

Using Diffusion Priors for Video Amodal Segmentation

Supplementary Material

In this appendix, we extend the discussion of our approach on video amodal segmentation. We first discuss additional setup details for our method (Sec. 6), and then cover more experimental analysis (Sec. 7), followed by examples of our method’s potential applications (Sec. 8). We also show more qualitative results from our method (Sec. 9). Please see the project page for a video version of all figures.

6. Additional setup details

6.1. Stable Video Diffusion modifications

First, we replace the input conditioning \mathbf{c} , originally an RGB image, with binary modal masks of shape $\mathcal{R}^{T \times 1 \times H \times W}$. By default, the variational autoencoder (VAE) [25] in SVD requires a 3-channel input. To address this mismatch in the number of channels, we replicate the binary mask three times, following the approach for single-channel VAE inputs in a recent work [22]. After encoding each (replicated) mask separately, we obtain a latent tensor of shape $\mathcal{R}^{T \times C \times \frac{H}{F} \times \frac{W}{F}}$. This latent representation, concatenated with a noise image of the same shape, forms the input to our backbone which is a spatio-temporal 3D U-Net [4, 44]. The final shape of this input becomes $\mathcal{R}^{T \times 2C \times \frac{H}{F} \times \frac{W}{F}}$. In contrast to the vanilla SVD, where the latent space of a single image is duplicated T times to align with the 3D U-Net’s input requirements, our 3D U-Net gets as input T unique frames of the modal mask sequence being used as conditioning.

6.2. Inference details

During inference with our video diffusion model, we follow common practices [3] by employing the stochastic sampler from EDM [20]. We simplify this process by omitting the second-order correction and keeping the explicit Langevin-like “churn” factors constant. The denoising process is performed over 25 steps. Specifically, when denoising the latents from z_t to z_0 for $i \in \{t, \dots, 1\}$, each denoising step can be expressed as:

$$\hat{\mathbf{z}}_{i-1} \leftarrow \hat{\mathbf{z}}_i + (\sigma_{i-1} - \sigma_i) \frac{(\hat{\mathbf{z}}_i - D_\theta(\hat{\mathbf{z}}_i; \sigma_i))}{\sigma_i} \quad (2)$$

Furthermore, we employ classifier-free guidance (CFG) [15] to balance the quality and diversity of the generated samples. During training, we randomly set the conditioning to zero with a probability of $\rho = 0.1$ to simulate the unconditional case. During inference, we combine the conditional and unconditional predictions

using a guidance scale of $s = 1.5$, as defined as:

$$\tilde{F}_\theta(\mathbf{z}, \mathbf{c}) = F_\theta(\mathbf{z}, \emptyset) + s(F_\theta(\mathbf{z}, \mathbf{c}) - F_\theta(\mathbf{z}, \emptyset)) \quad (3)$$

After denoising, the latent predictions are projected back into pixel space using the VAE decoder, which yields three-channel representations. To convert these into single-channel binary masks in the amodal segmentation stage, we sum the channel values (from 0 to 255) and binarize the predictions by thresholding. The threshold is chosen as a per channel pixel-value of 200. Finally, we take the union of the prediction with the input modal masks, ensuring modal masks remain a subset of amodal masks and are properly reflected in the output.

Regarding inference time, our method takes approximately 0.95 seconds per frame on a single RTX 3090 GPU, using around 8GB of VRAM with FP16 precision.

6.3. Baselines

In this section, we provide additional details of the image- and video-level amodal segmentation methods used for comparison.

For image-level amodal segmentation, ‘Convex’ [60] generates the geometric convex hull of modal masks, while ‘Convex^R’ [60] refines this by including only the convex hull within occluded regions predicted by ‘PCNet-M’. ‘PCNet-M’ [60] is a self-supervised regression method that recovers amodal masks within occluder areas based on frame-level object ordering recovery. ‘AISFormer’ [51] employs a transformer-based head appended to a modal segmentation backbone to directly predict all amodal bounding boxes and masks within an image. ‘pix2gestalt’ [36] is an image diffusion-based method that generates amodal content conditioned on the RGB image and modal masks of the objects.

For video-level amodal segmentation, ‘SaVos’ [57] employs a CNN-LSTM architecture that processes RGB and modal mask patches, along with optical flow, to predict amodal masks and motions. ‘EoRaS’ [9] proposes an object-attention encoder that incorporates Bird’s-Eye View (BEV) 3D information, relying on having access to groundtruth camera parameters. ‘C2F-Seg’ [11] leverages a vector-quantized latent space for coarse feature learning, refined with a convolutional module; though designed for image-level tasks, it extends to video segmentation using a spatial-temporal transformer block.

For generic video regression approaches, ‘Video-MAE’ [50] is a transformer-based autoencoder that we adapt for our task by setting the masking ratio to zero, ap-

Table 5. **Quantitative results on MOVi-B/D with uncropped input.** Enlarged modal region-cropped input limits the model’s ability to predict an amodal mask when an object is fully occluded. Using the entire image as input restores the model’s ability to complete amodal masks fully, especially when the modal area is small. This results in substantial metric improvements compared to Table 2 in the main paper. We copy over the results here for reference.

Input	Method	MOVi-B		MOVi-D	
		mIoU	mIoU _{occ}	mIoU	mIoU _{occ}
Modal cropped	VideoMAE [50]	78.74	42.86	70.93	32.78
	3D-UNet	82.16	49.81	75.65	40.86
	Ours (Top-1)	83.51	53.75	77.03	44.23
	Ours (Top-3)	83.93	54.56	77.76	45.6
Uncropped	VideoMAE [50]	85.35	49.53	79.13	42.41
	3D-UNet	84.24	46.17	76.90	36.69
	Ours (Top-1)	87.8	53.69	82.97	47.86
	Ours (Top-3)	88.43	54.64	84.04	49.43

plying supervised training, and using the decoder during inference. ‘3D-UNet’ [4], the backbone of our video diffusion model, contains interleaved residual and transformer blocks with spatial and temporal modules but is trained to perform one-step generation without any iterative denoising.

7. Additional experiments

Note that the video versions of all qualitative results in this and the following sections can be found directly on the project page.

Improved results on MOVi-B/D. All results we report till now on MOVi-B/D follow prior work in segmenting objects in a region which is defined as a 100% extension of the region enclosed by the input modal mask. Therefore, all images are cropped to this region before being sent as input to any of the methods. This is different from the standard protocol used in other datasets, where the *entire* image is sent as input to the methods (without any cropping). Here, we include results from training our model with the entire image as input on the MOVi-B/D datasets. As shown in Table 5, this fix significantly improves metrics, with our method achieving 4% and 6% gains in mIoU on MOVi-B and MOVi-D, respectively. Regression methods also benefit notably from this setting. We conclude that this is because MOVi-B/D include many instances of complete occlusions of objects, for which segmentation in a cropped region is not enough for predicting the amodal mask.

Qualitative evidence for pseudo-depth conditioning. The quantitative advantage of pseudo-depth conditioning was demonstrated in Table 3 of the main paper. Here, we provide qualitative evidence to illustrate the source of this improvement. As shown in Figure 10, pseudo-depth conditioning encourages our method to segment areas *closer* to

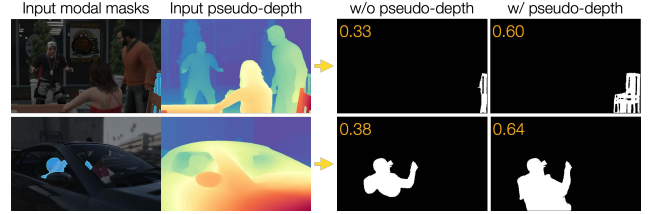


Figure 10. We show **how pseudo-depth aids amodal segmentation.** Object’s surrounding regions with lower depth values, i.e., closer to the camera, act as potential occluders. In the top row, the occluders are the person and chair to the left of the object; in the bottom row, the occluder is the car door below the person. Depth information implicitly guides our method to complete these occluded regions.

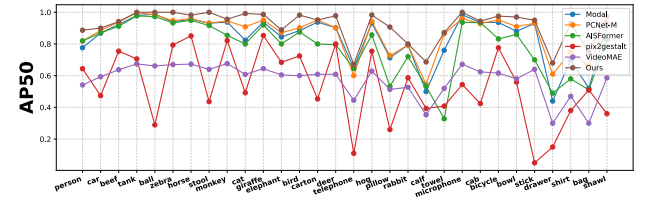


Figure 11. **Comparison across diverse categories on TAO-Amodal.** On a subset of the most frequent super-categories, our method consistently outperforms others under the AP50 metric. The overall trend is aligned with the quantitative results in Table 1. We attribute the strong generalization ability of our model to the SVD priors and its effective utilization of temporal information.

the camera, suggesting that depth serves as an implicit indicator of potential occluders and therefore, gives information about which occluded boundary to extend in order to predict the amodal mask.

Ablation on weights initialization. We leverage the real-world priors learnt by large-scale diffusion models by utilizing pretrained SVD checkpoints [3]. Here, we evaluate the importance of this initialization. In Table 6, we compare the performance of our model and the 3D U-Net with and without pretrained weights. Results show that excluding the checkpoint leads to a performance drop for both models, with a more pronounced decline for ours. These results underscore the importance of the SVD priors.

Building an end-to-end segmentation and completion model. Unlike our two-stage method, which first performs amodal segmentation and then inpaints content, the image diffusion-based method pix2gestalt [36] adopts a one-stage approach to directly generate amodal content and derive masks. A similar one-stage approach can be extended for our video setting. However, as shown in Table 7, our two-stage method demonstrates clear advantages over

Table 6. **Ablation of SVD priors.** We study the effect of using pretrained SVD weights as initialization for our training. We find that leveraging priors from large-scale pretraining of SVD enhances both our method and the 3D UNet baseline, with particularly substantial improvements observed for our method.

Method	pretrained ckpt?	SAIL-VOS		TAO-Amodal		
		mIoU	mIoU _{occ}	AP ₂₅	AP ₅₀	AP ₇₅
Ours	✗	68.89	26.96	93.73	79.45	57.87
Ours	✓	75.17	51.28	94.89	85.03	66.87
3D UNet	✗	70.85	32.66	94.88	83.81	59.75
3D UNet	✓	72.79	39.54	94.59	83.83	64.33

Table 7. **Ablation study on end-to-end amodal content completion.** We train an end-to-end version of our two-stage pipeline with a dataset of curated modal-amodal RGB training pairs from SAIL-VOS, in a similar fashion to pix2gestalt [36]. Compared to the two-stage results in Table 1 of the main paper, this approach shows a significant performance drop in both in-domain and zero-shot evaluations, highlighting the superiority of the two-stage method.

Method	SAIL-VOS		TAO-Amodal		
	mIoU	mIoU _{occ}	AP ₂₅	AP ₅₀	AP ₇₅
Two-stage	77.07	55.12	97.28	89.25	71.99
One-stage	66.15	40.31	70.65	57.51	37.22

the one-stage approach. We attribute this low performance of the end-to-end method to the lack of data available for training such a single-stage method. In contrast, the two-stage method benefits from breaking down the pipeline into video amodal segmentation and content completion. For the former, it is easy to find large-scale training data of modal-amodal mask pairs from synthetic datasets. For the latter, since the content completion task reduces to video inpainting, less amount of training data is sufficient for finetuning.

Generalizability across diverse categories. Our model demonstrates strong generalization ability in a zero-shot setting on the real-world TAO-Amodal dataset, which includes many previously unseen categories. TAO-Amodal is a collection of 7 different datasets, covering a wide range of in-the-wild scenarios. Specifically, it consists of 833 object categories, out of which only 20 categories are seen during training in SAIL-VOS. For more clarity, we include a performance breakdown on a subset of the most frequent super-categories in TAO-Amodal, as shown in Figure 11. The results further highlight our model’s capacity to generalize across diverse categories.

8. Examples of applications

4D reconstruction. Our method enables 4D reconstruction for occluded objects when used in conjunction with off-

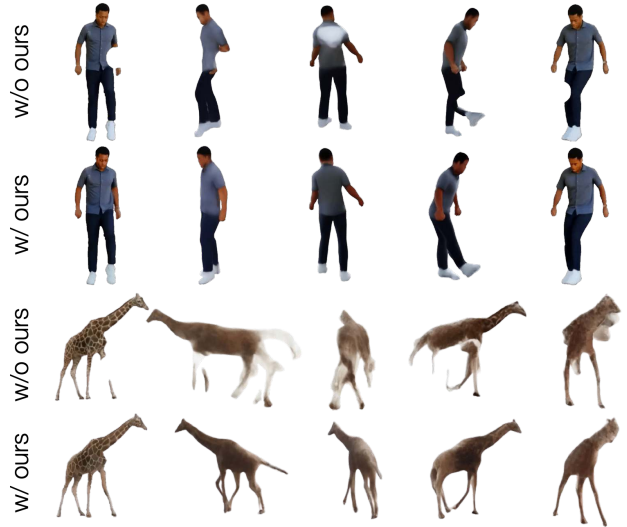


Figure 12. **4D reconstruction results.** Without amodal completion by our method, the 4D reconstruction exhibits blank regions and unrealistic artifacts in occluded areas, such as the person’s back and leg. The varying occluded portions over time confuse SV4D, disrupting its understanding of the object’s 3D structure. In contrast, using completed objects from our method significantly improves the reconstruction quality, producing more consistent and clear novel-views.

the-shelf SV4D [54]. In Figure 12, we compare reconstructions with and without completion. Without completion, blank regions appear in occluded areas, making it more difficult to hallucinate reasonable re-projections across different views. In contrast, our method allows SV4D to produce consistent and clearer 4D reconstructions.

Scene manipulation. With amodally completed objects in the scene, we can change their orderings and positions without exposing previously occluded regions. Figure 13 shows examples of scene manipulation, where our method facilitates manual re-composition of scenes by inpainting the occluded content of objects.

Pseudo-groundtruth for TAO-Amodal masks. TAO-Amodal [18] provides ground truth for amodal bounding boxes but lacks annotations for amodal *masks* due to the challenges of manual labeling of occluded objects in videos. We show that our method can be used to generate high-quality *pseudo*-ground truth masks for this dataset by using the information about ground-truth amodal bounding boxes, which define the extent of the amodal shape. We find that using the amodal bounding boxes to crop the input modal mask sequences, one can train a more accurate video amodal segmentation method exclusively on SAIL-VOS. This way, our approach significantly improves evaluation

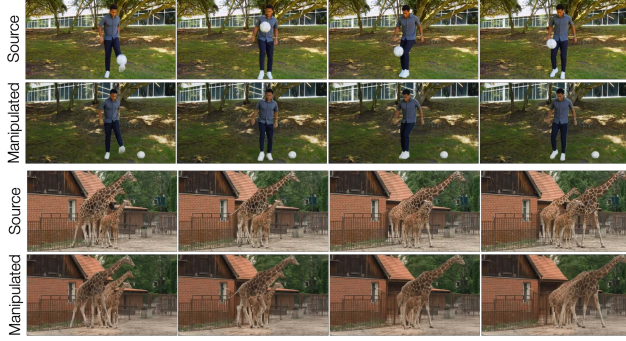


Figure 13. **Scene manipulation examples.** Using de-occluded objects from our method, we can reposition and reorder them to create new scenes. In the top rows, the relationship between the person and the soccer ball is altered, changing the scene from “the person is juggling” to “the person places the soccer ball aside and practices a juggling posture.” In the bottom rows, the middle giraffe is moved to the front and its position is adjusted.

Table 8. **Pseudo-groundtruths on TAO-Amodal.** We show that using the amodal bounding box prior from the TAO-Amodal dataset to specify the extent of the output amodal segmentation mask, can help improve the quality of video amodal segmentation. We use this version of our method to produce ‘pseudo-groundtruths’ for TAO-Amodal. We find that these pseudo-annotations can help improve the quantitative performance of baselines like VideoMAE. See text for more details

Input setting	SAIL-VOS		TAO-Amodal		
	mIoU	mIoU _{occ}	AP ₂₅	AP ₅₀	AP ₇₅
Uncropped	77.07	55.12	97.28	89.25	71.99
Amodal cropped	87.44	69.81	99.59	99.59	99.48

metrics and aligns precisely with the amodal bounding box extent, as shown in Table 8. Figure 14 further illustrates the qualitative results of the pseudo-ground truth masks which are high-fidelity across diverse object categories. Quantitatively, we find that using the pseudo-groundtruths for fine-tuning baselines like VideoMAE (which have already been pre-trained on SAIL-VOS), improves their performance on the TAO-Amodal dataset by around 25%, 25%, and 20% on AP₂₅, AP₅₀, and AP₇₅ respectively. Apart from this, the generated pseudo-groundtruths can be used to semi-automate the amodal mask annotation process as this is a challenging and inherently ill-posed problem.

Note that we do not include this data point in the main paper as at inference we cannot expect to have access to amodal bounding boxes but in order to produce pseudo-groundtruth, one can adopt this approach.

9. Qualitative results

A video version of all figures in this section are available on the project page.

Here, we present qualitative results from all datasets and additional, in-the-wild scenarios. Figures 16, 17, 18, 19 and 20 compare our amodal segmentation method with more baselines on SAIL-VOS, TAO-Amodal, and MOVi-B/D. Our method demonstrates superior performance in generating high-fidelity shapes in the occluded regions of objects. Figure 21 showcases additional in-the-wild content completion results, highlighting the photo-realistic quality and strong generalization capability of our method.

Failure cases. In Figure 15, we show four different kinds of failure cases. In the first case with a person swimming, our method does not successfully complete the person’s amodal region. This happens often if the object of interest is occluded throughout the extent of the video; our model is not able to understand if this is a completely visible object or a consistently occluded object. In the second case, the occluded object is a bow, which has never been seen before and is completely out-of-distribution from the set of objects in SAIL-VOS. Our method fails in this case. In the third and fourth case, our method incorrectly assumes the height of a completely visible man to be greater than what it is, and predicts a sitting person to be standing. Therefore, our method lacks contextual cues about what the scene is and how the modal region looks like in the first-stage.

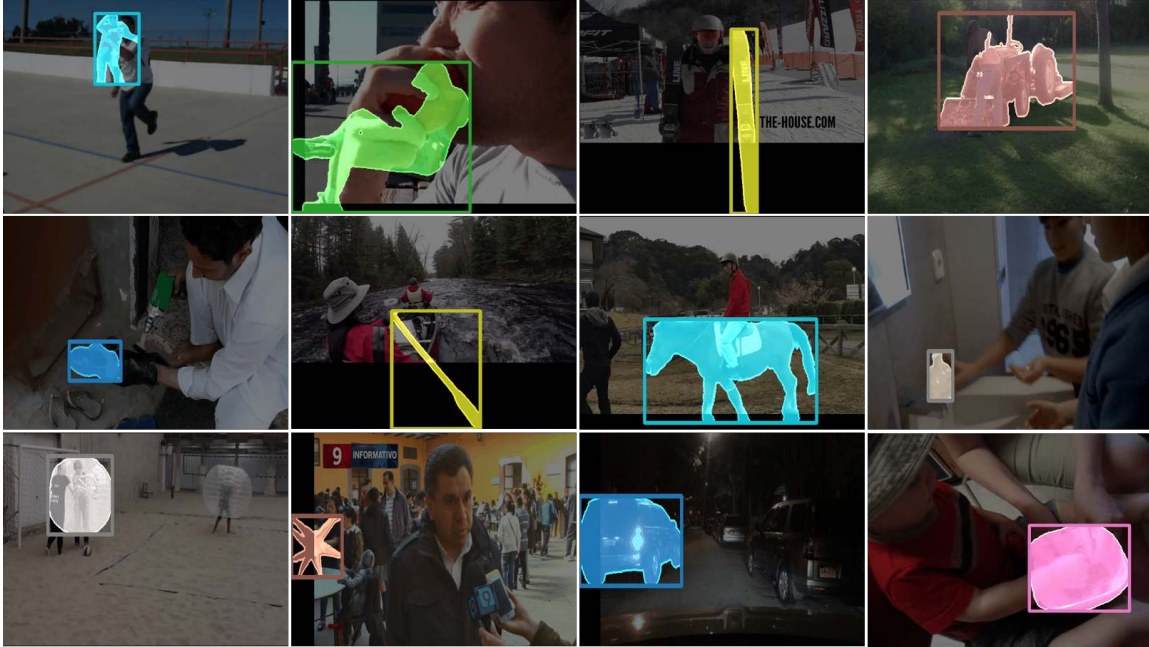


Figure 14. **Qualitative results for pseudo-ground truth of TAO-Amodal masks.** Leveraging the amodal bounding box as a strong prior, our method demonstrates versatility across diverse categories, such as person, tractor, and bottles, and generalizes well to unseen categories like snowboards and horses. This high-quality pseudo-ground truth can semi-automate the manual annotation of amodal masks in real-world videos.

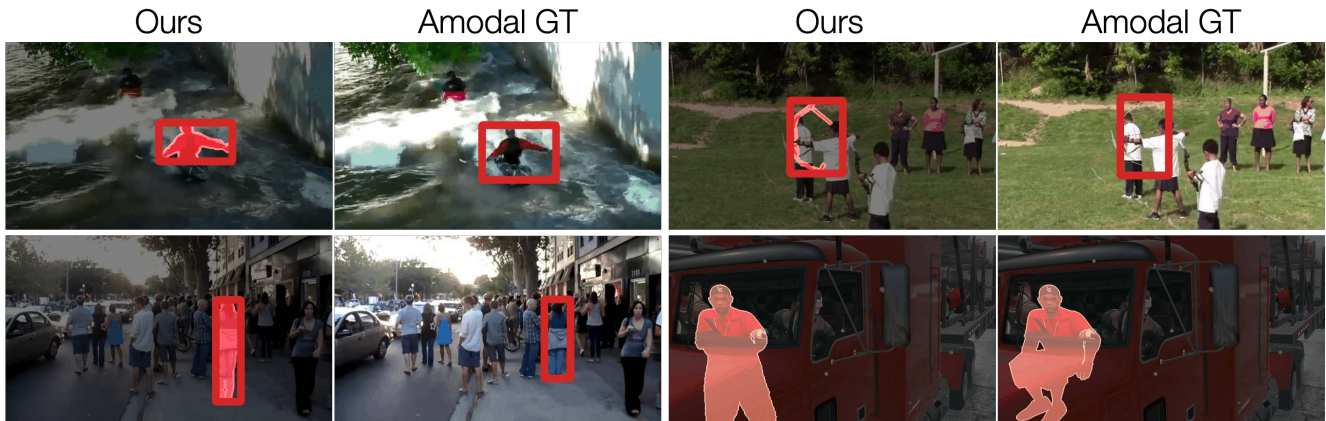


Figure 15. Qualitative analysis of failure cases of our method. See text for more details.

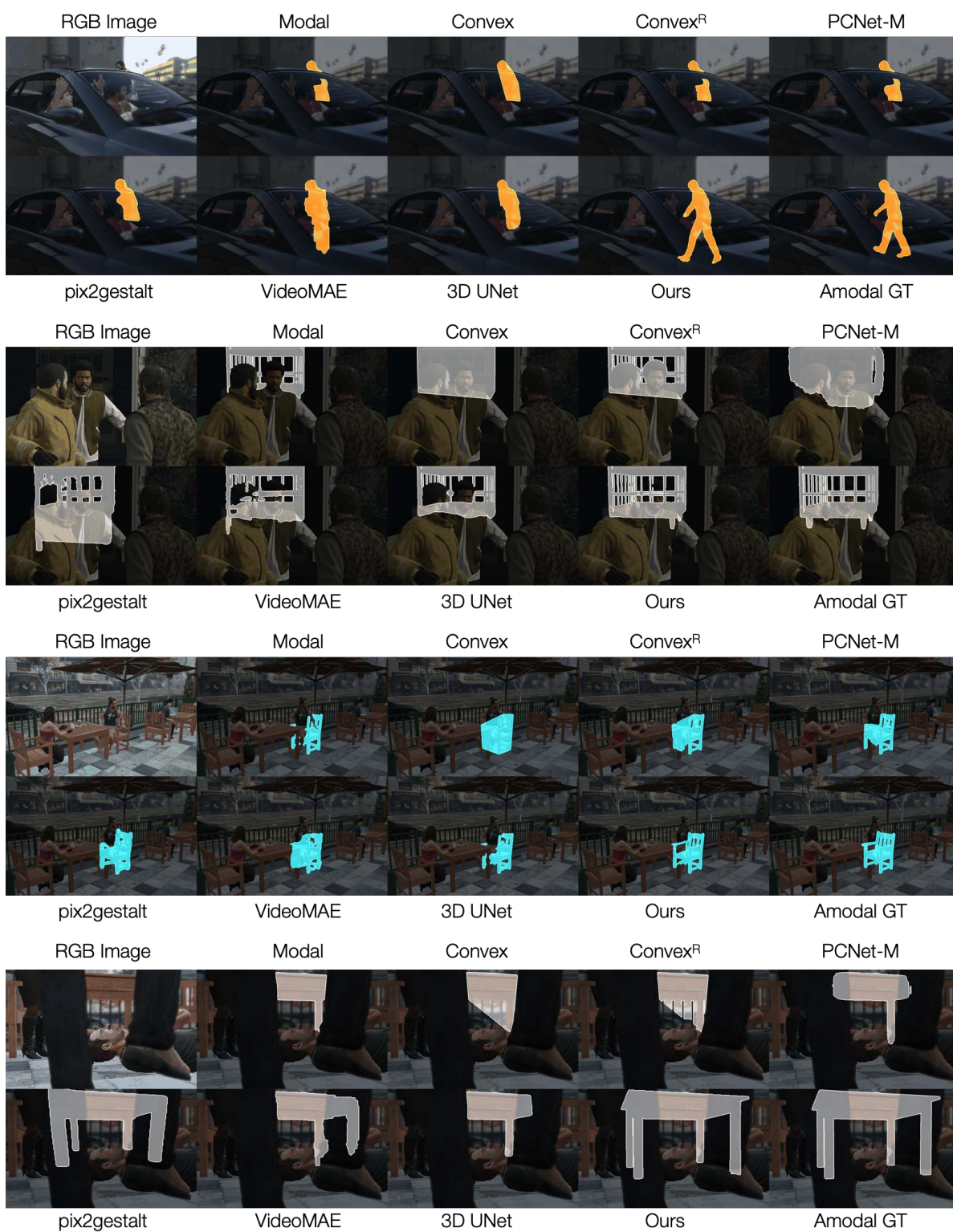


Figure 16. Qualitative results on SAIL-VOS. (1/2)

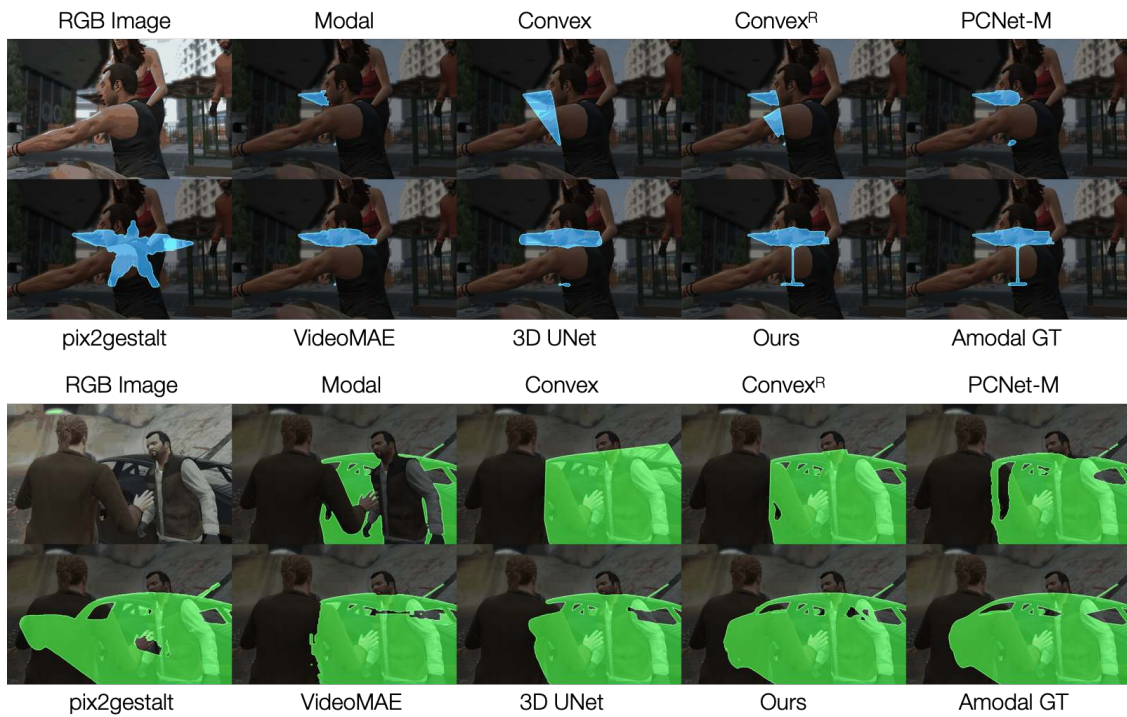


Figure 17. Qualitative results on SAIL-VOS. (2/2)



Figure 18. Qualitative results on TAO-Amodal. (1/2)

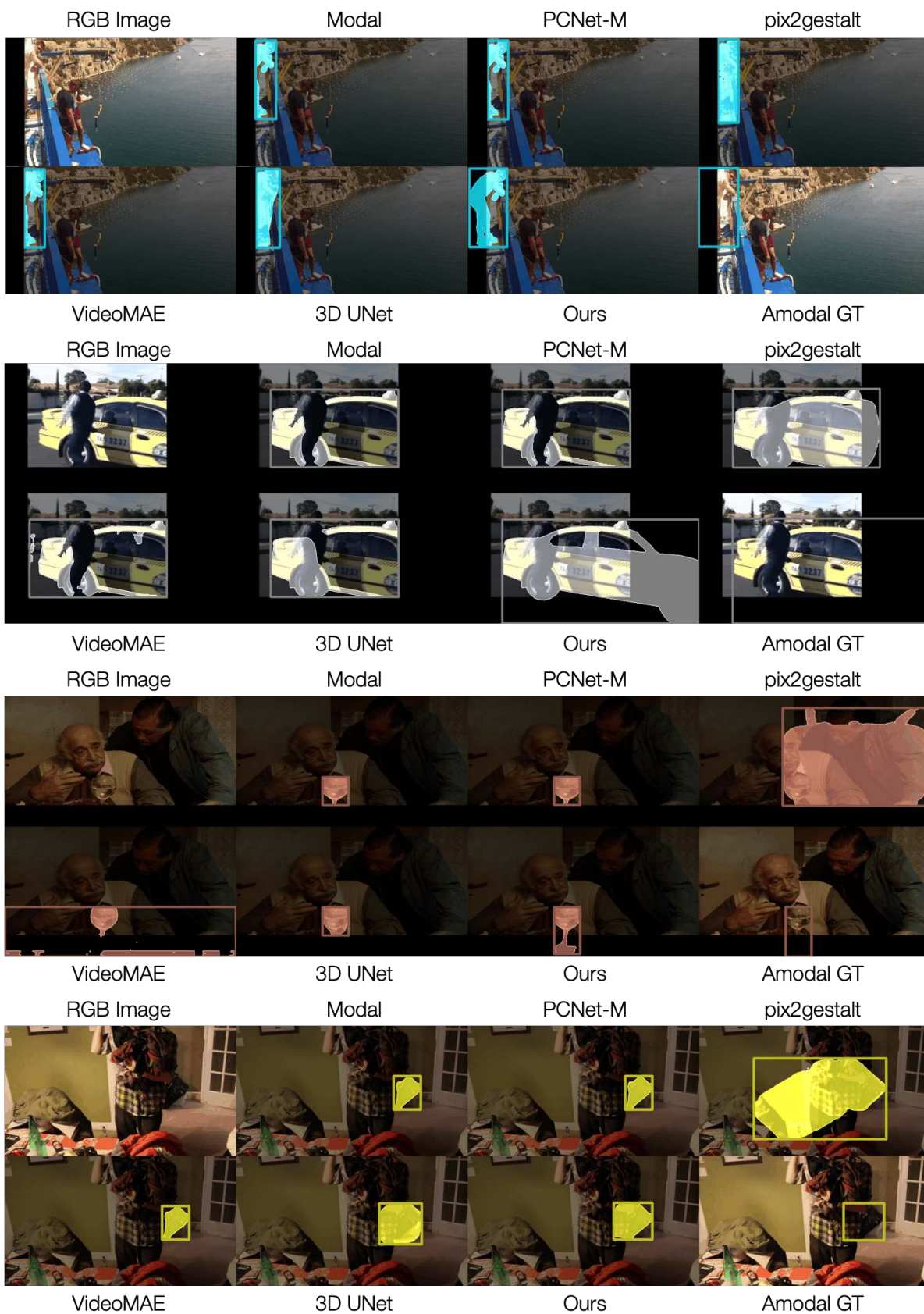


Figure 19. Qualitative results on TAO-Amodal. (2/2)

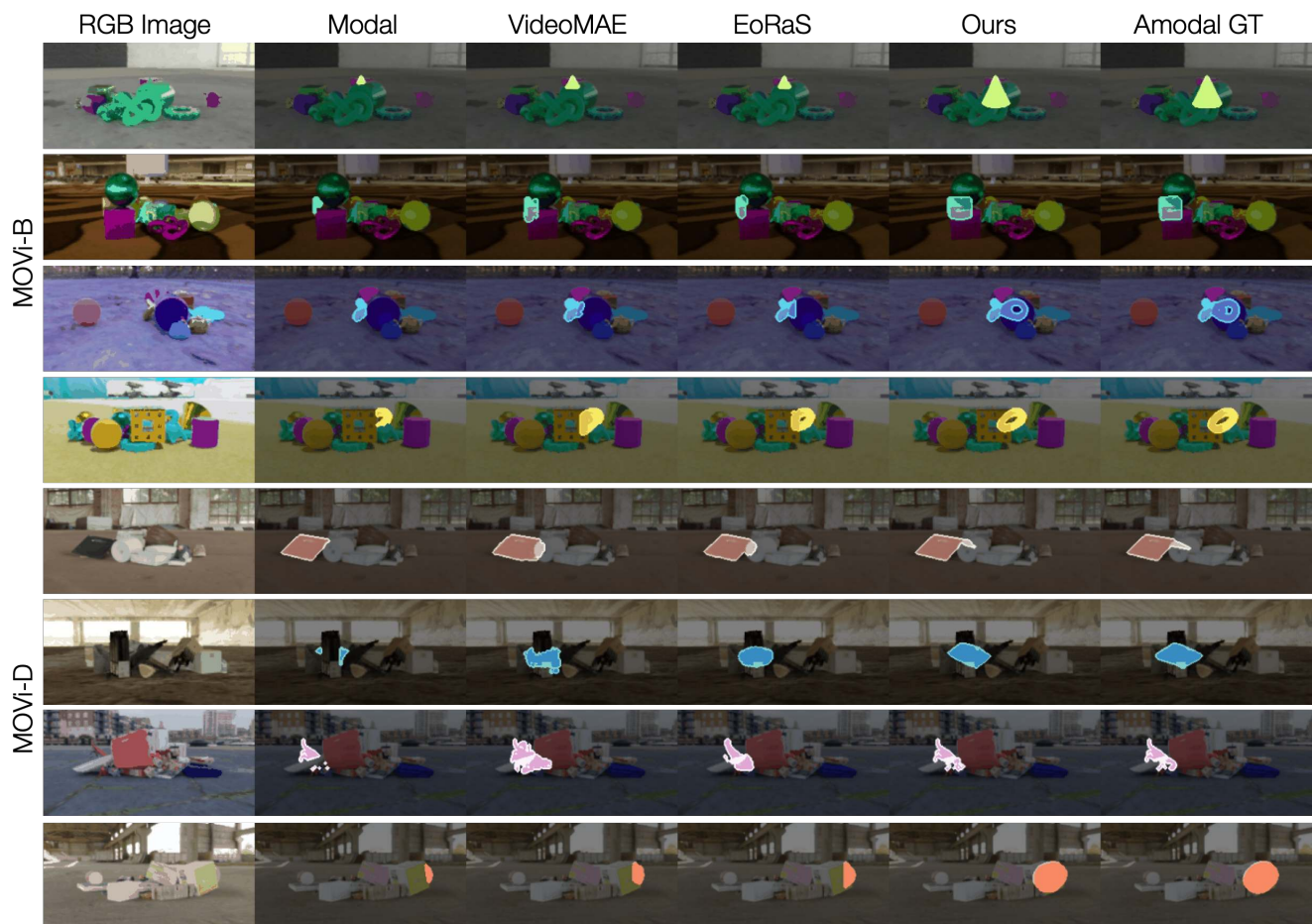


Figure 20. Qualitative results on MOVi-B/D.

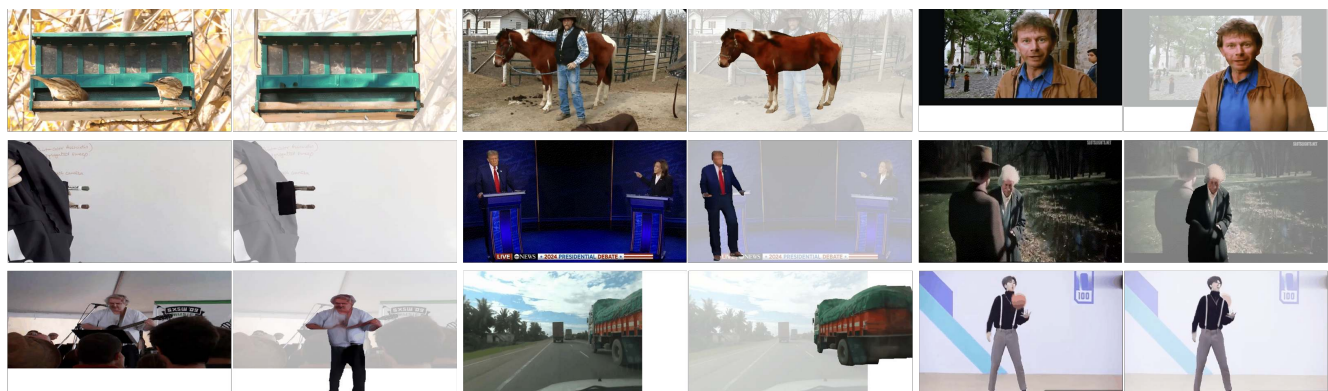


Figure 21. Qualitative results for amodal content completion for in-the-wild scenarios.