Unveil Inversion and Invariance in Flow Transformer for Versatile Image Editing

Supplementary Material

The supplementary is organized as follows:

- Additional detailed related work: 1. text-to-image flow-based models. 2. invariance control in diffusion-based image editing.
- Analysis of approximation error in Euler inversion.
- Versatile image editing with AdaLN invariance control.
- Investigation of attention-based invariance control in flow transformer.
- Additional non-rigid editing results.
- Failure case study.

8. Additional related work

We present the detailed discussion as the extension for the related work in the main text.

Text-to-Image Flow-based models. Flow-based models [1, 22, 23] interpolate the probability transition path between two data distributions via the ordinary differential equation (ODE) and learn the conditional velocity field of the transition path. Such a formulation also implies the diffusion models that use Gaussian probability paths. The advantage over diffusion models is to allow faster simulation of the probability flow ODEs, which induces fewer sampling steps. Later works further rectify and optimize the non-optimal transition paths [14, 25, 34, 48]. Driven by these theoretical benefits, recent large-scale text-to-image models such as Stable Diffusion 3 [6] and Flux implement flow matching with the diffusion transformer architecture (DiT) [32] and achieve new state-of-the-art text-to-image synthesis. However, recent image editing (especially tuning-free editing) approaches are mainly based on diffusion and U-Net [36] whereas flow matching and transformer lack exploration. This paper investigates the flow-transformer models as the foundation for tuning-free image editing. We analyze flow inversion and image invariance control based on transformer architecture in editing.

Invariance control in diffusion-based image editing. The invariance control preserves the original image's unedited contents in diffusion-based image editing. The instruction-based methods explicitly add the original image features as the condition and retrain the T2I model into a text-guided image-to-image (TI2I) model. For example, the InsP2P [4] and InsDiffusion [8] concate the latent of the original image with the noisy latent. The IP-adapter [52] and T2I-adapter [29] add the extra branch to the U-Net to inject features into the U-Net of diffusion. In the tuning-free paradigm, P2P found injecting the cross-attention corresponding to the text prompt can main the unedited contents. Furthermore, P2P-zero [11] and Plug-and-Play [43] explore the self-attention to preserve the invariance during editing. However, these attention-based methods struggle with non-rigid editing such as changing the layout. MasaCtrl [5] proposed the mutual-self attention and copied the K and V in the later diffusion process to adapt to the layout change but deteriorates object and style change. Similarly, InfEdit [49] combines cross-attention and mutual-self attention to mitigate the deficiency of rigid and non-rigid editing. However, this may degrade the fidelity of edited images. The third category of the refinement approach [2, 19] filters out components of the predicted noise corresponding to the non-target regions to preserve the non-target regions. However, these method require careful selecting hyperparameters of the filter. Moreover, these approaches are mainly based on the diffusion U-Net models. To fully leverage the generation prior of flow transformer, it is necessary to develop more flexible invariance control system to reconcile both rigid and non-rigid editing types based on the flow transformer for high fidelity and versatile editing.

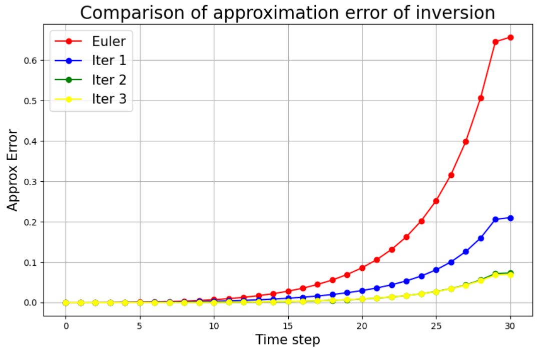

Figure 10. **Comparison of approximation error of inversion with fixed-point iteration at each time step**.

9. Analysis of approximation error of Euler

We compare the approximation error of the plain Euler inversion and the fixed-point iteration in Figure 10. Concretely, we calculate the MSE difference of two latents at the same time step in inversion $\mathbf{x}_t$ and reconstruction $\hat{\mathbf{x}}_t$. The result shows that the proposed inversion method can significantly reduce the approximation error at each time step during the reconstruction process. The numerical results also show that increasing the iteration number larger than 3 will lead to marginal improvement on the reconstruction quality. In practice, one iteration can significantly reduce the inversion error.

10. Versatile image editing with AdaLN invariance control

To validate the adaptability of our method on versatile editing types, especially for non-rigid editing, we categorize the editing results into 12 different editing types. For all editing experiments, we set the timestep S for AdaLN feature replacement as 1 and sampling steps as 30, and summarize the editing results in Figure 13 and Figure 15. The results show that our method can reconcile both rigid and non-rigid editing, fully taking advantage of the powerful generation ability of the flow transformer.

Concretely, our method supports the fine control for **visual text editing**. The editing results show that our method can manipulate the letter-level visual text such as 'cross' to 'crown'. This also validates that the feature replacement within AdaLN can discriminately connect the changed text to changed image semantics which is better than the non-discriminative self-attention replacement mechanism. As for the other non-rigid editings, we show results on **facial attributes, shape, pose, and quantity**. Such editing types change the object structure and layout in the image and demand the invariance control mechanism to have a more precise and flexible ability to manipulate the image semantics. Our method shows accurate and flexible editing results on these non-rigid editing types. For other rigid editing types, our method also shows good generalization performance and can support different levels of geometric changes. For example, our method can distinguish the foreground and background, replace the background with other contents in **Background change**, and change the whole image style in **Style**. Our method can also replace small-sized objects, such as 'torch to flowers' and 'add angels' in **Object replacement**. In conclusion, the proposed AdaLN invariance control mechanism can support versatile editing types.

11. Study of attention-based invariance control

In comparison to the proposed invariance control based on AdaLN, we also show investigation results based on the attention replacement in Figure 11. Since there is only self-attention in MM-DiT and the text and image features are processed in the attention jointly, we test different strategies of injecting Q, K, and V values of text and image features, instead of the attention map used in P2P [11] and MasaCtrl [5]. The conclusions are similar to the properties of self-attention in DM models. For image values, injection of the Q, K, and V values from the original image can preserve the contents of the original image but injecting attention values in more steps from the original image will make the edited image overwritten by the original image, which hinders the editing. In the Figure 11, we show that injecting the V or Q values more 20% time steps will hinder the non-rigid editing 'lying to standing', and make the edited image the same as the original image. Injection of K values does not inject the invariant contents of the original image to the target image.

For text values, the results of injecting TXT Q, K, and V do not follow the color and structure pattern. This shows that the individual injection of the Q, K, and V values from the text features cannot effectively preserve the structure of the original image and thus are not enough for invariance control. So, we further test the injection of the combination of KV, QK, and QV values. The results show that the injection of KV features may overwrites the edited image with the original image contents and hinders the non-rigid editing. The injection of QK does not hinder the editing but the edited image fidelity degrades. The injection of QV does not effectively control the invariance, and the edited images are very different from the original images. In conclusion, the attention-based invariance control in MM-DiT has similar properties in U-Net. It is not an effective tool for control invariance for non-rigid editing and overly injecting can influence the editing effect. Inappropriate injection position may degrade the fidelity.

12. Additional non-rigid editing results

To further validate the advantages of our method over non-rigid editing, we select 40 images for pose editing in the PIE dataset. The quantitative results are summarized in Table 3. Our method outperforms others in CLIP similarity indicating that our method has the better ability for non-rigid editing. The attention-based methods show a better ability to preserve the contents but cause a lower CLIP score, indicating that their editing abilities are hindered. This is because the images are not edited and remain as the before-images for non-rigid editing types. For example, in the case of '`bird to X`' in Figure 7. This confirms our argument that attention-based invariance control can hinder non-rigid editing ability. Besides, we also recall that in the full PIE benchmark Table 1, our method outperforms InfEdit in PSNR, LPIPS, MSE, SSIM, and Distance.

In contrast, PnP takes the second place in non-rigid editing but the preservation is much worse than ours. The PSNR, MSE, and SSIM are obviously worse than our method. In conclusion, our method can do good non-rigid editing while maintaining a reasonable ability to preserve the background contents.

13. Failure case study

We present the failure case study to better understand the limitations of our framework. So far, we already demonstrated the superior advantages of the invariance control based on the AdaLN. However, as discussed in the Limitation section, if the real image is not within the prior distribution of the model, the inversion is far from the authentic

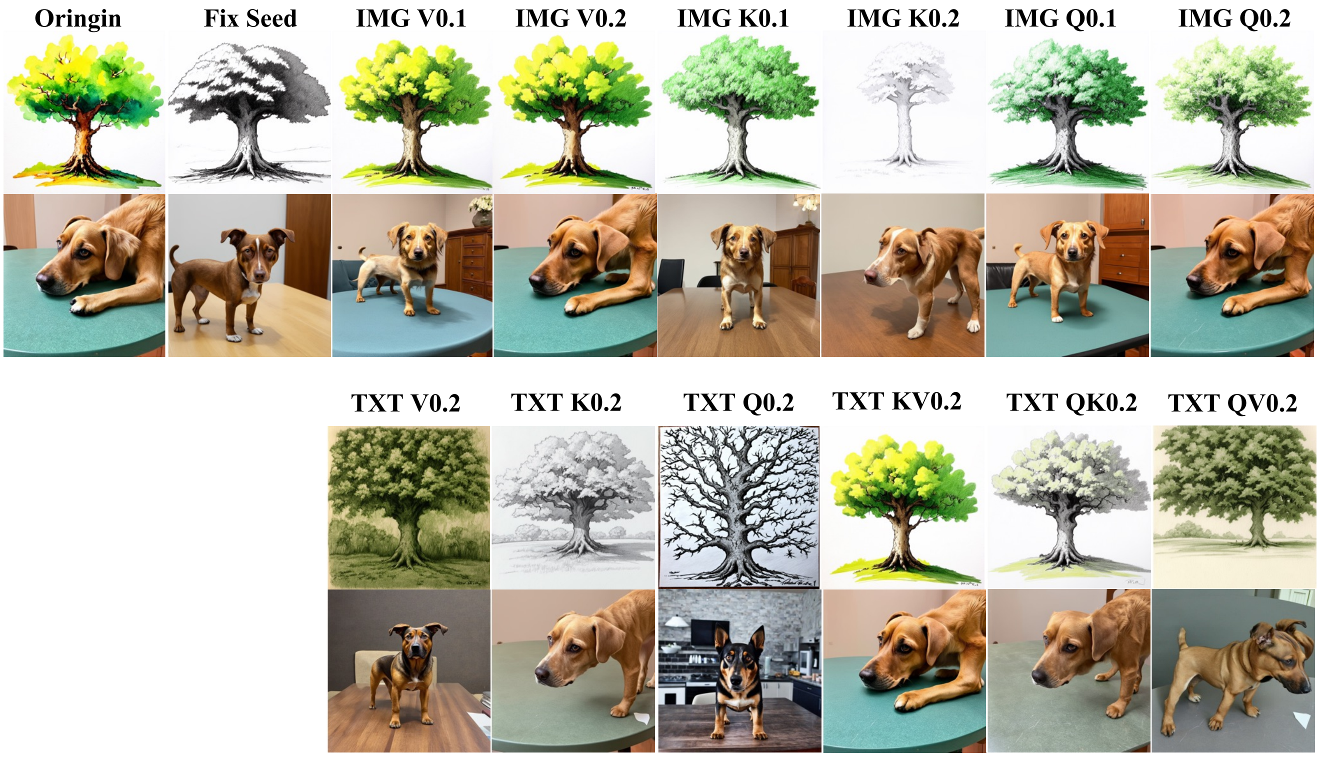

Figure 11. **Investigation of attention-based invariance control in MM-DiT**. Fix seed is the generation results with the same seed but without any invariance control. We inject the Q, K, and V values of text (TXT) and image (IMG) features with different time steps to evaluate the invariance preservation ability. For image features, after the 20% (IMG0.2) time steps, the injection makes the edited image the same as the original image. For text features, injection Q, K, and V more than 20% (TXT0.2) time steps do not effectively control the invariance. We further test the combination injection of KV, QK, and QV.

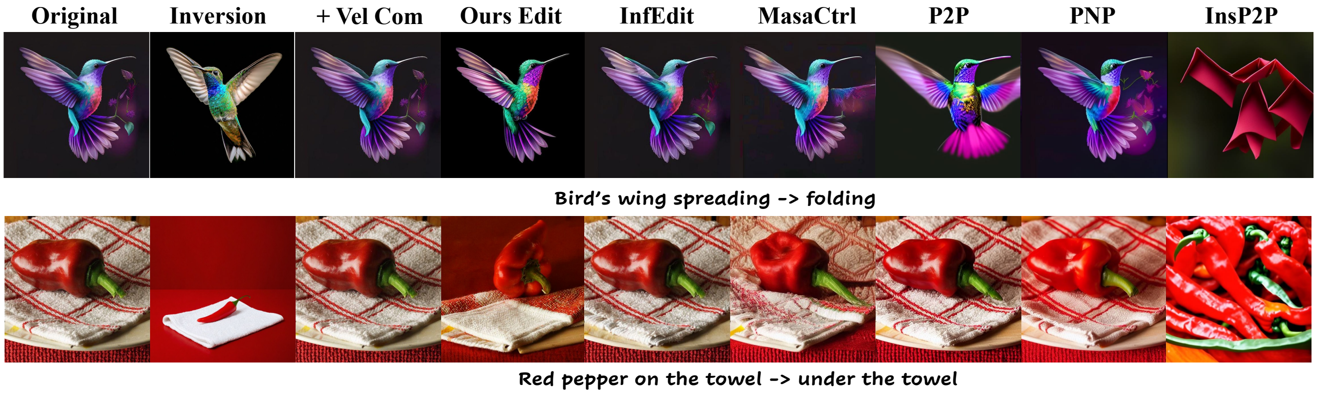
Original
Inversion
+ Vel Com
Ours Edit
InfEdit
MasaCtrl
P2P
PNP
InsP2P
Bird's wing spreading -> folding
Red pepper on the towel -> under the towel

Figure 12. **Failure case study**. We present two failure cases whose inversions are far away from the original image. The inversions presented are processed with 1 fixed-point iteration. Even though the velocity compensation can recover the original image, the editing still fails.

generation process and is quite different from the original image. Thus, the text-to-image alignment based on the reconstructed inversion trajectory is mismatched. In this case, changing the text prompt may also cause changes in non-target regions. We show the case in Figure 12, and the result shows that if the inversion (without the velocity compensation) seriously deviates from the original image, the output image also cannot be properly edited even if the image can be fully recovered with the velocity compensation. We also show the results of other methods, and none of the methods

Table 3. **Quantitative comparisons in non-rigid image editing.** Evaluated using the PIE benchmark. Different metrics are scaled.

	Structure	Background Preservation				CLIP Similarity	
Method	**Distance** ↓	**PSNR** ↑	**LPIPS** ↓	**MSE** ↓	**SSIM** ↑	**Whole** ↑	**Edited** ↑
PnP	20.63	22.76	116.36	79.66	77.58	26.64	22.82
P2P	9.42	26.63	59.28	32.94	83.64	26.56	22.43
MasaCtrl	20.53	22.47	91.84	89.11	79.79	26.59	22.38
InsP2P	53.38	20.82	165.33	243.23	72.18	23.35	20.13
InfEdit	**6.12**	**28.35**	**43.36**	**23.64**	85.35	25.98	22.18
Ours	20.39	24.02	103.73	53.70	**87.47**	**26.93**	**22.87**

successfully made the right edit since this image may not be within the domain of the model.

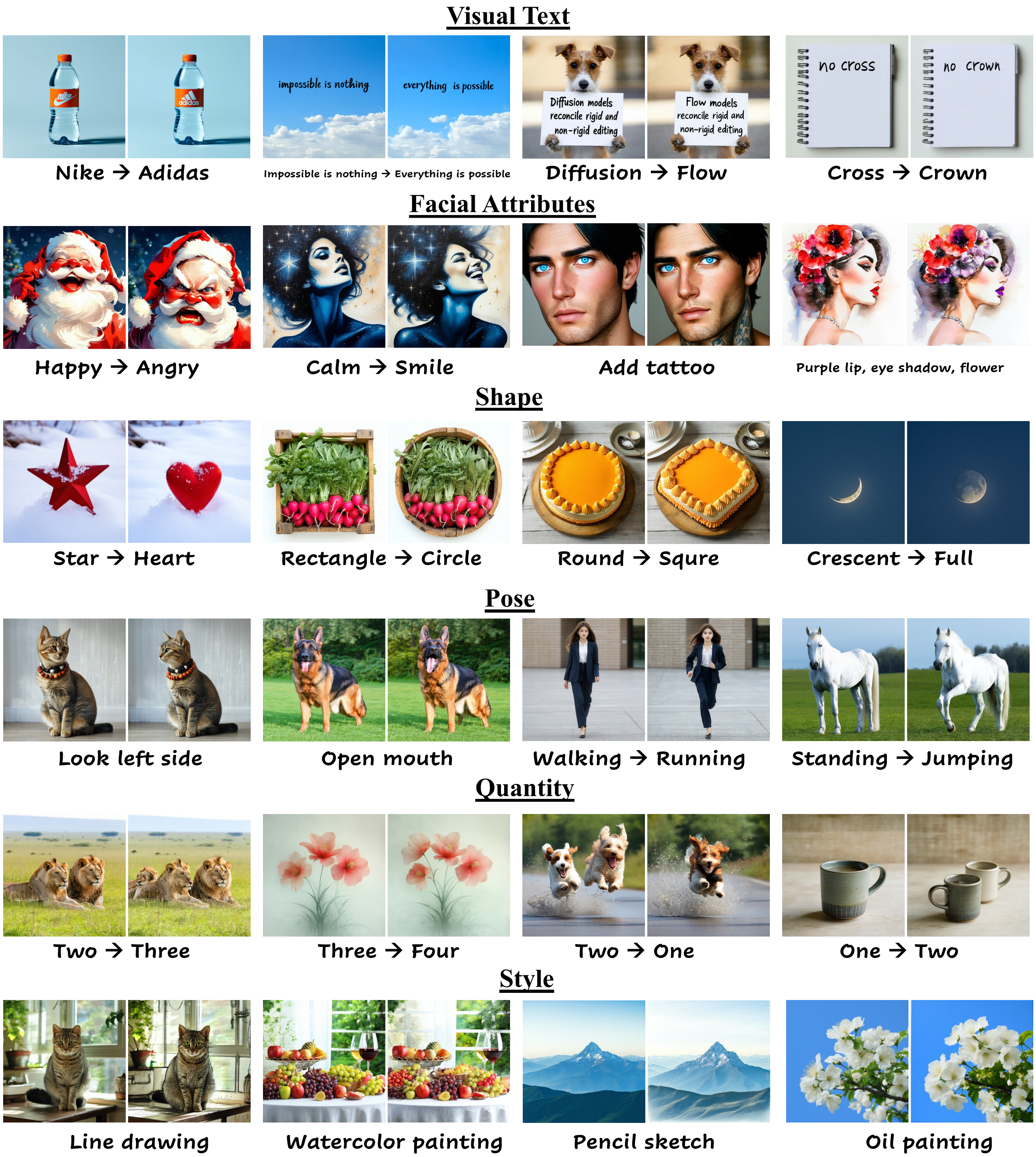

Figure 13. **Qualitative results on versatile editing types Part I**. Zoom in for details.

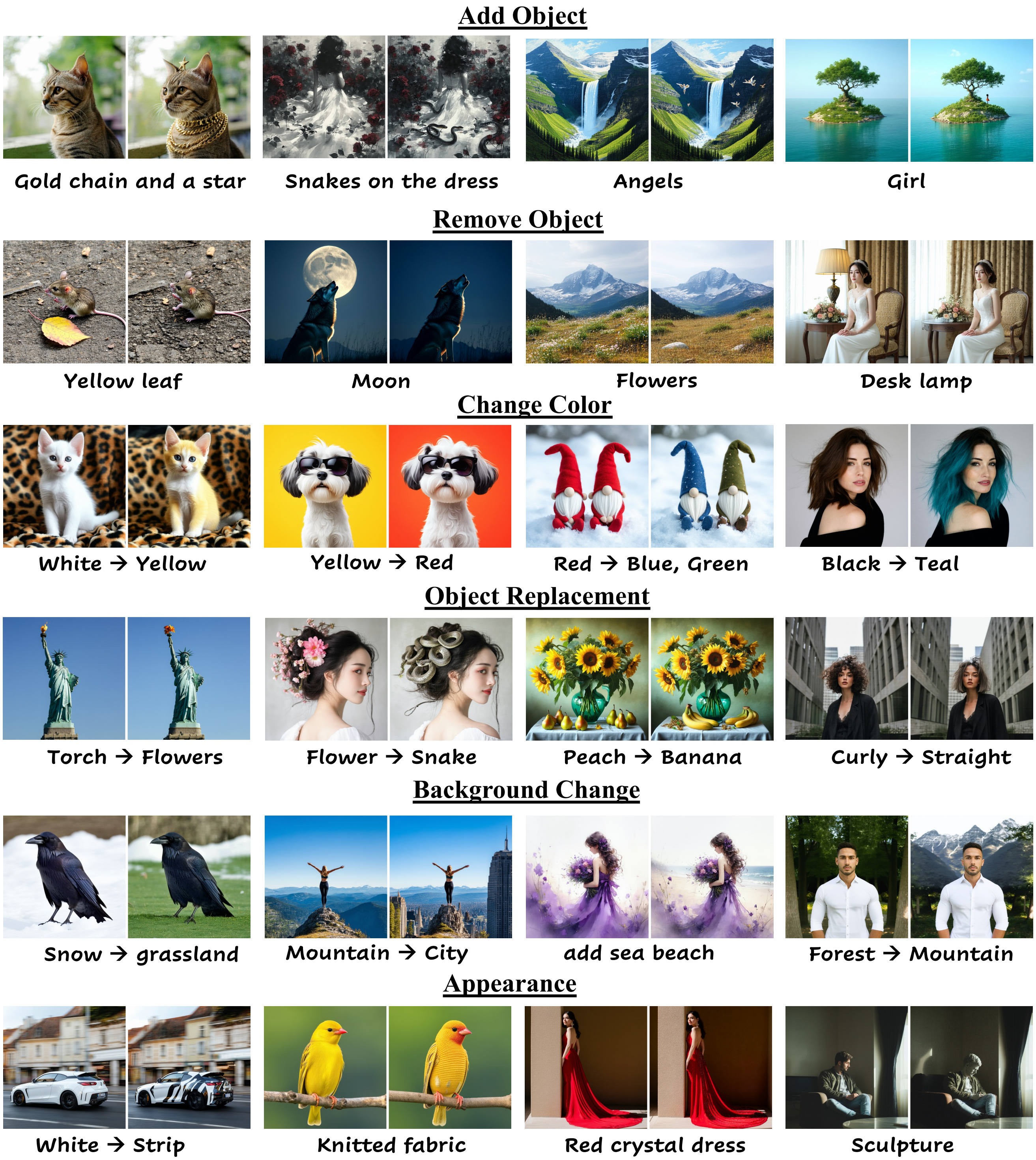

Figure 14. **Qualitative results on versatile editing types Part II**. Zoom in for details.

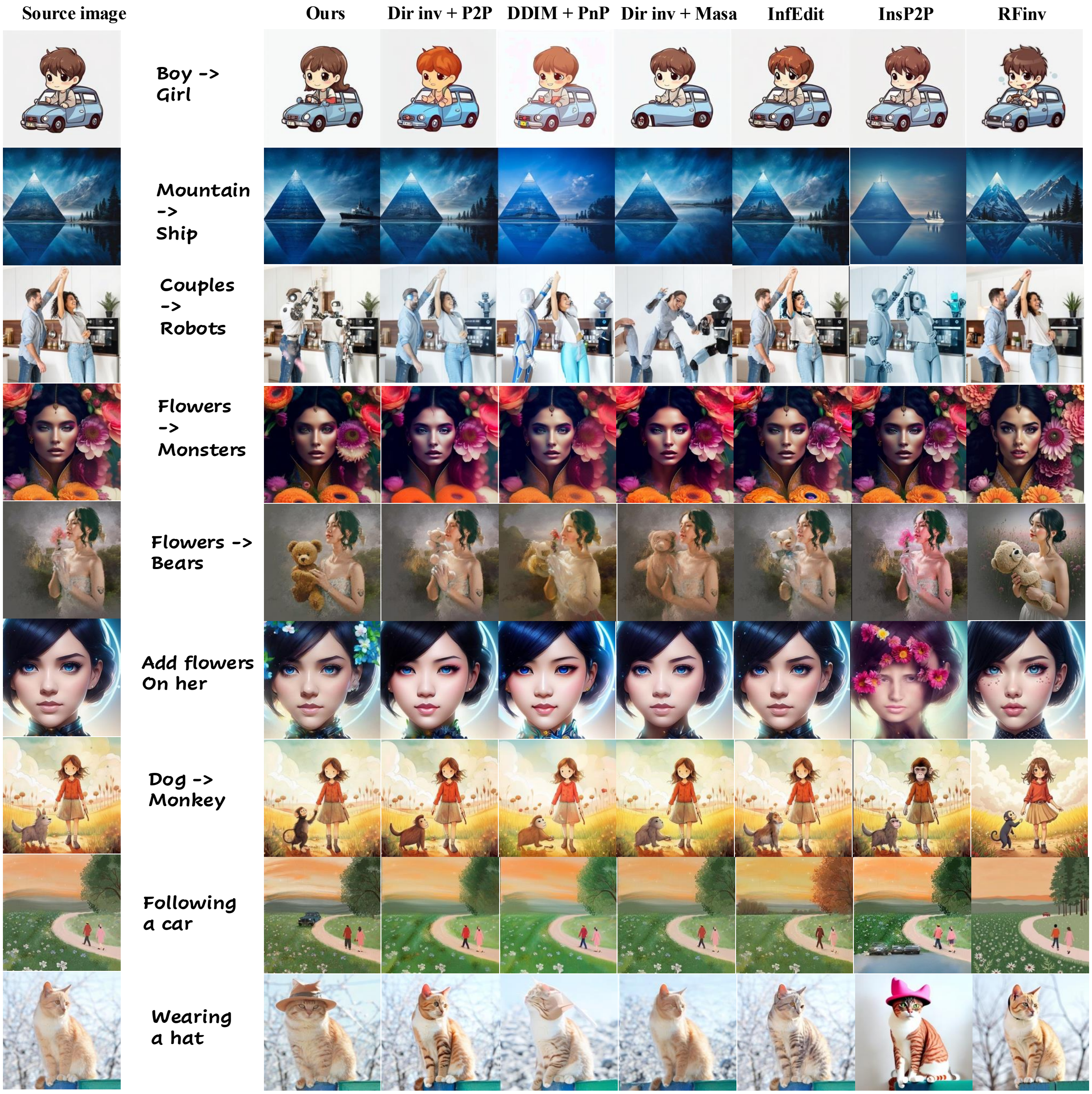

Figure 15. **Additional qualitative results on PIE benchmark**. Zoom in for details.