Image Guides Images: Consistent Video Amodal Completion with Rectified In-Context Exemplar Guidance

Supplementary Material

Contents

A. Results on TAO-Amodal

In this section, we compare the performance of various methods on the real-world TAO dataset [3]. Since no amodal ground-truth is available for quantitative evaluation, we present visualizations of different methods, as shown in Figs. S3–S8. It can be observed that our method achieves video amodal completion with stable consistency and high visual quality. As illustrated in Fig. S3, despite the significant motion variations of the athlete, our method can still reconstruct the complete human body for each movement while maintaining the consistency of visual elements (e.g., shoes, upper garments). In contrast, TACO [7] and SD-VAS (Diffusion-VAS) [2] fail to complete occluded human body parts. This deficiency stems from the limited generalization capability of their fine-tuned models, which is also demonstrated in Figs. S5 and S7. Regarding object completion, as shown in Fig. S8, our method maintains the consistency of object completion across different scales even when the object size changes due to camera movements.

Overall, by leveraging the prior of pre-trained image inpainting models, our method achieves high generation quality and ensures inter-frame consistency through rectified in-context learning. In contrast, previous SOTA VAC methods rely on fine-tuning video models, which not only degrades their generalization capability but also reduces their generation performance. As a result, their completion results may fail to provide effective support for downstream recognition tasks.

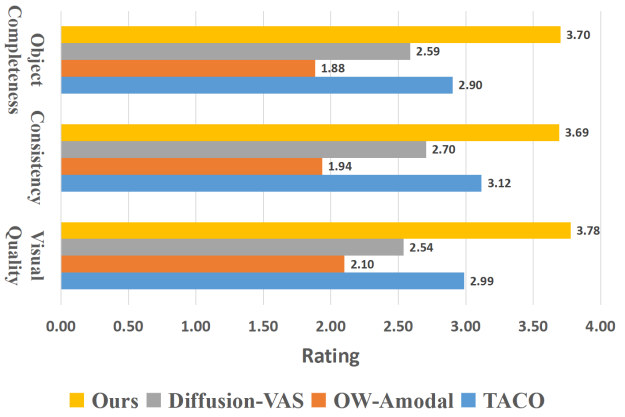


Figure S1. Statistics for user study.

	Exemplar Selection	PSNR(↑)	SSIM(↑)	LPIPS(↓)	IoU(↑)
𝕊-I	*First frame*	23.58	0.881	0.103	81.9
𝕊-II	*Random*	23.79	0.896	0.1206	79.2
Ours	*Biggest visible*	24.71	0.910	0.070	86.6
	Prompt Setting	**PSNR**(↑)	**SSIM**(↑)	**LPIPS**(↓)	**IoU**(↑)
𝕋-I	*Minimal*	24.70	0.911	0.071	86.1
𝕋-II	*Detailed*	24.36	0.905	0.076	84.2
𝕋-III	*Distracting*	24.15	0.905	0.081	83.6
Ours	*Object description*	24.71	0.910	0.070	86.6

Table S1. Influence of the setting of IC .

B. User Study

To evaluate the perceptual quality on real data, we conduct a user study with 44 participants using 28 real-world videos. We compare our method against three representative baselines, TACO [7], SD-VAS (Diffusion-VAS) [2], and OW-Amodal [1]. For each video, we present the amodal results of four methods side by side in a randomized order and ask the participants to compare them before assigning scores to each method along three dimensions: visual quality, consistency with the visible content, and object completeness. Each dimension is scored on a 5-point scale, where a score of 1 indicates very poor quality or a clear mismatch after comparison, and a score of 5 indicates high quality and good agreement with human expectation.

We collect all ratings and report the average score of each method on the three dimensions. Fig. S1 shows statistics for user study results. In particular, our method achieves the highest average scores on all three dimensions with margins that are clearly larger than those of the baselines, which further demonstrates the effectiveness of the proposed approach.

C. Exemplar Frame Selection Strategy

In this section, we discuss the strategy for exemplar frame selection. Our method adopts an intuitive strategy: select the frame with the largest visible region in the sequence as the exemplar, and a second frame sufficiently distant from the exemplar as the collaboration frame. To validate the impact of exemplar selection on performance, we compare with the following strategies:

- $\mathbb{S}$-I: Select the first frame as the exemplar and the second frame as the collaboration frame;
- $\mathbb{S}$-II: Randomly select two frames as the exemplar and collaboration frame.

As shown in the upper part of Table S1, exemplar selection strategies significantly affect the completion results. $\mathbb{S}$-I yields the worst performance due to insufficient complementary information between the two consecutive frames. For $\mathbb{S}$-II, the selected frames may have excessively small visible regions, leading to incomplete completions. In contrast, our strategy ensures the effectiveness of the exemplar, while the selected collaboration frame facilitates the completion of the exemplar—thereby enhancing the performance of IC learning.

D. Influence for Text Prompts

Additionally, we investigate the influence of text prompts on our algorithm, as presented in Table S1. Notably, compared with exemplar selection strategies, variations in prompt designs exert minimal impact on the results. We use object-specific descriptions for exemplar completion and test three prompt settings for subsequent frame completion:

- $\mathbb{T}$-I: Minimal Prompt – Only use "main object" as the completion prompt;
- $\mathbb{T}$-II: Detailed Prompt – "A quadriptych sharing the same object: " + object_description + "with the left two panels containing the object partially occluded by a masking object, and the right two panels showing the complete object without occlusion.";
- $\mathbb{T}$-III: Distracting Prompt – Randomly adopt object descriptions from unrelated sequences.

It can be observed that the performance of $\mathbb{T}$-I, $\mathbb{T}$-II, and our default prompt is nearly identical, while $\mathbb{T}$-III only exhibits a slight performance drop. This indicates that our method is robust to prompt variations, as our IC framework prioritizes visual cues from the exemplar over textual prompt information. This also aligns with human cognitive habits: once the complete appearance of an object is inferred, subsequent predictions can be made based solely on this visual impression.

E. Details for IC setting

Fig. S2 shows the visual illustrations for the discussion in Sec. 4.2 "IC Component Discussion." Consistent with the results in Tab. 3, the setting of IC inputs has a significant influence on the outputs. As shown in Fig. S2 (a), only providing object information without explicit guidance on the task, the model can not carry out the "incomplete-to-complete" transformation. Configuration P-II, as shown in Fig. S2 (b), attempts to provide the correct transformation logic by using the visible part of the object as the query and the completed object as the target. While it achieves better image quality compared to P-I, there is a significant deformation in the output, sacrificing the vital scene context by excluding the surrounding environment. The surrounding scene in the original frame serves as a crucial contextual anchor. Diffusion models rely on this comprehensive background information to determine the object's stable global position, scale, and environmental consistency across dynamically changing frames. Removing this context isolates the desired transformation but prevents the DiT backbone from stabilizing the object's features against background noise and movement. Consequently, when the visible region of the target object contains limited structural information, the resultant completions exhibit pronounced temporal instability and flickering.

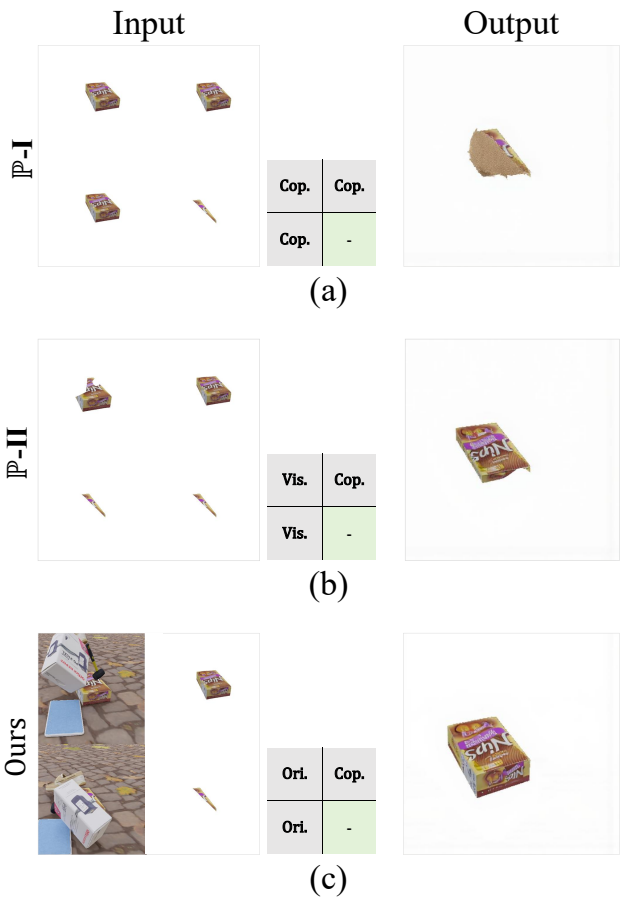


Figure S2. Visual results of Sec. 4.2 "IC Component Discussion".

F. Inpainting Mask Generation Pipeline

Following the paradigm of OW Amodal [1], the inpainting mask M_{ip} is generated through two sequential steps: generate all visible objects in the given images and analyze the spatial order between these objects. Proceed as follows, for more details please refer to [1].

1). We first applied RAM to the input image X to extract all recognizable visual concepts in X, as a set of tags $T = \{t_1, t_2, ..., t_M\}$, where t_m denotes the class label for the recognizable visual concept m_{th}. Then GroundingDINO [6] combined with SAM [4] is used to generate segmentation masks for all concepts in T, denoted as $S = \{s_1, s_2, ..., s_M\}$.

2). We analyze the spatial order between masks in S and the target visible mask M using InstaOrderNet [5]. For each s_i in S, we adopt InstaOrderNet to compute the binary occlusion indicator, assigning a value of $occ_i = 1$ if the segment occludes the target (masked by M) and 0 otherwise. Then we aggregate the masks in S that are spatially in front of M to obtain the final inpainting mask M_{ip} as:

$$M_{\text{occ}} = \bigcup_{\left(\substack{i=1 \\ \text{occ}_i=1}\right)}^{m} s_i. \tag{1}$$

G. SOTA Method Comparison Settings

Following TACO [7], we resize all results to 256×256 resolution when calculating the metrics. For the E^{2}FGVI baseline, to avoid the severely degraded performance when masking out all but the visible object area, we change the background to white and then use the same dilated bounding box of the ground-truth (GT) amodal mask as the inpainting mask. For PSNR, SSIM, and LPIPS values, since a significant portion of the synthesized frames is white, amodal object is cropped using a dilated bounding box of the GT amodal mask and compute the PSNR, SSIM, and LPIPS metrics within this specific region.

H. Details for Using Nano for IC

We first explored the viability of using a powerful commercial large image model, Nano Banana, for direct amodal completion tasks. Given the model's demonstrated capabilities in complex image understanding and editing, it was necessary to investigate whether such generalized pre-trained models could intrinsically handle the amodal challenge, thereby validating the fundamental premise of utilizing open-world priors for occluded appearance synthesis. However, practical application revealed significant limitations in adapting these tools to the inherently ill-posed nature of amodal completion tasks. As illustrated in Fig. S9 (a), we show multiple output results for a single input, revealing that when directly inputting only the visible region, the model produced relevant and plausible completion results with a certain probability only when the occluded object possessed a substantial visible area. This sporadic success reaffirms the feasibility of leveraging large pre-trained models for amodal completion. However, the model, heavily influenced by its prior knowledge, often generated objects visually inconsistent with the provided visible region. Conversely, completion for small visible regions (representing severe occlusion) failed completely. We further explored providing explicit scene context, feeding the full scene image alongside the visible part, as shown in Fig. S9 (b). While this approach slightly increased the consistency between the completed object and the visible boundaries for larger objects, the failure to complete heavily occluded (small visible area) instances—often resulting in the model outputting the unmodified scene image or generating irrelevant content—demonstrates the model's inability to consistently perceive the amodal task even with explicit context. Finally, we attempted a collaborative prompting strategy akin to our proposed in-context learning, providing two input images simultaneously to guide the completion, as depicted in Fig. S9 (c). Under this setup, the success rate for completing small-area objects improved, confirming the potential of the ICL paradigm for transferring the "incomplete-to-complete" transformation logic. Nevertheless, a pronounced flaw emerged: completions, especially for smaller occluded regions, suffered from severe spatial position inconsistency. This confirms that while generalized diffusion models retain rich appearance priors, their vanilla attention mechanisms fail to semantically anchor the completion structurally and temporally, particularly when distinguishing task-critical cues from irrelevant scene context, demanding explicit guidance mechanisms for robust VAC.

References

[1] Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A Ehinger. Open-world amodal appearance completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6490–6499, 2025. 1, 3

[2] Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using diffusion priors for video amodal segmentation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 22890–22900, 2025. 1

[3] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 436–454. Springer, 2020. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment

Figure S3. Qualitative comparisons on TAO.

anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[5] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21210–21221, 2022. 3

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3

[7] Ruijie Lu, Yixin Chen, Yu Liu, Jiaxiang Tang, Junfeng Ni, Diwen Wan, Gang Zeng, and Siyuan Huang. Taco:

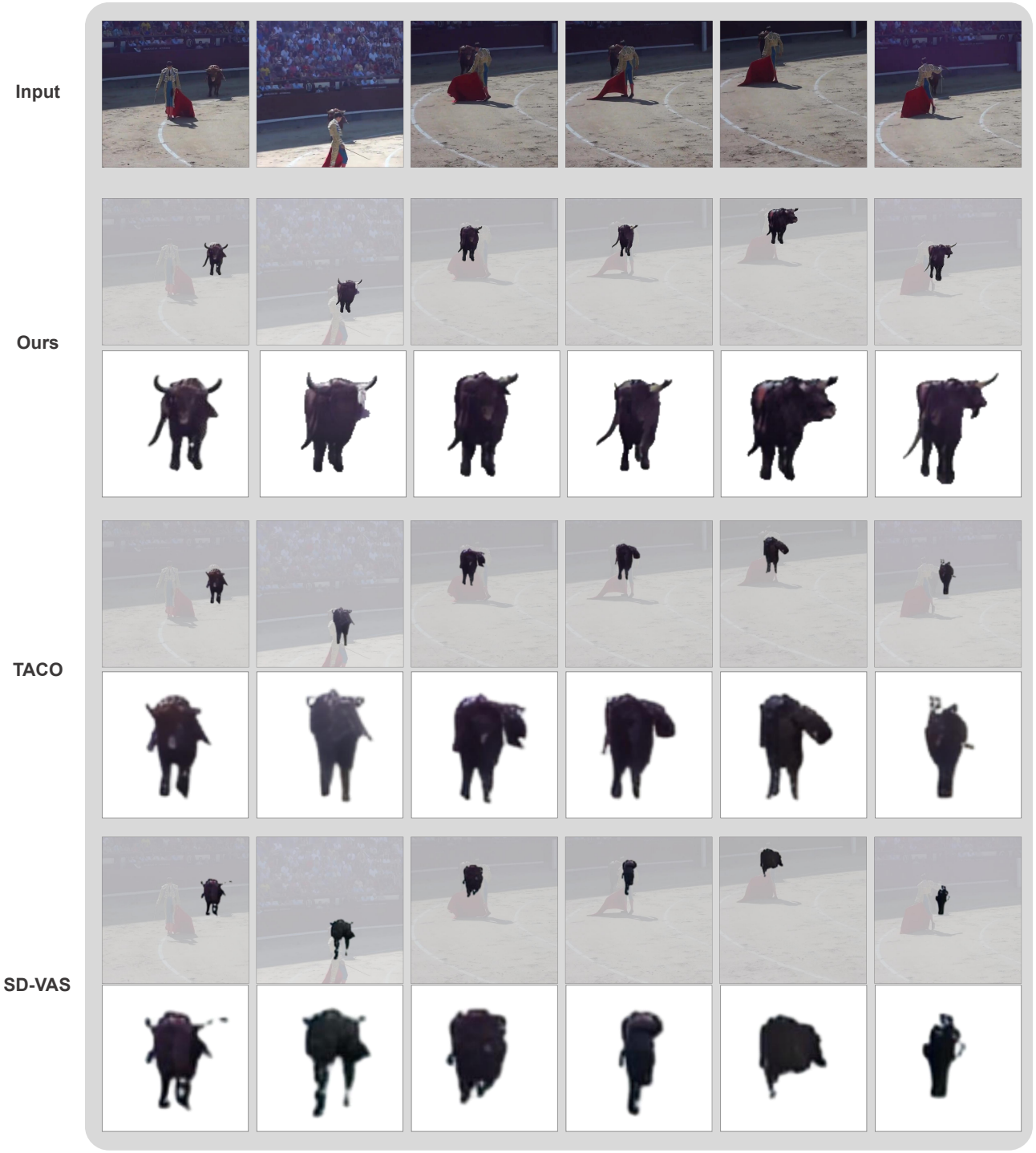


Figure S4. Qualitative comparisons on TAO.

Taming diffusion for in-the-wild video amodal completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13650, 2025. 1, 3

Figure S5. Qualitative comparisons on TAO.

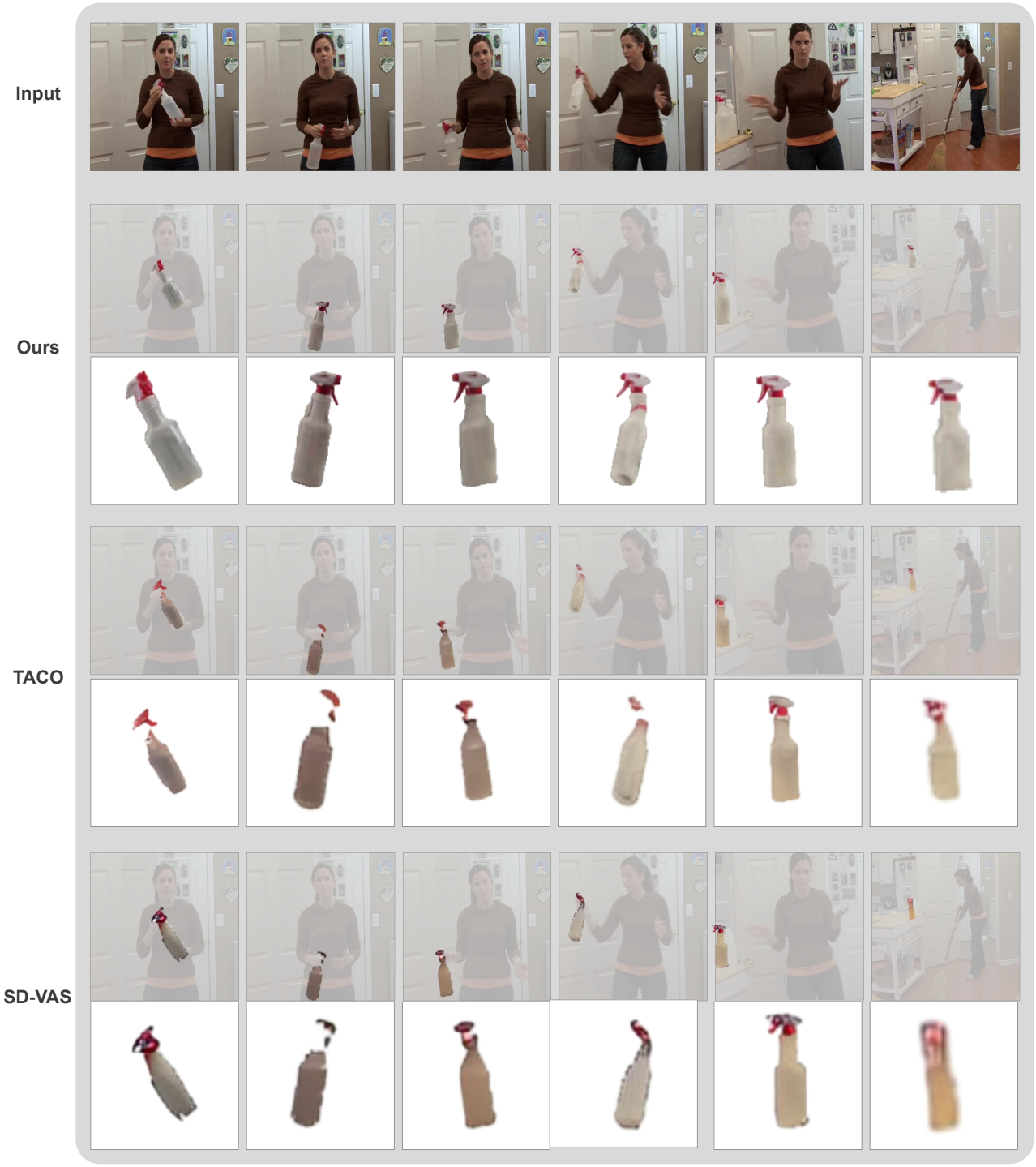


Figure S6. Qualitative comparisons on TAO.

Figure S7. Qualitative comparisons on TAO.

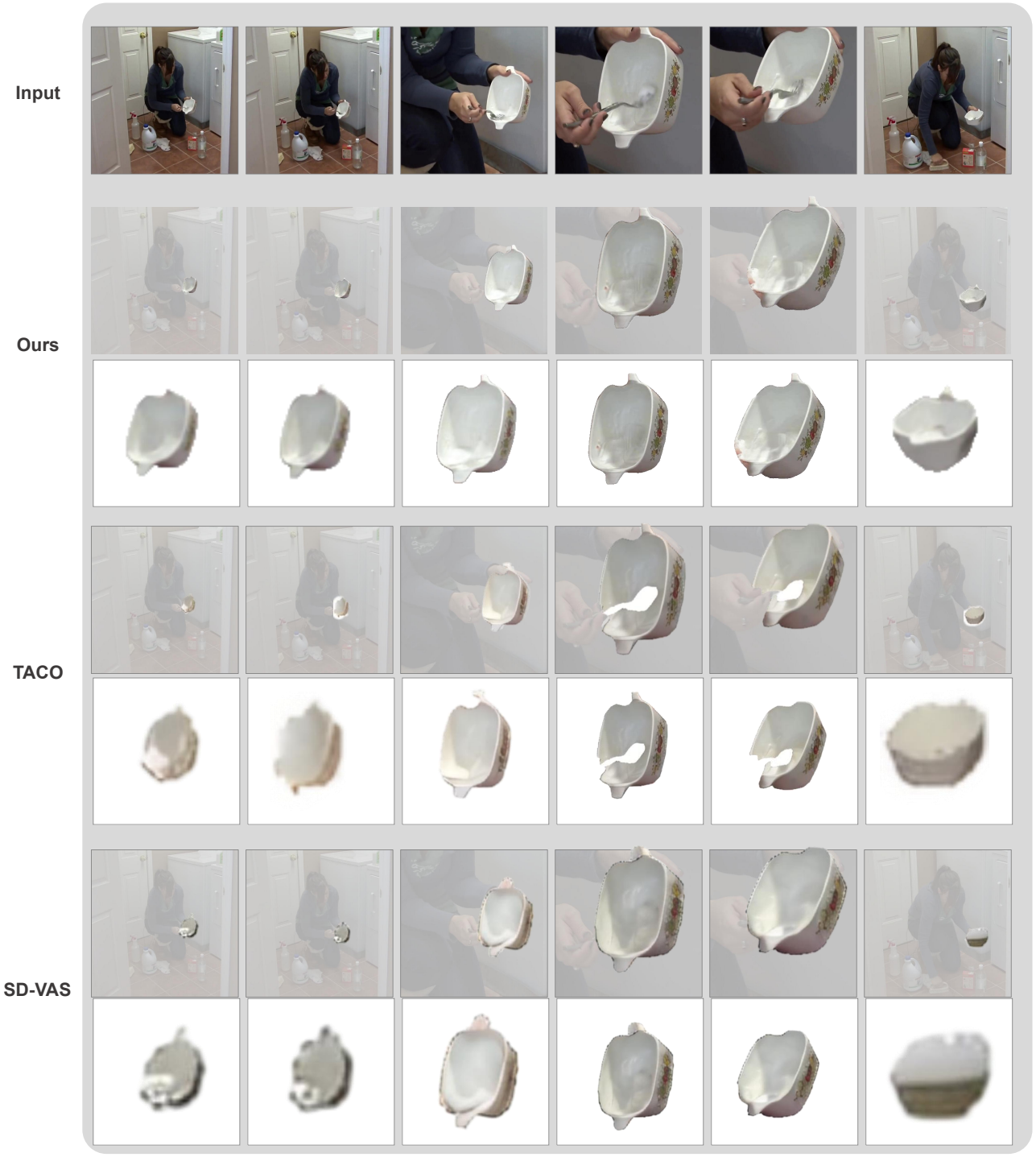
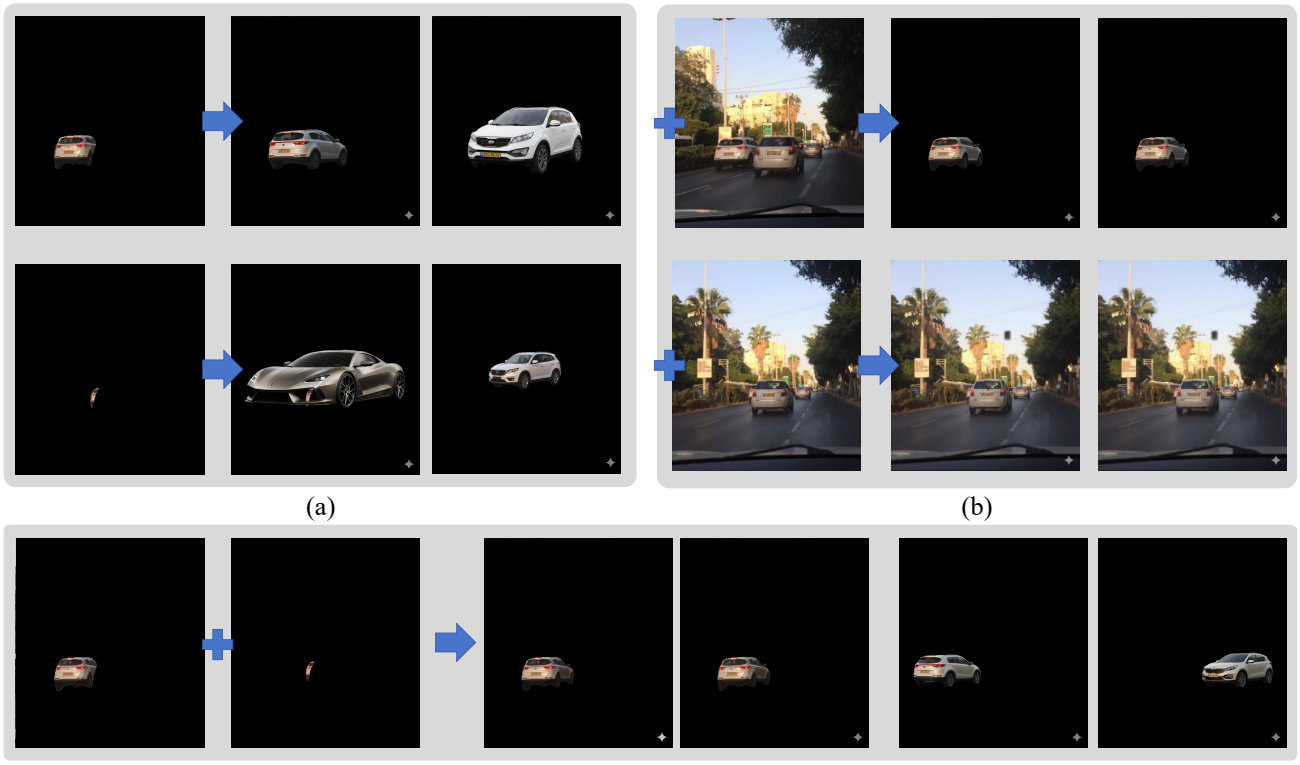


Figure S8. Qualitative comparisons on TAO.



(a) (b)

(c)

Figure S9. Details for Using Nano for IC.