MUSE: Multi-Subject Unified Synthesis via Explicit Layout Semantic Expansion

Supplementary Material

In this supplementary material, we provide more design details of our layout-controllable multi-subject synthesis (LMS) method MUSE, along with extensive experimental results. These include comparisons between our proposed MUSE and other LMS methods, evaluations of concatenated cross-attention (CCA) versus decoupled cross-attention (DCA), ablation studies on the progressive two-stage training strategy, and ablation experiments on the subject synthesis strength scale. Additionally, in the $MUSE\_AttnProcessor.py$ file, we provide the implementation of the final cross-attention operation that integrates both the CCA and DCA methods. This implementation is built using the diffusers [5] library.

A. More Details of Encoding Control Information

For the text for layout control and the images for subject synthesis, we use CLIP-ViT-L-14 and CLIP-ViT-G-14 [4] models for encoding, respectively.

For layout text features, we extract the class token from the CLIP model's final output. For a class text like "dog", the encoded feature has a size of [1,768]. Bounding box information is Fourier-encoded with a frequency of 16, producing a feature of size [1, 64]. These two features are concatenated along the feature dimension, resulting in a [1, 768+64] feature, which is further processed by an MLP to produce a fused grounding token of size [1,768]. The MLP consists of three linear layers with SiLU activation functions.

For image encoding, the straightforward approach is to use the same final class token output of CLIP encoding (size [1, 1280]) concatenated with bounding box features to create the grounding token. While sufficient for simple text class, this is inadequate for subject synthesis, as the class token lacks rich spatial information. Existing subject synthesis works like IP-Adapter [8], RealCustom [1], and InstantID [6] utilize shallower CLIP features, such as the last hidden states (size [256, 1664]), which offer ample spatial information. However, excessive spatial detail leads to redundancy in subject synthesis (*e.g.*, copy-paste artifacts), especially since SDXL [3] itself uses only 77 tokens for text prompts control.

To address this, we adopt IP-Adapter-Plus's approach: train 4 learnable tokens as queries, processed through four layers of perceiver attention, extracting compressed image features of size [4, 2048]. Directly concatenating bounding box information ([4, 2048+64]) into the tokens for MLP fusion degrades subject synthesis, as independently processed tokens may produce inconsistent results due to neural network black-box behavior. Instead, we independently encode the Fourier-transformed bounding box information into [1, 2048] through an MLP, and add this layout encoding to each image token, achieving coherent grounding tokens. Mapping the dimension of the image grounding token to 2048 is intended to initialize the mapping layer parameters in DCA using those from the pretrained model, thereby reducing the training difficulty of the model.

B. More Details of Experimental Setup

Since the data samples used in both training and testing contain multiple subjects, we set the number of subjects per sample to 10. For samples with more than 10 subjects, we select the 10 largest bounding box areas. For samples with fewer than 10 subjects, we use padding by introducing trainable empty tokens to maintain 10 subjects. These include text, image, and coordinate features. During training, we randomly drop captions for images and conditions (*e.g.*, subject texts, images and bounding boxes) for MUSE with a 0.1 probability to enhance robustness.

Regarding the loss function, we maintain consistency with the original pre-trained model. The diffusion network is trained to accurately predict added noise under additional multiple control conditions.

For each test experiment, we use five random seeds and report the average results.

C. More Qualitative Experiment Results

Fig. 1 provides more qualitative comparisons of our method against other LMS approaches, including GLIGEN [2] and MS-Diffusion [7], on the MS-Bench-Random dataset. Our approach consistently demonstrates superior performance in both layout control and multi-subject synthesis.

D. More Ablation experiments on Training Strategies

Fig. 2 presents qualitative results on the MS-Bench-Random dataset, comparing our progressive two-stage LMS framework with models trained using the full-DCA method and single-stage training combining CCA and DCA method. Our proposed framework achieves both accurate layout control and effective subject synthesis. The results confirm that our proposed progressive framework is more suitable for LMS tasks.

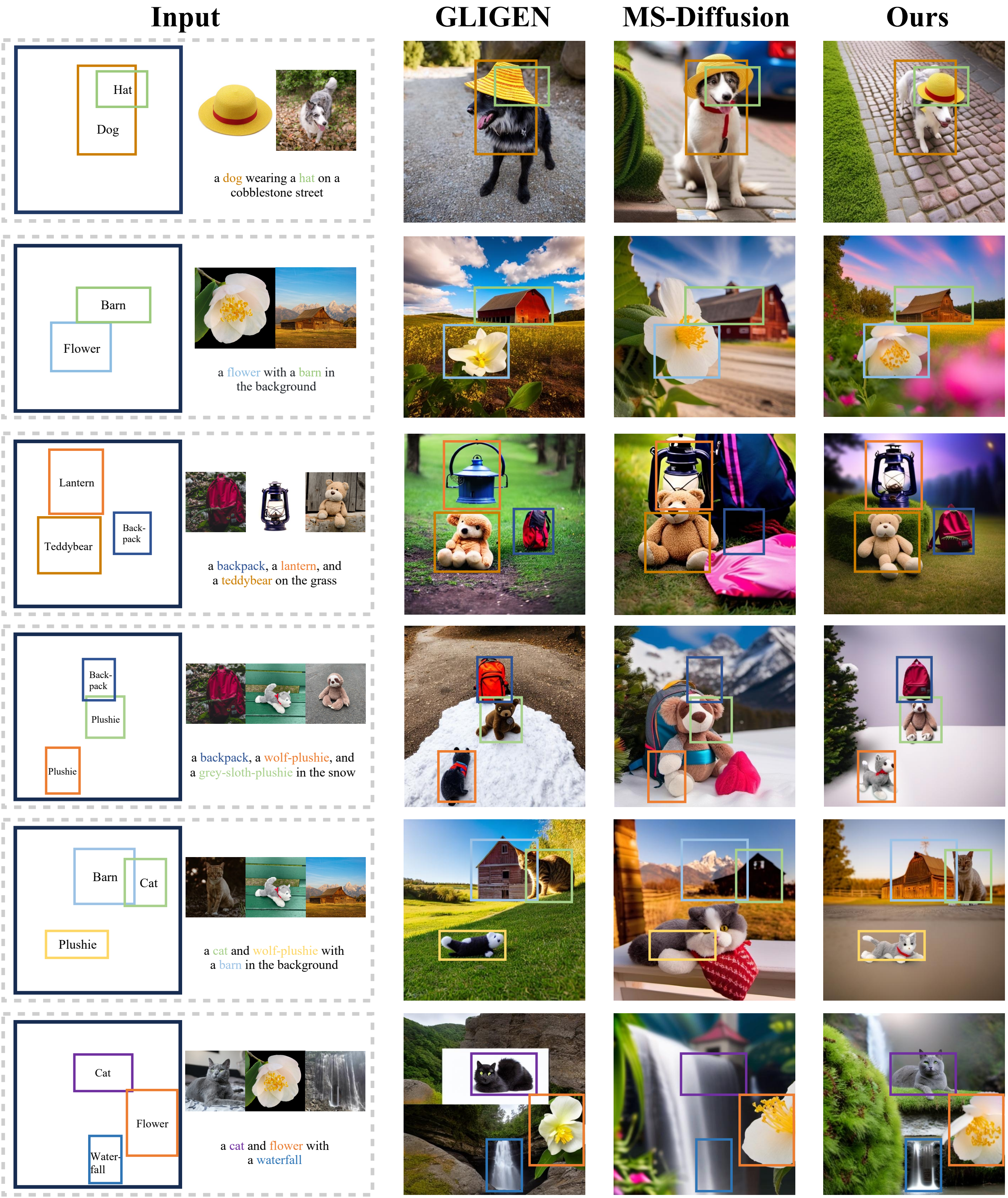
Input
GLIGEN
MS-Diffusion
Ours
Hat
Dog
a dog wearing a hat on a cobblestone street
Barn
Flower
a flower with a barn in the background
Lantern
Teddybear
Back-pack
a backpack, a lantern, and a teddybear on the grass
Back-pack
Plushie
Plushie
a backpack, a wolf-plushie, and a grey-sloth-plushie in the snow
Barn
Cat
Plushie
a cat and wolf-plushie with a barn in the background
Cat
Flower
Water-fall
a cat and flower with a waterfall

Figure 1. Qualitative experiments on MS-Bench-Random. Our method demonstrates strong LMS performance across various conditions, showing good practical applicability in real-world.

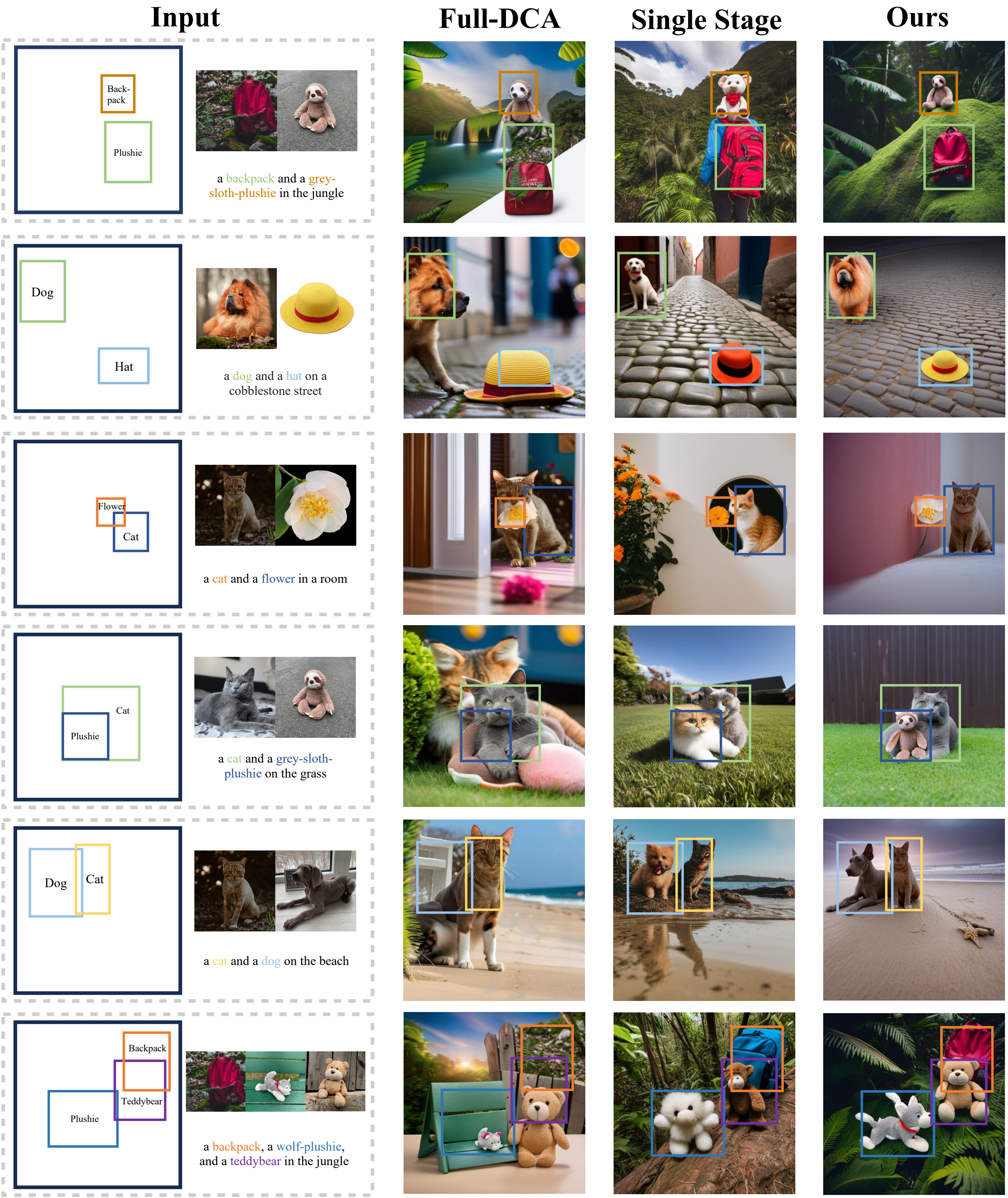
Input
Full-DCA
Single Stage
Ours
Back-pack
Plushie
a backpack and a grey-sloth-plushie in the jungle
Dog
Hat
a dog and a hat on a cobblestone street
Flower
Cat
a cat and a flower in a room
Cat
Plushie
a cat and a grey-sloth-plushie on the grass
Dog
Cat
a cat and a dog on the beach
Backpack
Teddybear
Plushie
a backpack, a wolf-plushie, and a teddybear in the jungle

Figure 2. Ablation study on training strategies conducted on MS-Bench-Random. It shows that using both CCA and DCA methods along with our proposed progressive two-stage training strategy significantly improves performance on the LMS task.

Scale	MS-Bench				MS-Bench-Random			
	CLIP-T	CLIP-I-l	SR-0.6	SR-0.65	CLIP-T	CLIP-I-l	SR-0.6	SR-0.65
0.6	0.328	0.811	0.878	0.790	0.327	0.769	0.871	0.712
0.8	0.323	0.827	0.890	0.819	0.321	0.779	0.894	0.755
1.0	0.313	0.832	0.889	0.822	0.310	0.782	0.897	0.763

Table 1. Ablation experiments on the subject synthesis strength scale. The evaluated metrics include CLIP-T, CLIP-I-local (abbreviated as CLIP-I-l), and LMS Success Rate (SR), determined using CLIP-I-local score thresholds of 0.6 and 0.65, referred to as SR-0.6 and SR-0.65, respectively.

E. Ablation experiments on Subject synthesis Strength Scale

We conducted ablation experiments on the subject synthesis strength scale λ. While the default sacle is set to 0.8, we provide results for $\lambda = 0.6$ and $\lambda = 1.0$ on the MS-Bench-Random dataset. Quantitative comparisons are shown in Tab. 1. The results indicate that high subject synthesis strength scale can weaken text-following ability, while low strength scale reduces subject synthesis quality. A balanced value achieves optimal performance.

References

[1] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: narrowing real text word for real-time open-domain text-to-image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2024. 1

[2] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1

[3] 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. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[5] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[6] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1

[7] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 1

[8] 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