalities, enabling robust handling of occlusions in single-view reconstruction.

C. Adding Controllability to Pretrained Generative Models

Here, we discuss three strategies for adding controllability to pretrained generative models. First, fine-tuning. Technologies like ControlNet [42] and adapters [43] allow fine-tuning a pretrained model to take in new modalities, usually spatial control signals such as edge maps and poses, without destroying the original representations. Second, training-free guidance, which does not require training the generative model with the condition. The guidance signal is formulated as an energy function to represent the compatibility between the generated states and the task-specific needs. In diffusion or flow-matching models, the goal is to find a delta step in each denoising iteration to steer the output to satisfy the task-specific needs. There has been extensive literature on diffusion guidance [44–46] and recently on flow-matching [47] as well. Third, we can also treat the controllable generation problem as an editing problem, where

uncontrolled generation is first performed, and then we apply generative editing [48, 49] to make the outcome follow our specifications. In this paper, we incorporate the idea of editing in building a training-free guidance for contact-conditioned 3D generation.

D. Physics-based Reasoning in 3D Reconstruction

As discussed in the introduction, physics-based reasoning helps reduce the ambiguity in visual signals and build a physically plausible reconstruction. Many efforts assume the objects to be stable and static and use optimization to prevent penetration and unstable falling [50–52]. In [4, 53, 54], contact is analyzed in dynamic scenes to improve 3D reconstruction. Estimated physics properties may also be used to generate physically realistic videos [55, 56].

III. PRELIMINARIES

A. 3D Generation: TRELLIS and Amodal3R

Amodal3R [3] is a 3D generation model taking occluded images and masks as input. It is fine-tuned on TRELLIS [24], a flow-matching model capable of image-to-3D and text-to-3D generation. Our framework uses Amodal3R as the base model to provide 3D priors. Thus we will first introduce TRELLIS and Amodal3R briefly.

TRELLIS contains two stages for 3D generation: sparse structure generation and structured latent generation, which are in charge of the generation of the overall geometry and the local details and appearance, respectively. The sparse structure is an occupancy voxel grid of resolution 64, $s \in \mathbb{R}^{64^3}$, representing the geometry in a unit cube. The structured latent is feature vectors attached to each occupied voxel, which can be decoded into 3D Gaussians or FlexiCube [57] representations of resolution 256^3 for mesh generation. The first stage output resolution is high enough for meaningful contact-induced guidance, thus we mainly focus on the first-stage guided generation in this paper, and keep the second stage as is. In the first stage, the latent space for sparse structure generation is a dense 3D voxel feature map with downsampled spatial resolution, $x \in \mathbb{R}^{16^3 \times 8}$ (where 16 is the spatial resolution and 8 is the feature dimension), which can be mapped from and to the space of s through a UNet voxel encoder $\mathcal{E}$ and decoder $\mathcal{D}$, respectively. The generation is through flow matching with Diffusion Transformers $\mathcal{T}$ (DiT) [58]. Amodal3R incorporates a visibility mask, which highlights the visible part of the object of interest, and an occlusion mask highlighting the occluder, as extra input. Please refer to the original papers for more details.

B. Vysics: Contact Estimation from Video

In our real-world experiments, contact information comes from Vysics [4], an optimization-based object tracking and 3D reconstruction framework that also infers contact points and associated contact force from a video of an object interacting with a robot and the environment, and uses the estimated contacts to assist in building the object model. The set of contact points serves as a sparse representation of regions of high contact force on the surface of the object during the interactions, which could include sliding point-to-surface contacts and surface-to-surface contacts (e.g., a table surface supporting the object). In other words, one should not interpret the number of contact points as the number of required isolated pokes on the object. The contact points can be obtained through more natural and brief interactions. In this paper, we use the estimated contact points from Vysics as the guiding signal for the 3D generation under occlusion.

C. Drag-based Generative Editing

The operation "dragging" originates from 2D image editing, where the goal is to move a point in the image to a different coordinate, while keeping the overall image semantically consistent and spatially reasonable. After the pioneering work DragGAN [59], the drag-based editing is also realized in diffusion models [48, 60]. On a high level, the drag-based editing in diffusion models can be done by adding noise to an image (as the reference) and denoising. In the denoising process, a drag-based loss encourages the features around the target point to resemble the reference features around the source point, and a content-preservation loss encourages features everywhere else to be consistent with the reference. We will build our guidance in a similar drag-based formulation.

IV. METHOD

A. Problem Formulation

We formulate the fusion of data-driven 3D priors and sparse contact point data for object reconstruction as a guided flow-matching problem. Flow-matching models define a vector field v_t that establishes a probability path $p_t(x_t)$ from an initial base distribution p_1 (noise) to a target distribution p_0 (3D objects). By drawing a sample $x_1 \sim p_1(x_1)$ and integrating the ordinary differential equation (ODE) $dx_t = v_t(x_t)dt$, we generate samples x_0 that conform to the target distribution $p_0(x_0)$. The flow-matching model learns the vector field $v_t(x_t) = \mathcal{T}(x_t; t)$.

We extend this to a guided setting where the goal is to steer the probability path toward a conditional distribution $p_t(x_t|c)$, where c represents the condition, which is the specified contact points in this problem. This is achieved by determining a guidance vector field $g_t(x_t)$ such that the updated vector field, $v'_t(x_t) = v_t(x_t) + g_t(x_t)$, directs the integration from p_1 toward a state compatible with the constraints. Following the principle of training-free energy-based guidance [45], we define an energy function $J(x_t; c)$ that represents the local discrepancy between the current state and the contact constraints. The value of J should be small when x_t is more compatible with the contact condition c, and large otherwise. The guidance vector is then computed as the gradient of this energy with respect to the current state:

$$g_t(x_t) = -\lambda_t \nabla_{x_t} J(x_t; c), \tag{1}$$

where λ_t is the step length, typically a hyperparameter. J can be instantiated in various ways. Considering that the x_t is a

noisy latent state with no clear geometric meaning, it may be tempting to map x_t to the noise-free occupancy s_0 to measure the compatibility with given contact points. However, the drag-based guidance formulation, which converts the point condition into a feature-similarity loss, offers more flexibility in designing the energy function J, as detailed in Sec. IV-B.

B. Drag-based Contact Guidance

The objective of contact guidance is to steer the generative process such that the resulting 3D geometry satisfies sparse physical constraints while remaining on the manifold of plausible shapes. We define the guidance energy J as the local discrepancy between the generated representation around a contact point $p^c \in c$ and a reference representation around some source location to "drag" from. To obtain this source, we first perform an unguided pass of the flow-matching model to generate a reference shape s_0^*, which should be a plausible shape given the image but unaware of the contact condition. Then we identify the nearest surface point p^s on the reference shape as the source point for each contact point target p^c.

Given the source and target point pairs and the goal of matching the two neighborhoods, we now need to decide what feature space to measure the similarity and instantiate the guidance. We explore three options, representing different trade-offs between spatial resolution, numerical stability, and computational complexity. Fig. 2 illustrates the three guidance options.

1) Voxel-Space Guidance: The most direct approach is to measure compatibility in the decoded occupancy space. The loss is computed as:

$$J_{\text{vox}} = \frac{1}{|\mathcal{N}_\epsilon|} \sum_{\Delta p \in \mathcal{N}_\epsilon} |\hat{s}_0(p^c + \Delta p) - s_0^*(p^s + \Delta p)|^2, \quad (2)$$

where $\mathcal{N}_\epsilon$ is a 3D neighborhood area, and $\hat{s}_0 = \mathcal{D}(\hat{x}_0) = \mathcal{D}(x_t - v_t t)$ is the decoded noise-free occupancy estimated from the current noisy latent state x_t.[1] While this offers high spatial resolution, the gradient path of backpropagation ($\hat{s}_0 \to \mathcal{D} \to \hat{x}_0 \to v_t \to \text{DiT} \to x_t$, see Fig. 2) is deep and prone to accumulating gradient noise.

2) Latent-Space Guidance with Sub-coarse-grid Shifting: To bypass the backpropagation through the decoder, one can guide the denoised latents $\hat{x}_0$ directly. However, the latent grid is coarser (16^3), meaning a standard drag would only be precise to integer coarse-grid units. To achieve "sub-coarse-grid" precision, we shift the reference occupancy s_0^* by integer units in the 64^3 fine grid such that the offset between the source and target points is of integer units in the coarse latent space. This shifted occupancy is re-encoded to a target latent x_0^{**}, and we apply a cosine-based drag loss:

$$J_{\text{lat}} = \frac{1}{|\mathcal{N}_\epsilon|} \sum_{\Delta p \in \mathcal{N}_\epsilon} D\left(\hat{x}_0(p^c + \Delta p), x_0^{**}(p^s + \Delta p)\right), \quad (3)$$

[1]Here we assume affine conditional probability path for the flow model, which is true for TRELLIS.

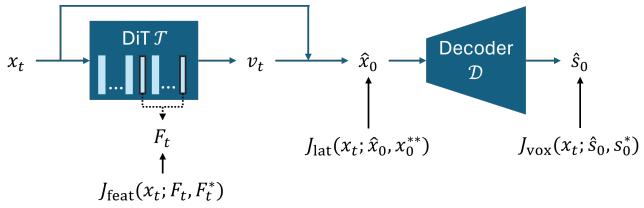

Fig. 2. The three guidance options visualized in one denoising step.

where $D(\cdot)$ is the inverse normalized cosine similarity:

$$D(x, y) = \frac{1}{0.5 \cdot \frac{x \cdot y}{\|x\|\|y\|} + 0.5}.$$

While this provides contact-point following of the same spatial-resolution as the voxel-space guidance, the backpropagation from $\hat{x}_0$ to x_t still introduces artifacts, and the re-encoding needs extra computation.

3) Feature-Space Guidance: Our recommended approach applies guidance directly to the intermediate feature layers F_t of the DiT $\mathcal{T}$ that takes x_t as input. We target the middle and final blocks of the DiT to capture a balance of semantic and geometric structure. This is similar to the practice in DragonDiffusion [48], which applies the drag-based image editing loss on the intermediate layers of the denoising UNet. Our guidance energy is defined as:

$$J_{\text{feat}} = \frac{1}{|\mathcal{N}_\epsilon|} \sum_{\Delta p \in \mathcal{N}_\epsilon} D\left(F_t(p^c + \Delta p), F_t^*(p^s + \Delta p)\right). \quad (4)$$

By guiding the features internally, we shorten the gradient path and avoid the instabilities associated with noise-free state prediction. Although limited to the 16^3 resolution, this manifold provides the highest stability and best overall 3D reconstruction accuracy, as shown later in Sec. V-E.

V. EXPERIMENTS

A. Shared setup for both experiments

The goal of our experiments is to demonstrate that sparse physical cues can significantly refine a generative 3D prior when visual observations are heavily occluded. We evaluate this on a large-scale synthetic dataset and a real-world robotic interaction dataset. In both settings, the model receives a single 2D image of the target object with partial occlusion as visual input for the 3D generation. In the synthetic experiment, the points are sampled from the ground truth object surface. In the real-world experiment, the points are estimated by Vysics [4].

We use the same set of hyperparameters for the experiments on synthetic data and real data. The flow-matching model takes 12 timesteps in total for inference. The 3D neighborhood $\mathcal{N}_\epsilon$ for drag-based contact guidance is set to a 21^3 cube in the fine voxel grid and 5^3 cube in the coarse feature grid, which roughly correspond to the same neighborhood area.

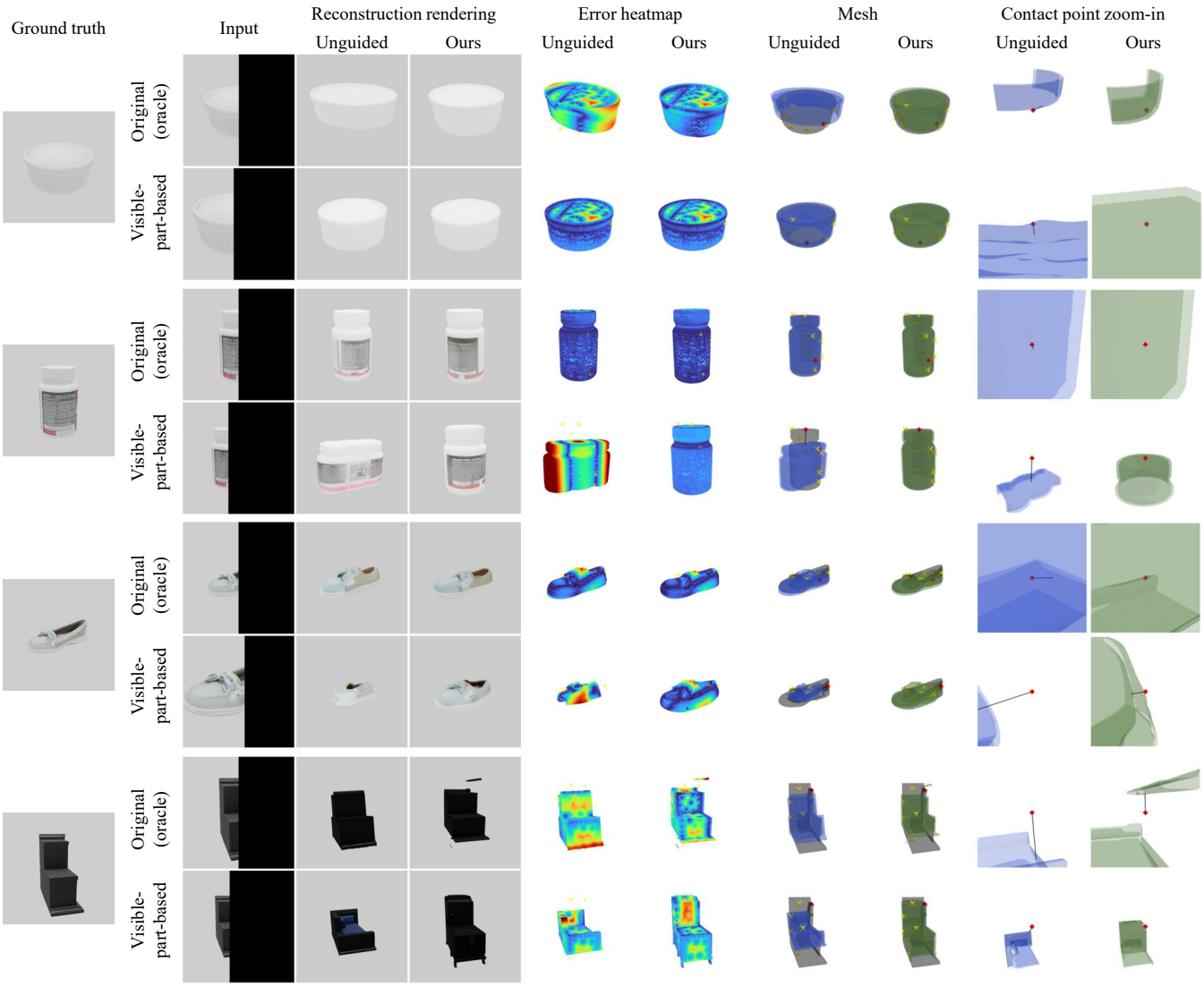

Fig. 3. Qualitative comparison of the geometry reconstruction. For each object, we show the reconstruction under two input cropping conditions (see Sec. V-C). Occlusion masks are shown in black. The error heatmaps depict the one-sided point-wise Chamfer distance from the predicted shape to the ground truth. The mesh view shows predictions in blue (unguided) and green (guided) overlaid on the ground truth mesh in gray. The contact points are shown in yellow crossings in both error heatmaps and mesh views, with one point in red zoomed-in in the right-most two columns with a black line connecting the contact point to the nearest point on the predicted surface.

Input condition	Guidance method	Chamfer Distance ↓		F@0.01 ↑		F@0.02 ↑		F@0.05 ↑		Contact Distance ↓		RGB-LPIPS ↓	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Original (oracle)	Unguided [3]	0.0390	0.0329	0.380	0.350	0.658	0.677	0.910	0.968	0.0169	0.0108	**0.154**	**0.135**
	Voxel (J_{vox})	0.0414	0.0380	0.360	0.323	0.629	0.624	0.901	0.955	0.0165	0.0112	0.303	0.302
	Latent (J_{lat})	0.0411	0.0381	0.354	0.306	0.626	0.599	0.904	0.948	**0.0149**	**0.0100**	0.175	0.163
	Feature (J_{feat}, *ours*)	**0.0375**	**0.0309**	**0.384**	**0.379**	**0.665**	**0.693**	**0.918**	**0.976**	**0.0149**	0.0114	0.156	0.137
Visible-part-based	Unguided [3]	0.0475	0.0429	0.330	0.288	0.585	0.563	0.868	0.923	0.0213	0.0152	**0.161**	0.145
	Voxel (J_{vox})	0.0473	0.0415	0.327	0.294	0.583	0.564	0.868	0.935	0.0207	0.0139	0.303	0.302
	Latent (J_{lat})	0.0467	0.0420	0.326	0.280	0.584	0.563	0.871	0.916	0.0185	0.0135	0.175	0.163
	Feature (J_{feat}, *ours*)	**0.0430**	**0.0374**	**0.345**	**0.302**	**0.613**	**0.608**	**0.892**	**0.960**	**0.0179**	**0.0127**	0.162	**0.139**

TABLE I

QUANTITATIVE EVALUATION OF 3D RECONSTRUCTION UNDER OCCLUSION ON THE GSO [61] DATASET, WITH TWO INPUT CONDITIONS. ↓ MEANS SMALLER IS BETTER, ↑ MEANS LARGER IS BETTER. THE BEST IS HIGHLIGHTED IN BOLD FONT. THE OBJECTS ARE NORMALIZED IN A UNIT CUBE.

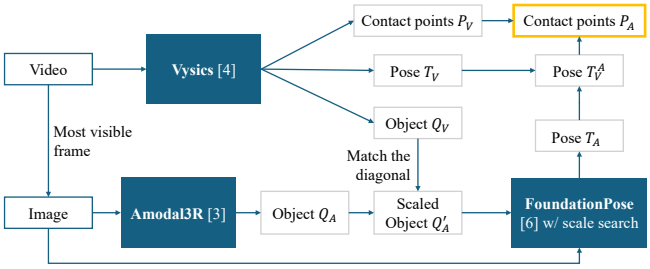

Fig. 4. The process of converting contact points estimated by Vysics to the prediction space of Amodal3R for contact-guidance in real-world experiments. T_V, T_A are the poses of objects Q_V, Q_A in the camera frame. T_V^A converts between the two object frames.

B. Implementation details on guidance

We empirically found that the voxel-space and latent-space guidance is more numerically sensitive and requires careful scheduling of the guidance step size λ_t. We split the 12 denoising steps into early, middle, and late stages, each with 4 timesteps. λ_t is set to 0.2, 1.0, and 0.5, for the early, middle, and late stages, respectively. And we run the guidance recurrently [44] three times at each timestep. For the feature-space guidance, we simply use $\lambda_t = 1$ and do not apply recurrent guidance.

C. Experiment on synthetic data

The synthetic experiment utilizes the Google Scanned Objects (GSO) [61] dataset, a collection of 3D-scanned household items. We randomly select 200 objects and render them with artificial occlusions by masking out half of the image. We generate synthetic contact points by randomly sampling 10 points from the unobserved parts of the object surface. Since TRELLIS [24] is trained on object-centered renderings, under heavy occlusion Amodal3R [3] tends to assume the full shape is centered too. Our masked GSO renderings are themselves object-centered, which lets the unguided baseline guess the occluded half correctly from this assumption alone, an advantage unavailable in practice, where the full extent is unknown beforehand. To control for this, we evaluate two input conditions on GSO: an *Original (oracle)* condition using the ground-truth object placement and scale, and a *Visible-part-based* condition, where the object is re-scaled and re-centered using only the visible half heuristically, mirroring what is achievable under real-world occlusion. Concretely, the crop scales the visible mask's maximal side length to a fixed ratio of the image, and centers it with a left offset, shifted as needed to keep the whole mask inside the frame. This heuristic not intended to be a research contribution, but a minimal engineering effort to crop reasonably without knowledge of the full shape.

D. Experiment on real-world data

Real-world validation uses the Vysics [4] dataset featuring a Franka robot interacting with tabletop objects through pushing, pivoting, and toppling. The data includes RGBD videos with partially occluded objects, the corresponding masks, the robot trajectories, and the ground truth object mesh for evaluation. There are multiple videos of interactions, each approximately 10 seconds long, for every object, and the inference is on each video independently. Contact points are derived from each video of interaction, while the generative model uses a single image frame from the video when the area of the visible object mask is the largest. We also crop the image around the visible mask using the same heuristic as in the synthetic experiment (Sec. V-C) to make the visible part of the object roughly centered and occupy the major part of the image. We do not have privileged information about the hidden part of the object when cropping the images.

In order to convert the contact points to Amodal3R's prediction frame, we first run the unguided Amodal3R to obtain an initial reconstructed mesh. The mesh scale is initialized by matching the diagonal length of the best-fitting oriented bounding box between Amodal3R's predicted mesh and Vysic's reconstructed mesh using Open3D [62]. Then we align the mesh to the RGBD image using FoundationPose [6] with scale refinement. Vysics also yields the object's tracked pose in each RGBD frame. Therefore, the contact points in Vysic's object frame can be converted to Amodal3R's object frame. We then run a second run of Amodal3R generation with the guidance of contact points. Vysics typically yields hundreds of contact points that could be redundant for explaining the contact dynamics. We subsample 20 points using farthest point sampling for Amodal3R's guidance. An overview of this process is depicted in Figure 4.

E. Results on synthetic data

Table I shows the quantitative results of 3D reconstruction under occlusion on the GSO dataset [61]. In terms of the 3D evaluation, we measure the Chamfer distance and the F-score of the points on the predicted shape at different distance thresholds. We also report contact distance as the distance from the specified contact points to the reconstructed mesh to show the adherence of the generation results to the contact condition. Furthermore, we report the LPIPS score [63] of the rendered RGB images, in order to evaluate the quality of the generated objects. Comparing the two input conditions, all methods achieve better 3D accuracy under the *Original (oracle)* condition than under the *Visible-part-based* condition, confirming that the position/scale prior leaks shape information when the object is rendered at its ground-truth placement. Among the three guidance strategies, feature-space guidance is consistently the most accurate, and it is the only strategy that outperforms the unguided baseline in both input conditions. Under the oracle condition, voxel- and latent-space guidance underperform the unguided baseline while the feature guidance slightly outperforms, since the baseline already benefits from the centering shortcut, leaving guidance little room to help. In the *Visible-part-based* condition, however, all three guidance strategies show a consistent improvement in Chamfer distance over the unguided baseline, showing that guidance refines the shape using the contact evidence itself, rather than relying on the centering prior. On the other hand, the unguided baseline

Categories	bakingbox	bottle	egg	oatly	milk	styrofoam	toblerone	all
BundleSDF [5]	3.84	2.65	3.70	2.45	3.17	2.55	2.44	2.98
Vysics [4]	1.83	1.36	1.05	1.25	1.53	1.45	1.02	1.45
Amodal3R [3]	1.12	**1.32**	0.85	1.13	**1.42**	1.26	0.89	1.18
Ours	**1.07**	1.54	**0.82**	**0.77**	1.60	**1.13**	**0.83**	**1.11**

TABLE II

CHAMFER DISTANCE OF THE MESHES RECONSTRUCTED UNDER OCCLUSION ON THE VYSICS [4] REAL-WORLD DATASET. UNIT: CM.

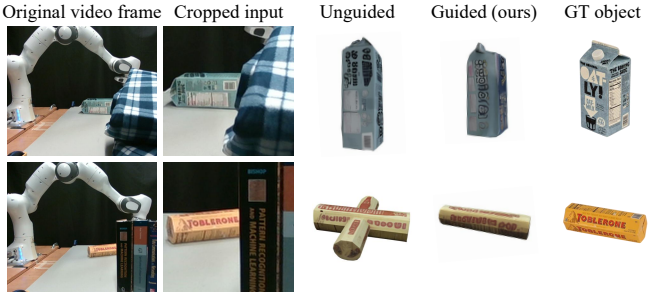

Fig. 5. Qualitative examples of 3D reconstruction on Vysics data. The top row is a successful case, and the bottom row is a failure case.

still tends to achieve better or comparable LPIPS scores than the guided results across both conditions, suggesting that our guidance can introduce visual artifacts, a limitation also visible for the last object in Fig. 3.

Fig. 3 shows qualitative examples of the 3D reconstruction on the GSO dataset under both input conditions, comparing the unguided and feature-guided predictions. The error heatmaps show that the guided reconstruction has a smaller error to the ground truth shape. The zoomed-in mesh views further show that the predicted surface moves closer to the specified contact points after guidance, illustrating the local geometric improvement induced by the contact guidance.

F. Results on real-world data

Table II shows the quantitative results of the 3D reconstruction under occlusion in the real-world data provided by Vysics [4]. The comparisons include BundleSDF [5], representing pure vision-based reconstruction; Vysics [4], representing joint optimization of vision and physics-based reasoning; Amodal3R [3], representing visual reconstruction with generative prior knowledge; and ours, which combines vision, generative priors, and physics-based insights of the contacts. We use feature-space guidance in this experiment. Our method achieved the overall best geometric accuracy in terms of Chamfer distance. See Fig. 5 for example cases. The guidance helps improve the aspect ratio of the reconstructed objects when a vision model alone cannot infer correctly under heavy occlusion. The guidance also alleviates the multi-head Janus problem [10, 64], which could be attributed to the asymmetry introduced by the contact points.

VI. CONCLUSIONS

We explored the integration of data-driven 3D shape priors and physics-driven contact information to tackle the challenging task of 3D object reconstruction under heavy occlusion. This approach highlights contact as a novel and exploitable modality that complements vision, harnessing the physical constraints of the environment to enable robots to build accurate object models for manipulation in cluttered real-world settings. Under the paradigm of guided flow-matching generation, we propose a drag-based contact loss function to guide the dense geometry with sparse contact points. Experimental results on synthetic data and real-world data both show that our framework improves the reconstruction accuracy by incorporating both sources of knowledge. We hope this work could inspire more explorations of joining the merits of big data and physics-based insights for building more powerful robot vision systems.

Our work still has limitations. The guided generation gets close to, but may not exactly land on, the specified contact points, and can introduce artifacts. The drag-based guidance loss assumes that the unguided results have a roughly correct structure, which may not always be true in practice. We will explore solutions to these problems in future work.

ACKNOWLEDGMENT

This work was supported by an NSF CAREER Award under Grant No. FRR-2238480 and the RAI Institute. M. Ghaffari was supported by AFOSR MURI FA9550-23-1-0400 and AFOSR YIP FA9550-25-1-0224.

REFERENCES

[1] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02 783 649 241 273 668, 2023.

[2] A. Brohan et al., "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[3] T. Wu et al., "Amodal3R: Amodal 3D Reconstruction from Occluded 2D Images," *arXiv preprint arXiv:2503.13439*, 2025.

[4] B. Bianchini et al., "Vysics: Object reconstruction under occlusion by fusing vision and contact-rich physics," in *Robotics: Science and Systems (RSS)*, 2025.

[5] B. Wen et al., "Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 606–617.

[6] B. Wen et al., "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.

[7] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[8] S. Szymanowicz et al., "Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8863–8873.

[9] Y. Xu et al., "Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] B. Poole et al., "Dreamfusion: Text-to-3d using 2d diffusion," in *The Eleventh International Conference on Learning Representations*, 2023.

[11] R. Liu et al., "Zero-1-to-3: Zero-shot One Image to 3D Object," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 9298–9309.

[12] P. Wang et al., "Imagedream: Image-prompt multi-view diffusion for 3d generation," *arXiv preprint arXiv:2312.02201*, 2023.

[13] M. Liu et al., "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 226–22 246, 2023.

[14] J. Li et al., "Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model," in *The Twelfth International Conference on Learning Representations*, 2024.

[15] X. Long et al., "Wonder3d: Single image to 3d using cross-domain diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 9970–9980.

[16] C. Lin et al., "DiffSplat: Repurposing Image Diffusion Models for Scalable Gaussian Splat Generation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[17] P. Mittal et al., "AutoSDF: Shape Priors for 3D Completion, Reconstruction and Generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 306–315.

[18] Y.-C. Cheng et al., "SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 4456–4465.

[19] A. X. Chang et al., "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[20] M. Deitke et al., "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 142–13 153.

[21] L. Zhang et al., "Clay: A controllable large-scale generative model for creating high-quality 3d assets," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–20, 2024.

[22] B. Zhang et al., "3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models," *ACM Transactions On Graphics (TOG)*, vol. 42, no. 4, pp. 1–16, 2023.

[23] Z. Zhao et al., *Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation*, 2025.

[24] J. Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 469–21 480.

[25] Z. Chen et al., "3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 576–26 586.

[26] L. Yushi et al., "Gaussiananything: Interactive point cloud flow matching for 3d object generation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[27] E. Ozguroglu et al., "Pix2gestalt: Amodal Segmentation by Synthesizing Wholes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 3931–3940.

[28] K. Chen et al., "Using Diffusion Priors for Video Amodal Segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 890–22 900.

[29] S. Xie et al., "Smartbrush: Text and shape guided object inpainting with diffusion model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 428–22 437.

[30] A. Lugmayr et al., "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[31] A. Agarwal et al., "SceneComplete: Open-World 3D Scene Completion in Complex Real World Environments for Robot Manipulation," *arXiv preprint arXiv:2410.23643*, 2024.

[32] X. Ju et al., "BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion," in *European Conference on Computer Vision*, Springer, 2024, pp. 150–168.

[33] J. Xu et al., "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024.

[34] K. Yao et al., "Cast: Component-aligned 3d scene reconstruction from an rgb image," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 4, pp. 1–19, 2025.

[35] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, 2024.

[36] G. Chou et al., "Diffusion-SDF: Conditional Generative Modeling of Signed Distance Functions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2262–2272.

[37] W. Yuan et al., "Pcn: Point completion network," in *2018 international conference on 3D vision (3DV)*, IEEE, 2018, pp. 728–737.

[38] X. Yu et al., "Pointr: Diverse point cloud completion with geometry-aware transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 498–12 507.

[39] Z. Chen et al., "Anchorformer: Point cloud completion from discriminative nodes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 581–13 590.

[40] M. Humt et al., "Shape completion with prediction of uncertain regions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 1215–1221.

[41] Y. Wang et al., "Learning generalizable shape completion with sim (3) equivariance," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[42] L. Zhang et al., "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[43] C. Mou et al., "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 2024, pp. 4296–4304.

[44] A. Bansal et al., "Universal guidance for diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 843–852.

[45] J. Yu et al., "Freedom: Training-free energy-guided conditional diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 174–23 184.

[46] H. Chung et al., "Diffusion posterior sampling for general noisy inverse problems," *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

[47] R. Feng et al., "On the guidance of flow matching," *Forty-second International Conference on Machine Learning*, 2025.

[48] C. Mou et al., "Dragondiffusion: Enabling drag-style manipulation on diffusion models," *arXiv preprint arXiv:2307.02421*, 2023.

[49] Y. Shi et al., "Dragdiffusion: Harnessing diffusion models for interactive point-based image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8839–8849.

[50] W. Agnew et al., "Amodal 3d reconstruction for robotic manipulation via stability and connectivity," in *Conference on robot learning*, PMLR, 2021, pp. 1498–1508.

[51] J. Ni et al., "Phyrecon: Physically plausible neural scene reconstruction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 25 747–25 780, 2024.

[52] K. Yao et al., "Cast: Component-aligned 3d scene reconstruction from an rgb image," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 4, pp. 1–19, 2025.

[53] W. Yang et al., "Twintrack: Bridging vision and contact physics for real-time tracking of unknown dynamic objects," *arXiv preprint arXiv:2505.22882*, 2025.

[54] C. Song et al., "Inferring 3d shapes of unknown rigid objects in clutter through inverse physics reasoning," *IEEE robotics and automation letters*, vol. 4, no. 2, pp. 201–208, 2018.

[55] B. Chen et al., "Physgen3d: Crafting a miniature interactive world from a single image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6178–6189.

[56] H. Jiang et al., "Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos," *IEEE International Conference on Computer Vision (ICCV)*, 2025.

[57] T. Shen et al., "Flexible isosurface extraction for gradient-based mesh optimization," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–16, 2023.

[58] W. Peebles et al., "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[59] X. Pan et al., "Drag your gan: Interactive point-based manipulation on the generative image manifold," in *ACM SIGGRAPH 2023 conference proceedings*, 2023, pp. 1–11.

[60] H. Chen et al., "Mvdrag3d: Drag-based creative 3d editing via multi-view generation-reconstruction priors," *arXiv preprint arXiv:2410.16272*, 2024.

[61] L. Downs et al., "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 2553–2560.

[62] Q.-Y. Zhou et al., "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[63] R. Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[64] Y. Shi et al., "Mvdream: Multi-view diffusion for 3d generation," in *The Twelfth International Conference on Learning Representations*, 2024.