Concept Weaver: Enabling Multi-Concept Fusion in Text-to-Image Models

Supplementary Material

A. Method Details

Details of Concept Bank Training. Given the model and image examples with custom concepts, we can fine-tune the components of the model to embed the single-concept into the pre-trained model. Textual Inversion [1] has been widely adopted; however, it suffers from undetailed expression of custom concept due to the limited degree of freedom. There is also Dreambooth [4], which requires fine-tuning of all the parameters of the model, making it time consuming to fine-tune to a large number of concepts. As we will leverage the self-attention layer and residual block features as a source for structural preservation, we chose framework of Custom Diffusion [2] following the score matching loss:

$$\mathbf{E}_{\epsilon,x,p,t}[||\epsilon - \epsilon_\theta(x_t, p, t)||], \tag{1}$$

where ϵ_θ is denoising network and ϵ is sampled noise from unit gaussian. t, p represents timestep and text condition, respectively. With the text condition $p \in R^{s\times d}$ and self-attention feature $f \in R^{(h\times w)\times c}$, the cross attention layer consists of $Q = W^q f, K = W^k p, V = W^v p$, and the attention output is represented as :

$$A(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V.$$

We only fine-tune the 'key' and the 'value' weight parameters, W^k, W^v, of the cross-attention layers. Also, we use modifier tokens [V*], which are placed ahead of the concept word (*e.g.*, [V*] dog) and operate as a constraint to general concepts.

Unlike the basic models of Custom Diffusion, our approach incorporates a robust augmentation strategy. This involves significantly varying the size and position of training images within the overall dataset. Such resizing and repositioning augmentations grant greater geometric freedom, or action expressiveness, to the generated outputs. Additionally, this method helps to minimize potential artifacts during the region-specific denoising phases, enhancing the overall quality and accuracy of the generated images.

We can also incorporate Low-Rank (LoRa) adaptation on our framework. In case of using LoRa-based adaptation, we fine-tune the Low Rank nodes on all of weights of query, key, and value of cross attention layers. More specifically, we only fine-tune low-rank bias $\Delta W^q, \Delta W^k, \Delta W^v$ to obtain new weights $W^{q-new} = W^q + \Delta W^q, W^{k-new} = W^k + \Delta W^k, W^{v-new} = W^v + \Delta W^v$. In our case, we used rank $r = 4$.

Details of Template Image Generation. In template image generation process, we use Stable Diffusion [3] model version $\geq$2.0 as the earlier version models often fail to generate images that contain multiple objects.

More specifically, when we use Stable Diffusion v2.1, we optionally used guided generation process in which to use multi-concept guidance prompt such as p_{mc} = *"photo of two animals in the same background"*, along with target prompt (e.g. p_{tg} =*"photo of a dog and a cat playing with a ball, mountain background"*). At each generation steps, we use the summed version of two score outputs from two prompts such as $\epsilon = \epsilon_\theta(z_t, t, p_{tg}) + \lambda\epsilon_\theta(z_t, t, p_{mc})$. If we use Stable Diffusion XL (SDXL), we did not used multi-concept guidance prompt. In practice, we recommend to use SDXL for high fidelity.

Details of Inversion and Feature Extraction. From the source image x_{src}, we generate the noisy latent space z_T with the DDIM [5] forward process:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}z_t + \left(\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\epsilon_\theta(z_t, t, p_{src}),$$

where we deterministically get the next step latent z_{t+1}. Here $\alpha := \Pi_{i=1}^{t}(1-\beta_t)$, and β_t is the variance schedule. From the inverted latent z_T, we can accurately reconstruct the source image using a reverse DDIM process [5]:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}z_t + \left(\sqrt{\frac{1-\alpha_{t-1}}{\alpha_{t-1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\epsilon_\theta(z_t, t, p_{src}).$$

During the reverse reconstruction process, we extract the features from the U-Net's l-th layer f_t^l at each timestep t.

Details of Implementation. Instead of using a densely annotated mask, we used dilated mask in which the mask region is expanded from the original area. Here we used a filter size of 21x21 for the mask dilation. If we used real concepts, we used original dilated masks. When we generated the images which contain unreal concepts such as animated characters, we found that using rectangular masks (e.g. in the second row of Fig. 3) shows better results.

For self-attention and residual layer feature injection, we only apply the injection to early timesteps. If our entire timesteps for sampling is T, we apply self attention injection to early timesteps such as $t > 0.6T$, and residual layer injection to $t > 0.5T$. For concept-free suppression, we used weight of $\lambda = 0.3$.

In our generation pipelines, we can filter out unsatisfied samples in mask generation steps. If we cannot obtain the proper concept-wise objects masks in the template images, we filter out the image and use other templates. We can automatically drop the sample if the overlapping regions of

two extracted masks are over 90 percent. Also, we randomly showed the generated outputs with CLIP text-image similarity scores higher than 0.3. For fair comparison, we applied same filtering protocol to the baseline of Mix-of-show. In case of early methods, we only applied the CLIP based filtering, as the methods suffer from severe concept missing.

B. Further Comparison

To further compare the generation process between our proposed method and Mix-of-show, we show the further comparison results. As both methods rely on region-wise guidance for multi-concept generation, we compare the difference between two methods in Fig. 3. In our proposed method, we start from generated template images and the object-wise segmented masks. With those conditions, we can translate the template images to concept-aware outputs. In case of Mix-of-show, the method relies on rectangular shape layout boxes, and also apply concept-wise sampling on each box region.

As observed in the figure, the output objects from mix-of-show only follow the approximated spatial conditions of given box regions, as it is much more sensitive to initial noise conditions. In our case, as we start from template images, the output concepts accurately follow the mask regions.

In order to show the comparison with more generated samples, we show the outputs in Fig. 4 and Fig. 5. For fair comparison, we show the outputs filtered with protocols elaborated in our implementation details. In case of Mix-of-show, we can see the generated concepts are properly places on some samples, but in many cases the concept is not properly applied. Also, if we generate the objects with complex actions or interactions (e.g. 'kissing', 'riding a boat'), the outputs from Mix-of-show often fails to reflect the text conditions or suffer from the two concepts mixing. Considering that baseline of Mix-of-show requires additional optimization for concept weight combining, our method shows superiority in both of generation quality and flexibility.

For more detailed comparison on perceptual quality, we show the detailed user study result in Table. 1. We conducted detailed user study using three different parts: background, human face, and real concepts. To evaluate the generation quality, we asked the users to score their preference with more detailed questions: 1) Inclusion of target background or human face concepts (Concept Match) , 2) Realism of generated background or human faces (Realism). Also, we asked same questions to users with showing the generated images on the real concepts. The results show that our proposed method outperforms our main baseline of Mix-of-show in all categories.

Method	Background		Human Face		Real Concept	
	C. Match↑	Realism↑	C. Match↑	Realism↑	C. Match↑	Realism↑
Mix-of-show	3.83	4.08	2.52	3.04	3.67	3.75
Ours	**4.29**	**4.46**	**4.34**	**4.05**	**4.58**	**4.42**

Table 1. **Human Preference Study**. We assess three different categories of Background, Human Face, and Real concepts. We collected answers from 12 different users each assessing 20 images.

C. More Qualitative Results

In order to further show the qualitative results on animated concepts and concepts in same category, we show the outputs in Fig. 6. Our method can generate multi-concept outputs even with animated characters. In the third row, we show the outputs with two concepts which are within same category. Even we use the custom concepts with the same class, we can generate the multi-concept aware results without concept mixing. In Fig. 7, we show more qualitative result using Low-Rank adaptation for single-concept customization.

In order to experiment the multi-concept personalized generation on local regions, we show the results of multiple concept fusion on single subject (e.g. human) in Fig. 1. The results shot that our proposed method works not only for multiple separated objects, but also to the local components of single object. The results further show the robustness of our proposed method.

D. Details of Evaluation

For image-alignment score calculation, since our generated images contain multiple concepts, we cannot use the whole image-wise similarity scores. Instead, we extracted the concept-wise images using text-guided segmentation model. For example, if we evaluate images which contain '[c1] dog' and '[c2] cat', we run a segmentation model with the text prompts of 'dog' and 'cat' to obtain segmented masks. Then we cropped the rectangular region which contain segmented masks from the image. Then we calculated the cosine similarity between the image embedding vectors from extracted images and the concept (training) images. As the baseline methods often fails to generated all concepts, we did not calculated the scores when the generated images fail to contain all foreground concept objects for fair comparison.

For human preference evaluation, we collected opinions from 20 participants from the age group of 20-49. We constructed 2 different survey sets, each of which contains 10 generated images per each baseline model and 10 questions. We use the generated outputs from baselines and ours : Textual Inversion, Custom Diffusion, Perfusion, Mix-of-show and ours. Therefore, each survey set contains 50 generated images. We divided the participants into two groups and gave them different survey set. For further explanation, we show the example of survey form in Fig. 8.

Figure 1. **Composing custom concepts into single object.** We showcase a successful generation of custom local concepts.

Figure 2. **Failure Cases.** If we use extremely complex or unrealistic text conditions, our method shows degraded generation performance.

E. Limitations and Societal Impacts

Limitations. Although our method shows great performance in multi-concept generation, our method still has limitations. If we give extremely difficult or unrealistic text conditions, our method still show limited performance in text-alignment such as in Fig. 2. Since this problem comes from the limited performance of pre-trained Stable Diffusion, we expect to solve the problem with using improved diffusion model backbones.

Societal Impact. Since our method can synthesize realistic custom concept images, our method can be maliciously abused if the privacy-sensitive concepts are used. To prevent this, there should be a proper filtering system to check if the training concept is free from ethics issue.

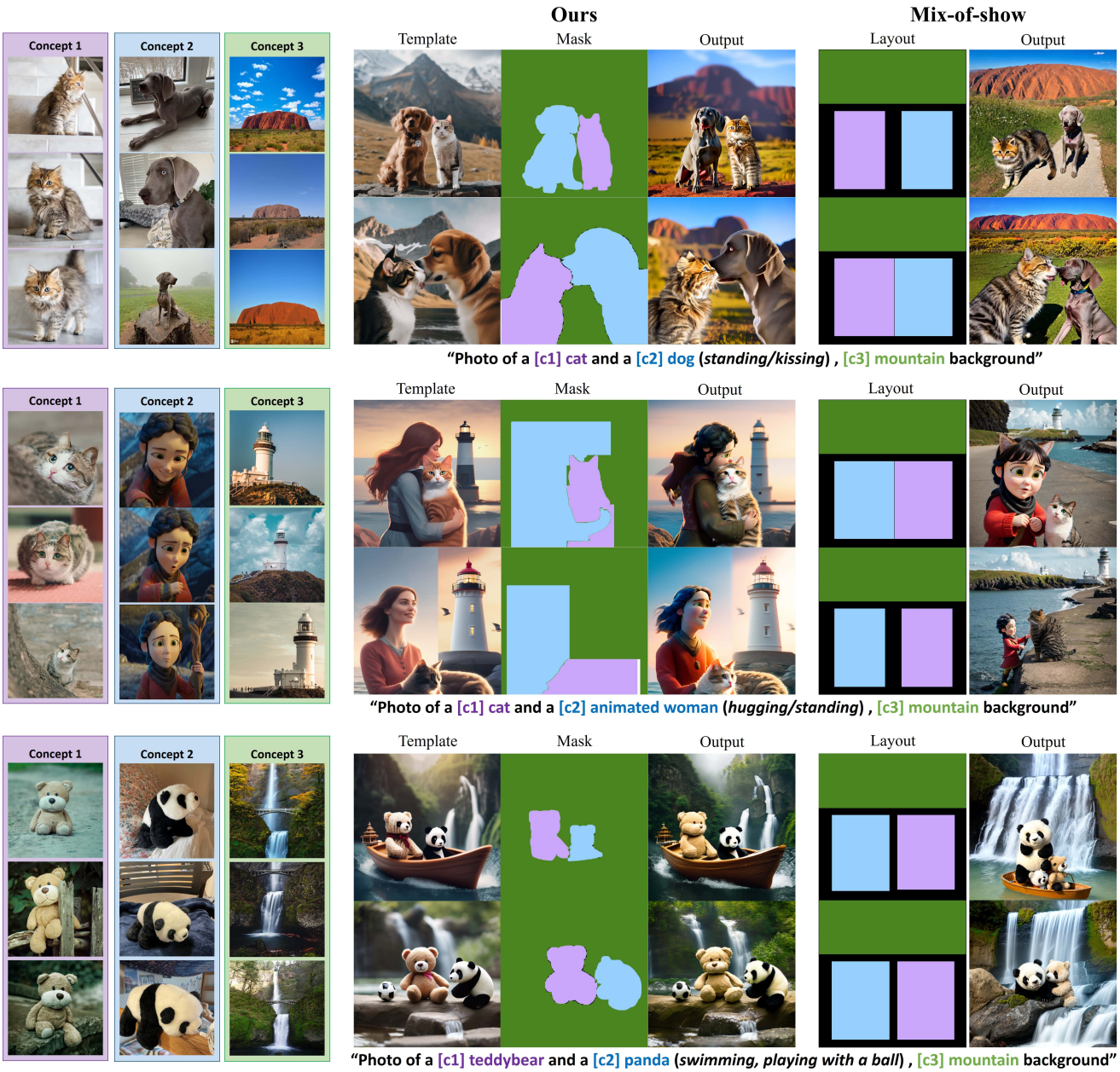


Figure 3. **Detailed Generation Outputs.** We show the detailed generation process of ours and the baseline method. In our proposed method, we use template image and concept-wise mask condition for generating accurate multi-concept images. For the baseline mix-of-show, the method use layout information for multi-concept generation.

Figure 4. **Further Comparison with Mix-of-show.** We show the comparison results with the baseline of Mix-of-show. Our method successfully generated the target concepts following the given text conditions while the baseline method suffers from concept mixing or misalignment with text conditions.

Figure 5. **Further Comparison with Mix-of-show.** We show the comparison results with the baseline of Mix-of-show. Our method successfully generated the target concepts following the given text conditions while the baseline method suffers from concept mixing or misalignment with text conditions.

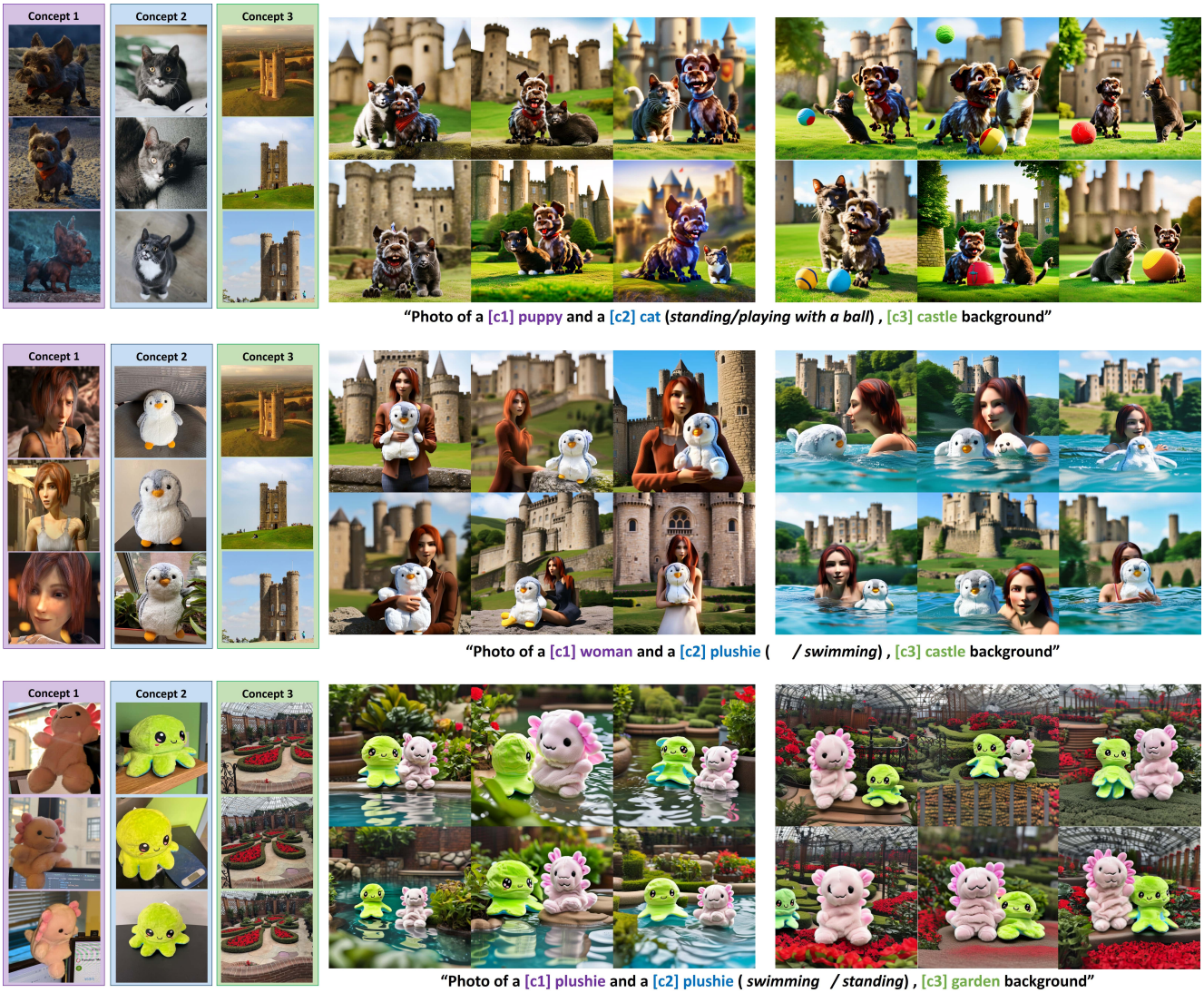


Figure 6. **More Qualitative Results.** We show more comparison results including animated concepts (the 1st, 2nd rows), and including two concepts within the same category (3rd Row), respectively.

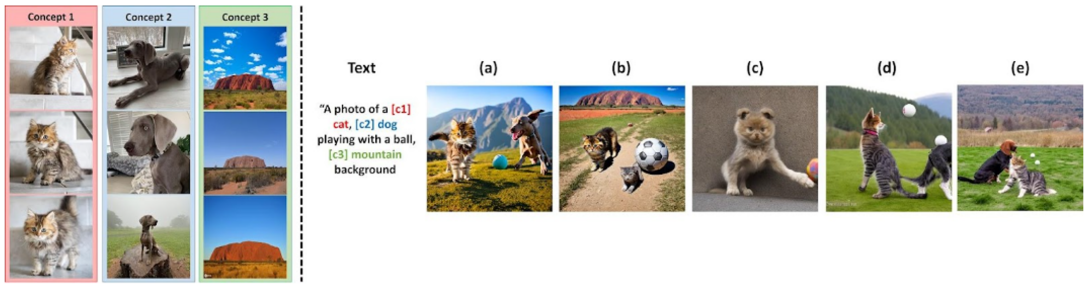
Figure 7. **More Qualitative Results on Low Rank adaptation.** We show more generated outputs from our method using Low-Rank adaptation based fine-tuning.

Question 1

The following images are synthesized images with given concepts (left side) and text condition. Please score the images (a) - (e) .

Do you think the images reflect the given text condition well?

(5 : strongly agree / 4: agree / 3: neutral / 2: disagree / 1: strongly disagree)



	1 point	2 point	3 point	4 point	5 point
(a)	○	○	○	○	○
(b)	○	○	○	○	○
(c)	○	○	○	○	○
(d)	○	○	○	○	○
(e)	○	○	○	○	○

Figure 8. **Human Evaluation Example.** We show the example question for human preference evaluation.

References

[1] 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

[2] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1

[3] 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. 1

[4] 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

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