MC^2: Multi-concept Guidance for Customized Multi-concept Generation

Supplementary Material

In Sec. 6, we show the details of the MC++ benchmark. In Sec. 7, we show more implementation details, including hyperparameter setting and corresponding ablation. In Sec. 8, we discuss the refinement stage mentioned in Sec. 4.7 of the main paper and more possible limitations. In Sec. 9, we show additional results. In Sec. 10 and Sec. 11, we discuss future work and societal impact of our work.

6. MC++ Benchmark

We adapt CustomConcept101 [26] to compositions of three and four subjects. The original CustomConcept101 is a dataset of 101 concpets with 3-15 images for each concept. For evaluating multi-concept customization, the CustomConept101 contains prompts for 101 compositons of two subjects. To facilitate the evaluation of multi-concept customization on more than two subjects, we collect prompts for 57 compositions of three subjects, and 14 compositions of four subjects from the original CustomConcept101. Each composition has 12 prompts. Basicly we follow the procedure of [26] to get the prompts. We first used ChatGPT [7] to propose the prompts then manually filter them to get the final set of prompts.

7. Implementation Details

We choose LoRA as our single-concept customization model. We adopt a popular community implementation of LoRA [48]. LoRA modules are trained for linear layers and 1×1 conv layers in text encoder and unet with rank set to 16. We do not finetune the token embeddings. All reference images are captioned as `photo of a <concept name>`. Each concept has a unique `<concept name>` defined by CustomConcept101, *e.g.* `pet_dog1`. All LoRAs are trained for 1000 steps with batch size set to 2, learning rate set to 1e-4. LoRA scale is set to 0.7 for merging the LoRA parameters into the diffusion model during inference.

Here we list the hyperparameters used in our method. In Eq. (7), α is set to 0.8. In Eq. (8), we schedule the learning rate λ linearly. It starts from 20 then decays to 10 linearly. We take a single gradient step per diffusion time step. In Eq. (9), $w_0, w_1, ..., w_n$ are set to $1.4, 5.6, ..., 5.6$ for customized multi-concept generation, set to $5.6, 1.4, ..., 1.4$ for compositional generation. In Eq. (10), α_1, α_2 are set to 0.5, 0.4. The MCG is performed at the first 25 steps of the diffusion process. The Gaussian filter used to smooth the cross-attention maps has a kernel size of 3 and a standard deviation of 0.5. Figure 9 shows ablation of the hyperparameters. The α controls the balance between $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$. The w controls the balance between the uncustomized model and the customized models. The red box marks our recommended hyperparameters. We also provide a quantitative ablation in Tab. 5. We set $\omega_0 + \omega_1 = 7$, which is a common CFG scale. And $\omega_1 = \omega_2$, which means the two customization modules are given equal importance. Higher ω_0 results in higher CLIP-T. And there is a trade-off between prompt fidelity and subject fidelity. The α affects the overall effect.

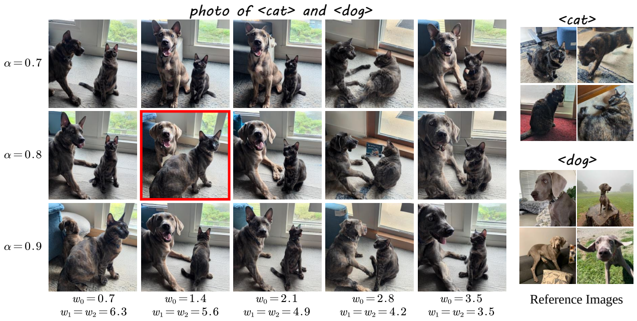
photo of <cat> and <dog>
$\alpha = 0.7$
$\alpha = 0.8$
$\alpha = 0.9$
$w_0 = 0.7$ $w_1 = w_2 = 6.3$
$w_0 = 1.4$ $w_1 = w_2 = 5.6$
$w_0 = 2.1$ $w_1 = w_2 = 4.9$
$w_0 = 2.8$ $w_1 = w_2 = 4.2$
$w_0 = 3.5$ $w_1 = w_2 = 3.5$
<cat>
<dog>
Reference Images

Figure 9. **Abltion of hyperparameters of our method.** Ablation of hyperparameters α, w_0, w_1, w_2.

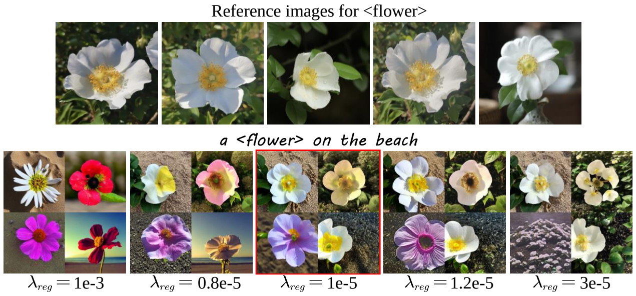

Figure 10. **Ablation of the parameter λ_{reg} in Cones 2 [31].** The λ_{reg} is the coefficient of the loss term $\mathcal{L}_{reg}$ for training a Cones 2 model.

About the implementation of the baselines, we basically follow their official repo except Cones 2 [31]. It was implemented based on SD 2.1 rather than SD 1.5. As our experiments are based on SD 1.5, we adapt the official code to SD 1.5. However, we found their default hyperparameter sub-optimal for SD 1.5. As shown in Fig. 10, when the λ_{reg} is set to default value 1e-3, the model failed to learn the `<flower>` concept. The λ_{reg} is the coefficient of the regularization term $\mathcal{L}_{reg}$ for training a Cones 2 model. A large λ_{reg} hinders the learning of the model. We set λ_{reg} to 1e-5 in the experiments.

	CLIP-T/CLIP-I/DINO			
	ω_0=0.7 ω_1=ω_2=6.3	ω_0=1.4 ω_1=ω_2=5.6	ω_0=2.1 ω_1=ω_2=4.9	ω_0=2.8 ω_1=ω_2=4.2
α=0.7	0.655/0.730/0.456	0.714/0.719/0.429	0.730/0.726/0.437	0.731/0.753/0.482
α=0.8	0.745/0.732/0.455	0.771/0.759/0.488	0.769/0.747/0.481	0.797/0.748/0.485
α=0.9	0.735/0.736/0.461	0.756/0.756/0.485	0.758/0.755/0.474	0.776/0.730/0.451

Table 5. **Quantitative ablation of hyperparameters.**

8. Limitations

Algorithm 1 Propose Masks Based on Cross-Attn Maps

```
Input: Cross-attention maps $A_1, A_2$ for the two subjects,
    overlap threshold $\theta_1$, binarization threshold $\theta_2$.
Output: Masks $M_1, M_2$ for the two subjects
 1: $A_1 \leftarrow \text{Gaussian}((A_1 - \min(A_1))/\max(A_1))$
 2: $A_2 \leftarrow \text{Gaussian}((A_2 - \min(A_2))/\max(A_2))$
 3: $M_1, M_2 \leftarrow A_1 > A_2, A_2 > A_1$
 4: $O \leftarrow (A_1 > \theta_1) \& (A_2 > \theta_1)$
 5: $M_1 \leftarrow M_1 \odot (1 - O \& M_2)$
 6: $M_2 \leftarrow M_2 \odot (1 - O \& M_1)$
 7: $M_1, M_2 \leftarrow M_1 > \theta_2, M_2 > \theta_2$
 8: $M \leftarrow \text{Dilate}(M_1 | M_2)$
 9: $M \leftarrow M - M_1 | M_2$
10: $D_1 \leftarrow \text{DistanceTransform}(1 - M_1)$
11: $D_2 \leftarrow \text{DistanceTransform}(1 - M_2)$
12: $M_1 \leftarrow M \odot (D_1 < D_2) + M_1$
13: $M_2 \leftarrow M \odot (D_2 < D_1) + M_2$
14: $M_1 \leftarrow \text{Gaussian}(M_1)$
15: $M_2 \leftarrow \text{Gaussian}(M_2)$
16: return $M_1, M_2$
```

As mentioned in Sec. 4.7 of the main paper, we add a refinement stage for the composition of multiple concepts that share similar features. When two concepts share similar features, it may be difficult to tell them apart based on the corresponding cross-attention maps. Here we propose an algorithm for extracting masks for the concepts based on the cross-attention maps. The goal is to propose a binary mask with soft boundary for each concept. Algorithm 1 shows how to extract masks for two subjects. We visualize the masks in Fig. 11. In Lines 1 and 2, we rescale and smooth the attention maps. In Lines 3 to 6, we detect the overlap of the two attention maps with a threshold. In Line 7, we binarize the attention maps to get the masks. From Line 8 to Line 13, we dilate the masks and assign the dilated area to the proper subject based on distance transform. Finally we perform Gaussian blur to get a soft boundary. The masks M_i are then used in Eq. (14) for merging the outputs of multiple diffusion models, which is identical to [5].

$$z_{t-1} = z^u_{t-1} + \sum_{i=0}^{n} w_i (M_i \odot z^i_{t-1} - z^u_{t-1}). \tag{14}$$

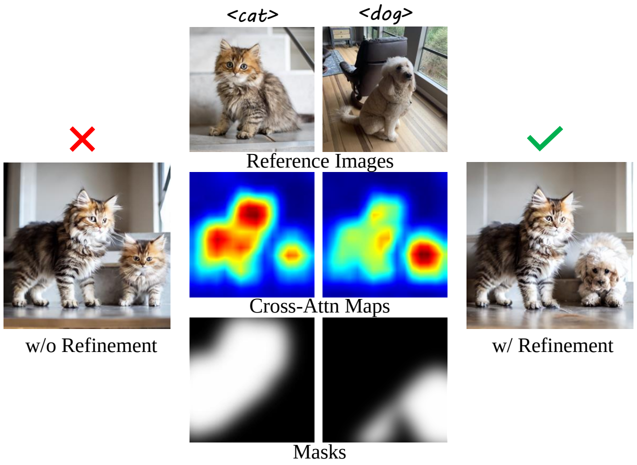

Figure 11. **Visualization of the refinement stage.** Here we show the cross-attention maps and masks involved in the refinement stage.

Note that the masks are generated on the fly and we replace the Eq. (9) in the main paper with Eq. (14). As shown in the Fig. 11, with the refinement stage, the dog is correctly generated.

Another limitation of our method is that the architecture of multiple parallel diffusion models requires relatively larger memory usage, particularly when composing multiple customized concepts. This is partly due to our implementation. We literally maintain multiple instances of the same diffusion model in memory for the sake of simplicity. Addressing this limitation involves optimizing memory utilization by storing only a single instance of the diffusion model in memory, thereby enhancing memory efficiency.

Despite MC^2 enables users to compose multiple separately trained, even heterogeneous customized models, the customized models should be trained from the same diffusion model. Such limitation is inherited from [29].

9. Additional Results

Figure 12 shows more qualitative comparisons of the proposed method and the baselines [16, 26, 31] on customized multi-concept generation. The baselines sometimes omit one of the specified concepts, *e.g.* the white chair for `<sofa> and <chair>` and the person for `<person> and <cat>`. Our method demonstrates higher fidelity to the reference images even compared to Custom Diffusion [26] that requires jointly training the two concepts, or Mix-of-Show [16] that requires training to merge the two single-concept customized models. Cones 2 [31] shows relatively low fidelity to the reference images, considering that it requires the least trained parameters. Our method demonstrates a more satisfying effect. Figure 13 shows qualitative comparisons of the customized multi-concept genera-

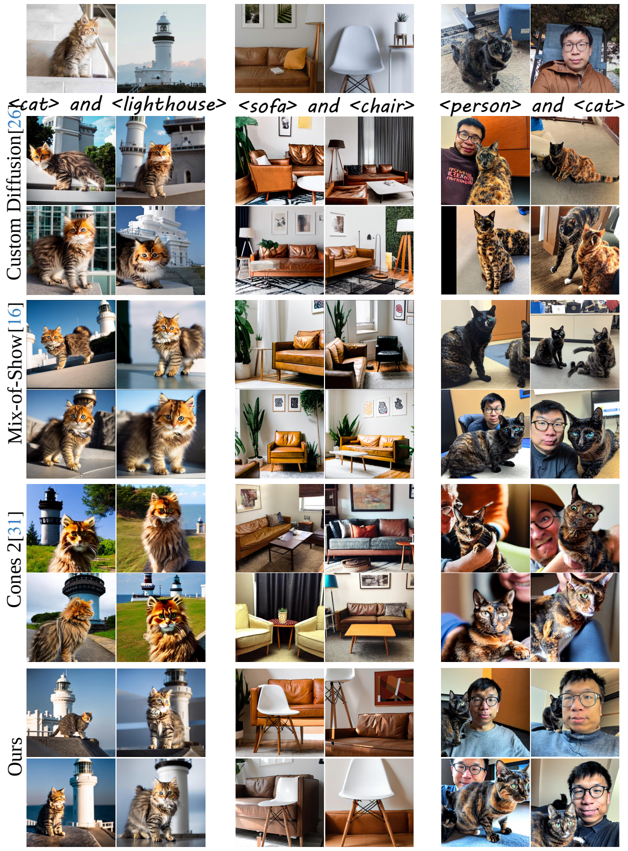

Figure 12. **Comparisons on generating two subjects.** Qualitative comparisons of customized multi-concept generation methods.

tion methods on composition of three concepts. Our method demonstrates more higher fidelity to the reference images.

Figure 14 shows more qualitative comparisons of the compositional generation methods. Our method demonstrates better consistency with the input text prompts compared to the baselines [8, 28, 29].

10. Future Work

In addition to addressing the limitations mentioned in Section 8, there exist several promising avenues for future research. [9, 51] delve into the realm of multi-concept customization for text-to-video generation. An intriguing prospect is to investigate the adaptability of our proposed MC^2 to the domain of text-to-video generation.

For compositional generation, an interesting avenue for exploration involves building upon our methodology. Our approach not only addresses current challenges but also opens up a novel design space for further investigation. This provides a foundation for the development of innovative methods to enhance compositional generation techniques.

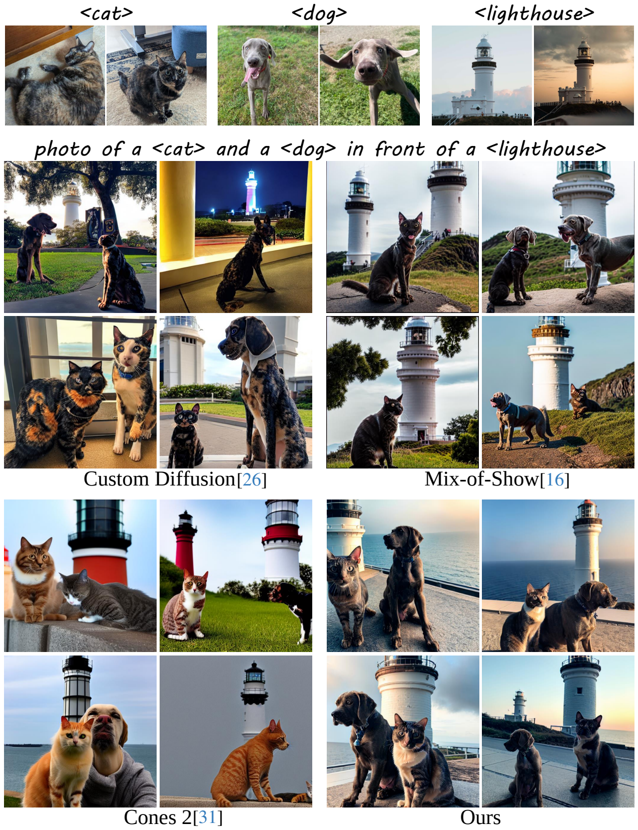

Figure 13. **Comparisons on generating three subjects.** Qualitative comparisons of customized multi-concept generation methods on three subjects.

11. Societal Impact

First and foremost, MC^2 empowers users to effortlessly generate visually captivating compositions reflecting their unique ideas and preference. Additionally, MC^2's ability to enhance the capabilities of existing text-to-image diffusion models opens up new avenues for artistic exploration and innovation, potentially inspiring broader adoption and engagement in creative endeavors. However, MC^2 may blur the lines between ethical and unethical image manipulation. Without proper guidance and ethical considerations, individuals may engage in harmful practices such as image forgery or digital impersonation. We advocate for the development of legal frameworks that address AI-generated content, including penalties for malicious use.

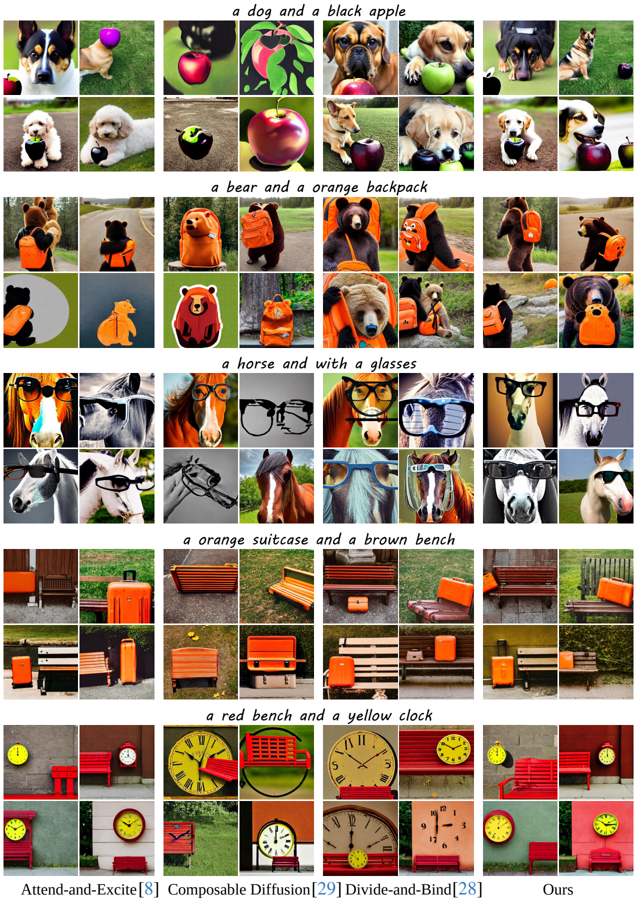
a dog and a black apple
a bear and a orange backpack
a horse and with a glasses
a orange suitcase and a brown bench
a red bench and a yellow clock
Attend-and-Excite[8]
Composable Diffusion[29]
Divide-and-Bind[28]
Ours

Figure 14. Qualitative comparisons of compositional generation methods.