GenHOI: Towards Object-Consistent Hand–Object Interaction with Temporally Balanced and Spatially Selective Object Injection

Supplementary Material

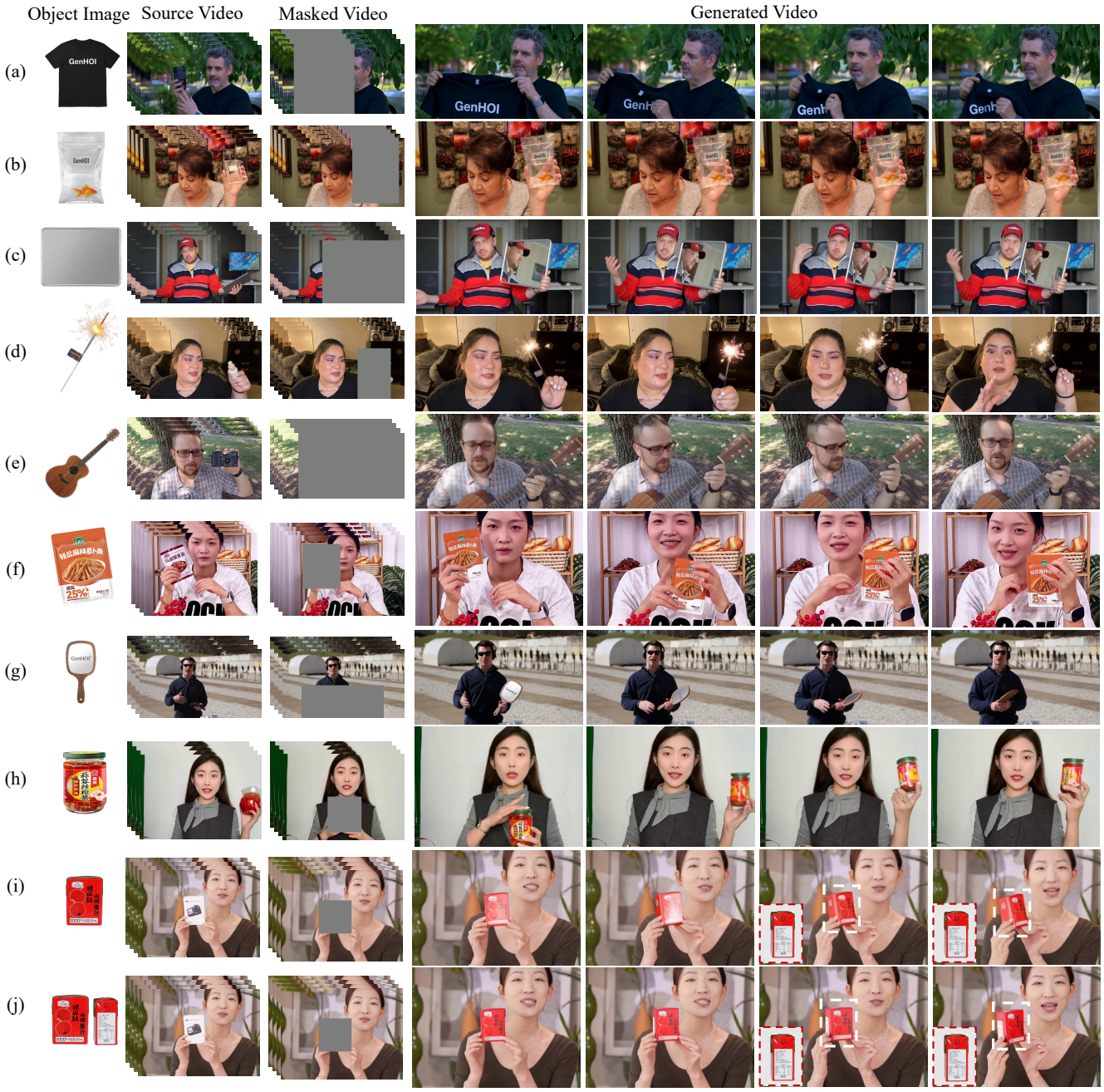

Figure 1. Additional qualitative results including deformable objects (a,b), dynamic physics (b,c,d), complex interactions (e,f), robustness to occlusions during generation (f), and object rotation (g-j). Better viewed by zooming in.

:

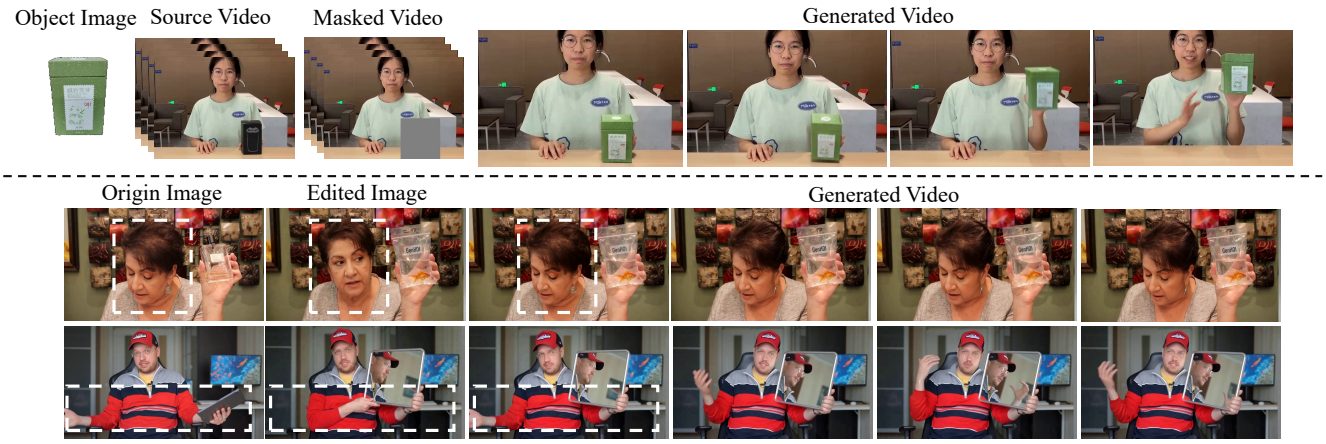

Figure 2. Top: Cross-reenactment results without an edited first frame, showing that our method remains robust and preserves object identity using only the reference image. Bottom: Robustness to first-frame edits—despite pose or position shifts in the edited first frame, the generated videos maintain consistent object appearance and interaction.

1. More Results and Analysis

1.1. Challenging Cases

Fig. 1 presents additional manipulation results beyond the AnchorCrafter evaluation dataset, including deformable objects (Fig. 1 (a,b)), dynamic physics (Fig. 1 (b,c,d)), complex interactions (Fig. 1 (e,f)), robustness to occlusions during generation (Fig. 1 (f)), and object rotation (Fig. 1 (g-j)). These cases highlight that our model maintains stable hand–object contact and preserves object identity under large deformation and dynamic motion. The results also indicate robustness to partial occlusions, where the object remains coherent across frames. It also handles viewpoint changes during rotation well, preserving the overall geometry even for unseen views.

1.2. Multi-view Condition & rotation

The proposed method can generate unseen object views using single-image condition, making it more user friendly. Fig. 1 (g,h) demonstrates its behavior under rotation, where the results are plausible and preserve object category and geometry with single object image. However, generating unseen views remains dependent on the model's generative priors, which may lead to texture inconsistencies with the target object. For example, in Fig. 1 (i), the synthesized textures in unseen views (white dashed boxes) differ from those of the original object (red dashed boxes). This limitation can be alleviated by introducing multi-reference conditioning, where multi-view inputs significantly improve texture consistency (Fig. 1 (j)).

1.3. Influence of first-frame editing quality.

The edited first frame mainly serves as auxiliary information, providing a scale cue for the object. This is because our inpainting-based design and the proposed Head-Sliding mechanism extract texture features directly from the object reference image rather than the edited first frame. Consequently, our method can generate reasonable videos **even without an edited first frame**, as shown in Fig. 2 (top). For the same reason, the final videos are robust to pose or position shifts introduced by the initial edit, as shown in Fig. 2 (bottom).

1.4. Mask boundary influence.

Our method is also robust to variations in mask precision. As shown in Fig.1(d,e), both precise and coarse masks produce high-quality results. Larger masks generally induce greater changes from the original video, but the mask should at least cover the original hand–object interaction region.

2. Limitation

Although GenHOI demonstrates strong performance in HOI scenarios, it still exhibits certain limitations that call for further study and improvement.

- In the first–last frame mode, the quality of the generated video largely depends on the physical plausibility of the given first and last frames. Existing image editing models struggle to guarantee this, especially when the hand pose in the first frame needs to transition naturally to the pose in the last frame within a very short time.
- For challenging manipulations or non-rigid objects, the output quality may vary across different runs, reflecting the inherent ambiguity of these scenarios. How to consistently produce satisfactory results in such complex cases remains an open problem.