Supplementary Material of MCA-Ctrl

1. Baseline Method

We compare MCA-Ctrl to the subject-driven editing and generation methods on DreamBench [10] and DreamEditBench [9] public datasets. This section provides a brief introduction to these methods:

- DreamBooth [10]: It's a method of fine-tuning for each subject, optimizing all U-Net parameters and placeholder embedding.
- Textual Inversion [5]: This method fine-tunes each subject, optimizing the placeholder embeddings to reconstruct the subject image. It takes 3,000 training steps to learn new concepts.
- Re-Imagen [2]: A tuning-free method that takes several images as input and then focuses on retrieval to generate new images.
- BLIP-Diffusion [7]: The model learns the multimodal subject representation step by step through the multimodal control capability of built-in BLIP-2, achieving a certain degree of zero-shot subject-driven generation.
- Customized-DiffEdit [3]: This is a method that needs fine-tuning. DiffEdit automatically generates the mask to be edited by contrasting predictions conditioned between the source and subject prompts. In this paper, we follow [9] and replace the diffusion model in DiffEdit with the DreamBooth fine-tuned model to implement subject editing. The generated image of this method is highly consistent with the condition image, but the foreground and connecting parts will appear stiff and have semantic incongruity.
- DreamEditor [9]: This method needs fine-tuning for each class. It is implemented based on Stable Diffusion, GLIGEN, or copy-paste, and refines the target subject through iterative generation.
- InstructPix2Pix [1]: A tuning-free instruction-driven editing method that takes the source image and editing instructions as input. Although it does not explicitly express the subject, it can be a novel representation of the subject by redefining the context. We make a qualitative comparison with this method.
- IP-Adapter [11]: A tuning-free method primarily designed for consistency-based generation.
- FreeCustom [4]: A tuning-free method that leverages attention control to achieve multi-concept composition.
- PHOTOSWAP [6]: A tuning-free method that enables subject swapping based on the input subject and condition images.
- TIGIC [8]: A tuning-free method that enables subject addition based on the input subject image, condition image, and localization mask.

2. Experimental Setting

2.1. Computational Efficiency

Our three parallel diffusion processes are implemented in code by concatenating operations in the batch size dimension, i.e. each time for inference, our input shape is [3, C, H, W]. In the Self-Attention layer, we obtain the features corresponding to the subject, condition, and target images by segmenting Q, K, and V matrices and carrying out SALQ and SAGI operations. This paper describes three parallel diffusion processes to display the interaction among the subject image, condition, and target image more clearly and intuitively. **Therefore, MCA-Ctrl does not cause redundant computing resource load, and its computational efficiency is the same as that of a single execution of Stable Diffusion under the same batch size.** Table 1 shows the specific computational efficiency comparison between MCA-Ctrl and Stable Diffusion baseline.

Table 1. Comparison of computational efficiency between the UNet2DConditionModel of MCA-Ctrl and Stable Diffusion.

	Model	#Params	FLOPs
Edit (batch size=3)	Stable Diffusion	859.3955M	2032.4952G
	MCA-Ctrl	859.3955M	2032.4952G
Generation (batch size=2)	Stable Diffusion	859.3955M	1354.9968G
	MCA-Ctrl	859.3955M	1354.9968G

2.2. Architecture of Subject Location Module

As described in Section 3, the Subject Location Module consists of an object detection model Grounding DINO and a segmentation model SAM that receives a multimodal image-text pair as input and outputs a prompt-specified mask. Table 2 lists the parameters of the Grounding DINO and SAM used in this document (All parameters that do not appear in the following table use the default parameters).

Table 2. Specific important parameters of the model used in the Subject Location Module.

Model	parameter	value
DINO	backbone	swin_B_384_22k
	position_embedding	sine
	enc_layers	6
	dec_layers	6
	hidden_dim	6
	nheads	8
	box_threshold	0.3
	text_threshold	0.25
SAM	checkpoint	sam_vit_h_4b8939.pth

2.3. Specific Parameters of SALQ and SAGI

As stated in Section 3.3 and Section 4.1, a total of six parameters are involved in the experiment in this paper, namely S_{GI}, E_{GI}, S_{LQ}, E_{LQ}, $Layer_{GI}$and $Layer_{LQ}$. Based on all the experimental verification, we set two default settings to make the model generation effect better: (1) SALQ is carried out continuously after SAGI operation, there is no gap between them, and the two operations do not overlap, so E_{GI}=S_{LQ}; (2) If SAGI is performed at a time step, it is performed at all layers in UNet, so $Layer_{GI}$=0. Based on the above assumptions, we mainly discussed the following four parameters: S_{GI}, E_{GI}, $Layer_{LQ}$, and E_{LQ}. These parameters can be adjusted for different classes to ensure more consistent editing and generation.

In Table 4 and Table 3, we supplement the specific parameter settings of Our (Uniform) and Ours (Specified) models mentioned in the presentation of quantitative results for subject generation and subject swapping to help the reader reproduce the results (uniform parameter settings are used for classes not mentioned in the table).

2.4. Analysis of E_{GI}

We illustrate the impact of E_{GI} on image generation in Figure 1. In complex scenarios, omitting SAGI can lead to challenges such as failing to localize the target and confusion in global features. As E_{GI} is delayed, subject features become increasingly distinct. However, beyond a certain point (empirically around 60% of the total denoising steps for most cases), further increasing the execution steps of SAGI has a diminishing effect on image quality.

2.5. More Visualization

We present enlarged versions of Figures 7, 8, and 9 from the main text in Figures 2, 4, and 3, respectively.

To further demonstrate the zero-shot generation capability of MCA-Ctrl, we provide additional results in Figures 6 and 5. As shown, MCA-Ctrl excels at customized generation for high fine-grained objects such as animals and characters, achieving remarkable text-image and image-image consistency in the results.

Table 3. Specific parameters of SALQ and SAGI (Swapping).

Ours (Uniform)				
Subjects	S_{GI}	$Layer_{LQ}$	E_{GI}	E_{LQ}
All	0	8	20	48
Ours (Specified)				
Subjects	S_{GI}	$Layer_{LQ}$	E_{GI}	E_{LQ}
backpack	0	0	15	48
backpack-dog	0	10	35	48
berry-bowl	0	10	17	48
can	0	8	10	48
colorful-sneaker	0	8	15	48
dog	0	10	10	48
dog2	0	10	25	48
dog5	0	8	15	48
dog6	0	10	10	48
dog8	0	10	10	48
duck-toy	0	8	15	48
fancy-boot	0	10	30	48
wolf-plushie	0	10	5	48

Table 4. Specific parameters of SALQ and SAGI (Generation).

Ours (Uniform)				
Subjects	S_{GI}	$Layer_{LQ}$	E_{GI}	E_{LQ}
All	0	0	35	48
Ours (Specified)				
Subjects	S_{GI}	$Layer_{LQ}$	E_{GI}	E_{LQ}
backpack	0	0	25	48
backpack-dog	0	0	30	48
berry-bowl	0	0	30	48
can	0	0	40	48
colorful-sneaker	0	0	40	48
cat	0	0	25	48
dog	0	0	30	48
dog2	0	0	30	48
dog5	0	0	30	48
dog8	0	0	30	48
duck-toy	0	0	25	48
fancy-boot	0	0	40	48
wolf-plushie	0	0	25	48

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.

Figure 1. Analysis of E_{GI}.The results above are generated with a total of 50 denoising steps. Cases with green borders represent those with better performance.

1

[2] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 1

[3] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1

[4] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. *IEEE*, 2024. 1

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[6] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, Hyun-Joon Jung, et al. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[8] Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. *arXiv preprint arXiv:2403.12658*, 2024. 1

[9] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen.

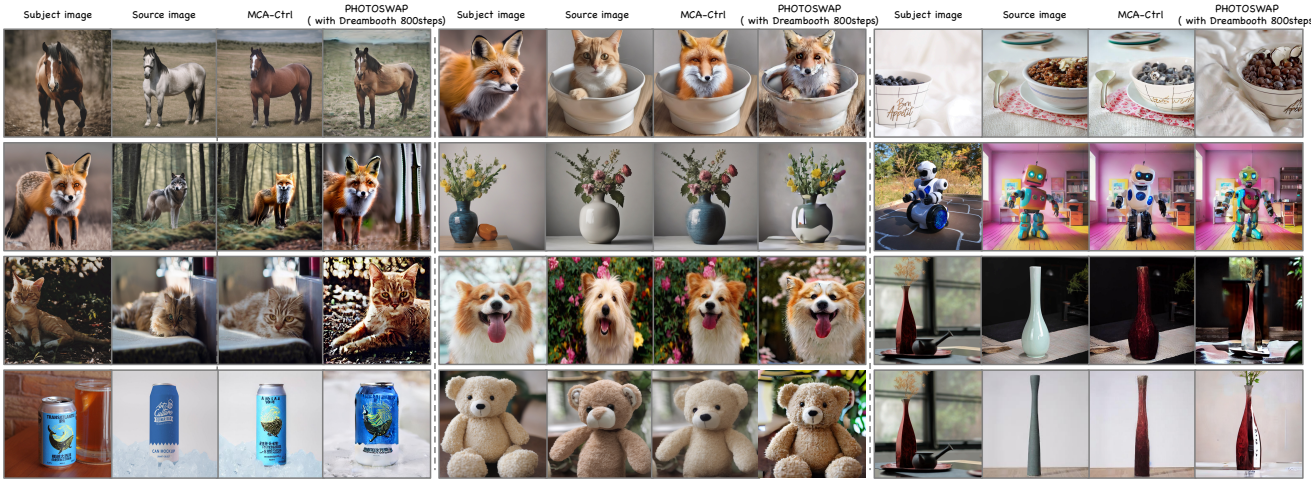


Figure 2. Enlarged version of Figure 7 in the main text.

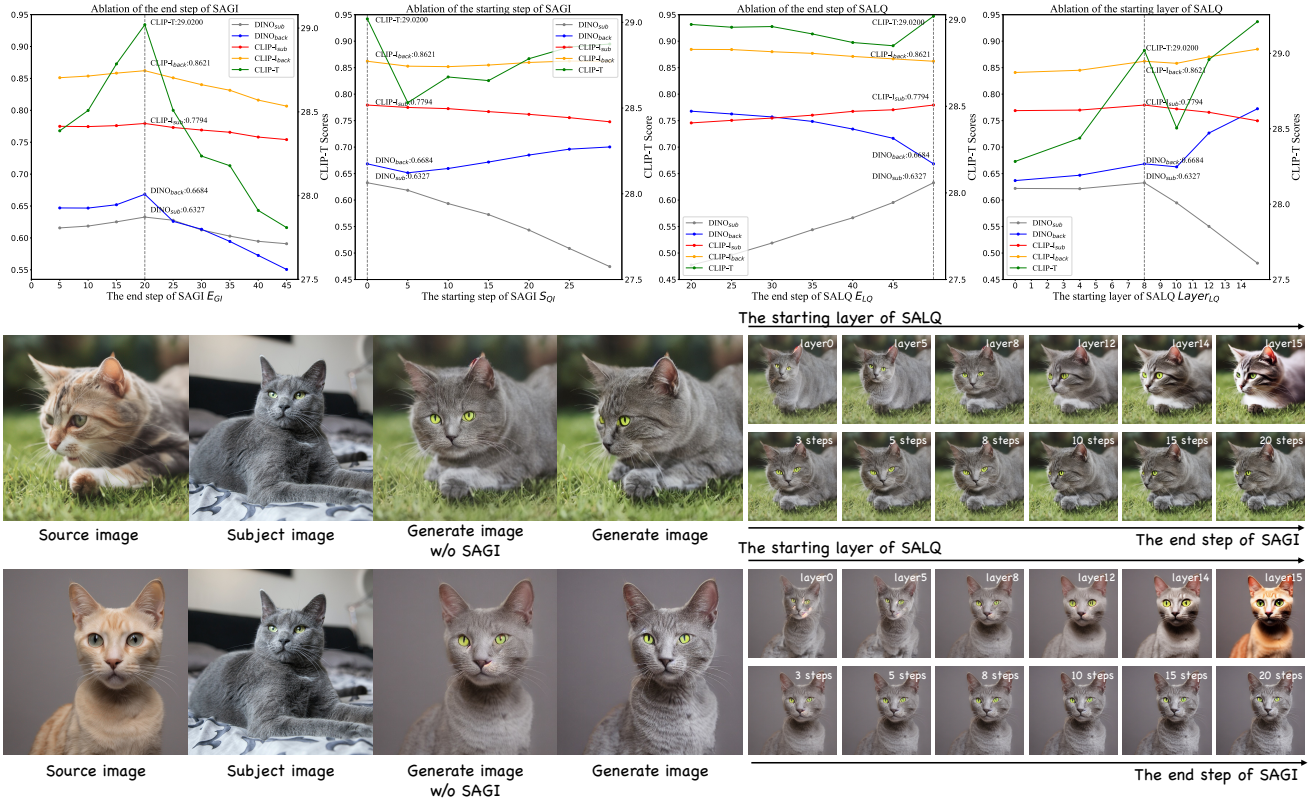


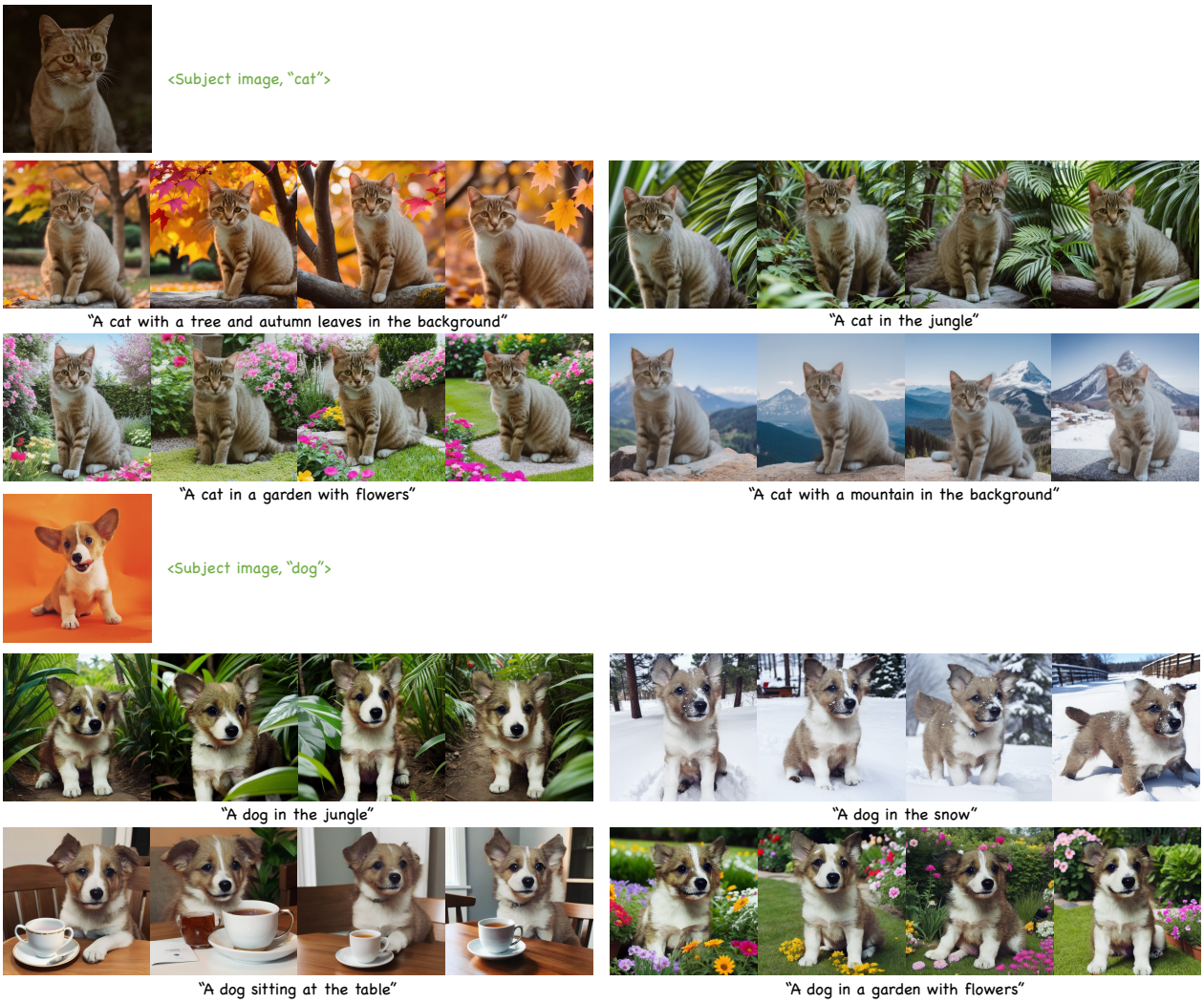
Figure 3. Enlarged version of Figure 9 in the main text.

Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023. 1

[10] 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*, pages 22500–22510, 2023. 1

[11] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

Figure 4. Enlarged version of Figure 8 in the main text.



"A cat with a tree and autumn leaves in the background"

"A cat in the jungle"

"A cat in a garden with flowers"

"A cat with a mountain in the background"

<Subject image, "dog">

"A dog in the jungle"

"A dog in the snow"

"A dog sitting at the table"

"A dog in a garden with flowers"

Figure 5. More customized generation results of MCA-Ctrl (1).

Figure 6. More customized generation results of MCA-Ctrl (2).