Multi-Subject Personalization

Arushi Jain, Shubham Paliwal, Monika Sharma, Vikram Jamwal, Lovekesh Vig
TCS Research, New Delhi, India
{j.arushi, shubham.p3, monika.sharma1, vikram.jamwal, lovekesh.vig}@tcs.com

Abstract

Creative story illustration requires a consistent interplay of multiple characters or objects. However, conventional text-to-image models face significant challenges while producing images featuring multiple personalized subjects. For example, they distort the subject rendering, or the text descriptions fail to render coherent subject interactions. We present Multi-Subject Personalization (MSP) to alleviate some of these challenges. We implement MSP using Stable Diffusion and assess our approach against other text-to-image models, showcasing its consistent generation of good-quality images representing intended subjects and interactions.

1. Introduction Text-to-image generation deep models [10, 9, 13] have found applications across various industries, delivering impressive visual assets for purposes such as advertising, marketing, entertainment, and creative content creation like story visualization. Fine-tuning through models such as LoRA [5], Textual Inversion [2], and Dreambooth [12] further allows personalization of style and objects. Nevertheless, when generating more than one personalized subject, they frequently produce images with unclear or hybrid characteristics of multiple subjects. Recently, efforts have been made to address this limitation, with models like Custom-Diffusion [7] and Subject-Diffusion [8], allowing for the generation of images with multiple personalized subjects. Nonetheless, these models still struggle to generate complex compositions with more than two customized subjects. To address these constraints, we present a novel approach for generating images featuring multiple personalized subjects based on Stable Diffusion [10], including humans, animals, and objects. Our method enables more precise control over the subject's appearance and positioning within the generated image, thus offering artists the enhanced ability to craft complex scene imagery tailored to specific requirements.

2. Proposed Method: MSP-Diffusion Our approach for Multi-subject Personalization, using diffusion-based image generation models, as shown in Figure 1, has the following four salient features:

(i) Fine-tuning the model for multiple personalized subjects: Our approach centers on embedding a specific subject instance within the output domain of a text-to-image diffusion model by associating each subject with a unique identifier. We achieve this by jointly fine-tuning the Dreambooth model using a limited set of sample images for each subject, represented by distinct modifier tokens, V^*, initialized with different rarely-occurring tokens. We incorporate training images featuring subjects in diverse angles and poses to enhance subject comprehension and, consequently, improve reverse diffusion-based generation.

(ii) Controlling the positioning and orientation of multiple subjects: We employ the Composite Diffusion technique [6], as shown in Figure 1, to precisely control the spatial placement of personalized subjects using subject-segment image in the diffusion process. The approach compels the diffusion process to create an appropriate subject in the corresponding segment region, effectively resolving issues related to missing and hybrid subjects while ensuring the correct number of subjects are generated. We further provide an implementation involving ControlNet [14], where we introduce control conditioning inputs in the form of Openpose[1] images, in addition to text-based conditioning. The subject-segment image defines individual subjects, while the Openpose information controls the subject's posture and scale.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

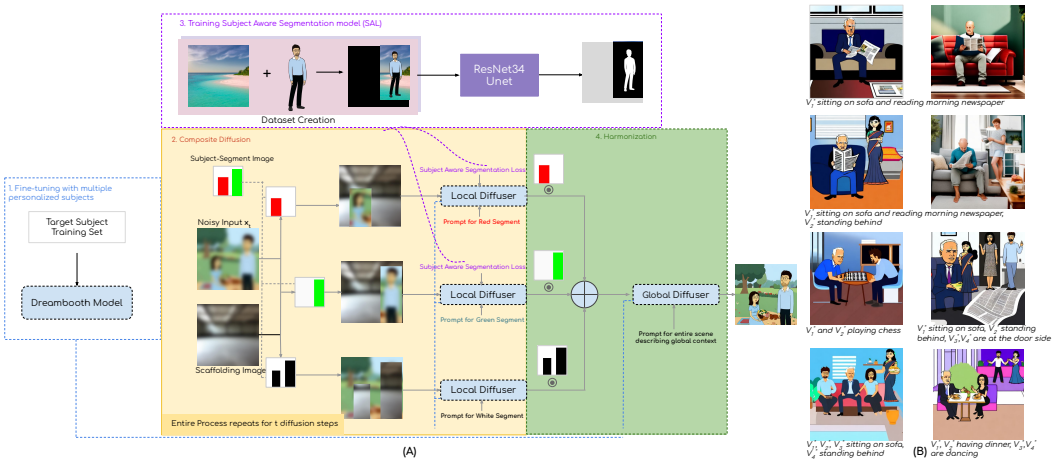

Figure 1: (A) Overview of the MSP-Diffusion architecture, (B) Sample generations from the method

(iii) Ensuring precise subject-appearance: In order to gain control over the desired subject's appearance in the generated image, we introduce a subject-aware segmentation loss (SAL) derived from a ResNet-34 based U-Net [11] segmentation model trained on the same dataset of target subjects used to fine-tune the Dreambooth model.

(iv) Harmonization of multiple subjects: Similar to Composite Diffusion [6], we use a two step generation process. During the initial k steps, when subjects in individual segments are constructed independently, the resulting composite often lacks harmony and smooth blending at the segment edges. To address this, harmonization is performed that allows each segment to develop in the context of the others. Additionally, we suggest the use of a global diffuser to capture the image's overall context, ensuring harmonization among the various segments containing personalized subjects.

3. Evaluation We evaluate MSP-Diffusion on our self-compiled Subject-Dataset11 featuring 11 diverse target subjects, including comic characters (male, female), human characters (male, female), pets, and still objects. We used three evaluation metrics for performance comparison: (CLIP-I) [2], (CLIP-T) [3], and image alignment using our subject-aware similarity (SAS). Table 1 shows that as the number of personalized subjects increases, Dreambooth and Textual Inversion exhibit reduced image alignment, whereas MSP-Diffusion without scaffolding image, achieves a CLIP-I score of 0.93. Notably, our method consistently outperforms previous models in terms of CLIP-I, CLIP-T, and SAS scores when dealing with scenarios involving more than one personalized subject. Next, in the case of text alignment, the performance of our method with scaffolding image deteriorates as the number of subjects increases due to the generation of hybrid or missing subjects. Therefore, ControlNet [14] based Openpose is incorporated in the MSP-Diffusion which successfully mitigates the issue, albeit introducing a dependency on Openpose. Please refer to Appendix for more details.

Table 1: Table showing comparison of proposed MSP-Diffusion on Subject-Dataset11. Here, n, GD, SAL and SAS refers to number of personalized subjects, Global Diffuser, Subject-aware loss and Subject-aware Similarity, respectively.

Methods	**Configuration**	**n = 1**			**n = 2**			**n = 3**			**n >3**		
		CLIP-I	CLIP-T	SAS	CLIP-I	CLIP-T	SAS	CLIP-I	CLIP-T	SAS	CLIP-I	CLIP-T	SAS
Dreambooth[12]		**0.93**	0.50	0.89	0.91	0.53	0.85	**0.92**	0.52	0.83	0.92	0.53	0.83
Textual Inversion [2]		0.92	**0.75**	0.83	0.90	0.74	0.81	0.91	**0.74**	0.78	0.92	**0.75**	0.79
Ours w/ scaffolding	GD=F, SAL=F	**0.93**	0.72	0.89	0.91	0.77	0.87	0.91	0.73	**0.86**	0.88	0.73	0.81
	GD=T, SAL=F	0.92	0.72	0.89	0.91	0.75	0.87	0.90	0.70	0.81	0.87	0.69	0.79
	GD=F, SAL=T	0.92	0.73	0.88	0.91	0.76	0.87	0.91	0.73	**0.86**	0.88	0.73	0.81
	GD=T, SAL=T	0.92	0.73	**0.90**	0.91	**0.77**	0.87	0.90	**0.74**	0.82	0.87	0.72	0.81
Ours w/o scaffolding	GD=F, SAL=F	0.92	0.72	0.88	**0.93**	0.66	**0.89**	**0.92**	0.63	**0.86**	**0.93**	0.68	0.83
	GD=T, SAL=F	0.92	0.72	0.88	**0.93**	0.65	**0.89**	0.91	0.63	**0.86**	**0.93**	0.68	**0.84**
	GD=F, SAL=T	0.92	0.73	0.89	**0.93**	0.65	**0.89**	**0.92**	0.63	**0.86**	**0.93**	0.68	**0.84**
	GD=T, SAL=T	0.92	0.73	0.88	0.92	0.66	**0.89**	0.85	0.63	0.82	**0.93**	0.68	0.83
Ours w/o scaffolding w/ control-net	GD=T, SAL=F	0.90	0.72	0.80	0.90	**0.82**	0.80	0.89	**0.82**	0.79	0.90	**0.83**	0.80

4. Conclusion We introduced MSP as an innovative approach for generating high-quality images featuring multiple personalized subjects. Our experimental results highlight the superior performance of our method compared to text-to-image models or vanilla personalization models such as Textual Inversion and Dreambooth.

Ethical Implications Text-to-image generation carries substantial ethical considerations. While these models offer valuable applications like creative content generation and marketing, they also open the door for misuse, enabling the creation of deceptive images. Another thing to mention here is that our evaluation dataset utilizes target subjects from Comicgen [1] and Alamy [2] strictly for research purposes.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1611.08050 [cs.CV]

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618 [cs.CV]

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV]

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[6] Vikram Jamwal and Ramaneswaran S. 2023. Composite Diffusion | whole $>= \Sigma$ parts. arXiv:2307.13720 [cs.CV]

[7] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*.

[8] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. arXiv:2307.11410 [cs.CV]

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV]

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242 [cs.CV]

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV]

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]

[1] Comicgen: `https://gramener.com/comicgen/`

[2] Alamy `https://www.alamy.com/`

A Appendix

In the appendix, we provide supplementary material to the paper due to space constraints in the main paper. It is organized as follows:

A.1 Dataset

We assess MSP-Diffusion on our self-compiled Subject-Dataset11[3] featuring 11 diverse target subjects, including comic characters (male, female), human characters (male, female), pets, and still objects. We obtain 11 training images corresponding to each target subject from online sources such as Comicgen [4] and Alamy [5]. Below, we present sample target images of the 11 subjects featured in our Subject-Dataset11 which we plan to release publicly strictly for research purposes. In Figure 2, you can observe the training images, each accompanied by unique identifiers and the corresponding subject class, used for fine-tuning Dreambooth model. For instance, "sks male cartoon" features "sks" as the unique identifier and "male cartoon" as the class to which the subject belongs. We employed 11 different instances of each target subject image, featuring variations in pose (sitting, standing, walking) and orientation (front, back, left, right). These images were instrumental in fine-tuning the Dreambooth model.

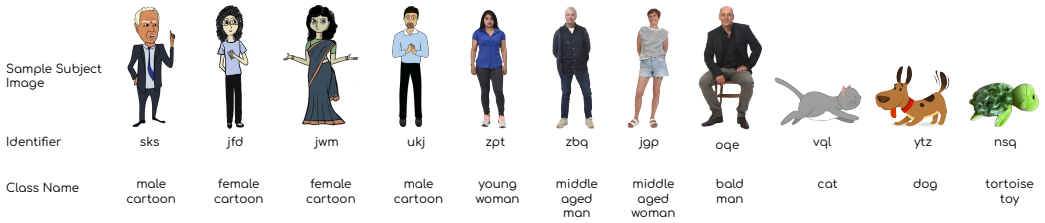

Figure 2: Figure displaying sample training images, each linked to a distinct subject, alongside the unique identifier and class name used in the fine-tuning process of the Dreambooth model.

Given the limited number of training images per target subject for fine-tuning Dreambooth and the UNet segmentation model, we augment the dataset with transformations. This includes zooming subjects (scaling factors 0.4 to 0.5), cropping subjects (0.65 to 0.95 of original size with 0.1 probability), and introducing random translations on a 512×512 image canvas. This augmentation results in a dataset of 1064 training images, 194 validation images, and 150 test images. To further enhance the UNet segmentation model's effectiveness in isolating target subjects for subject-aware loss and improved subject appearance during image generation, we introduce 100 diverse background scenes to these images. The process of cropping and merging backgrounds is illustrated in Figure 1.

The evaluation dataset includes 11 different compositions of the target subjects in various settings (1-subject, 2-subjects, 3-subjects, and >3 subjects) with 5 text prompts for each composition. We generate 4 images per text prompt and report the average evaluation metrics.

A.2 Problem with existing methods

Figure 3 showcases image quality comparisons between previous methods (Dreambooth [12] and Textual-Inversion [2]) and our MSP-Diffusion approach. Columns (A), (B), (C), and (D) represent image generation scenarios with 1, 2, 3, and more than 3 personalized subjects. For single subject personalization, Dreambooth and Textual-Inversion perform well. However, as the number of personalized subjects increases, issues such as missing subjects, incorrect appearances, and hybrid subjects with mixed characteristics become apparent, as seen in columns (B), (C), and (D) of rows 1 and 2 in Figure 3 for Dreambooth and Textual-Inversion.

Therefore, our primary goal is to accurately generate images with multiple personalized subjects based on a given text prompt. In Row 3 of Figure 3, we present the results of our proposed MSP-Diffusion method, where we have successfully addressed the limitations observed in previous approaches.

[3]Subject-Dataset11: `https://tinyurl.com/2j7jbzzx`

[4]Comicgen: `https://gramener.com/comicgen/`

[5]Alamy: `https://www.alamy.com/`

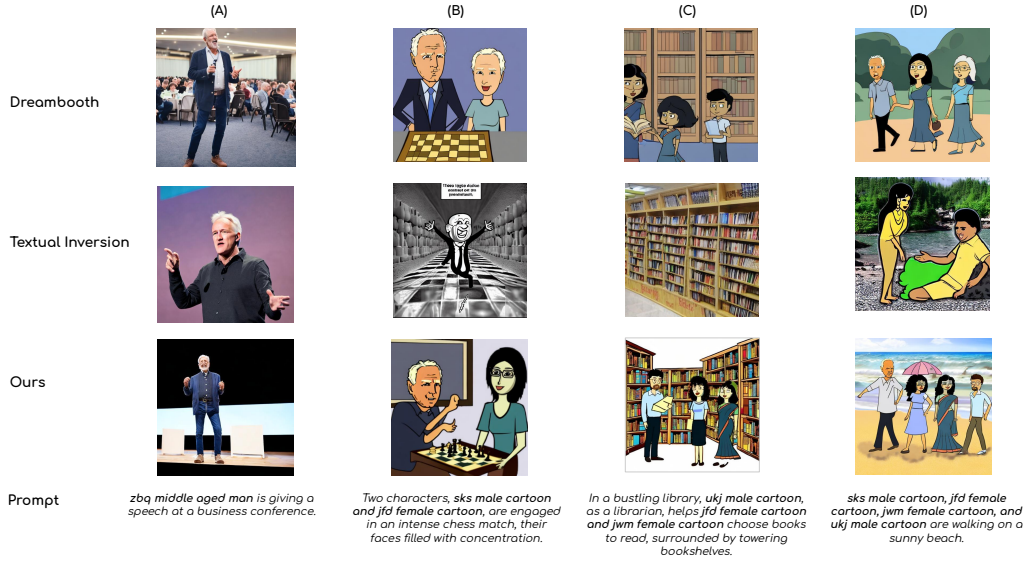

Figure 3: Qualitative comparison of MSP-Diffusion against Dreambooth [12] and Textual Inversion [2]. Columns A, B, C and D refer to the images having number of personalized subjects as 1, 2, 3 and more than 3, respectively. Here, it is clearly visible that models like Dreambooth and Textual-Inversion struggle to generate images with multiple personalized subjects and produce poor quality images (missing subjects, hybrid subject exhibiting the characteristics of multiple subjects and wrong subject's appearance).

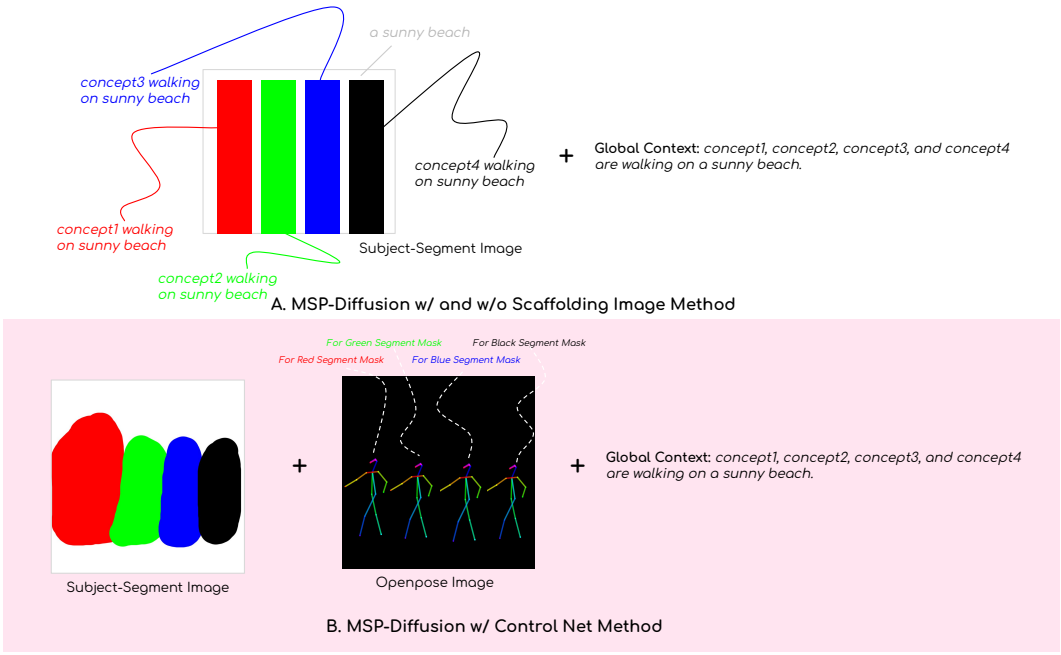

Figure 4: Figure illustrates a sample input to the MSP-Diffusion method during inference. (A) presents the input format for scenarios both with and without a scaffolding image, while (B) provides additional details regarding an Openpose image in the context of ControlNet conditioning with MSP-Diffusion. It's worth noting that *conceptI* can be substituted with the specific subject, such as "sks male cartoon," for instance.

A.3 Input Format of MSP-Diffusion

Figure 4 depicts the MSP-Diffusion inference input format. It consists of a colored subject-segment image with segments representing different personalized subjects, each accompanied by its respective local text prompt. Additionally, a global text prompt is provided, conveying the overall context for the generated image. This global prompt plays a vital role in harmonizing individual segments and capturing the intended subject interactions, aligning with the specifications in the text prompt. When we incorporate ControlNet's [14] openpose [1] into the MSP-Diffusion method, we also provide an additional Openpose image, as shown in Figure 4(B) describing the posture and size of the personalized subjects.

A.4 Subject-Aware Segmentation Loss (SAL)

Given an input image denoted as x_0, an image diffusion algorithm progressively introduces noise to the image, resulting in the creation of a noisy image referred to as z_t. Here, the variable t signifies the number of iterations at which noise is incrementally added. At each time step t, this image diffusion algorithm employs a neural network to acquire the ability to predict the noise components ϵ_u and ϵ_c that are applied to the noisy image z_t. This predictive process encompasses two distinct scenarios: one where no specific conditions are imposed, and the other involving the incorporation of particular conditions conveyed through textual prompts denoted as c. The forward process variances β_t can be learned by re-parameterization. Considering, reverse diffusion process for an input noisy image x_t, ϵ_u as unconditional noise prediction and ϵ_c as conditional noise prediction, and $\alpha_t := 1 - \beta_t$ [4], the final projected image $\hat{x_0}$ at time t is given by:

$$\hat{x_0} \leftarrow \frac{x_t}{\sqrt{\alpha_t}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\epsilon$$

$$where, \epsilon = \epsilon_u + 7.5 * (\epsilon_c - \epsilon_u)$$

Now, we define subject-aware segmentation loss L_{SAL} which is a self supervised binary-cross entropy loss indicating the distance of generated subject from the learned features of the target subject. The input to the subject segmentation model (M_{SAL}) is created by isolating the segment-specific region m^{seg_j} from the final projected image $\hat{x_0}$ represented by $(\hat{x_0} \odot m^{seg_j})$. Then, M_{SAL} provides the pixel-level score of the generated subject against its learned target features. For each segment j,

$$L_{SAL}(\hat{x_0} \odot m^{seg_j}) = 1 - BCE(M_{SAL}(\hat{x_0} \odot m^{seg_j}), M_{SAL}(\hat{x_0} \odot m^{seg_j}))$$

Now, we provide external guidance to the diffusion model by computing gradient from this loss, ∇^{seg_j}:

$$\nabla^{seg_j} \leftarrow \nabla L_{SAL}(\hat{x_0} \odot m^{seg_j})$$

A.5 Experimental Setup

The Dreambooth model is fine-tuned on a T4 12GB GPU, taking 4 hours with a batch size of 1 for a total of 22,000 training steps. Additionally, the ResNet-34 based UNet for subject segmentation was trained on T4 GPU with batch size of 4 and learning rate of $1e-4$ for 14 epochs. In the case of MSP-Diffusion, 80 diffusion steps were employed for two settings (with and without scaffolding image), while only 50 diffusion steps were used with ControlNet. MSP-Diffusion with a scaffolding image initially utilized the global diffuser for 8 steps, whereas MSP-Diffusion without scaffolding and with ControlNet required only 2 global diffuser steps. The inference for MSP-Diffusion (with and without scaffolding image) was performed on a 10GB MIG A100 GPU. Conversely, image generation via MSP-Diffusion with ControlNet was accomplished using Google Colab Pro, leveraging 12 GB T4 GPU.

A.6 Evaluation Metrics

We compare the performance of MSP-Diffusion against existing text-to-image generation methods such as Textual-Inversion [2] and Dreambooth [12] based on two essential criteria: subject similarity and context coverage, employing three distinct assessment methods as follows:

- **CLIP-I [2]**: It quantifies similarity by considering semantic distances in the CLIP-space. Specifically, CLIP-I calculates the cosine similarity between the generated images and their corresponding subject-specific source images.
- **Subject-aware similarity (SAS)**: We found that the CLIP scoring mechanism is not inherently aware of the target subjects in the embedding space. Consequently, even quite dissimilar subjects may yield a CLIP-I score of 0.8. To address this limitation, we propose an alternative evaluation method called Subject-aware similarity (SAS). It computes cosine similarity between subject-aware feature embeddings, leveraging the latent features extracted from the encoder of the Subject-aware UNet segmentation model.
- **CLIP-T [3]**: We measure context similarity using CLIP-Text Similarity, computed as the average similarity score for the best generations corresponding to each individual text-prompt across all compositions.

Figure 5: Figure showing the qualitative results of proposed MSP-Diffusion model in settings such as (A) w/ scaffolding image, (B) w/o scaffolding image and (C) w/o scaffolding image but w/ ControlNet conditioning. Here, n refers to the number of personalized subjects in the image-generation.

A.7 Qualitative Results

Here, we present a comprehensive qualitative analysis of the MSP-Diffusion method across three different settings: (A) with a scaffolding image, (B) without a scaffolding image, and (C) with ControlNet's Openpose control-conditioning. Figure 5 provides a visual representation of the performance in generating multiple personalized subjects, clearly demonstrating the effectiveness of our proposed method. In the case of MSP-Diffusion without a scaffolding image, although the background is occasionally missing, the number and appearance of personalized subjects surpass the method with a scaffolding image. Conversely, MSP-Diffusion with the incorporation of ControlNet conditioning yields the best results. However, it's important to note that this method is computationally

Figure 6: Ablation study on adding Global Diffuser (GD) and Subject-aware Loss (SAL) in MSP-Diffusion. (A) demonstrates that the generated image becomes more harmonized with the integration of GD, (B) illustrates that the subject's appearance in the generation looks more like the corresponding training images with the integration of SAL.

intensive and relies on an additional dependency, requiring an Openpose image as input during inference.

Additionally, we present a qualitative ablation analysis that underscores the effectiveness of integrating a Global Diffuser and Subject-aware Loss (SAL) into the MSP-Diffusion method. In Figure 6(A), it is evident that without the inclusion of a global diffuser, the interactions and overall context of the generated image lack harmony. However, upon incorporating the global diffuser, the personalized subjects in the generated images exhibit harmonized interactions and a consistent background.

Moreover, as depicted in Figure 6(B), it becomes evident that the appearance of personalized subjects was inaccurate without the inclusion of SAL, resulting in subjects that differed from their

Figure 7: Qualitative results of our proposed MSP-Diffusion model for generating images featuring multiple personalized subjects. Here n represents the number of personalized subjects in the generated image.

corresponding training images. To maintain control over the intended subject's appearance, we introduced SAL, ensuring that the subject's appearance was faithfully retained.

Now, we provide some more qualitative results of MSP-diffusion for generating images with multiple personalized subjects in Figure 7. This showcases that we are able to generate good quality harmonized images having multiple personalized subjects with the intended interactions between subjects.

A.8 Limitations

Our research paper, titled "MSP-Diffusion," presents compelling findings, as illustrated in Figure 7. Upon closer examination of the generated content, it is evident that the quality of generations involving distinct subject categories, such as comics featuring real human characters, falls short of our expectations. We are actively engaged in ongoing efforts to enhance this aspect of our work.

Furthermore, we have observed occasional instances where the background scene in the image generated using MSP-Diffusion without a scaffolding image appears to be missing. We are currently working on resolving this issue. Additionally, we are committed to reducing our reliance on Openpose for the generation process, exploring alternatives to further improve our methodology.

A.9 Key Contributions

Our paper contributes in the following ways:

1. A novel multi-subject aware personalized text-to-image generation approach named MSP-Diffusion, leveraging Stable Diffusion techniques to consistently produce high-quality images featuring multiple personalized subjects.
2. We utilize a limited number of training samples per personalized target subject, facilitating the fine-tuning of the base Dreambooth model.
3. We employ Composite Diffusion [6] alongside our subject-aware loss for personalized subjects within an image segment, allowing control over subject appearance.
4. We introduce a global diffuser capturing the global context of the image to ensure harmonization among various image segments containing personalized subjects.
5. A comprehensive comparison with existing state-of-the-art text-to-image generation models on a self-curated dataset of textual descriptions involving multiple personalized subjects.