DirectDrag: High-Fidelity, Mask-Free, Prompt-Free Drag-based Image Editing via Readout-Guided Feature Alignment

Supplementary Material

1. Supplementary Material

1.1. Improper Mask and Prompt

Figure 1 illustrates how poorly drawn mask or irrelevant prompt can cause distortion or semantic failure. Designing such inputs is often nontrivial and error-prone. This motivates our manual mask-free and prompt-free approach.

1.2. Unexpected Benefits Without Mask and Prompt

In rare cases, removing the mask or prompt surprisingly improves results. Figure 2 shows edits that are better or more natural without mask and prompt. This experiment is conducted using the GoodDrag [3] method.

1.3. Readout Network Architecture

Our readout module directly follows the architecture introduced in Readout Guidance [1] (see Figure 3). We use the same aggregation network to extract intermediate features from the decoder, where each decoder feature is first passed through a bottleneck layer to standardize the channel size. These bottleneck layers are made timestep-conditional by adding projected timestep embeddings, obtained from the pretrained U-Net's timestep encoding. The standardized features are then aggregated via a learned weighted sum. We adopt this architecture without structural changes, using it as a guidance signal for feature alignment in our drag-editing task.

1.4. Comparison with InstantDrag

See Figure 4 for visual comparisons with InstantDrag [2].

1.5. Extended Qualitative Comparison

We present additional comparisons with prior drag-based methods to highlight differences in fidelity and accuracy. See Figure 5 and Figure 6.

1.6. Extended Qualitative Examples

We showcase more qualitative results produced by DirectDrag across diverse scenes and manipulation tasks. See Figure 8.

1.7. Qualitative Results of the Ablation Study

To better understand the role of each component, we visualize editing results under different ablation settings. See Figure 9 and Figure 10.

1.8. Limitations and Example

In some cases, our method may over-preserve visual fidelity, resulting in insufficient deformation. Additionally, strong geometric warping can occasionally cause texture detail loss. Furthermore, our method is still less aligned with human intent compared to manual dragging. See Figure 7.

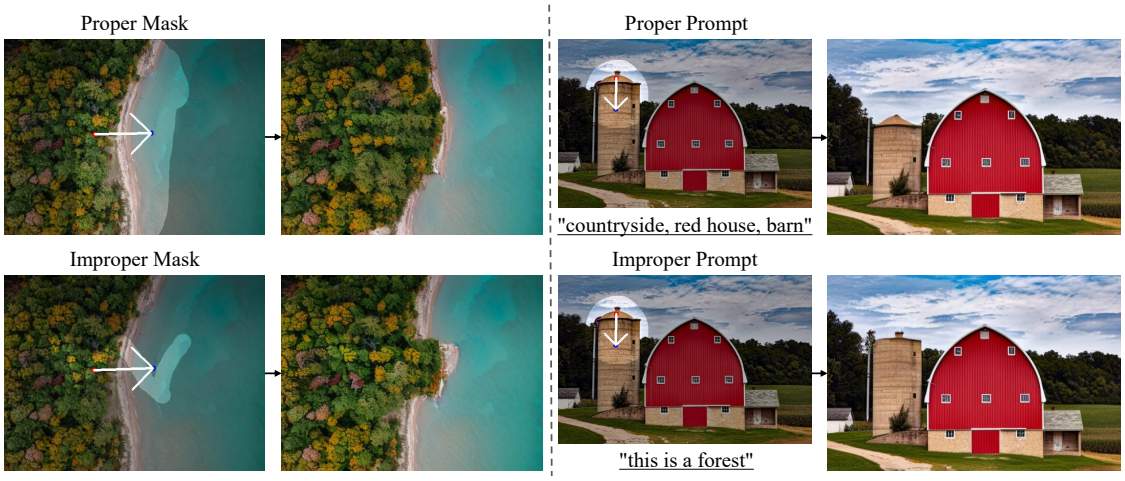

Figure 1. **Improper Mask and Prompt**

Figure 2. **Unexpected Benefits Without Mask and Prompt**

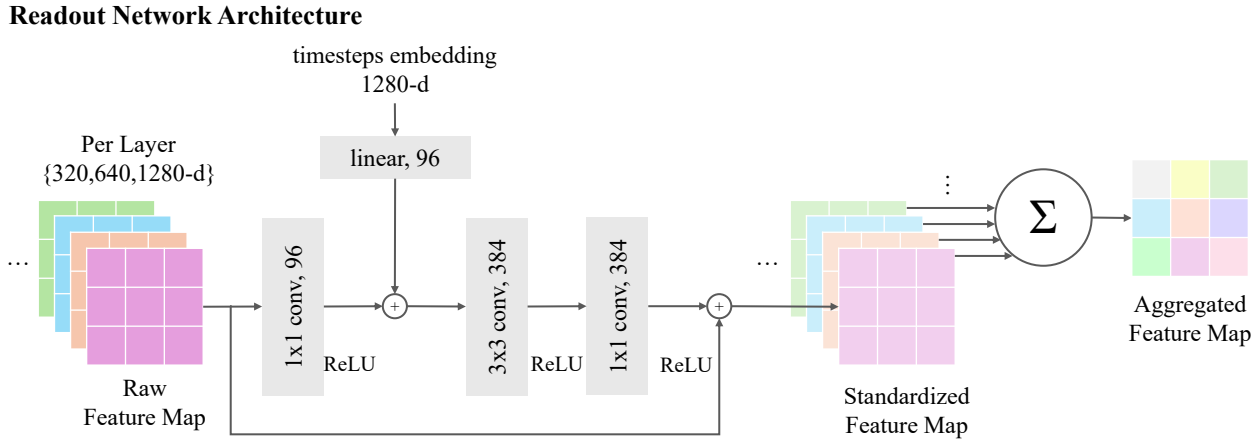

Figure 3. **Readout Network Architecture**

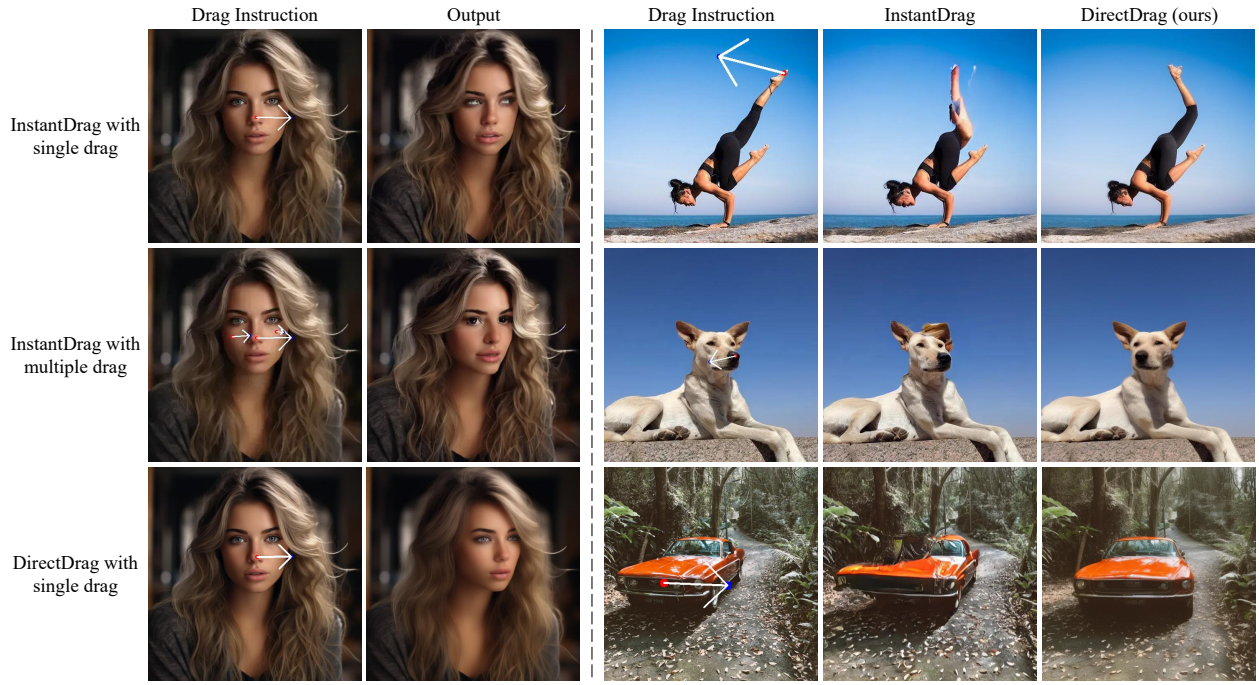

Figure 4. **Comparison with InstantDrag [2]** Left: InstantDrag requires multiple drag instructions to rotate the face, while DirectDrag achieves similar results with a single drag instruction. Right: InstantDrag often produces unstable or distorted results, while our method yields more faithful and coherent outputs.

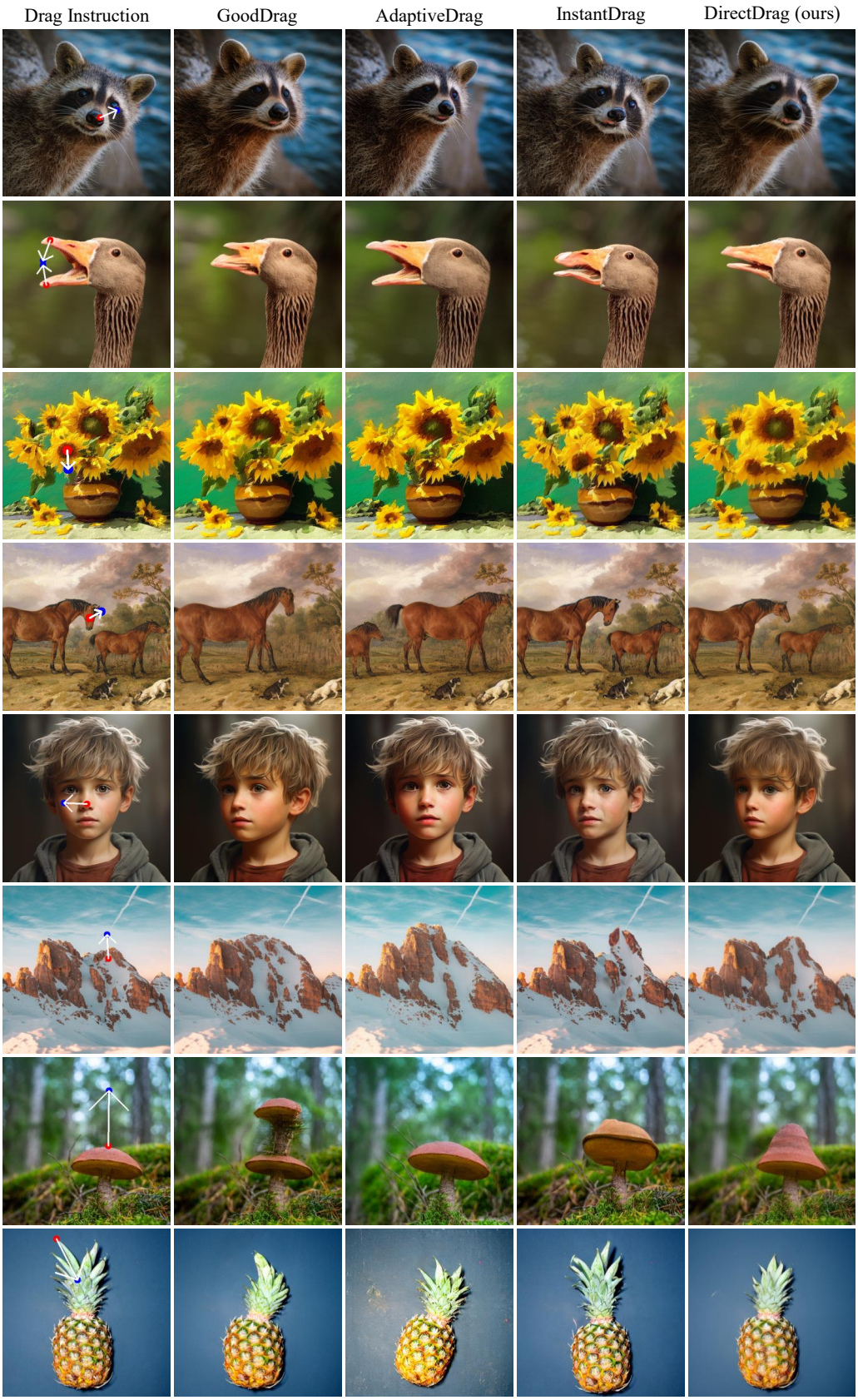

Figure 5. **Extended Qualitative Comparison**

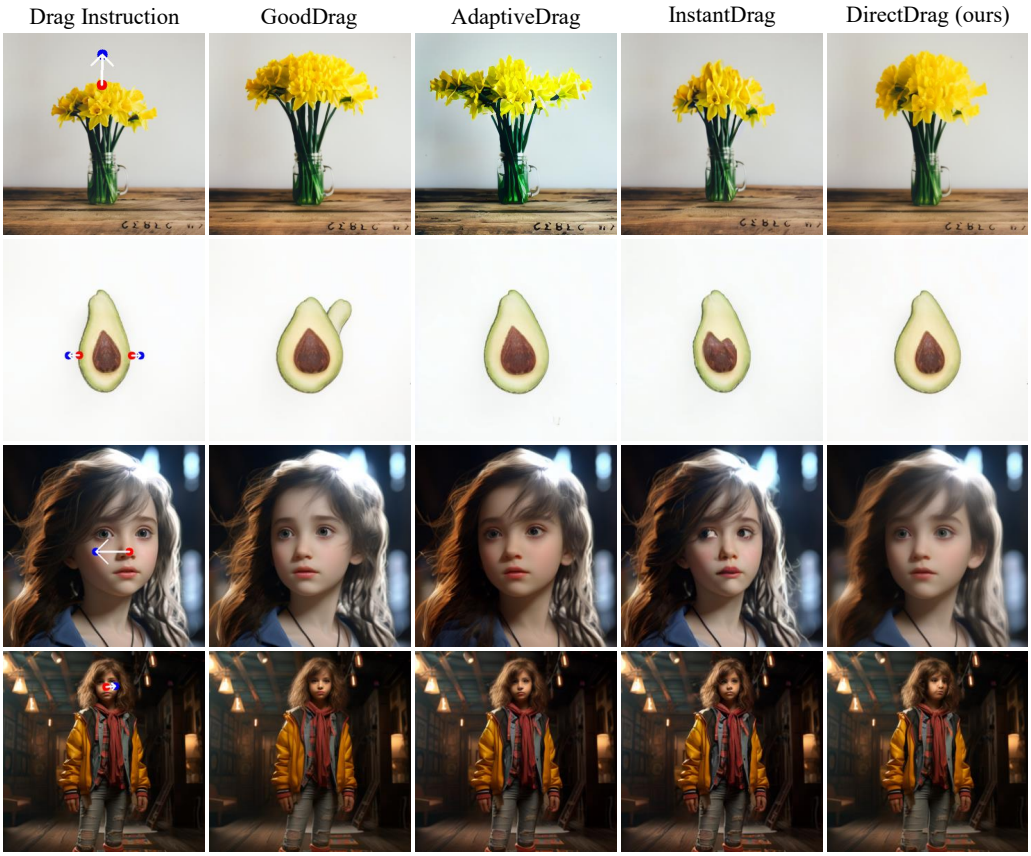

Figure 6. **Extended Qualitative Comparison**

Figure 7. **Qualitative Results of Limitations**

Figure 8. **Extended Qualitative Examples**

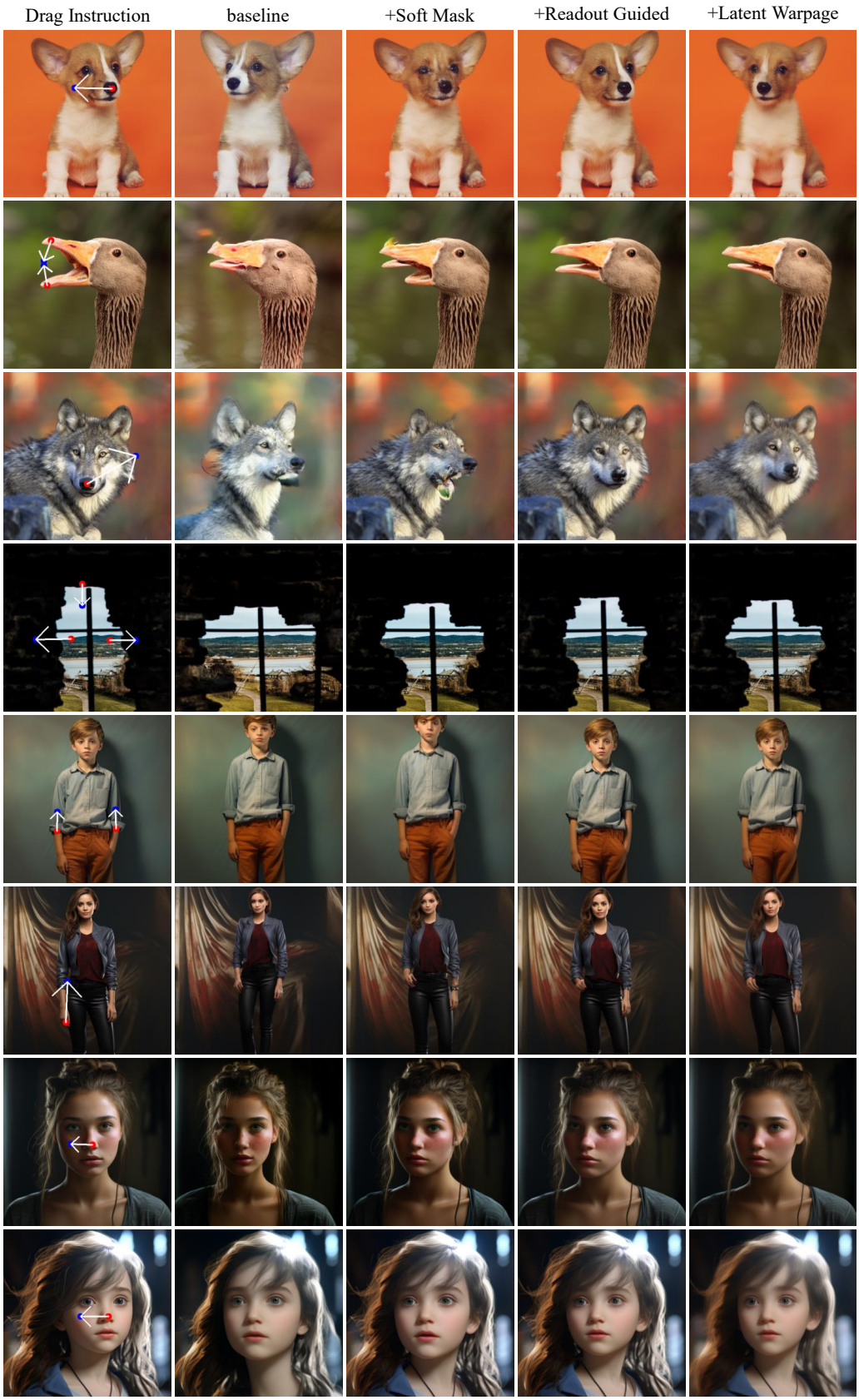

Figure 9. **Qualitative Results of the Ablation Study**

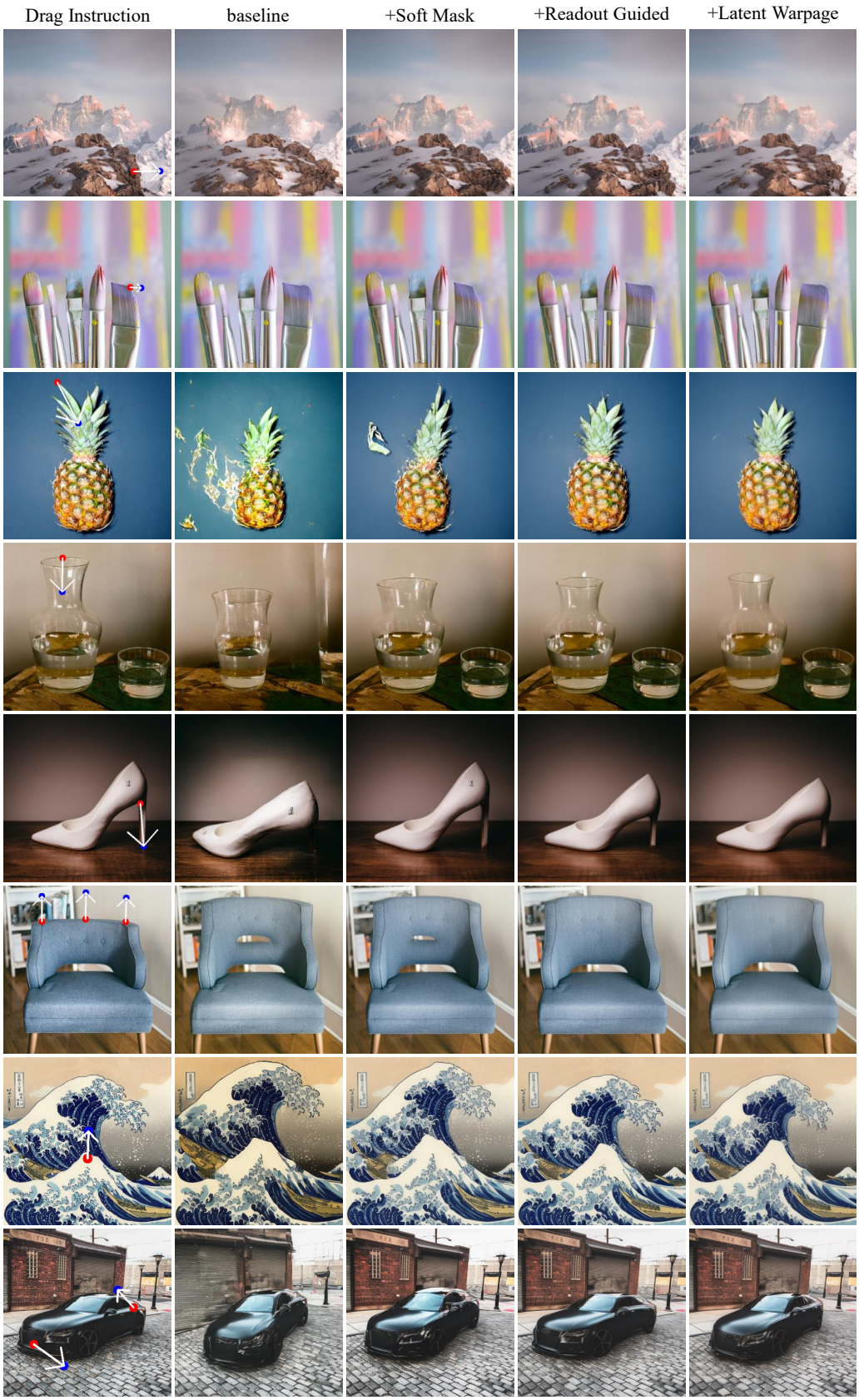

Figure 10. **Qualitative Results of the Ablation Study**

References

[1] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 1

[2] Joonghyuk Shin, Daehyeon Choi, and Jaesik Park. Instantdrag: Improving interactivity in drag-based image editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 1, 3

[3] Zewei Zhang, Huan Liu, Jun Chen, and Xiangyu Xu. Gooddrag: Towards good practices for drag editing with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. 1