Supplementary Materials for MasaCtrl: Tuning-free Mutual Self-Attention Control for Consistent Image Synthesis and Editing

Mingdeng Cao[1,2*] Xintao Wang[2✉] Zhongang Qi[2] Ying Shan[2] Xiaohu Qie[2] Yinqiang Zheng[1✉]
[1]The University of Tokyo [2]ARC Lab, Tencent PCG
https://github.com/TencentARC/MasaCtrl

A. Additional Background: Diffusion Models

Diffusion models [1, 6, 4] are generative models that can synthesize desired data samples from Gaussian noise via iterative denoising. A diffusion model defines a forward process and a corresponding reverse one. The forward process adds the noise to the data sample x_0 to generate the noisy sample x_t with a predefined noise adding schedule α_t at time-step t:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}}x_0, (1-\bar{\alpha}_t)I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. At step T, the data sample x_0 is transformed into Gaussian noise $x_T \sim \mathcal{N}(0, 1)$. The reverse process tries to remove the noise and generate a cleaner sample x_{t-1} from the previous noisy sample x_t:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t), \tag{2}$$

where μ_θ and σ_t are the corresponding mean and variance. The variance is a time-dependent constant, and the mean $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \epsilon\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}})$ can be solved by using a neural network $\epsilon_\theta(x_t, t)$ to predict the noise ϵ. To train such a noise estimation network ϵ_θ, the object is a simplified mean-squared error:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}(\|\epsilon - \epsilon_\theta(x_t, t)\|). \tag{3}$$

Therefore, by sampling x_{t-1} iteratively, a data sample x_0 can be synthesized from random Gaussian noise x_T. In addition, text prompts P can be conditioned into the predicted noise $\epsilon_\theta(x_t, t, P)$ so that the diffusion models can synthesize text-complied images.

B. Additional Ablation Study of MasaCtrl

Mutual Self-Attention Control. Our method aims to combine the image layout initialized by the desired prompt P with the image contents from the source image I_s. To achieve this, we propose to perform mutual self-attention in the later denoising steps (not at the premature beginning) and the decoder part of U-Net. After several iterative denoising steps S, the layout of the desired target image can be roughly formed, as shown in Figure 4(a) in the main paper. In the decoder part of U-Net, the formed target structure is much clearer than in the encoder part, as shown in Figure 4(b) in our main paper. We additionally provided the results of different strategies that perform mutual self-attention control with dense denoising steps and layers in U-Net, as shown in Fig. 1. We found that performing mutual self-attention in the earliest steps (earlier than the time step 4) and layers can synthesize a more consistent image with the source image, but fails to comply with the target modified prompt P, since the layout of the target image has not yet been formed in these too-early denoising steps and layers. In contrast, performing mutual self-attention control at late time steps (later than time step 25) and U-Net layers generates an image that highly complies with the target text description but loses the content information of the source image, since the image contents are already determined and cannot be changed significantly. In our method, we start performing mutual self-attention in the moderately earlier time step and the layers in the U-Net decoder part. This allows us to synthesize an image that complies with the target image and consists of similar contents from the source image. More results of consistent synthesis and real image editing can be found in Fig. 6.

Mask-guided Mutual Self-Attention. We observed the synthesis/editing using the proposed method would fail since the object and background are too similar to be confused in the query feature space (shown in Figure 2 in the main paper). To tackle this problem, we introduce a mask-guided mutual self-attention mechanism that performs the attention in the restricted regions for foreground objects and background separately. The detailed pipeline is shown in

*Work done during an internship at ARC Lab, Tencent PCG.

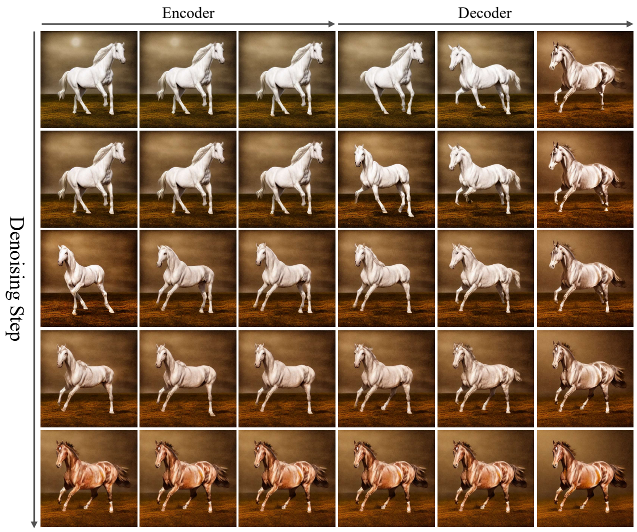

Figure 1: Additional ablation results of the start of the timestep and U-Net layer index to perform self-attention control.

Figure 2: Example of the multi-object confusion problem and the results of mutual self-attention.

Fig. 3. The object only queries image contents from the foreground object region in the source image rather than the whole object to avoid confusion.

Meanwhile, we also found that confusion problems may occur when editing multiple objects with MasaCtrl, especially when the objects are in the same class. As shown in Fig. 2, their dresses are exchanged when MasaCtrl is directly utilized to edit the source image with one boy and one girl. In this case, when we utilize the mask-guided strategy to restrict the query regions, this problem can be effectively alleviated. However, our strategy cannot tackle this problem perfectly, and some details in the edited image still differ from that in the source image.

C. Limitations and Discussion

Our method inherits most of the limitations of Stable Diffusion in generating desired images, and suffers from the following main aspects. First, since our method heavily relies on the image layout synthesized from the target prompt P, it would fail if the SD model could not generate a desired layout or shape, as shown in Fig. 4(1). Although recently proposed controllable strategies [3, 7] can alleviate this on the pre-trained SD model with various guidance, it still may fail. In addition, even if the SD model can generate the corresponding image layout, our method will fail when the target image contains unseen content or the target image layout/structure changes drastically. As shown in Fig. 4(2), the SD model can synthesize the target layout that complies with the target prompt P while with different contents (*i.e.*, the identity of the person and the back-

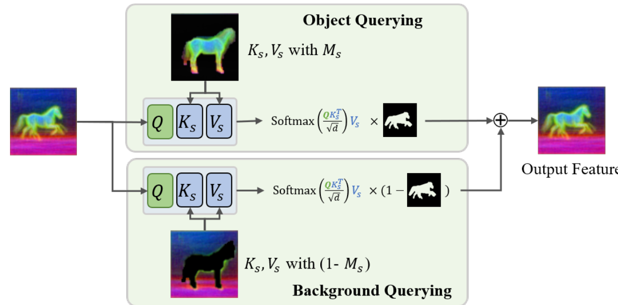

Figure 3: Pipeline of the proposed mask-guided mutual self-attention.

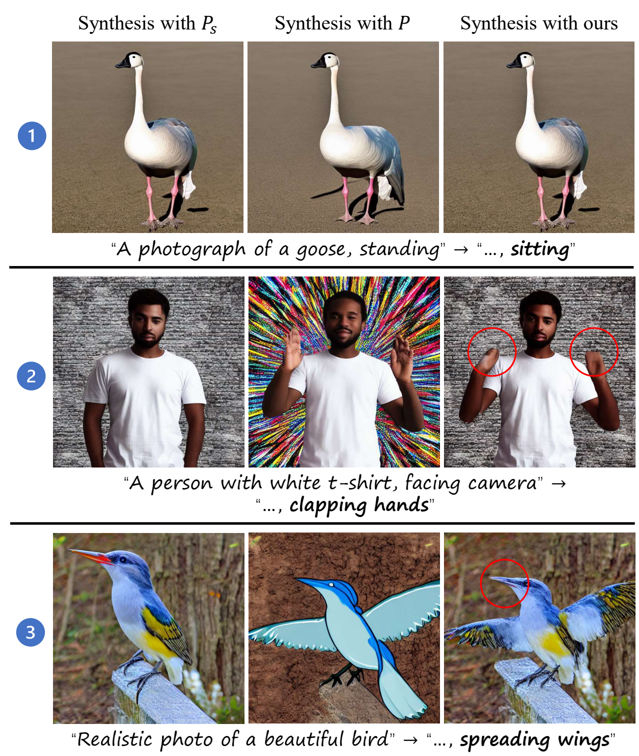

Figure 4: Different types of failure cases.

ground) from the source image. MasaCtrl can generate an image consistent with the source image but suffer from the artifact (the palm marked by the red circle). This is because the source image does not contain any contents related to the palm, thus the desired image cannot query the contents of the palm. Meanwhile, as shown in Fig. 4(3), although our method can synthesize desired image that is highly similar to the source image, we also found there still are some slight differences (the color of the bird's beak marked by the red circle) between the source image and the edited image. How to tackle these problems is our future work.

D. More Results with T2I-Adapter

Recently proposed controllable methods for diffusion methods, like T2I-Adapter [3] and ControlNet [7], can synthesize images with various guidance (*e.g.*, pose, sketch, depth). Therefore, we can apply these methods to synthesize the layout of the desired image and then utilize our method to query image contents from the source image to generate content-consistent images. Here, we show the consistent editing results of MasaCtrl integrated into T2I-Adapter [3] shown in Fig. 5. We can achieve similar editing results as Imagic [2] without fine-tuning .

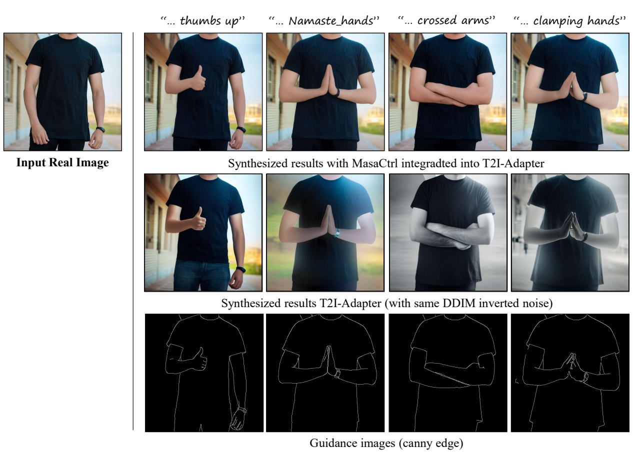

Figure 5: Real image editing results of proposed MasaCtrl integrated into T2I-Adapter [3]. We can perform non-rigid editing by preserving the image contents while changing its structure.

E. More Results on Other Models

Stable Diffusion XL. We also apply the proposed method in the recent Stable Diffusion XL (SDXL) model [5]. SDXL further enlarges the denoising U-Net three times compared to the previous SD model, achieving higher-quality synthesis images with novel model designs and conditioning strategies. The results of MasaCtrl on SDXL are shown in Fig. 7. We see that the proposed method can also generalize well on such a powerful model and synthesize high-quality consistent images.

Anything-V4. We also apply our method to the domain-specific models, *i.e.*, the amine-style model Anything-v4. Fig. 8 shows the synthesis results of our method and the model with fixed random seeds. The proposed method MasaCtrl can faithfully synthesize images while preserving the object identity and background in original anime-style images, further demonstrating the generalizability of the proposed method. Meanwhile, we further perform consistent image synthesis in Fig. 9. We can control the pose and action, even expression, with the proposed method by directly modifying the text prompt accordingly, demonstrating the consistent synthesis capability of MasaCtrl.

F. More Video Synthesis Results

Since MasaCtrl can synthesize content-consistent images, we can further achieve temporal consistent results by applying MasaCtrl to existing controllable models with dense guidance. We provide more video synthesis results shown in Fig. 10, and the video results can be found on our project page.

G. User Study

An example question in our user study is illustrated in Fig. 11.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1

[2] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3

[3] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3

[4] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. 1

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 5

Note that for each edit, Imagic requires 60 minutes for finetuning the diffusion model and optimizing the textual embedding.

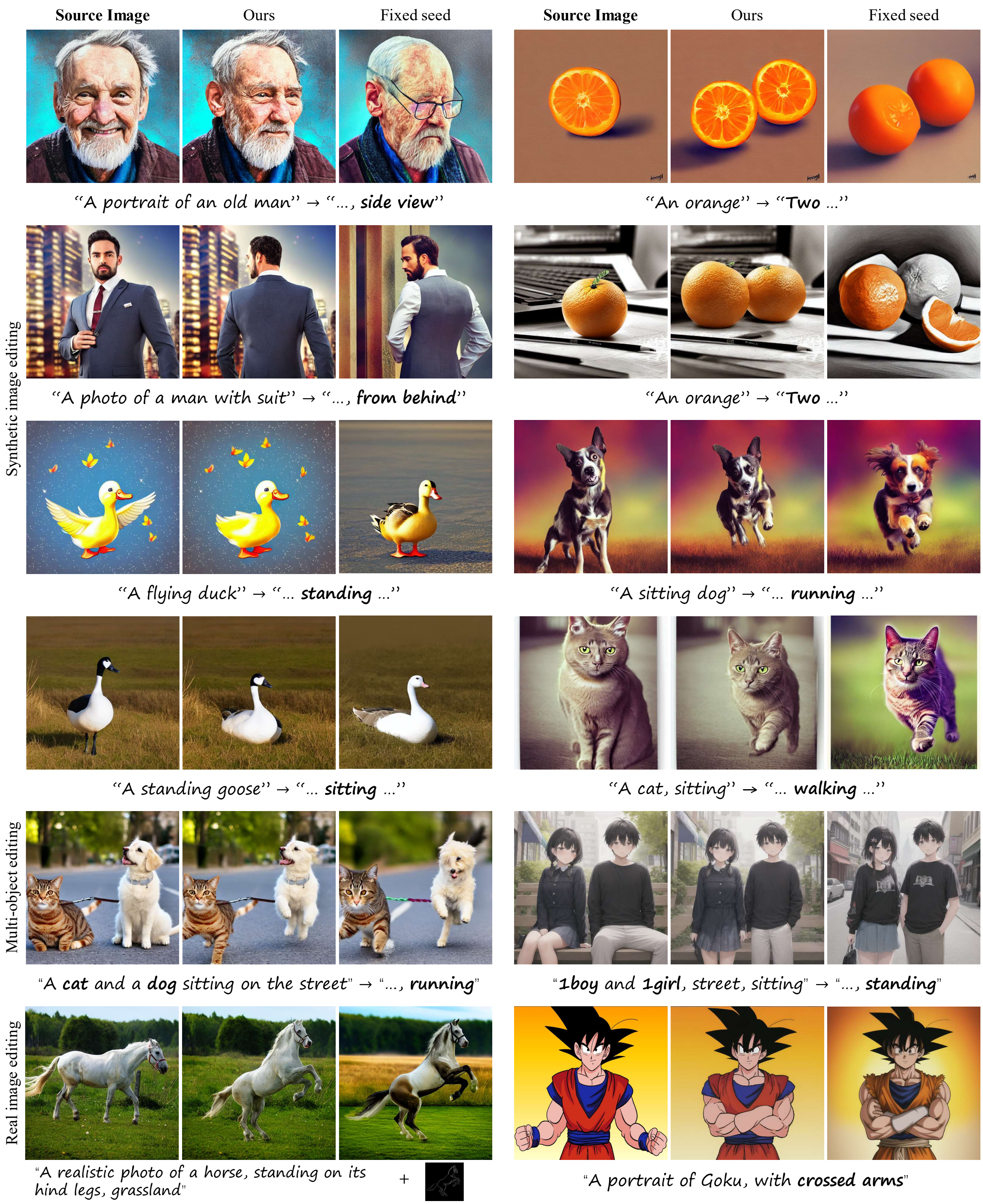

Figure 6: Additional editing results of the proposed MasaCtrl.

Figure 7: Synthesis results on Stable Diffusion XL (SDXL) model [5]. Consistent images can be synthesized by directly modifying the text prompts with the proposed MasaCtrl.

Figure 8: Synthesis results on the anime-style Anything-V4 checkpoint. Consistent images can be synthesized by directly modifying the text prompts with the proposed MasaCtrl.

[6] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[7] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3

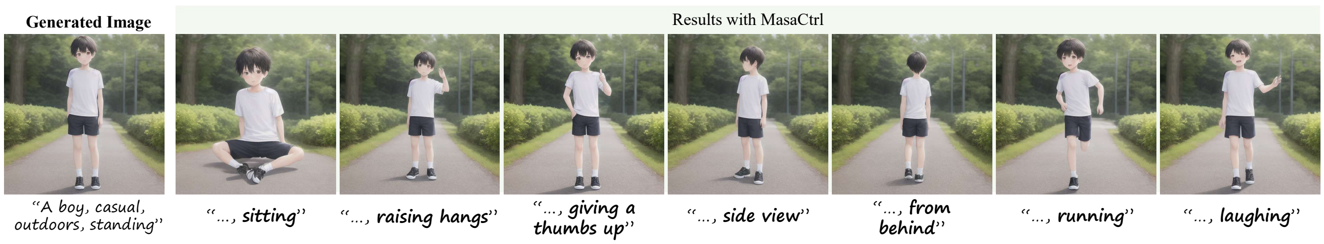

Figure 9: Multiple consistent synthesis results with proposed MasaCtrl on Anything-V4 checkpoint.

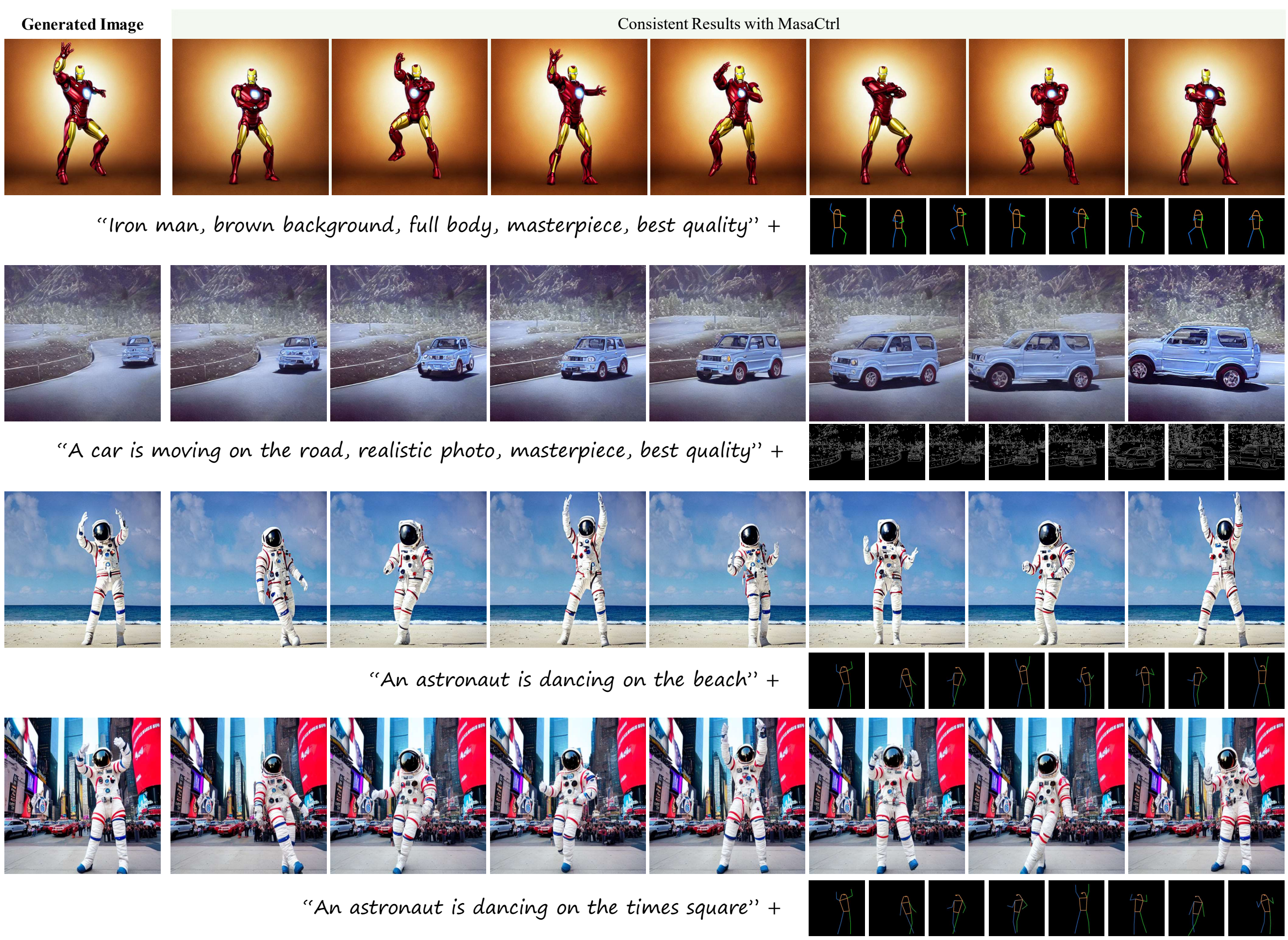

Figure 10: Additional video synthesis results with MasaCtrl.

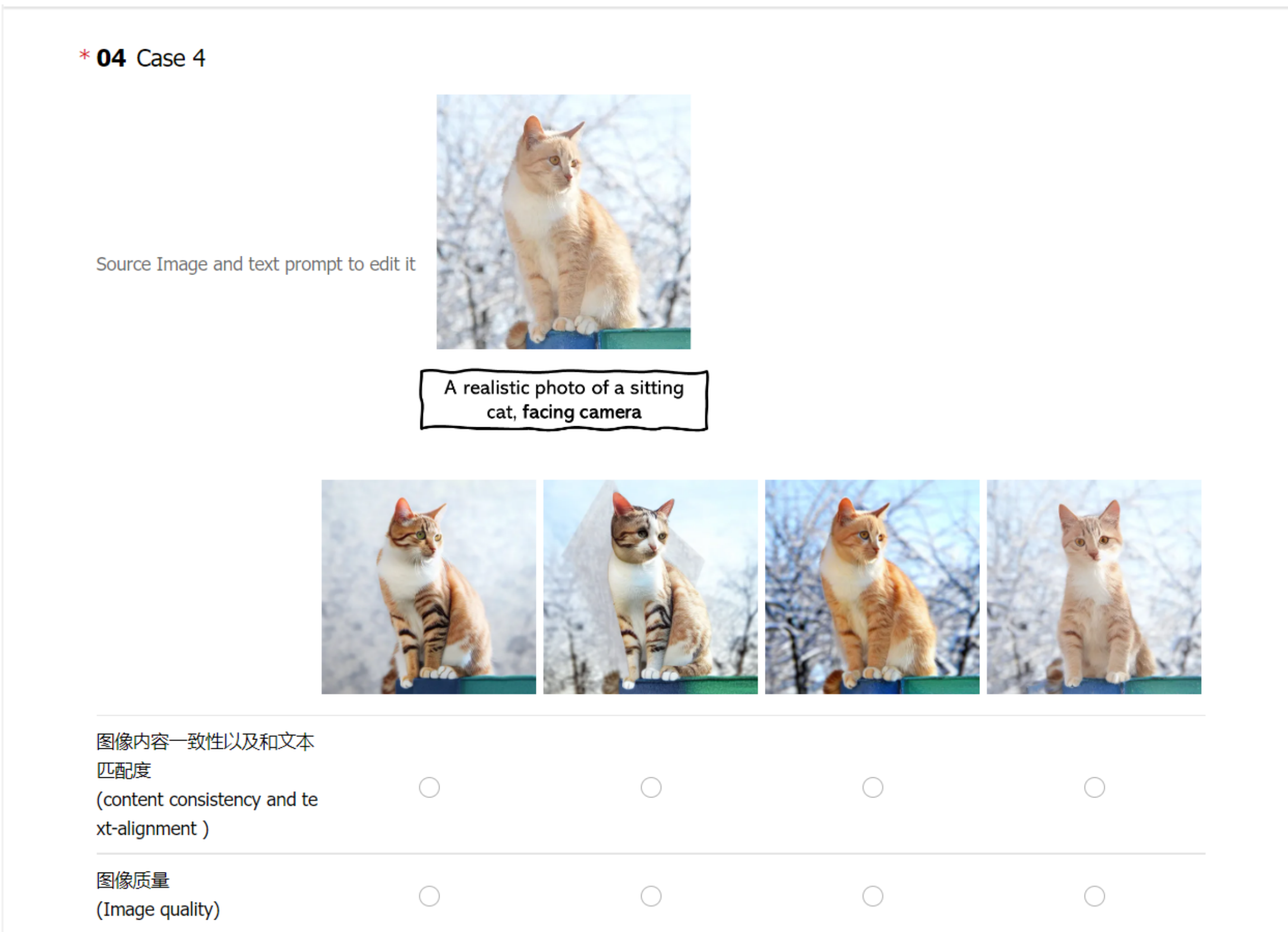

Figure 11: Illustration of the user study.