Contact-Aware Amodal Completion for Human-Object Interaction via Multi-Regional Inpainting

Supplementary Material

Region	BEHAVE	InterCap
Primary Region	48.07 %	35.05 %
Secondary Region	6.74 %	2.83 %

Table 5. Average percentage of occluded pixels in the primary and secondary regions for the BEHAVE and InterCap datasets.

A. Additional Details

We use the default parameters for all baselines and pre-trained models unless specified otherwise.

A.1. Occluded Pixel Ratios in Multi-Regions

Table 5 presents the percentage of occluded pixels within the primary and secondary regions. The percentage is computed based on the 2D area as follows:

$$\frac{|M_{\text{obj}}^{\text{full}} \cap M_{\text{region}}|}{|M_{\text{region}}|}, \tag{9}$$

where $M_{\text{obj}}^{\text{full}}$ denotes the projection of the fully rendered 3D object in image space, M_{region} corresponds to either the primary or secondary region, and $||$ represents the area of the mask, calculated by summing the binary mask values along the width and height axes.

In the BEHAVE dataset, the primary region effectively covers the inpainting area, with 48.07% of the primary region containing occluded parts. In contrast, the secondary region accounts for only 6.74%, emphasizing the need for careful handling of the secondary region.

A.2. Data Selection

For both the BEHAVE [2] and InterCap [8] datasets, we filter out images where the object occlusion is either less than 10% or greater than 70%, as these extremes provide limited value for evaluating occlusion handling. Additionally, we exclude frames where the visible area of the object is less than 5% of the human mask, ensuring sufficient detail for reliable analysis. These criteria maintain a balanced and robust dataset for evaluating our methods.

A.3. Implementation Details

Dataloader For the BEHAVE dataset, we utilized the dataloader provided by the HDM [36] GitHub repository (https://github.com/xiexh20/HDM). Based on this BEHAVE dataloader, we preprocess the InterCap [8] dataset to follow the same structure as the BEHAVE dataset, ensuring compatibility with minimal modifications to the original dataloader from HDM.

Baselines

- **Pix2Gestalt** [22]: We borrow the code and pre-trained model from https://github.com/cvlab-columbia/pix2gestalt and adapt it to be compatible with our dataloader implementation. Pix2Gestalt requires only the segmented image for amodal completion.
- **Xu et al.** [37]: To ensure a fair zero-shot comparison, we made several modifications to the code borrowed from https://github.com/k8xu/amodal. Since the original code was designed for 83 specific object classes, we replaced its InstaOrder [14] module with ground-truth depth ordering, supplied explicit occluder/occludee segmentation masks, and constrained its multi-iteration scheme to a single pass.
- **LaMa** [31]: We utilize the code from https://github.com/enesmsahin/simple-lama-inpainting. LaMa requires the original image and the occluder mask to perform inpainting.
- **Inst-Inpaint** [40]: We borrow the code and pre-trained model from https://github.com/abyildirim/inst-inpaint. Inst-Inpaint requires the original image and a text prompt specifying the object to remove. For example, "remove the person in the center."
- **Naive Outpainting** [25]: We employee the SD-inpaint model from https://github.com/huggingface/diffusers, which requires a segmented image and an inpaint mask. Here, the inpaint mask is defined as the remaining area outside the segmented image.

Application To demonstrate that our amodal completion method enhances downstream tasks like 3D reconstruction, we explored human-object interaction reconstruction, consisting of animatable human avatar creation and 3D object reconstruction.

We conducted both tasks on the BEHAVE dataset, which provides sequences with four synchronized views, ground truth SMPLH poses, and object poses for each timestamp. For simplicity, we used only a single view in both tasks.

For animatable human avatar creation, we followed the approach of GaussianAvatar [7]. Using single-view data and the provided ground truth SMPLH poses as input, we trained the human avatar model.

For 3D object reconstruction, we applied 3D Gaussian Splatting (3DGS) [12] to reconstruct moving objects from a single view. Since our setting involves a fixed camera with moving objects—unlike the original 3DGS setup with a static scene and moving camera—we adapted 3DGS by

treating the object's pose as the inverse of the camera's pose.

Comparing results using the original occluded images versus the amodally completed images in both tasks demonstrated the effectiveness of our amodal completion method in enhancing 3D reconstruction as shown in Appendix C.2.

A.4. Pseudo Code for Multi-Regional Inpainting

We present the pseudo code for Multi-Regional Inpainting in Algorithm 1, which outlines the key steps for handling multiple regions with varying occlusion levels. This approach ensures accurate and context-aware reconstruction by prioritizing regions based on occlusion characteristics. For full technical details and reproducibility, the complete implementation is included as an attached file.

Algorithm 1 Multi-regional Inpainting

procedure MULTI-REGIONAL INPAINT($p, I_{\text{in}}, M_p, M_s, r, T, S$)
  Input: $\mathcal{P}$ (text prompt), I_{in} (segmented input image),
  M_p (primary mask), M_s (secondary mask), r (strength),
  T (maximum timestep), S (scheduler)
  Output: Generated inpainted image I_{out}
  Step 1: Prepare Latents
  Initialize latent variable ℓ using I_{in} and random noise η
  Generate masked latent ℓ_{M_p} using M_p
  Generate masked latent $\ell_{M_p \cup M_s}$ using M_p and M_s
  Set $T' = \text{int}(T \times r)$ as the maximum timestep for M_s
  Calculate timesteps $\mathcal{T}$ based on T and r
  Step 2: Denoising Process
  for each $t \in \mathcal{T}$ **do**
    $\ell_{\text{input}} = \ell$
    Step 2.1: Scale Latent Model Input
    Scale ℓ_{input} using scheduler S with current timestep t
    Step 2.2: Concatenate Inputs for UNet
    if $t > T'$ **then**
      $\ell_{\text{input}} = \text{concat}(\ell_{\text{input}}, M_p, \ell_{M_p})$
    else
      $\ell_{\text{input}} = \text{concat}(\ell_{\text{input}}, M_p \cup M_s, \ell_{M_p \cup M_s})$
    end if
    Step 2.3: Predict Noise Residual
    $\eta' = \text{UNet}(\ell_{\text{input}}, t, \mathcal{P})$
    Step 2.4: Modify Latent Variable
    Update ℓ using guided noise prediction η' and scheduler S
  end for
  Step 3: Decode and Post-process
  Decode ℓ to generate final image I_{out}
  return I_{out}
end procedure

B. Additional Analysis on Amodal Completion

B.1. Human Amodal Completion

While our method is applicable to both human and object amodal completion, we introduce a refined approach specifically for human completion. Leveraging recent advancements in human mesh recovery techniques such as [20, 44], we can accurately delineate occluded regions of human. For human amodal completion, these occluded areas are localized by computing the intersection between the SMPL [18] body model's projection and the segmentation mask of the interacting object. This targeted approach enables efficient extraction of primary occluded regions, formalized as follows:

$$I_{\text{out}} = F_{T \to 0}\left(I_{\text{in}}^{\text{human}}, M_{\text{smpl}} \cap M_{\text{obj}}, \mathcal{P}\right), \qquad (10)$$

where $I_{\text{in}}^{\text{human}}$ represents the segmented image of the visible human parts, M_{smpl} is the SMPL body model projection, and M_{obj} denotes the visible object segmentation. This formulation enables precise identification of occluded human regions, allowing for focused and efficient inpainting within the primary occlusion areas.

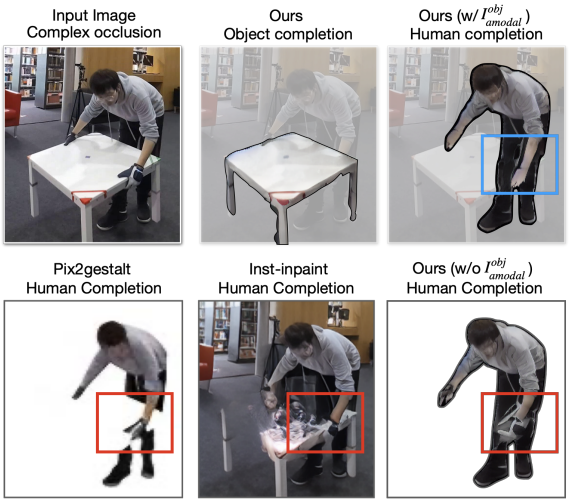
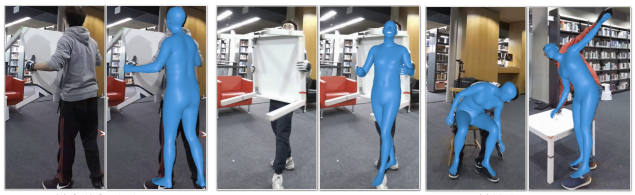


Figure 9. Mutual occlusion frequently occurs during HOI due to the dynamic nature of interactions. Baseline models often fail to produce plausible results, as highlighted in the red box. In contrast, our method generates more coherent results by progressively complete the object and human as shown in the upper row.

Complex Occlusion Scenarios Despite recent advancements, the dynamic nature of human-object interactions often introduces complex occlusions that challenge the quality of amodal completion results. For instance, in Fig. 9, when a person interacts with a table, the person's hand and arm occlude parts of the table, while the table simultaneously occludes parts of the person's legs. Such interactions complicate the accurate reconstruction of occluded human regions, even with topological priors, underscoring the challenges inherent in Human-Object Interaction (HOI) scenarios. Our observations indicate that repainting the entire region of intersection between the completed object and SMPL projection, rather than inpainting only the occluded areas, frequently yields more coherent and visually plausible results. This approach is captured in the formulation below:

$$I_{\text{out}} = F_{T \to 0}(I_{\text{in}}^{\text{human}}, M_{\text{smpl}} \cap \text{Seg}(I_{\text{amodal}}^{\text{obj}}), \mathcal{P}), \qquad (11)$$



(a) Self-Occlusion (b) Occlusion by an Object (c) Various poses

Figure 10. SMPL overlay images obtained by Multi-HMR [1] on the BEHAVE.

Contact Methods	SMPL MPJPE	Obj. Amodal CLIP ↑	Obj. Amodal mIoU ↑	Human Amodal CLIP ↑	Human Amodal mIoU ↑
Hand4Whole [19]	84.1 mm	26.59	74.54%	27.18	91.35%
DPMesh [44]	72.8 mm	26.73	76.24%	27.20	95.23%
Multi-HMR [1]	68.9 mm	26.91	77.64%	27.21	96.79%
GT-contact	–	27.07	80.15%	27.27	98.11%

Table 6. Experimental results w/o ground truth on BEHAVE. **Bold** denotes the result reported in main paper.

SMPL	Contact	Obj. mIoU ↑
Multi-HMR	-	74.80%
Multi-HMR	DECO	75.02%
Multi-HMR	VLM	77.64%
GT	**GT**	80.15%

Table 7. Different contact estimation methods.

where $I^{\text{obj}}_{\text{amodal}}$ represents the amodal completion image of the object, and $\text{Seg}(\cdot)$ represents a segmentation model. In our work, we utilized the Segment Anything Model (SAM) [24] as the segmentation model. This formulation enables more coherent inpainting by incorporating both the SMPL projection and object segmentation within the amodal completion framework.

B.2. Additional Details and Analysis on in-the-wild

Fig. 4 presents a pipeline without ground-truth annotations. Table 6 reports human mesh recovery accuracy in terms of MPJPE on the BEHAVE dataset, along with amodal completion results using predicted SMPL models and a Vision-Language Model for contact estimation. Notably, Multi-HMR [1] shows a MPJPE less than 70mm and achieves performance comparable to ground truth annotations in both object and human completion. Multi-HMR proves to be robust in occluded environments. We also illustrate the SMPL estimation results in Fig. 10.

Binary Contact Map. To improve practicality, we introduce a pipeline that does not rely on GT annotations. Although we discuss existing contact estimation methods (e.g., DECO [32]) in Sec. 6, these methods often fail to detect the presence of contact points, offering only marginal performance gains (see Tab. 7). Hence, we illustrate a VLM-based pipeline in Fig. 9. A prompt-engineered VLM [17, 21] takes an image and outputs both a textual description) Each ID corresponds to one point, and the estimated SMPL parameters then designate these joints as contact points. Similarly, for an image Fig. 10-(c) *left*, the VLM will produce a description "a man is sitting on a chair" and the hips joint IDs. Conversely, for a description such as "a person stands in front of a table," the VLM will not output any joint ID. As a result, combining Multi-HMR [1] with VLM approach achieves performance comparable to GT annotations, with a 2.5% gap as shown in Tab. 7. We plan to release the pipeline w/o GT.

SMPL accuracy Although imperfect SMPL estimation can cause challenges for object and human completion, Fig. 10 and Tab. 7 show that current SOTA models generally provide robust SMPL parameters in HOI scenarios, yielding sufficiently accurate contact estimates for our method. We achieve an mIoU of 96.79% for human completion. Even when SMPL parameters are misaligned due to occlusion, restricting the inpainting region to the intersection between the object segmentation mask and the projected human mesh effectively limits errors.

C. Additional Qualitative Results

C.1. Amodal Completion

Baseline Comparison To illustrate the strengths of our method compared to existing approaches, additional results are provided in Fig. 11. These examples showcase our pipeline's ability to handle complex occlusion scenarios while preserving finer details. By comparison, baseline methods often fail to deliver coherent and detailed completions under similar conditions, underscoring the effectiveness of our approach.

Diverse Outputs The diverse outputs generated by our pipeline, visualized in Fig. 12, highlight the flexibility of our approach in producing multiple plausible amodal completions for a single input. However, this diversity also exposes a limitation: the lack of consistency between these outputs. Addressing this challenge could drive future research, focusing on improving coherence across diverse completions to achieve more reliable and unified results, particularly for downstream tasks like 3D reconstruction.

Failure Cases We visualize failure cases in Fig. 13 to analyze the limitations of our approach, categorized into three types: 1. *Object Orientation Errors*: Misinterpreted object direction, often due to ambiguous visual cues, causes misalignment. 2. *Shape Completion Errors*: Challenges in predicting occluded regions, especially for complex geometries, result in unrealistic shapes. 3. *Segmentation Errors*: Inaccurate masks lead to flawed reconstructions, affecting amodal completion and 3D reconstruction. Segmentation errors can be mitigated by user-driven manual corrections, while shape errors can be addressed by adjusting the parameter r in our pipeline. However, resolving orientation errors requires further research and is left as a direction for future work.

C.2. 3D Reconstruction

The comparison of 3D reconstruction results in Fig. 14 highlights the effectiveness of using amodally completed images over original occluded images. These results demonstrate that our amodal completion method significantly enhances the quality of 3D reconstructions, validating its role as a vital preprocessing step for complex 3D tasks. Additionally, we provide videos showcasing novel-pose synthesis with human-object interaction in the attached file.

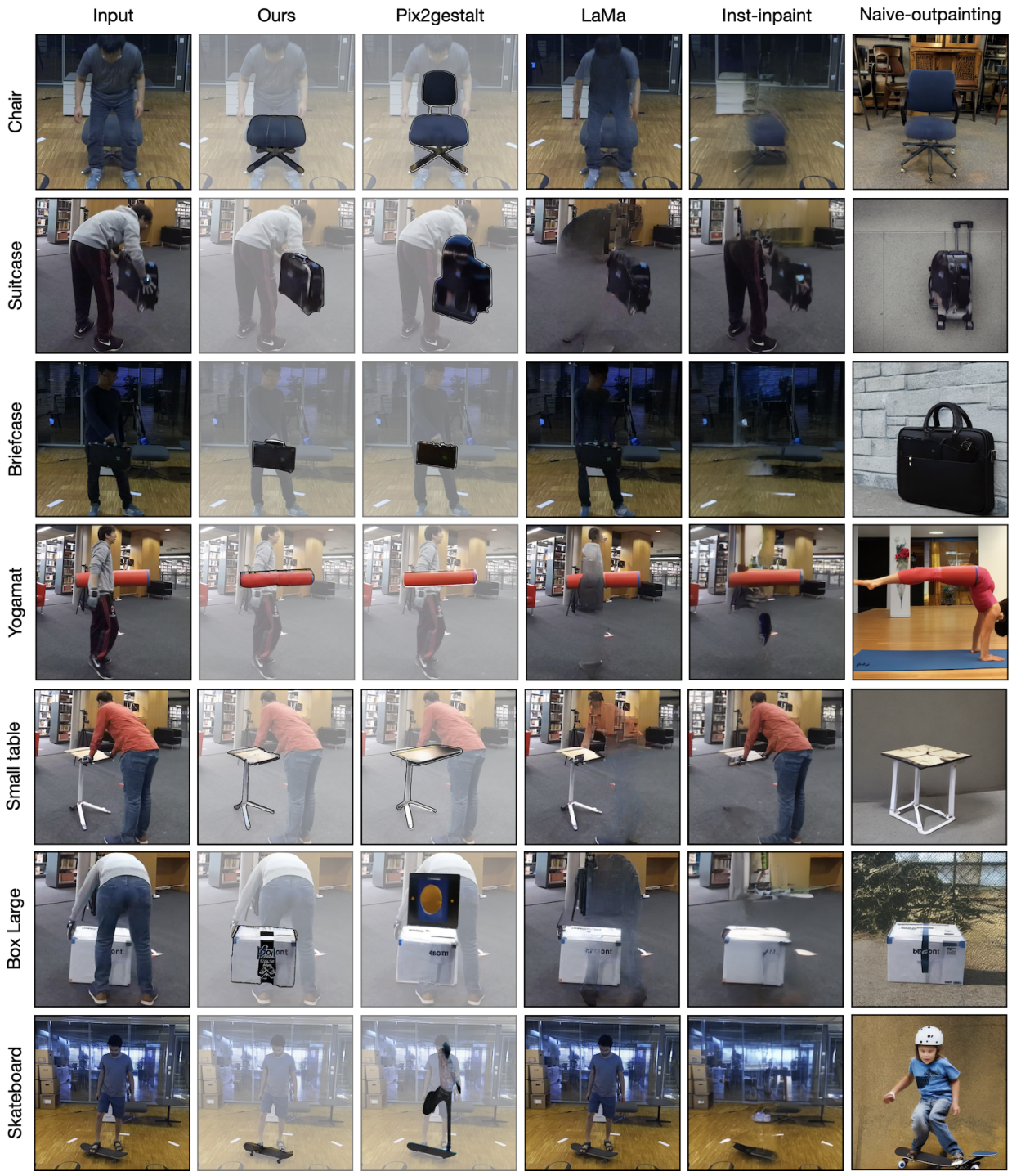


Figure 11. Qualitative comparison between ours and baseline models.

C.3. User study

Recognizing that CLIP score and mIoU have limitations in fully representing amodal completion quality, we conducted a user study. A total of 223 sample pairs were presented, with each pair evaluated by an average of 10 users. For each pair, users were asked to select the more accurate and realistic amodal completion result. This study focused exclusively on object amodal completion. Instructions and examples for the user study are provided in Fig. 15.

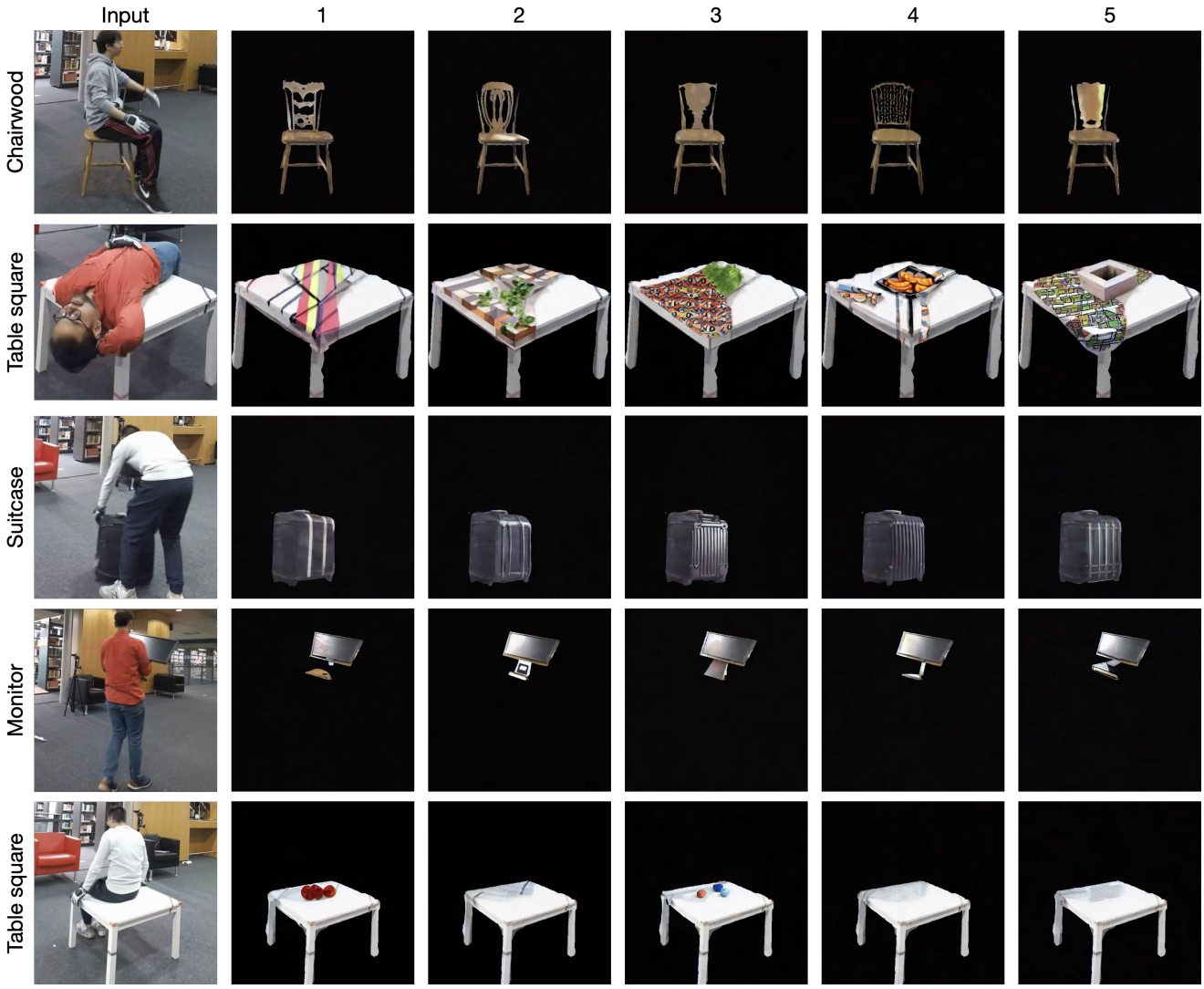


Figure 12. Diverse outputs generated by our pipeline. The visualization includes results from 5 different samples.

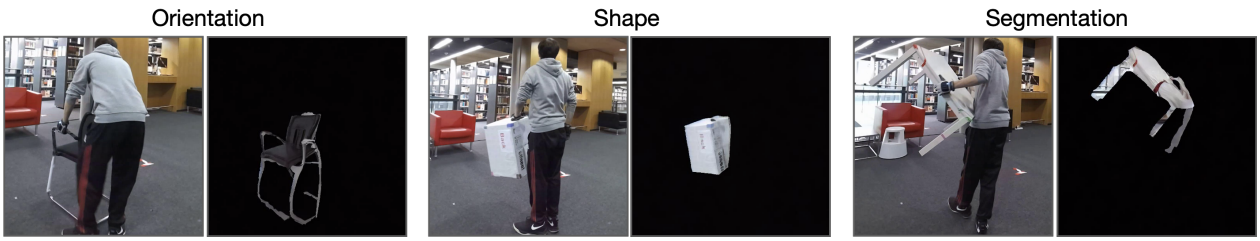


Figure 13. Failure cases from our pipeline, categorized into orientation errors, shape errors, and errors caused by poor segmentation.

Figure 14. Additional qualitative results of 3D-GS with and without amodal completion.

Section 1 of 10

Amodal Completion User Study: 151-200

In this task, you will see a series of image sets.

- The **left image** is the **Reference Image:** It shows a person interacting with an object, but part of the object is hidden behind the person.
- The **two images on the right** (labeled A and B) are **different guesses** of what the full object might look like if the person wasn't blocking it.

What You Need to Do:

- Look at the **object** the person is interacting with in the **Reference Image.**
- Decide which image **(A or B)** shows the object in the most realistic and accurate way. Here, please evaluate the quality only on the **object** the person is using, and ignore the person and background.
- Choose the one that best matches what you think the full object should look like based on what's visible in the Reference Image.

Question 152

Question 169

Figure 15. User study instruction and examples.