Pick Your Influencer: Being Selective is Good for Personalization

Ashutosh Ranjan, Vivek Srivastava, Shirish Karande
TCS Research
Pune, India
{ashutosh.ranjan2, srivastava.vivek2, shirish.karande}@tcs.com

Abstract

Diffusion models have shown exceptional capability to personalize the subjects with very few reference images of the subjects. However, state-of-the-art personalization techniques based on diffusion models suffer from some major limitations and generate images with distortion, identity mixing, and repetition of the subjects. U-Net blocks in the diffusion models are known to capture the information for diverse attributes such as the color, style, layout, objects, etc. In this work, we present a novel technique for personalization based on selective U-Net influence (*SelUT*) where we control the influence of the trained U-Net blocks during inference with the text-conditioned diffusion model. Furthermore, we present an ensemble selection technique to select the best generated image with *SelUT* based on the Characteristic Objects Method (COMET) considering quantitative evaluation metrics as the criterion. We observe that our approach helps address the limitations and shows significant improvement against the state-of-the-art techniques in quantitative and qualitative evaluation.

1 Introduction

With the advent of large foundational models such as text-conditioned diffusion models [21, 23], building the real-world solutions for visual storytelling is gaining popularity [19, 36]. For effective storytelling, it becomes crucial to personalize multiple characters in novel and unseen contexts consistently throughout the story [4, 10]. This capability is particularly transformative in fields such as marketing [5, 16, 17], gaming [35, 29], and education [8, 34], where tailored visual content can significantly enhance user engagement and experience. By enabling the generation of images based on diverse subject inputs, these models can empower storytellers to create intricate scenes that reflect complex interactions among characters.

Pioneer works on text-to-image personalization such as DreamBooth [24] and textual inversion [6] facilitates re-contextualization of a given subject by training with just 3-5 images of that subject. They present novel techniques to learn the association between the visual concepts of the subject with the unique identifier in the text prompt and further leverage this learned association to personalize the subject in novel and unseen contexts. But these approaches struggle to personalize multiple subjects simultaneously in an image. Recently, several follow-up works have attempted the task of multi-subject personalization (MSP) [12, 13, 32, 7, 14, 18, 20] (see Appendix for the related work). However, we observe that these techniques struggle to generate high-quality output with proper subject disentanglement while showing high fidelity to the text prompt (see examples in Table 1 and Appendix). We note that they frequently generate images with distortion, identity mixing, and repetition of subjects.

Several previous works [30, 3, 28] have highlighted the significance of diffusion model's U-Net in capturing and controlling information about diverse attributes such as color, style, objects, etc. In

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

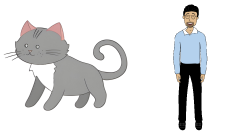
Limitation	Multi-Subject Reference	Inference Prompt	MSP Method	MSP Output	Ours
Distortion		[V1*] bald man and [V2*] girl dancing together	DreamBooth		
Identity Mixing		[V1*] cat watching television and [V2*] Ethan talking on phone	FastComposer		
Repetition		[V1*] old man and [V2*] cat dancing together	Mix-of-Show		

Table 1: Limitations of diffusion model based techniques for multi subject personalization. They struggle to disentangle the subjects provided in the reference images and fail to preserve the identity of all the subjects in the generated image. Distortion, inappropriate mixing of identity, and repetition of subjects is a common phenomenon observed in the generated images.

this work, we posit that the different U-Net blocks of the diffusion model capture varied information relating to MSP such as subject identity, multi-subject fidelity, aesthetic, etc. Controlling the influence of these blocks would provide us greater handle to address the limitations of the current techniques for MSP. To this extent, we present a novel selective U-Net influence (SelUT) technique with diffusion models. Specifically, we control the influence of the trained U-Net blocks during model inference (refer to Sections 2.1 and 2.2). It provides us a greater handle on interpreting the contribution of individual U-Net blocks across quality aspects such as identity disentanglement, image aesthetic, human preference, etc. Selectively influencing the U-Net blocks results in generating multiple images. We present a novel ensemble selection strategy to select the best generated image based on the Characteristic Objects Method (COMET [26]) which is a rank-reversal free multi-criteria decision-making strategy [15, 11, 25]. We model the ensemble selection of the image as the multi-criteria based ranking problem leveraging the scores from two quantitative evaluation metrics as the criterion i.e. prompt fidelity and image reward (refer to Section 2.3). We demonstrate the effectiveness of our approach in addressing the limitations of existing approaches with experiments performed on two different datasets i.e. Subject-Dataset11 and Synthetic Anime (refer to Sections 3 and 4).

2 Our Approach

2.1 Diffusion Model Training for MSP

In the first step, we perform the MSP training of the Stable Diffusion XL (SDXL) model (see Figure 1). We use the DreamBooth objective for our MSP training which is:

$$L(\theta) = \mathbb{E}_{x,c,\epsilon,\epsilon',t}\Big[w_t \left\|\hat{x}_\theta\left(\alpha_t x + \sigma_t \epsilon, c\right) - x\right\|_2^2 + \quad \lambda w_{t'} \left\|\hat{x}_\theta\left(\alpha_{t'} x_{pr} + \sigma_{t'} \epsilon', c_{pr}\right) - x_{pr}\right\|_2^2\Big] \tag{1}$$

The first term is the original loss function of the diffusion models which quantifies the distance between predicted image and the original image. Whereas the second term is prior preservation loss which quantifies the distance between predicted image and the original class image. The λ controls the weight of the prior preservation loss. Similar to the original DreamBooth training for single-subject personalization, we provide instance images of the multi-subject brought together in a single image with a white background. We provide 5 such instance images along with the text prompt *[V1*]<subject1> and [V2*]<subject2>, simple background*. Similar to [27, 13], we employ LoRA-based DreamBooth training and obtain the U-Net state dictionary. Formally,

$$Ut_{MSP} = \beta^{Ut}(\Gamma^{DB}(M_{SDXL}, L(\theta))) \tag{2}$$

where, Γ^{DB} denotes the LORA-based Dreambooth training of SDXL (M_{SDXL}) with the training objective given in eq 1. Whereas, β^{Ut} denotes obtaining the U-Net state dictionary (Ut_{MSP}) from the trained model (M_{MSP}).

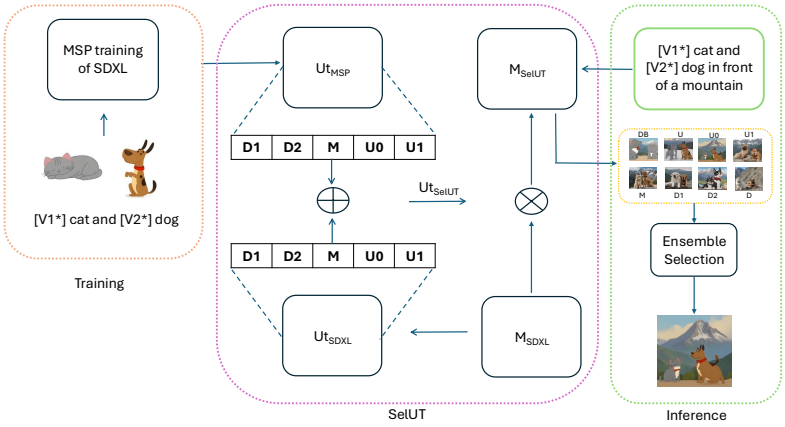


Figure 1: Architecture of the proposed selective U-Net influence (SelUT) with the ensemble selection technique. The blocks $D1$ and $D2$ (collectively denoted as D) are the down-blocks and $U0$ and $U1$ (collectively denoted as U) are the U-Net up-blocks. The block M is the middle block of the U-Net.

2.2 Selective U-Net Influence (SelUT)

It has been observed that diffusion model's U-Net plays a significant role in capturing the different aspects of image generation [3, 30] such as layout, color, and style. In this work, we posit that some of the major limitations with MSP techniques can be addressed by controlling the influence of the trained U-Net. To this extent, we selectively influence the U-Net blocks of a pre-trained SDXL model with the MSP trained U-Net blocks. The block set (B) in the U-Net comprises of $\{D1, D2, D, M, U0, U1, U\}$. Formally,

$$Ut^b_{SelUT} = Ut^b_{SDXL} \bigoplus Ut^b_{MSP}, \forall b \in B \tag{3}$$

$$M^b_{SelUT} = M^b_{SDXL} \bigotimes Ut^b_{SelUT}, \forall b \in B \tag{4}$$

where, Ut and M represent the U-Net and the diffusion model respectively. The superscript b in Ut and M denote the selected U-Net block from the set B. The operation $\bigoplus$ signifies updating the weights Ut^b_{SDXL} with the weights Ut^b_{MSP} while freezing the other blocks of Ut_{SDXL}. The operation $\bigotimes$ signifies replacing the U-Net of the pre-trained SDXL with the U-Net created with selective influence.

2.3 Inference and Ensemble Selection

With this proposed approach, we create multiple models and images corresponding to each selective U-Net block influence operation. Hence, it becomes crucial to determine the best image among all the generated images. We denote the generated set of images for the prompt P as:

$$I(P) = \bigcup_{b \in B} \Omega(M^b_{SelUT}, P) \tag{5}$$

where Ω denotes the inference with the model M^b_{SelUT} for the prompt P. We model the ensemble selection as the multi-criteria decision making problem. We leverage the quantitative evaluation metric scores as the criterion for the ensemble selection. These metrics evaluate the generated image on two different dimensions i.e. prompt fidelity (PF) and image reward (IR) (refer to section 3). Specifically, we use Characteristic Objects Method (COMET) where the preference of each alternative is obtained on the basis of the distance from the nearest characteristic objects. Formally,

$$ES(I(P)) = argmin_{b \in B}(\delta_{COMET}(I_m(PF, IR), P)) \tag{6}$$

where, I_m is a matrix of size B x 2 where we have two decision-making criteria for each generated image corresponding to the prompt P. Whereas δ_{COMET} calculates the rank of each generated image in the set $I(P)$. We select the image with the minimum rank.

	Subject-Dataset11										Synthetic Anime									
	Prompt Fidelity					Image Reward					Prompt Fidelity					Image Reward				
	SAP	DAP	S1P	S2P	ACP	SAP	DAP	S1P	S2P	ACP	SAP	DAP	S1P	S2P	ACP	SAP	DAP	S1P	S2P	ACP
DB	82.77	80.96	80.99	81.36	81.24	-0.31	-0.14	-0.31	-0.40	-0.29	84.93	81.21	86.11	89.42	82.40	0.60	0.32	0.089	0.11	0.67
MoS	72.48	71.32	73.51	70.19	66.36	-0.56	-0.40	-0.72	-0.89	-0.51	60.12	58.30	60.31	61.3	60.18	0.33	0.32	0.33	0.21	0.33
FC	73.82	70.47	67.94	68.15	71.65	-0.59	-0.48	-0.75	-0.78	-0.56	80.06	77.46	71.51	71.79	79.45	0.49	0.26	-0.01	0.017	0.55
U	82.45	79.12	79.80	79.27	82.16	0.51	0.08	1.08	0.66	0.31	84.44	86.52	88.53	89.59	82.16	0.51	0.31	1.49	1.08	0.31
U0	83.29	79.90	80.2	80.4	82.51	0.60	0.14	1.15	0.79	0.41	85.16	87.21	88.61	89.62	82.51	0.60	0.36	1.51	1.14	0.41
U1	86.50	83.58	83.18	80.92	87.58	0.59	0.21	1.16	0.90	0.73	87.81	86.91	88.46	88.49	87.58	0.59	0.13	1.44	0.90	0.73
M	86.06	83.23	82.57	80.93	87.57	0.677	0.29	1.257	0.97	0.77	87.41	88.36	88.85	87.48	87.57	0.69	0.37	1.55	1.08	0.77
D1	86.68	84.17	82.73	80.86	87.49	0.63	0.301	1.20	0.90	0.68	87.72	87.20	88.83	87.92	87.27	0.63	0.26	1.41	0.82	0.68
D2	86.64	83.95	83.52	81.57	87.25	0.671	0.300	1.255	0.88	0.67	87.24	85.86	88.30	87.88	87.15	0.62	0.07	1.45	1.04	0.67
D	85.25	83.72	82.58	81.47	85.93	0.54	0.24	1.16	0.87	0.57	86.25	83.90	87.32	87.19	85.76	0.54	0.04	1.37	0.92	0.57
ES w/o DB	**91.03**	88.26	86.65	84.93	92.30	1.12	1.02	**1.57**	1.34	**1.27**	95.31	91.85	92.97	93.41	**93.83**	1.16	0.93	**1.79**	1.33	1.11
ES with DB	90.73	**88.30**	**87.00**	**85.34**	**92.37**	**1.13**	**1.05**	1.56	**1.39**	1.26	**95.52**	**91.95**	**93.25**	**93.75**	93.64	**1.18**	**0.97**	1.78	**1.73**	**1.13**

Table 2: Quantitative performance evaluation of MSP methods. We highlight the best scores in each prompt category for the SelUT models (without ensemble selection (ES)). Furthermore, we show the **overall best** and the second best metric scores in the given prompt category. For the baselines, DB, MoS, and FC, we generate seven images for a prompt (equivalent to the size of U-Net blocks set B) and consider the score for the best generated image after ensemble selection. We consider all the seven images generated with DB along with the seven images from SelUT in *ES with DB* method. Whereas, we only consider the seven images generated with SelUT in *ES w/o DB* method.

3 Experiments

In this work, we use two datasets: (a) *Subject-Dataset11* [1, 12] which consists of 11 diverse characters in different pose and (b) *Synthetic Anime* which includes 5 anime comic characters synthetically generated with FLUX.1 [2]. The *Subject-Dataset11* dataset includes 6 comic characters and 5 real human characters. Out of these 11 characters, we prepare a multi-subject dataset of 25 character pairs by placing two characters on a white background. We perform the character pairing such that the real characters are paired with the real characters and vice-versa. Similarly, we create 10 character pairs in the *Synthetic Anime* dataset. We train the SDXL model and perform the DreamBooth LoRA fine-tuning with a common training prompt for a subject pair.

For evaluation of various aspects of multi-subject personalization, we consider 5 prompt categories: (i) **Same action prompts (SAP)**: both the subjects are performing the same action, (ii) **Different action prompts (DAP)**: both the subjects are performing the different actions, (iii) **Subject1 prompts (S1P)**: only subject1 is performing an action and there is no mention of subject2 in the prompt, (iv) **Subject2 prompts (S2P)**: only subject2 is performing an action and there is no mention of subject1 in the prompt, and (v) **Accessory and clothing prompts (ACP)**: subjects are prompted to wear different accessories and clothing types such as necklace, headphones, bandana, t-shirt, and suit. For each prompt category, we create 15 unique and diverse prompts for all the subject pairs. In the first four prompt categories, we consider the actions such as painting, dancing, watching television, etc. We consider the following three methods as the baseline to compare our MSP approach: DreamBooth (DB) [24], Mix-of-Show (MoS) [7], and FastComposer (FC) [32]. We keep the comparison with other technique such as PortraitBooth [20] as future work due to the lack of reproducibility and computational resource constraints at our end. However, the chosen methods show diversity in their approach and achieve the SOTA performance on the MSP task which helps us bring out the clear comparison of diverse state-of the-art methods with our approach.

4 Results and Analysis

In this section, we present the experimental results and our analysis. For evaluation, we use the two quantitative evaluation metrics: prompt fidelity measured with CLIP-T [22] and image reward [33]. Given that we do not have the ground truth reference images in our evaluation dataset, we rely on the reference-free evaluation metrics. We compute the alignment between the generated image and the text prompt with the CLIP-T score. The score ranges from 0 to 100 and the higher score implies higher prompt fidelity. Furthermore, we use *Image-Reward* which rates generated image based on the human preference across parameters such as text-image alignment, aesthetic, toxicity, and bias. A higher score signifies that the generated image is highly preferred by humans as compared to the other images generated using the same text prompt. We present the performance evaluation with these metrics in Tables 2, 3, and 4. We observe that SelUT with blocks U0 and U1 (in *Synthetic Anime*)

	Subject-Dataset11					Synthetic Anime				
	SAP	DAP	S1P	S2P	ACP	SAP	DAP	S1P	S2P	ACP
DB	7.90	8.74	14.57	**23.42**	3.20	11.33	4.16	12.85	17.77	4.00
U	2.73	2.91	4.57	4.28	4.00	14.00	13.88	12.85	15.55	7.00
U0	3.95	3.49	5.42	8.85	4.80	**16.66**	**20.13**	5.71	**37.03**	9.00
U1	17.62	16.32	13.42	11.41	16.0	**16.66**	13.19	**20.71**	5.92	19.0
M	**21.27**	16.61	16.57	11.71	18.80	12.00	**20.13**	15.71	7.40	**33.00**
D1	15.80	**21.57**	12.85	12.85	17.59	12.66	10.41	12.14	4.44	13.00
D2	16.71	16.61	**20.57**	13.42	**21.6**	11.33	9.72	15.0	8.14	8.0
D	13.98	16.61	12.0	14.28	14.00	5.33	8.33	5.0	3.70	8.0

Table 3: Percentage of prompts for which the generated image with the corresponding model is selected as the best image for *ES with DB* method (see Table 2). We highlight the **best** and the second best prompt percentage in each prompt category.

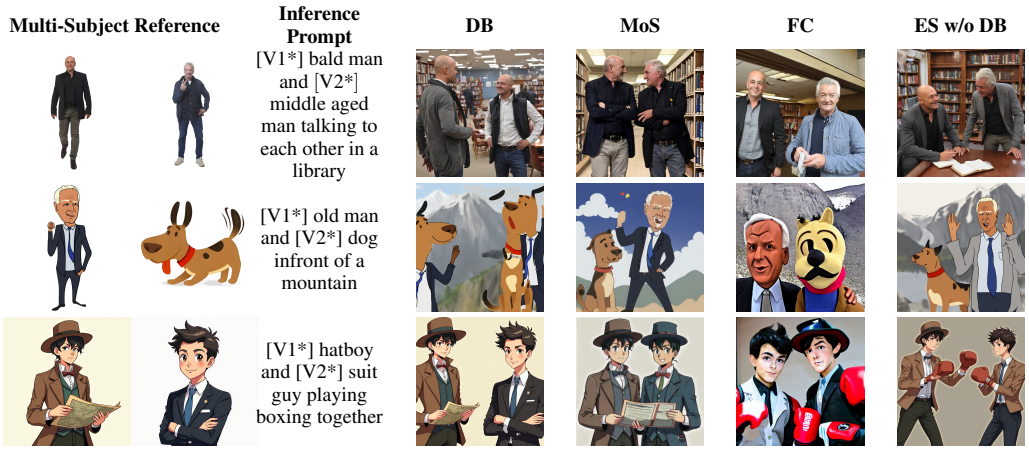


Table 4: Qualitative comparison of the personalized images generated with different methods. The existing approaches struggle to generate images that aligns with the inference prompt while also retaining the identity of the subjects that are provided with the multi-subject reference.

and D1 and D2 (*in Subject-Dataset11*) tends to generate images that align well with the prompt as compared to the full U-Net influence in DB. Furthermore, SelUT with blocks U and D achieves lower metric scores as compared to the individual sub-blocks within them (e.g. U0, D1, etc.). This further highlights the importance of selective influence of U-Net where influencing higher degree of U-Net weights tends to degrade the overall quality of the generation. We further observe that SelUT with block M tends to control the image aesthetic and generate images that are more likely to be preferred by the humans. From Table 3, we observe that even though we introduce 50% images generated by DB (7 out of 14) to the set $I(P)$ for ensemble selection, the majority of the best selected images across prompt categories are from SelUT with different U-Net blocks. We qualitatively compare different methods in Table 4 and observe that our method helps addressing the key limitations of the existing approaches (see more examples in Appendix). The ensemble selection of the images help to select the best personalized image generated for a prompt and the multi-subject reference. The proposed ensemble selection technique will prove useful across tasks where we generate multiple images for a prompt and seek to select the best image based on multiple quality criterion.

5 Conclusion

In this work, we present a novel technique for personalization with diffusion models based on selective U-net influence and present an exploration into the contribution of U-Net blocks for the MSP task. We observe that the U-Net blocks play a crucial role in image generation by influencing the different quality aspects of the generation. We further present a multi-criteria ensemble selection technique which will prove useful in several downstream tasks. We do acknowledge the higher inference cost due to the multiple models created with SelUT. However, we believe that future research would help shed more light into the interpretability, effectiveness, and the cost-efficiency of the solution.

References

[1] Make your own comic. https://gramener.com/comicgen/. Accessed: 2024-08-30.

[2] Announcing black forest labs. https://blackforestlabs.ai/announcing-black-forest-labs/. Accessed: 2024-08-30.

[3] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *arXiv preprint arXiv:2311.11919*, 2023.

[4] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[5] Doaa Farouk El-Desouky. Visual storytelling in advertising: A study of visual storytelling as a marketing approach for creating effective ads. *International Journal of Humanities Social Sciences and Education (IJHSSE)*, 7(10):118–127, 2020.

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[7] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

[8] Ariel Han and Zhenyao Cai. Design implications of generative ai systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, pages 470–474, 2023.

[9] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

[10] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899*, 2024.

[11] Alessio Ishizaka and Markus Lusti. How to derive priorities in ahp: a comparative study. *Central European Journal of Operations Research*, 14:387–400, 2006.

[12] Arushi Jain, Shubham Paliwal, Monika Sharma, Vikram Jamwal, and Lovekesh Vig. Multi-subject personalization. *arXiv preprint arXiv:2405.12742*, 2024.

[13] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

[14] 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.

[15] Pekka Leskinen and Jyrki Kangas. Rank reversals in multi-criteria decision analysis with statistical modelling of ratio-scale pairwise comparisons. *Journal of the Operational Research Society*, 56(7):855–861, 2005.

[16] Preeti Mehra and Pooja Kansra. Fostering gratifying customer experiences through the art of visual content and storytelling. In *Multidisciplinary Applications of Extended Reality for Human Experience*, pages 401–423. IGI Global, 2024.

[17] JR Ashlin Nimo, K Ravishankar, and Navaneetha Krishnan Rajagopal. Ai for character creation and storytelling in marketing. In *Balancing Automation and Human Interaction in Modern Marketing*, pages 39–58. IGI Global, 2024.

[18] Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024.

[19] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024.

[20] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024.

[21] 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. In *The Twelfth International Conference on Learning Representations*, 2023.

[22] 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.

[23] 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.

[24] 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.

[25] Wojciech Sałabun. The characteristic objects method: a new approach to identify a multi-criteria group decision-making model. *Intl J Comput Tech Appl*, 5(5):1597–1602, 2014.

[26] Wojciech Sałabun. The characteristic objects method: A new distance-based approach to multicriteria decision-making problems. *Journal of Multi-Criteria Decision Analysis*, 22(1-2): 37–50, 2015.

[27] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[28] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024.

[29] Yuqian Sun, Zhouyi Li, Ke Fang, Chang Hee Lee, and Ali Asadipour. Language as reality: a co-creative storytelling game experience in 1001 nights using generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 425–434, 2023.

[30] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

[31] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[32] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[33] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. Storydrawer: a child–ai collaborative drawing system to support children's creative visual storytelling. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–15, 2022.

[35] Yijun Zhao, Yiming Cheng, Shiying Ding, Yan Fang, Wei Cao, Ke Liu, and Jiacheng Cao. Magic camera: An ai drawing game supporting instantaneous story creation for children. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, pages 738–743, 2024.

[36] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024.

A Appendix

Related Work

The recent works on personalization with text-to-image diffusion models have shown impressive abilities to recontextualize a single subject in novel contexts with only a very few images of that subject [24, 14, 9, 31]. Textual Inversion [6] and DreamBooth [24] are the pioneer works on text-to-image personalization with diffusion models. Recently, we also observe works on multi-subject personalization with diffusion models [12, 13, 32, 7, 14, 20]. For instance, Custom Diffusion [14] proposes a fine-tuning technique where they identify a small subset of model weights and fine-tune them to update the model with the new concept. To prevent model forgetting, they use a small set of real images with similar captions as the target images. Mix-of-Show [7] proposes a framework comprising of an embedding decomposed LoRA (ED-LoRA) and gradient fusion to address the concept conflict and identity loss in decentralized multi-concept customization. Contrary to these approaches, FastComposer [32] enables inference-only personalization of multiple-subjects across diverse contexts. To address the identity mixing problem, they propose cross-attention localization which is supervising cross-attention maps of subjects with segmentation masks during training. Furthermore, to avoid identity overfitting they introduce delayed subject conditioning, preserving the subject's identity while following text instructions. Mixture of Attention [18] proposes a novel method which is prior preserving, fast, and layout-free. It distributes the generation between personalized and non-personalized attention pathways. It is designed to retain the original model's prior by fixing its attention layers in the prior (non-personalized) branch, while minimally intervening in the generation process with the personalized branch. In addition, a routing mechanism manages the distribution of pixels in each layer across these branches to optimize the blend of personalized and generic content creation. However, these techniques struggle to generate high-quality personalized output (see Table 5) demanding a thorough investigation of the diffusion models and the MSP training strategies to address these limitations.

Experimental Settings and Additional Results

As mentioned in Section 2, we use Characteristic Object Method (COMET) to select the best generated image for a given prompt. COMET is a five step method. We provide a broad overview of the approach below. For an elaborate discussion, we redirect the reader to the method originally discussed in [26, 25]. The five steps are:

Step 1: Define the space of the problem using multiple criterion. In our case, we use prompt fidelity and image reward as the criterion.

Step 2: Generate the characteristic objects which is a Cartesian Product of triangular fuzzy numbers cores for each criteria.

Step 3: Rank and evaluate the characteristic objects by using the Matrix of Expert Judgment (MEJ). We use the MEJ function as originally defined in [25].

Step 4: Create the fuzzy rule base where each characteristic object and value of preference is converted to a fuzzy rule.

Step 5: Inference in a fuzzy model and final ranking for the alternatives.

Also, in Table 6, we show the prompts from all the five prompt categories. We show the qualitative results for ACP prompts in Table 7 (We couldn't upload the pdf with all the prompt categories given the size limitations with OpenReview). Our observations are consistent with the analysis presented in Section 4. The existing methods fail to generate the high-quality images. Whereas, the selective influence of U-Net blocks significantly improve the quality of the image personalization and generation.

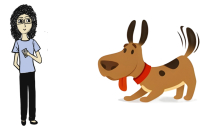
Limitation	**Multi-Subject Instance**	**Prompt**	**DreamBooth**	**Ours**
Distortion		[V1*] cat painting and [V2*] dog on watching television		
Distortion and Repetition		[V1*] old man and [V2*] dog sitting on camel back		
Distortion and Repetition		[V1*] bald man and [V2*] middle aged woman jumping in a playground		
Identity Mixing and Repetition		[V1*] old man and [V2*] deenuova dancing together		
Identity Mixing and Distortion		[V1*] bald man and [V2*] tortoise on a bridge		
Identity Mixing, Distortion, and Repetition		[V1*] deenuova and [V2*] dog on a bridge		

Table 5: Some more examples related to the limitations of DreamBooth. The limitations shown earlier in Table 1 do not necessarily occur in isolation. We also observe that these limitations can occur simultaneously in a single generated image.

SAP	DAP	S1P	S2P	ACP
[V1*] S1 and [V2*] S2 watching television together	[V1*] S1 cycling and [V2*] S2 sitting on bench	[V1*] S1 cycling	[V2*] S2 cycling	[V1*] S1 and [V2*] S2 wearing black suit.
[V1*] S1 and [V2*] S2 standing in front of a church	[V1*] S1 cooking food and [V2*] S2 reading a book	[V1*] S1 cooking food	[V2*] S2 cooking food	[V1*] S1 and [V2*] S2 wearing red tshirt
[V1*] S1 and [V2*] S2 painting together	[V1*] S1 working on laptop and [V2*] S2 painting	[V1*] S1 working on laptop	[V2*] S2 working on laptop	[V1*] S1 and [V2*] S2 wearing white shirt
[V1*] S1 and [V2*] S2 dancing together	[V1*] S1 inside a car and [V2*] S2 running outside	[V1*] S1 inside a car	[V2*] S2 inside a car	[V1*] S1 and [V2*] S2 wearing a track suit
[V1*] S1 and [V2*] S2 jumping together	[V1*] S1 lying on a beach and [V2*] S2 playing volleyball on beach	[V1*] S1 lying on a beach	[V2*] S2 lying on a beach	[V1*] S1 and [V2*] S2 wearing blue jumpsuit
[V1*] S1 and [V2*] S2 hugging each other	[V1*] S1 painting and [V2*] S2 watching television	[V1*] S1 painting	[V2*] S2 painting	[V1*] S1 and [V2*] S2 wearing boxing gloves
[V1*] S1 and [V2*] S2 playing boxing together	[V1*] S1 reading a book and [V2*] S2 dancing	[V1*] S1 reading a book	[V2*] S2 reading a book	[V1*] S1 and [V2*] S2 wearing a hat
[V1*] S1 and [V2*] S2 sitting on a camel back	[V1*] S1 talking on phone and [V2*] S2 watching television	[V1*] S1 talking on phone	[V2*] S2 talking on phone	[V1*] S1 and [V2*] S2 wearing headphones
[V1*] S1 and [V2*] S2 working on the laptop	[V1*] S1 playing with ball and [V2*] S2 running	[V1*] S1 playing with ball	[V2*] S2 playing with ball	[V1*] S1 and [V2*] S2 wearing necklace
[V1*] S1 and [V2*] S2 in front of a mountain	[V1*] S1 painting and [V2*] S2 working on a laptop	[V1*] S1 painting	[V2*] S2 painting	[V1*] S1 and [V2*] S2 wearing bandana
[V1*] S1 and [V2*] S2 on a bridge	[V1*] S1 sitting on a bench and [V2*] S2 cycling	[V1*] S1 sitting on a bench	[V2*] S2 sitting on a bench	[V1*] S1 and [V2*] S2 wearing spectacles
[V1*] S1 and [V2*] S2 sitting on a park bench	[V1*] S1 running outside and [V2*] S2 inside a car	[V1*] S1 running outside	[V2*] S2 running outside	[V1*] S1 and [V2*] S2 wearing hazmat suit
[V1*] S1 and [V2*] S2 cycling together	[V1*] S1 playing volleyball on beach and [V2*] S2 lying on a beach	[V1*] S1 playing volleyball on beach	[V2*] S2 playing volleyball on beach	[V1*] S1 and [V2*] S2 black helmet
[V1*] S1 and [V2*] S2 jumping together in playground	[V1*] S1 dancing and [V2*] S2 reading a book	[V1*] S1 dancing	[V2*] S2 dancing	[V1*] S1 and [V2*] S2 wearing a sombrero
[V1*] S1 and [V2*] S2 talking to each other in a library	[V1*] S1 watching television and [V2*] S2 talking on phone	[V1*] S1 watching television	[V2*] S2 watching television	[V1*] S1 and [V2*] S2 wearing raincoat

Table 6: Inference prompts for the five prompt categories. [V1*] and [V2*] are the unique identifiers for the subjects S1 and S2 respectively.

Table 7: The generated images for the prompt from the ACP prompt category: [V1*] and [V2*] wearing red t-shirt. We show the multi-subject reference ([V1*] and [V2*]) at the top. You can observe that the methods like Dreambooth, FastComposer and Mix-of-Show are unable to capture subject identity especially FastComposer and Mix-of-Show. Also, influence of the U-Net blocks tends to be dynamic as you can observe in some cases U block tends to perform better in others while in others U0.