Group Editing : Edit Multiple Images in One Go

Supplementary Material

Contents

A. Rationale

Having the supplementary compiled together with the main paper means that:

- The supplementary can back-reference sections of the main paper, for example, we can refer to introduction;
- The main paper can forward reference sub-sections within the supplementary explicitly (*e.g.* referring to a particular experiment);
- When submitted to arXiv, the supplementary will already included at the end of the paper.

To split the supplementary pages from the main paper, you can use Preview (on macOS), Adobe Acrobat (on all OSs), as well as command line tools.

B. Realted Work

Generative models for image editing. Image editing has seen remarkable progress driven by diffusion-based generative models [1–3, 5–7, 10–12, 14, 18, 19, 22–24]. Existing techniques can be broadly divided into inference-time zero-shot methods that edit images by manipulating the diffusion process itself (such as PnP [18], Prompt2Prompt [12], and MasaCtrl [3]), and training-based methods, which achieve editing by fine-tuning latent diffusion models (represented by ControlNet [25] and T2I-Adapter [13]). However, these methods remain tailored to single-image editing. When applied to a group of related images, they often fail to maintain coherence in appearance and structure, resulting in inconsistencies. Existing efforts to enforce consistency, whether by propagation [19] or attention-based correspondences [3, 12], are limited to small inputs and break down under complex geometric variation, in part due to the scarcity of suitable paired training data. In response, we formalize the problem of *Group-Image Editing* and introduce *GroupEditing*: a trainable framework that views related images as pseudo video frames to inherit implicit consistency priors from video models, while additionally incorporating an explicit correspondence module to ensure reliable alignment.

Video prior for editing task. Video generative models [4, 16, 17, 26, 27] provide powerful temporal consistency priors that can be effectively leveraged for image editing. Existing studies generally follow two directions: utilizing video data for training data curation and leveraging video models for inference-time guidance. For the former, Bagel [9], UniReal [8], and OmniGen [21] sample temporally coherent frames from video data to create high-quality training sequences, a strategy that implicitly injects structural and appearance continuity into the resulting image models. For the latter, Frame2Frame [15] utilizes a pre-trained video diffusion model to synthesize a sequence of frames and select an intermediate frame, thereby enforcing temporal smoothness and structural continuity directly during inference. ChronoEdit [20] leverages pretrained video generative models to reframe image editing as a video generation task, using the input and target images as video endpoints. While effective for enhancing single-image quality or ensuring short-range consistency, these approaches do not solve the fundamental challenge of Group-Image Editing across diverse, static views. Our approach differs by re-framing the image group as a pseudo video sequence, allowing us to explicitly inherit the powerful spatio-temporal coherence and geometric priors of large-scale video models for robustly unified editing.

C. More cases

D. Demo

We provide the demo video and project page in the file, please watch it for better illustration.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. S1

[2] Kaitong Cai, Jusheng Zhang, Yijia Fan, Jing Yang, and Keze Wang. Racot: Plug-and-play contrastive example generation mechanism for enhanced llm reasoning reliability, 2025.

[3] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. S1

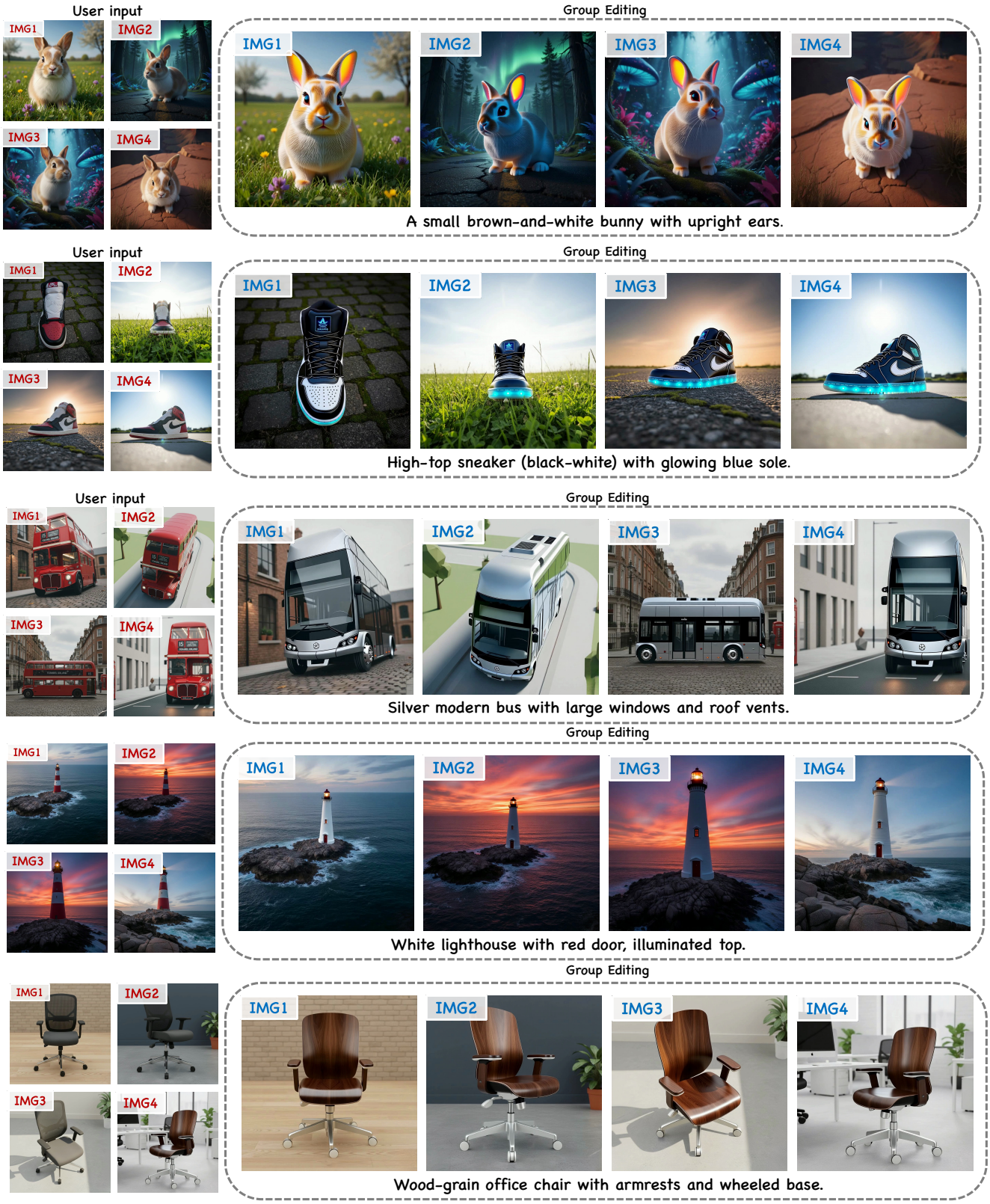

Figure 1. **More cases**.

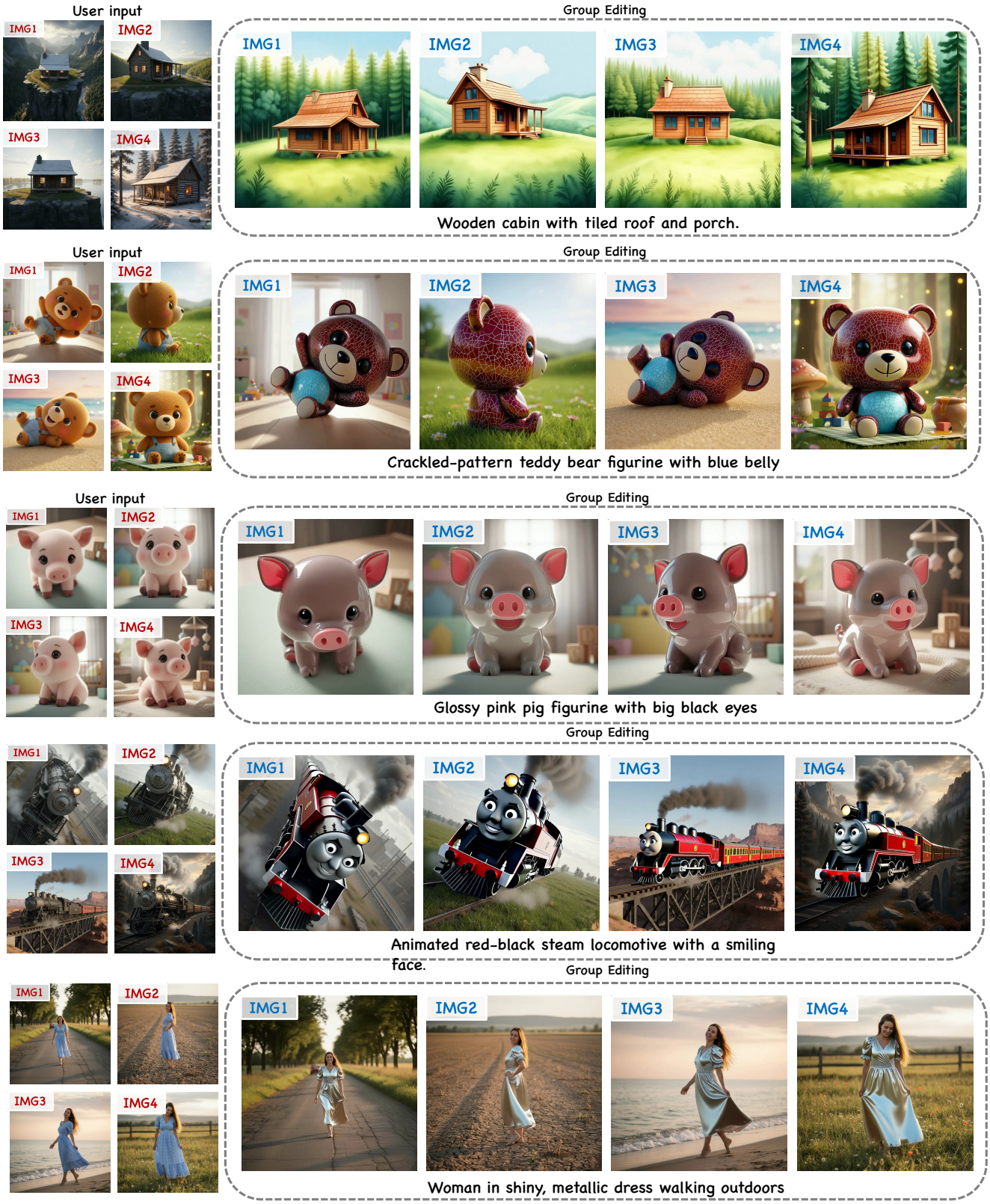

Figure 2. **More cases**.

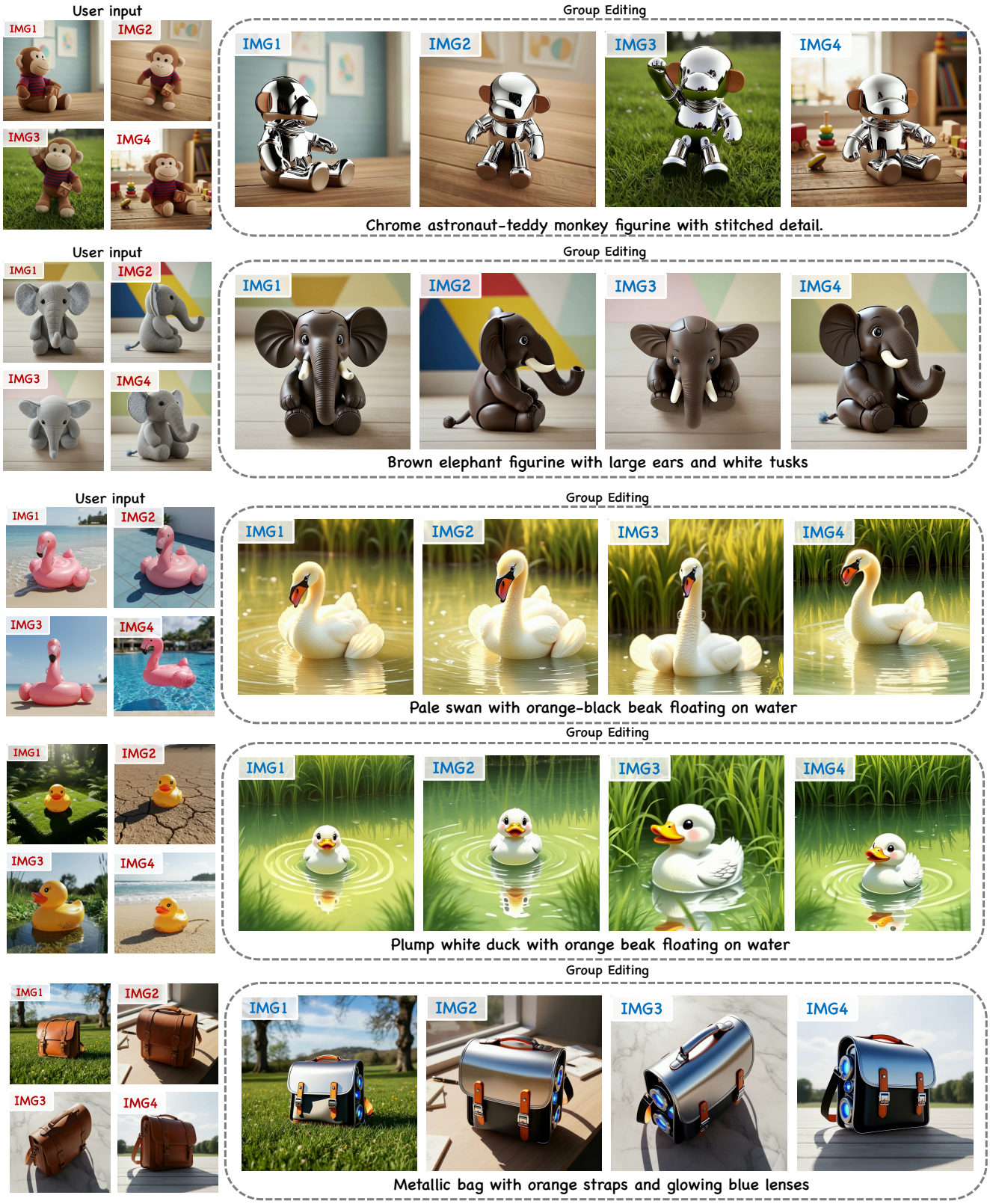

Figure 3. **More cases**.

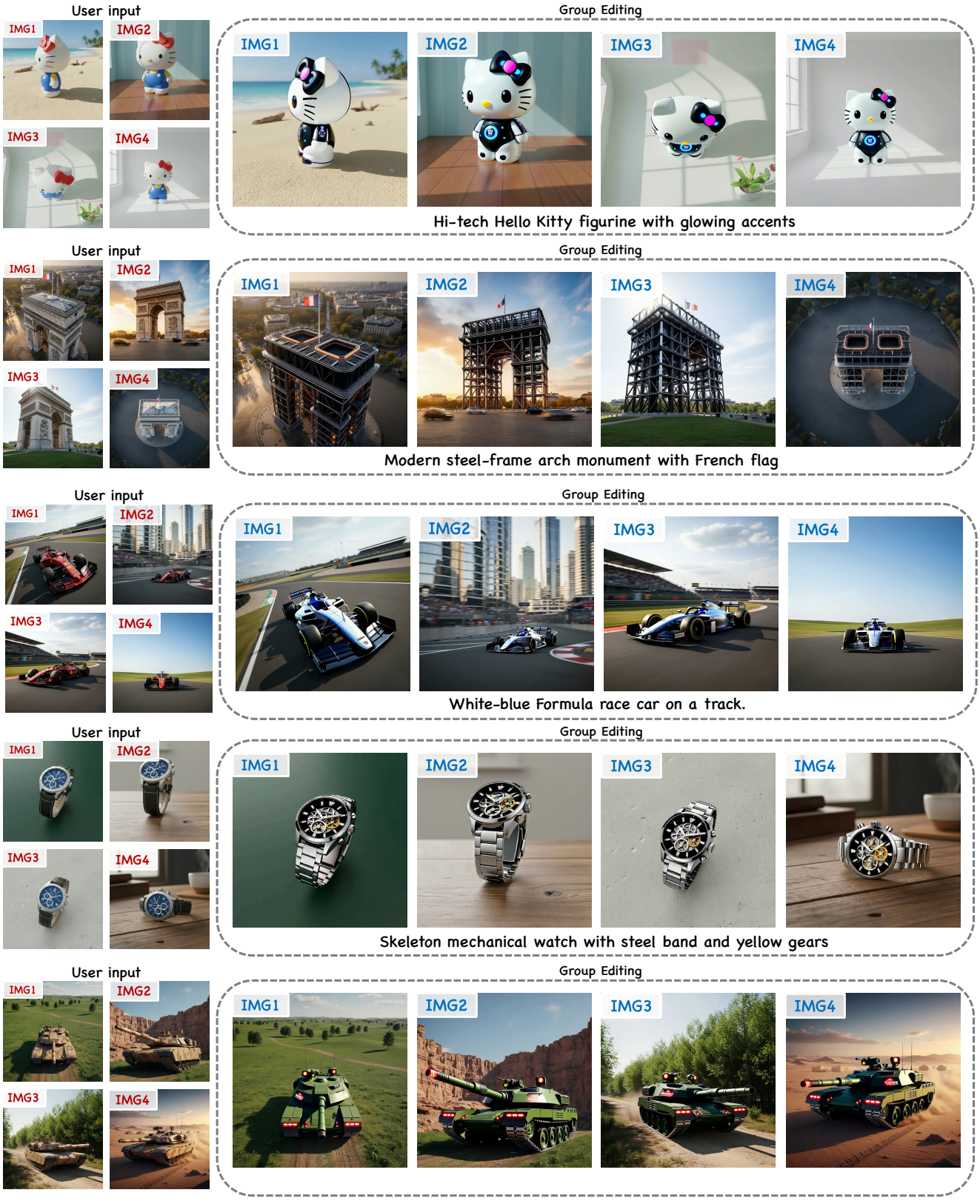

Figure 4. **More cases**.

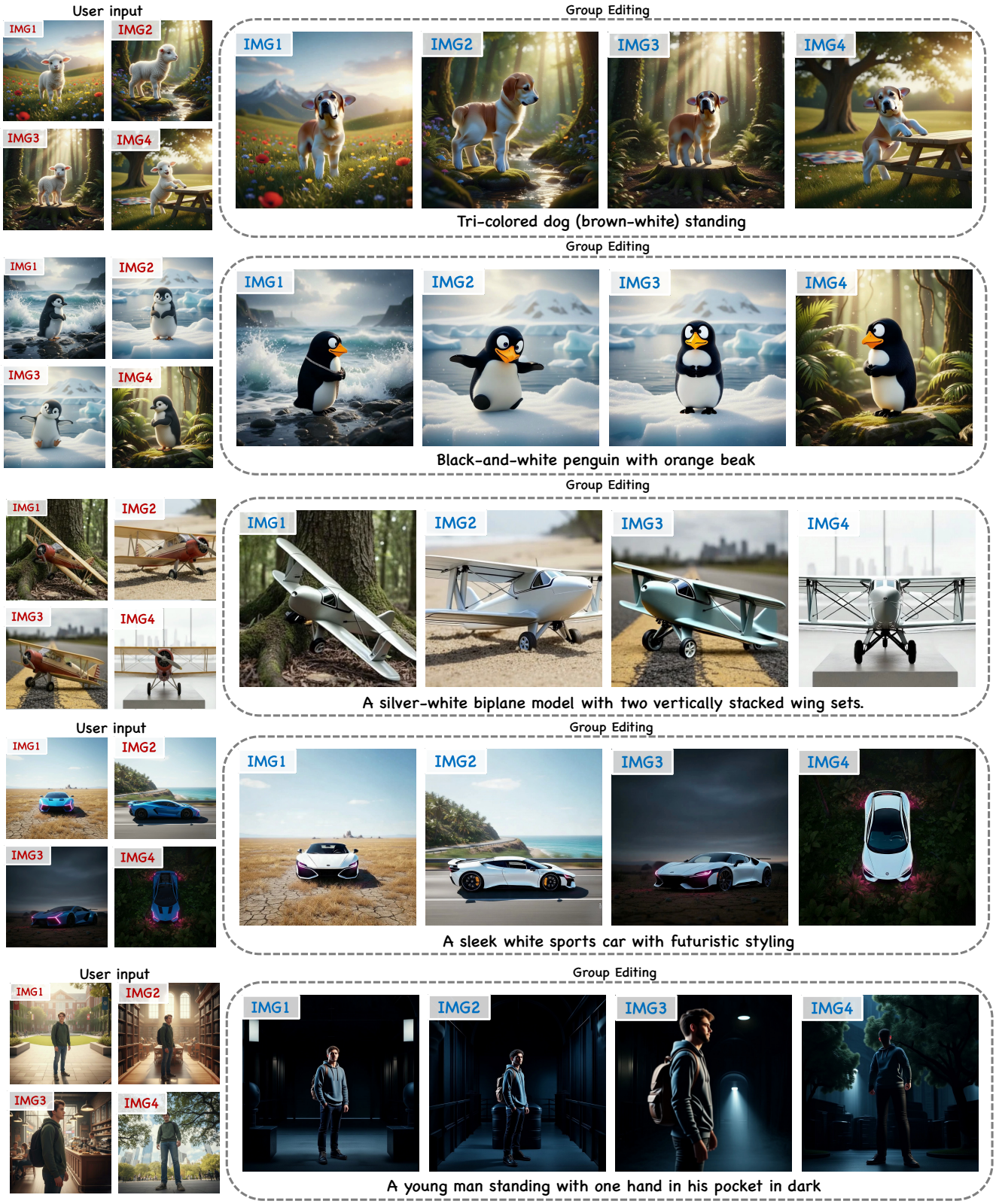

Figure 5. **More cases**.

[4] Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, et al. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. *arXiv preprint arXiv:2512.24146*, 2025. S1

[5] Chubin Chen, Jiashu Zhu, Xiaokun Feng, et al. S^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models. *arXiv preprint arXiv:2508.12880*, 2025. S1

[6] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *Advances in Neural Information Processing Systems*, 37:84010–84032, 2024.

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. S1

[8] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. S1

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. S1

[10] Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction, 2025. S1

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. S1

[13] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. S1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. S1

[15] Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaid, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7857–7866, 2025. S1

[16] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. *arXiv preprint arXiv:2502.01105*, 2025. S1

[17] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025. S1

[18] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023. S1

[19] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pages 112–129. Springer, 2024. S1

[20] Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Tianshi Cao, Kai He, Yifan Lu, Ruiyuan Gao, Enze Xie, Shiyi Lan, Jose M Alvarez, et al. Chronoedit: Towards temporal reasoning for image editing and world simulation. *arXiv preprint arXiv:2510.04290*, 2025. S1

[21] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. S1

[22] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm:counterfactual vision-language fine-tuning, 2025. S1

[23] Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[24] Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In *Forty-second International Conference on Machine Learning*, 2025. S1

[25] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. S1

[26] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. S1

[27] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. S1