Edicho: Consistent Image Editing in the Wild

Supplementary Material

Appendix

A. Overview

We first illustrate the additional implementation details in Sec. B. Additional ablation studies are included in Sec. C to validate the effectiveness of the designed components. We also provide additional correspondence prediction results in Sec. D. To systematically differentiate our approach from existing attention-based paradigms, we present structural comparisons in Sec. E. The results of self-reference editing by simultaneously editing two same images in Sec. F, as well as user studies in Sec. G, further support the effectiveness of the proposed method. Sec. H presents the additional qualitative comparisons with MimicBrush [3] and results of the proposed method.

B. Implementation Details

We adopt DDIM [7] and perform denoising for 50 steps at the resolution of 512. The proposed correspondence-guided denoising strategy is applied from 4^{th} to 40^{th} steps of the denoising process and from the eighth attention layer. λ and γ are respectively set to be 0.8 and 0.9. BrushNet [6] and ControlNet [9] are adopted as the reference network for the task of consistent local and global editing. We conduct our experiments on a single A6000 GPU.

C. Additional Ablation Studies

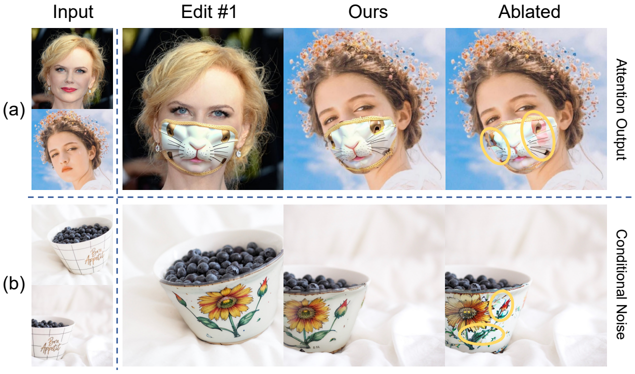

Figure S1. Additional ablations on the correspondence-guided attention (upper) and CFG (lower).

To further validate the effectiveness of the proposed components, we conduct experiments to ablate the correspondence-guided attention (upper) and CFG (lower). As in Fig. S1 (a), we first modify the correspondence-guided attention manipulation to warp the attention outputs instead of the queries. The distorted and inconsistent textures demonstrate that warping attention queries could better preserve the generative prior and achieve consistent results of high quality. Fig. S1 (b) demonstrates ablation results on the correspondence-guided CFG, where we guide the CFG in both conditional and unconditional noisy latents (instead of the unconditional only). The generation result of the ablated version turns out to be unnatural and has a fragmented look with chaotic textures, suggesting the superiority of our design in correspondence-guided CFG, which avoids deteriorating the prior by merely manipulating the unconditional latents.

D. Additional Correspondence Visualization

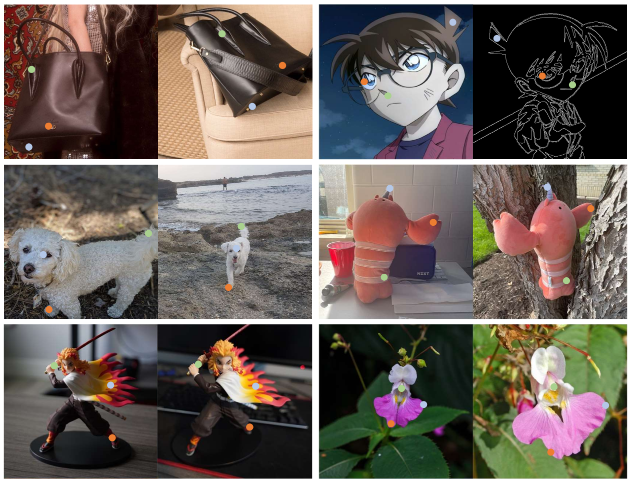

Figure S2. Additional explicit correspondence prediction results.

To demonstrate the performance of the explicit correspondence prediction method, we present additional results across a diverse range of scenarios. As in the main submission, the explicit results are predicted with DIFT [8]. As shown in Fig. S2, the explicit correspondence predictor is capable of accurately handling various challenging conditions. In the first row, we observe the correspondence robustness on objects with different lighting conditions and chromatic ranges. The handbag images illustrate that the correspondence predictor can handle variations in illumination and background textures, maintaining accurate correspondence predictions despite these changes. Additionally, the anime character and its corresponding line drawing showcase that, correspondences between colored and black-and-white images could be successfully predicted. We also provide samples under various camera poses and

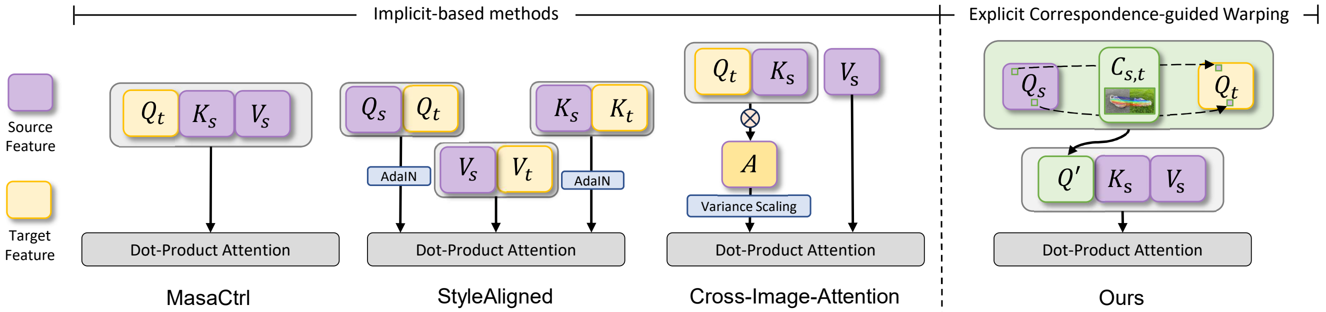

Figure S3. Attention manipulation comparisons between the prior implicit-based methods and the proposed explicit method, where A denotes attention weights and C indicates the correspondence. Unlike prior approaches that rely on attention-driven implicit representations to maintain consistency, our method introduces a paradigm leveraging pre-computed explicit correspondences to ensure consistent editing.

background conditions in the remaining rows, which also suggest correspondence robustness, forming a solid foundation for our subsequent consistent image editing tasks.

E. Attention Manipulation Comparisons

We visualize the attention manipulation comparisons between the prior implicit-based methods including MasaCtrl [2], StyleAligned [4], and Cross-Image-Attention [1], and the proposed explicit method in Fig. S3, where A denotes attention weights and C indicates the correspondence. MasaCtrl adopts the Query feature Q from the target, and Key & Value $K\&V$ from the source to perform attention computation. StyleAligned first performs AdaIN [5] for Q_s, Q_t and K_s, K_t to mix the features and then conduct dot-product attention, where the subscripts s and t denote that the attention features originate from the source or target image, respectively. Compared to MasaCtrl, Cross-Image-Attention chooses to apply variance scaling to the attention weights A computed from Q_t and K_s, to achieve contrast enhancement by feature standardization. Unlike these aforementioned approaches that rely on attention-driven implicit representations, our method introduces a paradigm leveraging pre-computed explicit correspondence, which achieves better consistent editing. We warp features of Q_s to Q_t and compute attention with the warped Query Q', K_s, and V_s. Besides this design, we also modify the classifier-free guidance (CFG) by incorporating the pre-computed correspondence and manipulate the unconditional noisy latents, aligning the edit more closely with the desired edits while maintaining high image quality.

F. Self-reference Results

To further evaluate the consistency of the proposed method, we conduct a self-reference experiment where two identical images are edited simultaneously. In each pair in Fig. S4, the first image is edited with the diffusion model, while the second one is obtained with our correspondence-guided attention and CFG to borrow features from the first during denoising. The results demonstrates remarkable consistency across diverse scenarios. For instance, the second chair and shoe receives highly matching floral embellishments as the first one does. This consistency highlights the robustness of our method in consistent editing when dealing with various objects and scenes, as a supplement for the main results.

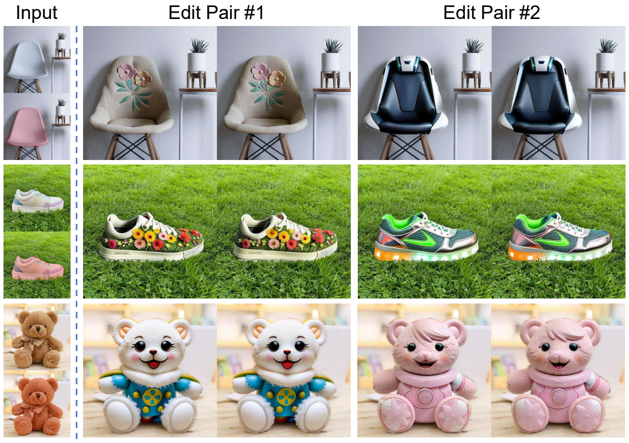

Figure S4. Self-reference results acquired by simultaneously editing two same images with the proposed correspondence-guided operations, which suggest the capability of the proposed method in maintaining editing consistency.

G. User Studies

As mentioned in the main submission, we conduct user studies to obtain results for user preference. For the task of local and global editing respectively, the 30 individuals are asked to finish up to 20 questions where they choose the best option among the four provided based on overall consistency, generation quality, and instruction-following, resulting 500 votes for each task. As in Fig. S5, our proposed solution has garnered significantly more preference compared to existing alternatives. In both evaluated tasks, over 60% of the participants opted for our approach. This

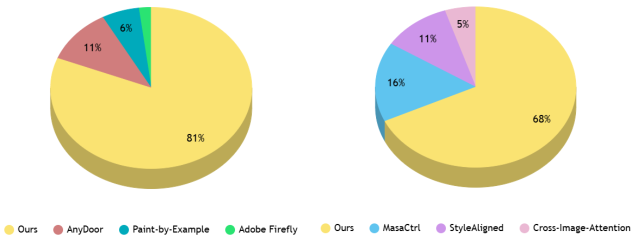

Figure S5. User study results of consistent local editing (left) and global editing (right).

endorsement from the user validates the practical value of our method as well as highlights its potential impact in real-world applications.

H. Additional Results

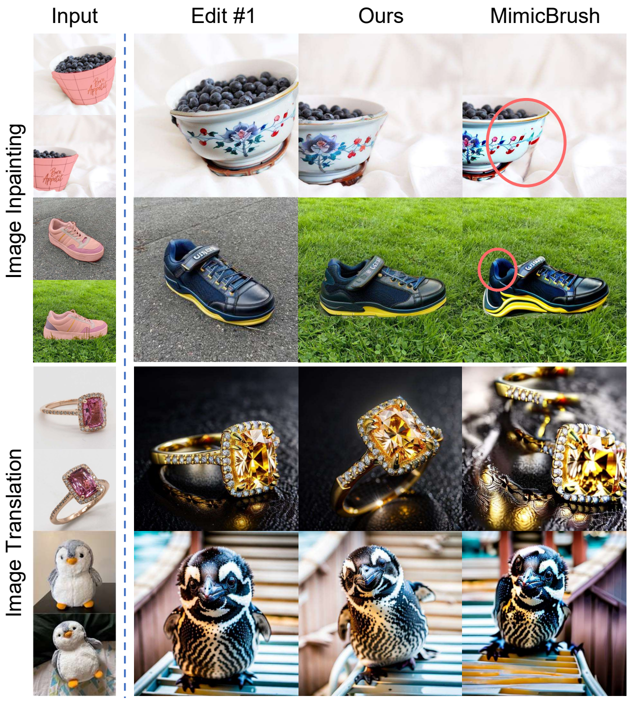

Figure S6. Comparisons with MimicBrush [3] on the local and global editing tasks.

For better understanding, we also incorporate additional comparisons with MimicBrush [3] in Fig. S6. The proposed method generates more consistent and natural images compared to MimicBrush, especially in the global editing task. Additional qualitative results for consistent local and global editing are also provided in Fig. S7. Editing results from the same initial noise are also provided in the figure, indicated as "Fixed Seed". The inpainted regions for local editing are indicated with light red color.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Int. Conf. Comput. Vis.*, 2023. 2

[3] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *arXiv preprint arXiv:2406.07547*, 2024. 1, 3

[4] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, 2017. 2

[6] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Eur. Conf. Comput. Vis.*, 2024. 1

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 1

[8] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Adv. Neural Inform. Process. Syst.*, 2023. 1

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, 2023. 1

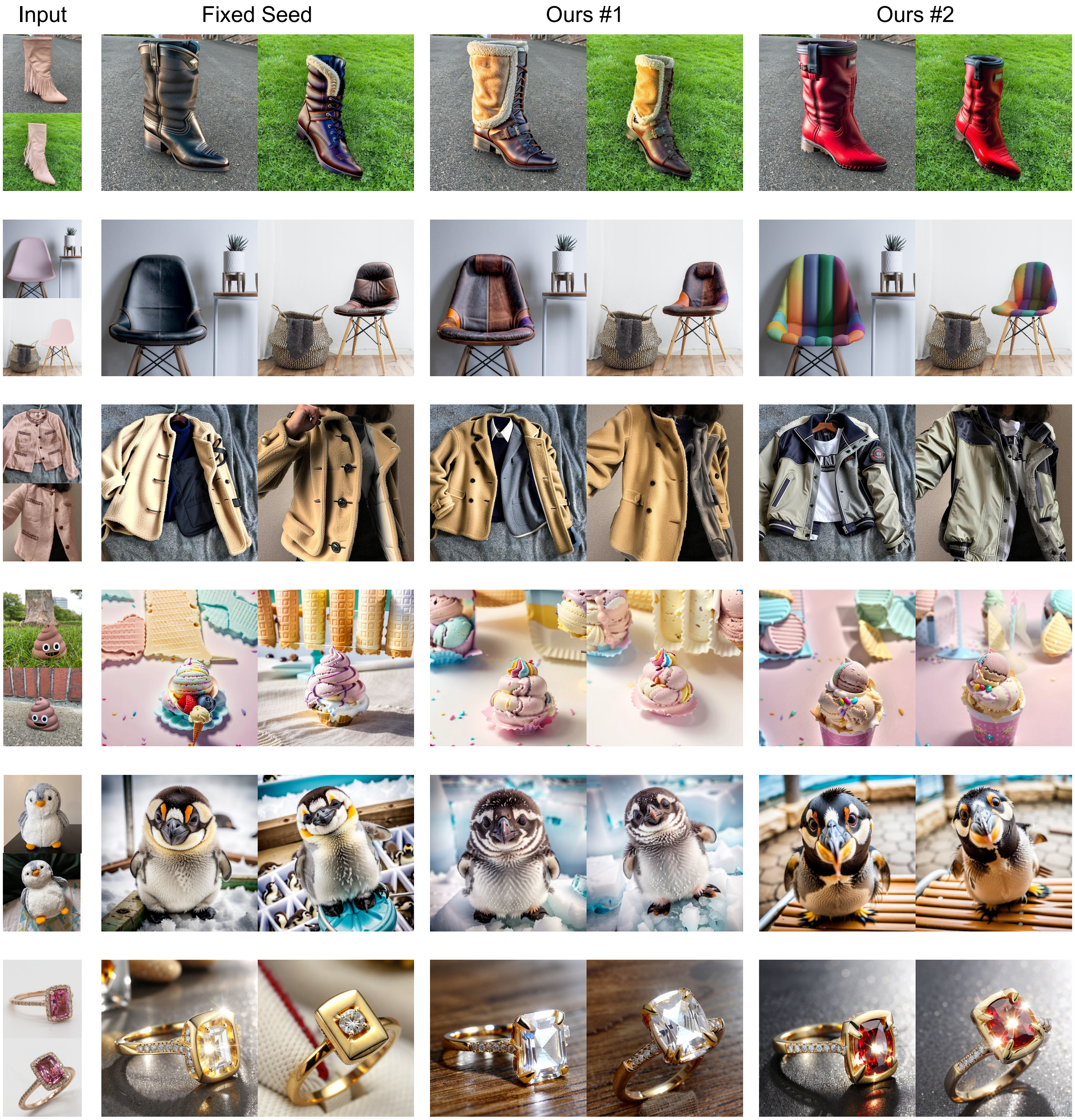

Figure S7. Additional qualitative results of the proposed method for local (upper three) and global editing (lower three ones). The inpainted regions for local editing are indicated with the light red color. “Fixed Seed” indicates editing results from the same random seed (the same initial noise).