Gen4Gen: Generative Data Pipeline for Generative Multi-Concept Composition

Chun-Hsiao Yeh*[1]
daniel_yeh@berkeley.edu
Ta-Ying Cheng*[2]
ta-ying.cheng@cs.ox.ac.uk
He-Yen Hsieh*[3]
heyenhsieh@g.harvard.edu
Chuan-En Lin[4]
chuanenl@andrew.cmu.edu
Yi Ma[1,5]
yima@eecs.berkeley.edu
Andrew Markham[2]
andrew.markham@cs.ox.ac.uk
Niki Trigoni[2]
niki.trigoni@cs.ox.ac.uk
H.T. Kung[3]
kung@harvard.edu
Yubei Chen[6]
ybchen@ucdavis.edu

[1] UC Berkeley
California, USA
[2] University of Oxford
Oxford, UK
[3] Harvard University
Massachusetts, USA
[4] Carnegie Mellon University
Pennsylvania, USA
[5] University of Hong Kong
Hong Kong SAR, China
[6] UC Davis
California, USA

Abstract

In this paper, we identify two major gaps in personalizing text-to-image diffusion models, i.e., placing personalized concepts into generated image: 1) Creating a high-quality multi-concept personalized dataset with detailed and aligned text descriptions is challenging. 2) There lacks comprehensive metrics to evaluate multiple personalized concepts in an image. To overcome these challenges, we propose ***Gen4Gen***, a novel generative data pipeline for creating a benchmark dataset (***MyCanvas***) that combines personalized concepts into complex compositions aligning with detailed text descriptions, aiming to benchmark and improve multi-concept personalization. In addition, we introduce comprehensive metrics (CP-CLIP / TI-CLIP) for evaluating the performance of multi-concept personalization models more effectively. Finally, we provide a simple yet effective baseline built on top of several personalization methods with empirical prompting strategies for future researchers to evaluate on *MyCanvas* benchmark. By improving data quality, we can significantly increase the multi-concept image generation quality without changing the model architecture or training algorithms, and we show our work can be simply plug in to personalization approaches. We suggest that leveraging strong foundation models for dataset generation could benefit various computer vision tasks. Code and benchmark dataset are available at https://danielchyeh.github.io/Gen4Gen/.

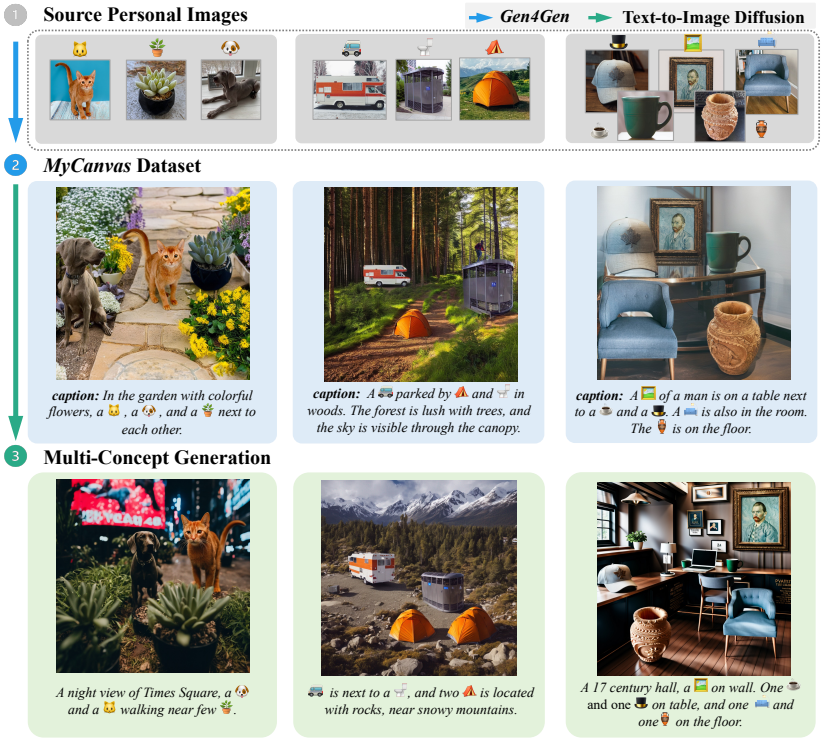

Figure 1: Starting from a few source images representing different concepts (each object illustrated by an icon), we propose *Gen4Gen*, a generative pipeline to compose complex multi-concept scenes paired with detailed text descriptions. Training with the resulting benchmark dataset (*MyCanvas*) significantly boosts multi-concept personalization performance without modifying model architectures or training strategies.

1 Introduction

Recent advances in text-to-image diffusion models [9, 12, 14, 27, 30, 34, 35, 37] have enabled users to personalize generation with minimal sets of concept images (e.g., their pets or recently bought houseplant) to generate new scenes incorporating these personal concepts (e.g., their pets in a night view of Times Square as shown in Figure 1). These efforts improve control over generation [1, 16, 19, 25, 36], but challenges remain, particularly in accurately handling *multiple* concepts in a single image.

As noted by [19], the pretrained stable diffusion [35] struggles to disentangle and represent multiple similar concepts (e.g., dog and cat) within one image. This limitation often carries over to fine-tuned personalization models. We believe this issue stems from mismatches between the text-image pairs in the pre-training datasets (e.g., LAION [38]) that emphasize single-object scenes with simplified captions. This misalignment complicates multi-concept personalization.

Rather than pursuing purely model-driven solutions, we develop a proof-of-concept benchmark dataset focused on multi-concept personalization. We propose *Gen4Gen*, a novel generative pipeline that leverages foundation models in foreground extraction [31], object

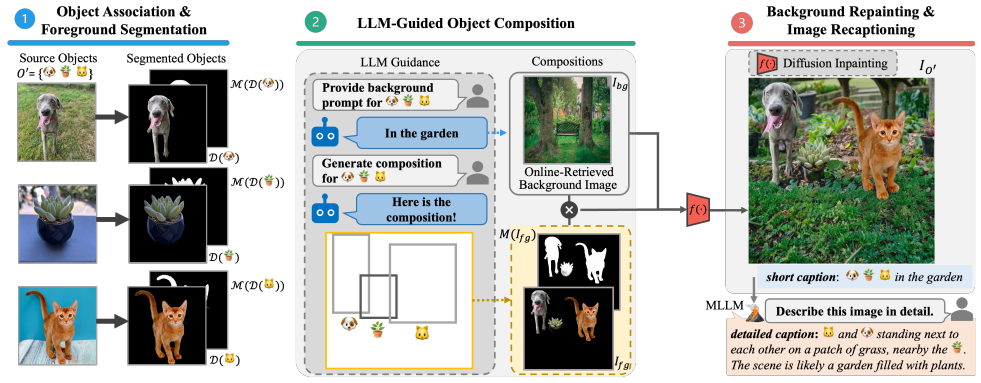

Figure 2: **Overview of the *Gen4Gen* Pipeline for Creating the *MyCanvas* Data.** (1) We use a category-agnostic saliency detector to segment foreground objects in a composition O'. (2) GPT is prompted to propose plausible bounding box layouts and background scene descriptions for O'. (3) A diffusion inpainting model embeds the foreground I_{fg} into a background image I_{bg}, generating the final image $I_{O'}$. To enhance textual diversity while maintaining image-text alignment, a subset of $I_{O'}$ is re-captioned using an MLLM (LLaVA).

composition [22], diffusion inpainting [30], and MLLMs [23] to synthesize complex multi-object images and rich text captions. Using this pipeline, we generate and filter over 10k images to create the *MyCanvas* benchmark.

Existing evaluation methods [16, 19, 36] mainly rely on user studies or small-scale testing. To provide a more comprehensive benchmark, we adopt principles from [3, 8, 10, 17, 18, 29, 37] and propose two new evaluation metrics: the Composition-Personalization-CLIP (CP-CLIP) and Text-Image alignment CLIP (TI-CLIP) scores, which jointly assess personalization fidelity and scene composition quality. In summary, our contributions are:

- **Integrating AI foundation models:** *Gen4Gen* demonstrates the power of cascaded foundation models in generating high-quality multi-concept image and text pairs.
- **Data quality matters:** Our *MyCanvas* benchmark dataset highlights that well-aligned image-text pairs substantially improve multi-concept personalization.
- **A new benchmark is needed:** We introduce a benchmark and two evaluation metrics (CP-CLIP and TI-CLIP) that jointly assess personalization, composition, and text-image alignment.

2 Related Works

Personalized Text-to-Image Generation. Personalization aims to adapt a pre-trained diffusion model to generate new scenes containing a user-specified concept based on a few reference images. Early methods such as Textual Inversion [13] and DreamBooth [36] pioneered this task by learning new token embeddings or fine-tuning the entire model, respectively. Later works enhanced fidelity and multi-concept personalization through regularized fine-tuning [1, 2, 6, 7, 9, 15, 16, 19, 20, 24, 40, 41, 42, 43]. For instance, Custom Diffusion [19]

fine-tunes cross-attention layers, SVDiff [16] modifies singular values, and MuDI [18] uses segmentation strategies. Different from prior model-centric approaches, we demonstrate that enhancing the data alone can substantially boost multi-concept personalization.

Text-to-Image Datasets and Benchmarks. Large-scale datasets [4, 30, 35, 37] have fueled diffusion models' success but often suffer from weak text-image alignment, especially for complex scenes [32, 33, 38, 39]. Our work shows that even a smaller, carefully curated dataset with detailed captions for multi-object compositions can significantly improve personalization. Furthermore, existing benchmarks like DrawBench [37], T2I-CompBench [17], and HRS [4] evaluate general generation ability, whereas we introduce the comprehensive benchmark focused on multi-concept personalization.

3 *Gen4Gen*: A Data-Driven Approach

Multi-concept personalization aims to synthesize images combining multiple user-provided concepts (e.g., dog, cat, houseplant) across diverse scenes. Prior work [19] highlights that this task becomes increasingly challenging as the number of concepts grows. While prior work [16, 19, 24] focuses on training strategies, we show that improving training data quality alone significantly boosts multi-concept generation.

3.1 Benchmark Dataset Creation Objectives

Existing datasets like LAION [38] often suffer from poor text-image alignment and low-quality multi-object scenes. To address this, *Gen4Gen* is designed with three principles: ***i)*** **Detailed text-image alignment** covering both foreground and background, ***ii)*** **High resolution** to enable high-quality personalization, and ***iii)*** **Realistic object layout and background generation**. We retain plausible but uncommon object combinations (e.g., lion and cat) to create a more challenging benchmark, while filtering out logically impossible scenes.

3.2 *Gen4Gen* Pipeline

Figure 2 illustrates our *Gen4Gen* pipeline, consisting of three key stages: (1) Object association and segmentation, (2) Object composition, and (3) Background repainting and image recaptioning. While automation is ideal, we incorporate human oversight to ensure a robust benchmark, as current models [22, 30, 31] can still introduce artifacts.

1) Object Association and Foreground Segmentation. We begin with a set of k objects $O = \{o_i\}_{i=1}^{k}$, where each object o_i is represented by a set of n images $X_{o_i} = \{x_j\}_{j=1}^{n}$ from DreamBooth, Custom Diffusion, and copyright-free databases. We identify object groups $O' = \{o_a, o_b, ...\}, O' \in O$ that can naturally co-exist (e.g., dog, cat, houseplant; see Figure 2).

For each object in O', we sample one image from X' and apply DIS [31] to extract the foregrounds $\mathcal{D}(X')$ and masks $\mathcal{M}(\mathcal{D}(X'))$. Since DIS is category-agnostic, it generalizes well across diverse objects. Importantly, objects with high latent similarity often fail in Custom Diffusion and stable diffusion, making our dataset a particularly challenging benchmark for multi-concept personalization.

2) Object Composition. We use the zero-shot capabilities of GPT [28] to generate background prompts P and bounding box suggestions for objects O' [22]. For each composition, GPT suggests plausible scenes (e.g., *"in a garden"*) and bounding box layouts, guiding the

Figure 3: **Examples from *MyCanvas*.** Our benchmark dataset contains multiple personalized objects in complex compositions with high-quality images and text.

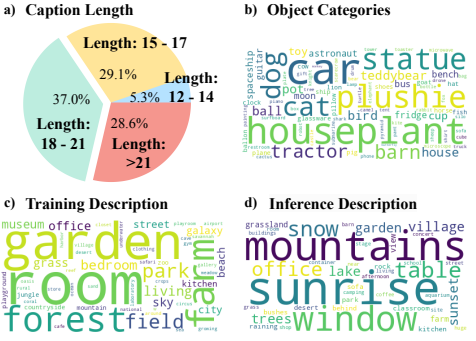

Figure 4: ***MyCanvas* Statistics.** a) Pie chart showing 30% of captions >20 words. b) Word cloud of object categories. c,d) Description word clouds from training/inference.

placement of objects within $\mathcal{D}(X')$ to create the composite foreground image I_{fg} and its mask $\mathcal{M}(I_{fg})$.

However, the above-described method would occasionally leads to object scaling problem (e.g., a sheep appearing larger than a house). To address this, we prompt ChatGPT with the following: `[Given a list of object names, your task is to generate a reasonable scale ratio for these objects in real-world terms, where the ratio for the largest object is set to 1.0,...]`. These scale ratios can reflect real-world proportions, ensuring a more accurate and logical arrangement in the generated layouts. (The detailed analysis of refining and not refining scale in Appendix).

3) Background Repainting and Image Recaptioning. While direct inpainting with models like Stable-Diffusion XL [30] is possible, we found that starting from a high-resolution background image and repainting yields better quality (see Appendix). Specifically, given an inpainting model f, we select a background I_{bg} from copyright-free sources based on prompt $p \in P$, and generate the final image: $I_{O'} = f(I_{fg}, \mathcal{M}(I_{fg}), I_{bg})$. We apply a 5×5 smoothing to $\mathcal{M}(I_{fg})$ to better integrate foreground and background.

To create a comprehensive benchmark dataset, we enhance the diversity of text descriptions while ensuring they closely match the images, even for longer prompts. For richer and accurate text descriptions, we use an MLLM [23] to automatically caption some images with the prompt: *"Describe what you see in this image in detail. Limit the description to 30 words"*. This ensures compatibility with CLIP's 77-token limit [32]. Recaptioning is performed on ten compositions O' within *MyCanvas*. These steps are repeated to construct the final *MyCanvas* benchmark dataset. Examples are shown in Figure 3.

3.3 Statistics of *MyCanvas*

For the *MyCanvas* benchmark dataset, we collected 150 objects (with one or multiple images each) and created 41 compositions (O'), generating over 10K images, later filtered to around 3K high-quality samples. Our benchmark dataset scale is more than sufficient for concept personalization, which typically requires only a few images for fine-tuning.

Figure 4 summarizes *MyCanvas*: **(a)** shows that captions average 17.7 words, with about 30% exceeding 20 words. **(b)** illustrates the diversity of objects, surpassing CustomConcept101 and DreamBooth. **(c, d)** highlight the variety in training and inference prompts, ensuring broader coverage and more complex compositions than prior benchmarks.

3.4 Enhancing Training-Time Text Prompts

On top of designing a well-aligned prompt with the images within the dataset, we also take a step further in exploring what the best prompt design is during training. We share some of the empirical findings and its intuitions below:

Global Composition Token. Previous arts like DreamBooth have shown that they can learn to map a new token to very difficult, challenging concepts (e.g., an abstract style like Monet art). We adapt this concept to complex compositions. By introducing a global token alongside individual tokens for each object, our model gains enhanced capabilities in describing detailed scene arrangements, leading to more realistic and coherent image generation.

Repeating Concept Tokens. We notice in a lot of cases where a complex composition involving multiple concepts could often lead to one or two concepts missing [5, 44]. This could be due to the model sometimes forgetting the details given a very long prompt. Thus, we employ a strategy of repeating concept token prompts during training. This encourages the model to ensure the presence of each specified concept in the generated images, enhancing overall object persistence and completeness.

Explicit Background Prompts. We observe an issue where backgrounds are inadvertently learned with the object identity in the token feature space. As an effort to disentangle background and concept compositions, we make sure that background has to be stated within the training prompt to encourage concept tokens learning only the object identity.

3.5 Personalized Composition Metric

As personalization complexity increases with more objects, models often struggle to capture details or overfit, a challenge not captured by prior benchmarks due to the lack of complex datasets like *MyCanvas* and overfitting risks.

To address this, we propose two metrics inspired by [3, 17]: Composition-Personalization-CLIP (CP-CLIP) for composition and fidelity evaluation, and Text-Image alignment CLIP (TI-CLIP) for assessing background generalization and overfitting.

To automate the full evaluation framework, we begin with a state-of-the-art model for open-vocabulary detection, OWL-ViT [26]. Given a generated image I_{gen} aiming to contain all objects in the set O', we obtain a set of cropped images $B_{pred} = \{b_{pred_1}, b_{pred_2}, \dots\}$ predicted by OWL-ViT, where $B_{pred} = \text{OWL}(I_{gen}, l_{O'})$, and $l_{O'}$ are the labels of objects within O' used as target vocabularies for detection.

For every cropped image $b_{pred_i} \in B_{pred}$ we obtained, we compute an average clip score $S_{i,j}$ against the image set $X_{o_j}, o_j \in O'$ as the following:

$$S_{i,j} = \frac{\sum_{x \in X_{o_j}} C(b_{pred_i}, x)}{|X_{o_j}|}, \tag{1}$$

where $C(\cdot)$ computes the dot product between two normalized image features. The final personalization CLIP score for b_{pred_i} is defined as $S_i = \max(\{S_{i,j}\}_{o_j \in O'})$, where we take the maximum similarity across all target objects o_j in O'.

Table 1: **Comparison of Personalization Performance.** We report CP-CLIP and TI-CLIP scores under three training settings. CD denotes Custom Diffusion, and Ours refers to CD with our prompting strategy. Best CP-CLIP scores are highlighted in **bold**. TI-CLIP indicates text-image alignment and is expected to remain *roughly stable* across methods.

	<= 3 Objects		4 Objects		5 Objects	
	CP-CLIP↑	**TI-CLIP**	**CP-CLIP↑**	**TI-CLIP**	**CP-CLIP↑**	**TI-CLIP**
CD + Source Images	0.26	0.16	0.21	0.13	0.23	0.17
CD + *MyCanvas*	0.41	0.17	0.47	0.17	0.50	0.15
Ours + *MyCanvas*	**0.51**	0.17	**0.55**	0.16	**0.57**	0.14
vs. baseline	+0.25	-	+0.34	-	+0.34	-

If there is more than one bounding box corresponding to the same o_j, we remove all except the one with the highest score from B_{pred} so the size $|B_{pred}|$ properly reflects how many personalized objects prompted by the text is reflected in the generated image. Finally, we obtain an overall CP-CLIP score per image:

$$\text{CP-CLIP}_{pred} = \frac{\sum_{b_{pred_i} \in B_{pred}} S_i}{|O'|}. \tag{2}$$

Note that the denominator is the number of objects within O' and not the number of bounding boxes; this acts as a penalty when a particular personalized object is not reflected in the image I_{gen}. We do not penalize when there are more bounding boxes than intended, as the generative model should be able to freely generate more objects than requested as long as it follows the text guidance.

Text-Image Alignment. To measure the amount of overfitting quantitatively, we calculate the TI-CLIP as a CLIP score between I_{gen} and the prompt p_{gen} that was used to generate I_{gen}. Note that while the formulation of TI-CLIP is very similar to CP-CLIP (i.e., one may think of TI-CLIP as a special case of the personalization clip score with the bounding box of the entire image and target of personalization being the text), it is evaluating an orthogonal concept of model's generalization quality and should thus be measured as a separate metric. On a high level view, TI-CLIP measures the background prompt (without the objects) with the whole generated image; there is no reason to believe that the background is improved during personalization, so a maintenance in TI-CLIP should be what we are aiming for when increasing the CP-CLIP score. This shows that the model is not overfitting to training set backgrounds.

Score Interpretability. Although CP-CLIP and TI-CLIP are theoretically bounded between 0 and 1, perfect scores are impractical. CP-CLIP averages similarity across multiple objects, and even identical objects from different angles yield scores around 0.6–0.7. TI-CLIP compares the background prompt to the full image; larger foreground objects may lower the score. Thus, a good model should **increase CP-CLIP while maintaining TI-CLIP**.

4 Experiments

4.1 Implementation Details and Quantitative Analysis

We evaluate *Gen4Gen* under three settings: 1) using individual source images, 2) using composed *MyCanvas*, and 3) applying our prompting strategy with *MyCanvas*. [19] was

Figure 5: **Qualitative Results for Multi-Concept Composition.** We show examples of increasing composition difficulty. Using Custom Diffusion [19], our *MyCanvas* improves disentanglement of visually similar concepts (e.g., statues, tractors) and better preserves object identities. With our prompting strategy, caption alignment improves further. **All results are based on SDXL [30].** Due to the space limit, additional improved results on DreamBooth [36], Break-A-Scene [2], GLIGEN [21], and I2VGen-XL [45] are in Appendix.

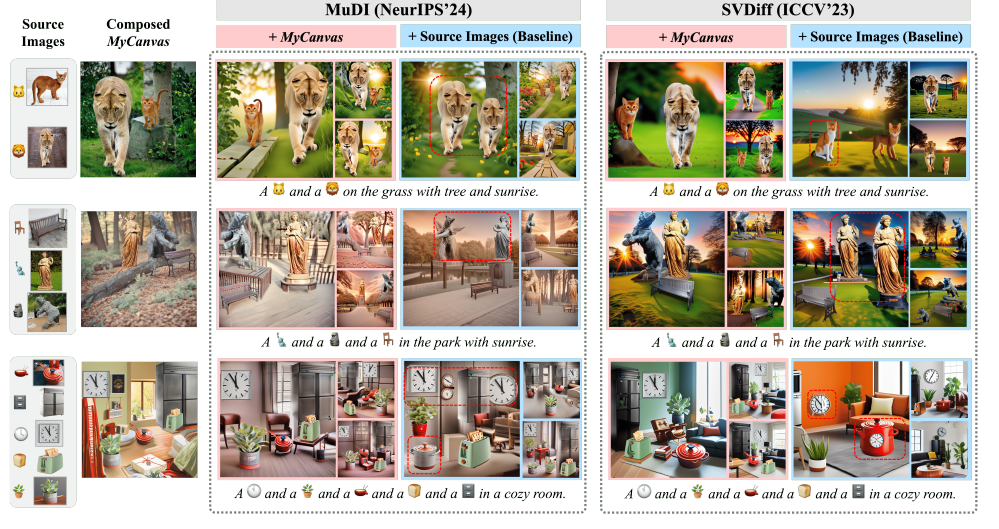

Figure 6: **Qualitative Comparison with Most Recent Baselines: MuDI [18] (SDXL-Based) and SVDiff [16] (SD2.1-Based).** Our *MyCanvas*, generated via *Gen4Gen*, enhances the performance across 2 to 5 concepts of more advanced baselines. Specifically, we can see that previous methods (without using our data) often duplicate the same concept multiple times in an image. By using *MyCanvas*, the generated images adhere much closer to the prompt in terms of counts.

chosen for its reproducibility and strong baseline performance. For each setting, models were trained on various compositions using SDXL [30], and evaluated using best checkpoints

with prompts different from those used during training to assess generalization. We use ViT-B-32 [11] as the backbone for OWL-ViT detection and CP-CLIP / TI-CLIP calculations.

Table 1 presents the outcomes across all compositions, organized by the number of objects. We use 41 text prompts, with 6 samples per prompt for each composition, resulting in a total of 246 generated images. It is evident that Custom Diffusion, when learning with the original source images, exhibits a ∼50% decrease in performance compared to its counterpart utilizing our composed *MyCanvas* dataset. By applying our prompting strategy to Custom Diffusion further amplifies the CP-CLIP score. Notably, our TI-CLIP score, indicative of background generalization, maintains consistency across all methods, ensuring that the observed increase in composition accuracy is not a consequence of overfitting.

4.2 User Preference Study

We conduct a user study with 30 participants who rate models from 1 to 10 across two criteria: *1) Image composition alignment* (scene quality and object arrangement) and *2) Text-to-image alignment* (consistency between image and caption). Table 2 shows that Custom Diffusion trained with *MyCanvas* significantly outperforms baselines in both aspects.

Table 2: **User Preference Study (Score: 1 to 10).** Users prefer our approach over the baseline methods for both image and text alignment, across $\leq$ 3 to 5 concepts.

	Prompting(Ours) + *MyCanvas*		**CD + *MyCanvas***		**CD + Source Images**	
	Image Alignment	Text Alignment	Image Alignment	Text Alignment	Image Alignment	Text Alignment
<= 3 Concepts	**8.4** (± 1.1)	**8.2** (± 2.6)	**6.8** (± 1.6)	**7.1** (± 2.1)	**3.5** (± 1.8)	**3.9** (± 2.6)
4 Concepts	**8.6** (± 1.4)	**8.9** (± 2.7)	**7.3** (± 1.8)	**7.5** (± 2.7)	**3.7** (± 1.9)	**3.4** (± 2.8)
5 Concepts	**8.8** (± 1.1)	**8.6** (± 2.7)	**6.6** (± 1.6)	**6.4** (± 2.5)	**3.9** (± 1.8)	**3.3** (± 2.7)

4.3 Qualitative Comparisons

Comparison with Personalization Methods. We primarily benchmark *MyCanvas* and our training-time prompting strategies on Custom Diffusion due to its simplicity and strong generalization. Figure 5 compares: 1) Custom Diffusion with source images, 2) Custom Diffusion with *MyCanvas*, and 3) with added prompting strategies. Using *MyCanvas* improves composition by ensuring all subjects are present, while our prompting strategy further enhances fidelity (e.g., better-preserved structures) and reduces missing elements (e.g., barns).
Generalization to Other Methods. To assess broader applicability, we compare against MuDI [18] and SVDiff [16] in Figure 6. Across all examples, training with *MyCanvas* consistently improves object composition and reduces missing or duplicated concepts. Results adapting to DreamBooth [36], Break-A-Scene [2], and GLIGEN [21] are in Appendix.

4.4 Ablation Study

Evaluating *MyCanvas* Quality. We developed a filtering tool (described in Appendix) to assess the quality of 800 images generated by our *Gen4Gen* pipeline. We evaluate each image based on: 1) the inclusion of personalized concepts, 2) their appropriate placement, and 3) the exclusion of visual artifacts, ranking them from 1 to 5. Subsequently, we aggregate these rankings to analyze the score distribution. Only images rated 4/5 were added to the *MyCanvas*

Table 3: **Quality Evaluation of *MyCanvas* (Rank: 1 to 5).** Our evaluation criteria include: *1) inclusion of personalized concepts, 2) accuracy of their placement, and 3) visual artifacts. MyCanvas* includes images with rank 4 and 5 (highlight in green).

	Rank: 1	Rank: 2	Rank: 3	Rank: 4	Rank: 5	Total Images
<= 3 Concepts	**9** (3.4 %)	**43** (16.3 %)	**72** (27.3 %)	**84** (31.8 %)	**56** (21.2 %)	**264**
4 Concepts	**16** (6.0 %)	**53** (19.8 %)	**112** (42.0 %)	**54** (20.2 %)	**32** (12.0 %)	**267**
5 Concepts	**19** (7.1 %)	**63** (23.4 %)	**127** (47.2 %)	**42** (15.6 %)	**18** (6.7 %)	**269**

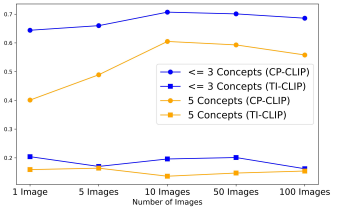

Figure 7: **Training Performance Based on Dataset Size.** For compositions with ≤ 3 concepts, fewer images are sufficient. However, stable performance for > 4 concepts requires 10 to 50 images.

Figure 8: **Failure Cases in Dataset Creation.** Failures stem from 1) unrealistic object placement by LLM causing identity distortion (e.g., cup $\rightarrow$ lamp), and 2) artifact introduction during inpainting.

dataset. Our findings in Table 3 indicate that generating high-quality images becomes more feasible with fewer than four concepts involved.

Training Data Size vs. Number of Concepts. We provide an analysis illustrated in Figure 7, training with varying number of images (1 to 100). While it is sufficient with very few image when training the compositions for ≤ 3 concepts, the training stabilizes between 10 to 50 images when there are more than 4 concepts. This shows that our dataset size is more than enough to obtain stable performance.

5 Conclusion

We introduce *Gen4Gen* and *MyCanvas*, a high-quality dataset with wellaligned image and text descriptions, as a benchmark for multi-concept personalization. We present extensive studies on our dataset, along with some training prompt amendments and a holistic metric, to show that improving data quality can lead to significantly better image generation for complex compositions. We hope that our contributions serve as a foresight to the possibilities of personalized text-to-image generation and automated dataset creation.

Limitations. As depicted in Figure 8, our current data creation pipeline still contains defects, particularly in challenging scenarios. These challenges stem from the LLM offering impractical guidance on object positions, and the diffusion inpainting introducing artifacts to objects. For now, weresort to a semi-automated screening process to address these issues. Future work could focus on automating the filtering process and assessing dataset quality. In addition, with the new MLLMs having rich multi-modal understanding, we can include additional visual guidances for better bounding box generation.

References

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023.

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023.

[3] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Lia, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[6] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023.

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023.

[8] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.

[9] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2024.

[10] Tan M Dinh, Rang Nguyen, and Binh-Son Hua. Tise: Bag of metrics for text-to-image synthesis evaluation. In *European Conference on Computer Vision*, pages 594–609. Springer, 2022.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[13] 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.

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022.

[15] 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. *arXiv preprint arXiv:2305.18292*, 2023.

[16] 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 (ICCV)*, pages 7323–7334, October 2023.

[17] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023.

[18] 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.

[19] 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.

[20] Yuheng Li, Haotian Liu, Yangming Wen, and Yong Jae Lee. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*, 2023.

[21] 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.

[22] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[24] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023.

[25] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023.

[26] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[28] OpenAI. Gpt-4 technical report, 2023.

[29] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022.

[30] 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.

[31] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

[32] 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.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21 (1):5485–5551, 2020.

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[35] 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.

[36] 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.

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[39] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023.

[40] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024.

[41] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15943–15953, October 2023.

[42] Zijie Wu, Chaohui Yu, Zhen Zhu, Fan Wang, and Xiang Bai. Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. *arXiv preprint arXiv:2310.08094*, 2023.

[43] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024.

[44] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[45] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.