

# OMG: Occlusion-friendly Personalized Multi-concept Generation in Diffusion Models

Zhe Kong<sup>1</sup>, Yong Zhang<sup>2</sup>, Tianyu Yang<sup>3</sup>, Tao Wang<sup>4</sup>, Kaihao Zhang<sup>5</sup>, Bizhu Wu<sup>6</sup>, Guanying Chen<sup>1</sup>, Wei Liu<sup>2</sup>, and Wenhan Luo<sup>1,7</sup>

<sup>1</sup>Shenzhen Campus of Sun Yat-sen University, <sup>2</sup>Tencent AI Lab, <sup>3</sup>International Digital Economy Academy, <sup>4</sup>Nanjing University, <sup>5</sup>Harbin Institute of Technology, Shenzhen, <sup>6</sup>Shenzhen University, <sup>7</sup>The Hong Kong University of Science and Technology

Homepage: <https://kongzhecn.github.io/omg-project/>  
Code: <https://github.com/kongzhecn/OMG/>



**Fig. 1:** We present OMG, an occlusion-friendly method for multi-concept personalization with strong identity preservation and harmonious illumination. The visual examples are generated by using LoRA models downloaded from civitai.com.

**Abstract.** Personalization is an important topic in text-to-image generation, especially the challenging multi-concept personalization. Current multi-concept methods are struggling with identity preservation, occlusion, and the harmony between foreground and background. In this work, we propose OMG, an occlusion-friendly personalized generation framework designed to seamlessly integrate multiple concepts within a single image. We propose a novel two-stage sampling solution. The first stage takes charge of layout generation and visual comprehension information collection for handling occlusions. The second one utilizes the acquired

visual comprehension information and the designed noise blending to integrate multiple concepts while considering occlusions. We also observe that the initiation denoising timestep for noise blending is the key to identity preservation and layout. Moreover, our method can be combined with various single-concept models, such as LoRA and InstantID without additional tuning. Especially, LoRA models on civitai.com can be exploited directly. Extensive experiments demonstrate that OMG exhibits superior performance in multi-concept personalization.

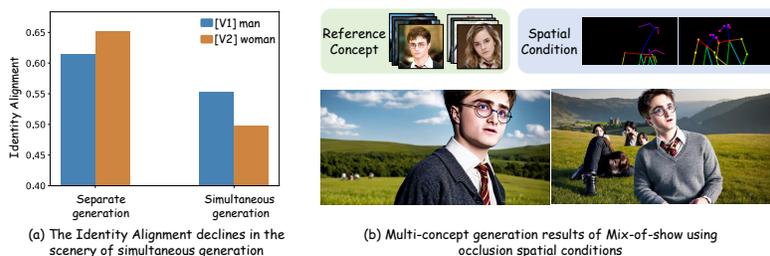
**Keywords:** Image Generation · Image Customization · Diffusion Model

## 1 Introduction

Personalized text-to-image generation is a promising path to realize identity-consistent story visualization. Numerous methods have been proposed for single-concept personalization, such as DreamBooth [33], Textual Inversion [12] and LoRA [22], showcasing their efficacy in achieving high-quality results. While excelling in single-concept personalization, these methods encounter challenges related to identity degradation when tasked with generating a single image encompassing multiple concepts, as shown in Fig. 2 (a).

Several multi-concept personalization methods have been proposed [15, 25, 28, 41], but they still encounter identity degradation problems when generating multiple concepts. Mix-of-show [16] can generate multi-concepts with realistic identity, but it cannot handle occlusion between concepts. Specifically, the method [16] adopts a regionally controllable sampling method, where each timestep injecting region prompts through regional-aware cross-attention. In cases where the concept regions experience occlusion, the final prediction results for these occluded regions are determined by a straightforward linear addition of the cross-attention results from multiple local sample regions. This simplistic approach leads to inaccurate predictions within the occluded regions, resulting in layout conflicts and identity degradation, as shown in Fig. 2 (b). Besides, there is disharmony between the foreground and background, leading to unnatural illumination in the image. Additionally, methods [16, 25] aim at merging two concepts into one diffusion model, which is computationally inefficient.

To address the aforementioned issues, we propose OMG, an occlusion-friendly personalized image generation framework designed to seamlessly integrate multiple concepts within a single image. Unlike other customization methods, our two-stage approach employs latent-level and attention-level layout control to tackle occlusion issues during multiple concept customization. The first stage generates an image with coherent layouts based on user-provided text prompts, without considering personalization. During this stage, additional visual comprehension information such as attention maps and concept masks is acquired through the first stage of sampling. In the second stage, concepts are injected into specific regions by leveraging the preserved visual comprehension information. During sampling, as illustrated in Fig. 2 (a), simultaneously generating two concepts in one image results in significant identity degradation. To address



**Fig. 2:** Existing methods face identity degradation and occlusion problems. (a) Given two text prompts with identifiers, “A [v1] man” and “A [v2] woman”, we generate 100 images for the two concepts separately (separate generation) and calculate the Identity Alignment between generated images and reference images. Subsequently, we employ another text prompt, “A [v1] man and a [v2] woman”, to randomly generate 100 images containing both concepts simultaneously (simultaneous generation) and calculate Identity Alignment. We find that the simultaneous generation of two concepts leads to the decline of Identity Alignment, resulting in identity degradation. (b) Given spatial conditions with occlusion between concepts, the Mix-of-show [16] cannot generate an integrity image and encounters an identity degradation problem.

this limitation, we propose a concept noise blending strategy to merge multiple noises from different single-concept models during sampling. In each timestep, different single-concept models only control the generation of one specific region, effectively mitigating identity degradation problems during the multiple-concept sample process. Additionally, we find that the disharmony problem can be solved by controlling the initiation timestep of concept noise blending. Differing from Custom Diffusion [25] and Mix-of-show [16], which require additional training or model optimization to merge multiple concepts into one model, the proposed OMG method can generate an image with multiple concepts directly by utilizing multiple single-concept models derived from the community (*e.g.*, civitai.com) in a plug-and-play manner, without additional tuning. It is computationally efficient and significantly alleviates the time-consuming problem. Extensive experiments and comparison results with other methods demonstrate its superiority. Our contributions are summarized as follows:

- We propose a novel two-stage framework for multi-concept customization. Our approach can generate an occlusion-friendly personalized image with strong identity preservation and harmonious illumination.
- We propose a Concept Noise Blending strategy to merge multiple noises from different single-concept models at both latent and attention levels. It mitigates identity degradation of the multi-concept generation and can be easily combined with different personalization frameworks such as LoRA or InstantID in a tuning-free plug-and-play manner.
- Extensive evaluations demonstrate the effectiveness of our proposed method.

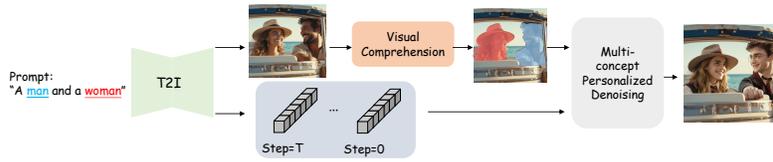
## 2 Related Work

**Text-to-Image (T2I) Synthesis.** Text-to-image synthesis involves the task of generating realistic and diverse images from text prompts. Recently, diffusion models [21, 39] have demonstrated remarkable progress, attributed to large-scale training datasets like Laion-400M [36] and Conceptual-12M [7]. Several text-to-image models, including SDXL [50], Imagen [35], and DALL·E 3 [5], have shown significant performance improvements.

**Single-Concept Customization.** Early image personalization approaches focus on expanding or fine-tuning the language vision dictionary of T2I diffusion models to associate new concepts with a limited set of subjects, achieved through the fine-tuning of pre-trained T2I models. Optimization-based methods, such as diffusion model based ones [6, 10, 18, 19, 33, 34, 38], or special textual embeddings [1, 12, 30, 42, 43, 48, 49], learn new concepts to describe target concepts. To reduce the trainable parameters, recent advancements have seen the adoption of Low-Rank Adaptation (LoRA) methods [22, 40] in concept customization. Moreover, studies [2, 8, 9, 13, 14, 27, 29, 37, 37, 44, 45, 47, 50, 51] have recently explored training additional modules for mapping concepts to textual representations while keeping the core pre-trained T2I models frozen. This significantly expedites the personalization process. For instance, in InstantID [44], an IdentityNet is designed to integrate facial images with textual prompts, successfully steering image generation in various styles using just a single facial image.

**Multi-Concept Customization.** Existing methods conduct joint training on multi-concept datasets with additional losses or extra optimization efforts to merge multiple models. Several approaches [3, 15, 17, 28] employ cross-attention maps to avoid the entanglement of multiple concepts. In Custom Diffusion [25], the proposition involves joint training or constrained optimization of multiple models. Notably, the work [16] introduces gradient fusion to minimize identity loss during concept fusion, along with the proposal of regionally controllable sampling to address attribute binding in multi-concept personalization. Modular Customization [31] disentangles customization concepts into orthogonal directions, streamlining the integration of multiple fine-tuned concepts, while preserving the integrity of each concept. [46] employs subject embeddings from an image encoder to enhance generic text conditioning in diffusion models. This augmentation empowers personalized image generation without the necessity for additional training when facing new concepts.

In contrast to the aforementioned methods, our approach diverges by obviating the need for extensive pre-training of additional network models or the optimization required for merging multiple models. Through a simple modification of the sampling process, our method seamlessly integrates multiple concepts into a single image using multiple models, thereby eliminating the necessity for model merging or additional tuning. Furthermore, our method exhibits robust generalization and can be effortlessly combined with various single-concept methods, such as LoRA [22] and InstantID [44], in a plug-and-play manner.



**Fig. 3:** Overviews of the proposed OMG, which contains two stages during sampling. The first stage takes charge of layout generation and visual comprehension information for handling occlusions. Leveraging the acquired information, the identities of concepts can be injected in multi-concept personalized denoising with the proposed latent-level and attention-level noise blending in the second stage.

### 3 Method

We propose a two-stage multi-concept customization framework to integrate multiple concepts into a single image. Unlike previous works, the proposed method can address identity degradation, occlusion, time-consuming fusion, and illumination disharmony problems. The overall framework of our proposed paradigm is illustrated in Fig. 3, which contains two stages during sampling.

#### 3.1 Preliminary

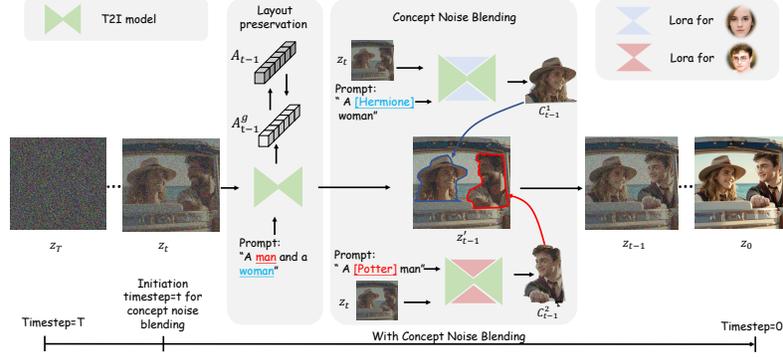
Latent diffusion model [21, 32, 39] belongs to a class of generative models containing a diffusion process and a reverse process in the latent space. In the diffusion process, an image  $x$  is firstly projected to latent space by an encoder  $\mathcal{E}$ :  $z_0 = \mathcal{E}(x)$ . Then random Gaussian noises are gradually added to the data sample  $z_0$  to generate the noisy sample  $z_t$  with a predefined noise adding schedule  $\alpha_t$  as timestep  $t$ :  $q(z_t|z_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}z_0, (1 - \bar{\alpha}_t)I)$ , where  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ . In the reverse process, a U-Net  $\varepsilon_\theta$  is trained to directly perform denoising in the latent space. The overall training objective is defined as

$$\mathcal{L} = E_{z_0, \epsilon, t} \|\epsilon - \varepsilon_\theta(z_t, t, c)\|_2^2, \quad (1)$$

where  $c$  is the embedding of the conditional text prompt and  $z_t$  is a noisy sample of  $z_0$  at timestep  $t$ .

#### 3.2 Stage 1: Visual Comprehension Information Preparation

Existing methods, such as Mix-of-show [16], encounter layout conflict challenges. As depicted in Fig. 2 (b), when the regions of two concepts occlude, [16] is incapable of generating an image with a coherent layout, resulting in identity degradation and compromise of the concept’s integrity. Given that cross-attention layers are effective in controlling the spatial layout and appearance [20], the modification of pixel-to-text interaction within these layers allows for preserving the content and spatial layout of the original image while adhering to the target prompt. By selectively modifying predefined regions in an image using a unique



**Fig. 4:** Overviews of the Multi-concept Personalized Denoising. This stage utilizes the acquired visual comprehension information and the designed concept noise blending method to integrate multiple concepts while considering occlusions.

identifier, while maintaining the content and structure of other regions, we can effectively mitigate the challenge of concept occlusion. Hence, the first stage aims to acquire visual comprehension information for multi-concept customization.

As illustrated in Fig. 3 (a), a textual prompt  $p$  describing multiple objects of an image is the input of a T2I model. It is imperative to emphasize that the text prompt  $p$  exclusively contains the class name (e.g., “man” or “woman”), deliberately excluding the introduction of the unique identifier (e.g., “[ $v$ ] man” or “[ $v$ ] woman”) at this point. Consequently, a non-customized image  $x_{ncus}$  with a coherent layout is generated through

$$x_{ncus} = T2I(p). \quad (2)$$

We employ the publicly available SDXL model as our T2I model. The denoising UNet network is composed of self-attention layers followed by cross-attention layers. In the denoising process, the fusion of embeddings from visual and text features occurs through cross-attention layers, generating cross-attention maps for each textual token in the U-Net. The cross-attention map  $A$  is calculated as

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right). \quad (3)$$

Here,  $Q$  represents a query matrix projection of intermediate features  $\varphi(z_t)$ , and  $K$  is a key matrix projection of text tokens  $\phi(p)$ , obtained through two learnable linear projections  $W_Q$  and  $W_K$ , respectively.  $Q$  and  $K$  are defined as

$$Q = W_Q \cdot \varphi(z_t), K = W_K \cdot \phi(p). \quad (4)$$

At each denoising step  $t$ , following the input of  $p$  to the T2I model, the cross-attention maps  $A_t$ , comprising  $N$  attention layers with corresponding spatial attention maps  $\{A_t^1, A_t^2, \dots, A_t^N\}$ , are acquired. It is imperative to retain all these obtained attention maps for identity injection in the second stage.



**Fig. 5:** Effect of the initiation timestep for concept noise blending. The initiation timestep for concept noise blending influences both the image layout and illumination. When the initiation timestep is 0, there is no concept noise blending operation during sampling, resulting in the same generation result for both stages.

To prepare for concept noise blending, it is necessary to locate the modified region in  $x_{ncus}$ . Relying on the robust image understanding capabilities [24] of visual comprehension, we can derive concept masks  $M$ . By inputting both the generated image  $x_{ncus}$  and the class name (*e.g.*, “man” or “woman”) from  $p$ , concept masks  $M$  corresponding to  $k$  class  $\{M_1, M_2, \dots, M_k\}$  can be derived.

### 3.3 Stage 2: Multi-concept Personalized Denoising

Upon obtaining a non-customized image  $x_{ncus}$  with acquired visual comprehension information, we inject the identity of concepts in the second stage. In previous works, such as [20], image editing is achieved by injecting the input text with an edit text prompt. Personalized multi-concept generation could adopt a similar approach by triggering concept generation through the identifiers in text prompts. However, it may face two drawbacks. Firstly, making a text prompt capable of generating multiple concepts necessitates the merging of multiple single-concept models into one like [16, 31], which requires additional network optimization and is inherently time-consuming. Additionally, as illustrated in Fig. 2 (a), employing a single prompt for a multi-concept generation often results in identity degradation. In contrast, we propose a Concept Noise Blending strategy to address the aforementioned issues. The overall architecture of the multi-concept personalized denoising is depicted in Fig. 4.

**Concept Noise Blending.** To mitigate the additional optimization costs associated with network merging, the proposed method directly leverages multiple single-concept models during inference, circumventing the need for network merging. Moreover, each single-concept model is solely responsible for generating a specific concept, effectively addressing the challenge of identity degradation.

During the multi-concept personalized denoising, the input global text prompt  $p$  and initiation noise remain consistent at the first stage. In the second stage, the objective is to generate a customized image containing multiple concepts leveraging the acquired visual comprehension information. Suppose we aim to generate an image  $x_{cus}$  containing  $k$  concepts  $\{C^1, C^2, \dots, C^k\}$ . Let  $T2I_c^i$  represent the  $i$ -th single-concept model designed to generate the concept  $C^i$  through



**Fig. 6:** Comparison of OMG with other methods on the single-concept customization. In both character customization and object customization, OMG exhibits superior identity alignment with reference images when compared to other methods.

concept text prompt  $p^i$ . The  $p^i$  encapsulates a special identifier that can be input to  $T2I_c^i$  for generating concept  $C^i$ . At timestep  $t$ , given text prompt  $p^i$  of concept  $i$ , the corresponding predicted noise  $C_{t-1}^i$  is obtained through

$$C_{t-1}^i = T2I_c^i(z_t, p^i, t). \quad (5)$$

Additionally, the  $T2I$  model is the same as the first stage. By inputting a global text prompt  $p$  at timestep  $t$ , the corresponding global output  $z'_{t-1}$  is obtained through the  $T2I$  model with occlusion layout preservation. The generated  $z'_{t-1}$  represents a non-customized noise. To inject the identity of the concept  $C^i$  into  $z'_{t-1}$ , specific regions in  $z'_{t-1}$  are overwritten with the corresponding concept noise  $C_{t-1}^i$  based on concept masks  $M$  through:

$$z_{t-1} = (1 - \bigcup_{i=0}^k M_i) * z'_{t-1} + \sum_{i=0}^k M_i * C_{t-1}^i, \quad (6)$$

where  $M_i$  denotes the mask for concept  $C^i$ . Through noise-level concept blending, the identity of concepts can be injected into one noise at each timestep.

MultiDiffusion [4] similarly incorporates noise fusion during sampling, by binding together multiple diffusion generation processes with a shared set of parameters or constraints to generate high-quality and diverse images that adhere to user-provided controls. In contrast, the proposed Concept Noise Blending does not necessitate multiple crops. Instead, different regions are calculated by distinct models. Ultimately, the results from various regions are fused based on the concept mask, eliminating the need for additional optimization steps.

**Occlusion Layout Preservation.** The initiation stage yields a non-customized image  $x_{ncus}$  with a coherent layout. In the second stage, despite the global prompt and initiation noise being identical to those in the first stage, the generated noises at each timestep are completely different due to Concept Noise Blending. We utilize the cross-attention maps  $A$  stored in the first stage to uphold the layout consistency of the generated image with  $x_{ncus}$ . This operation ensures the production of an occlusion-friendly multi-concept customized image.

In each timestep, we ensure that the layout is preserved in the generated image by modifying the cross-attention maps within the UNet during the  $T2I$

model sampling. For instance, at the  $t$  step,  $z_t$  is fed into the  $T2I$  model alongside the global prompt  $p$  and timestep  $t$ . Cross-attention maps play a crucial role in controlling the structure and geometry of an image. To maintain an occlusion-friendly layout, we overwrite the generated attention map in each timestep within the UNet with the stored maps. This process can be formulated as:

$$z'_{t-1} = T2I(z_t, p, t)\{A_t^g \leftarrow A_t\}, \quad (7)$$

where  $A_t^g$  denotes the generated attention map in the second stage of the  $T2I$  model and  $A_t$  denotes the stored attention map from the first stage.

### 3.4 Denoising Timestep of Concept Noise Blending

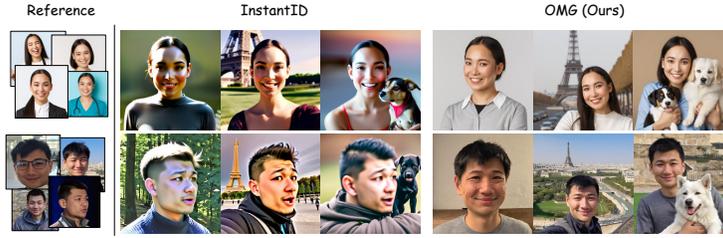
The initiation timestep for concept noise blending holds significant influence over both the image layout and illumination of the generated image. To elucidate the impact of different concept noise blending starting points, we present generated images at various timesteps in Fig. 5. Leveraging DDIM, the series comprises a total of 50 steps, with the leftmost image representing the outcome when Concept Noise Blending begins at step 50, indicating that concept noise blending operations are active throughout the entire sampling process. The rightmost image represents the result starting from step 0, indicating that no concept noise blending operations occur during the entire sampling process. Hence, when the concept noise blending operation starts at timestep 0, the generated image is the same as stage one.

Commencing concept noise blending at an early step may introduce layout conflicts in the composition and shape of objects within the generated image. However, with the increase in concept noise blending steps, the content information becomes more coherent and stable, effectively preserving the identity of the object. After iterative denoising, as the concept noise blending step approaches 0, the identity of the character diminishes gradually, resulting in a synthesized image resembling the first stage. This highlights the early stage of sampling governs the image layout, while the identity of concepts unfolds in later timesteps. The optimal step of concept noise blending is approximately 35.

Moreover, we observe that the illumination disharmony between the foreground and background is notable in the earlier steps. With increasing timesteps, the illumination gradually becomes consistent, suggesting a potential association between illumination and image layout information.

## 4 Experiments

**Datasets.** To evaluate OMG method, we collect a dataset that encompasses 15 distinct concepts. This dataset comprises 7 real-world characters, 3 anime characters, and 5 real-world objects, all annotated automatically by Blip-2 [26]. **Experimental Setup.** We implement OMG employing the SDXL model [50]. The multi-concept customization approach we propose can be seamlessly combined with various single-concept customization methods, such as LoRA [22] and



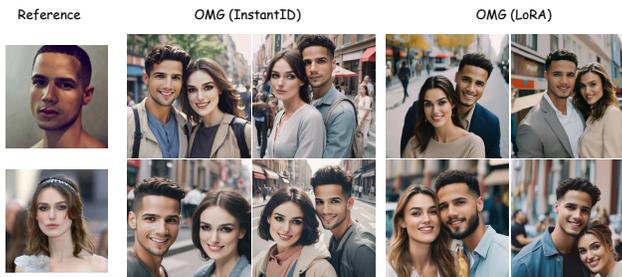
**Fig. 7:** Comparison with InstantID [44] in single-concept customization. OMG emerges as the superior method by generating images with more natural colors. This showcases the prowess of OMG over InstantID in single-concept customization. For LoRA and InstantID, we all adopt SDXL-base-1.0 as the base model for a fair comparison.



**Fig. 8:** Comparison of OMG with other methods on the same spatial condition on multi-concept customization. To make a fair comparison, all the comparison methods utilize the same spatial condition in each row. The proposed OMG can achieve the best performance in identity preservation in multi-concept customization.

InstantID [44]. For LoRA [22], we integrate the LoRA layer into the linear layer in all attention modules of the text encoder and Unet, with a rank of 256. We use the Adafactor optimizer with a constant learning rate for all experiments, setting the learning rate for the text encoder to  $3e^{-5}$  and for Unet to  $3e^{-3}$ . Single-concept fine-tuning requires approximately 2 hours on one A100 GPU. Regarding InstantID [44], we leverage the officially provided pre-trained Image Adapter model and IdentityNet model. We utilize the Antelopev2 model for face detection and face ID embedding extraction. When combining the proposed method with InstantID for multi-concept customization, only forward inference is needed during concept image generation, without any additional training.

**Evaluation Metrics.** Following [16], we evaluate our method using Image Alignment, which measures the visual similarity of generated images with the target concept using similarity in the CLIP image feature space [23]. Additionally, we adopt Text Alignment, which measures the similarity of generated images



**Fig. 9:** Comparison of OMG with InstantID [44] in multi-concept customization. OMG stands out by generating images with enhanced realism, characterized by a more extensive and vibrant color spectrum.

**Table 1:** Quantitative comparison on single- and multi-concept personalization. OMG achieves state-of-the-art performance in single-concept customization and achieves better identity preservation than other methods in multi-concept customization.

Method	Character						Object					
	Text Alignment			Identity Alignment			Text Alignment			Image Alignment		
	Single	Multiple	$\Delta$	Single	Multiple	$\Delta$	Single	Multiple	$\Delta$	Single	Multiple	$\Delta$
DreamBooth [33]	0.677	0.658	-0.019	0.456	0.480	0.025	0.713	0.717	0.004	0.805	0.800	-0.005
Textual Inversion [12]	0.673	0.673	0.000	0.292	0.294	0.002	0.693	0.697	0.004	0.784	0.781	-0.003
Custom Diffusion [25]	0.629	<b>0.704</b>	0.075	0.370	0.322	-0.048	0.695	0.755	0.060	0.840	0.778	-0.061
Mix-of-show [16]	0.675	0.639	-0.036	0.422	0.436	0.015	0.724	0.731	0.007	0.791	0.780	-0.011
OMG (Ours)	<b>0.693</b>	0.696	0.003	<b>0.514</b>	<b>0.510</b>	-0.004	<b>0.730</b>	<b>0.762</b>	0.032	<b>0.842</b>	<b>0.810</b>	-0.033

with given prompts using text-image similarity in the CLIP feature space [23]. However, for face images, Image Alignment may not accurately evaluate the similarity between the generated face and the real face. To address this, we use Identity Alignment to further illustrate the identity-preserving capabilities by measuring the ArcFace score [11] at which the target human identity is detected in a set of generated images. Consequently, we adopt Text Alignment and Image Alignment for objects, and for characters, Text Alignment and Identity Alignment are employed to measure the performance of methods.

#### 4.1 Quantitative Comparison

We compare OMG with several concept customization methods, including DreamBooth [33], Textual Inversion [12], InstantID [44], Custom Diffusion [25], and Mix-of-show [16]. All the methods except InstantID are training-based customization requiring multiple reference images. In contrast, InstantID [44] achieves personalized generation with just one reference image.

Following Custom Diffusion [25], we utilize 20 text prompts and 50 samples per prompt for each concept. Hence, a total of 1000 images are ultimately generated. For a fair comparison, all the comparison methods adopt DDIM sampling with 50 steps and a classifier-free guidance sample across all methods. Our evaluation spans various categories of concepts, including characters and objects. We use a single-concept tuned model to assess the identity-preserving effect of



**Fig. 10:** Qualitative ablation study of OMG. (a) Generating images with layout preservation can preserve reasonable structure and enhance realism in the generated images. (b) Concept Noise Blending can generate images with a more coherent image layout and harmonious illumination. (c) The proposed OMG can achieve multi-concept customization with an increasing number of concepts.

our method through a set of prompts. The experimental results including single-concept and multi-concepts are detailed in Tab. 1.

For single-concept, we achieve the best results in Text Alignment, Image Alignment, and Identity Alignment for characters and objects. We adopt LoRA for single-concept fine-tuning, which proves the effectiveness of LoRA in capturing the complex concepts’ identity. For multi-concept, the proposed method exhibits superior performance with the input images for object customization. For characters, our method performs better on Identity Alignment than other methods, which proves the superiority of our method in identity preservation.

In our comparative analysis, we compare the proposed method with InstantID [44]. Notably, InstantID [44] achieves image customization requiring only a single reference image for inference, while ours leverages multiple reference images for fine-tuning. To ensure an equitable comparison, we align the number of reference images used by InstantID with our approach and calculate the average mean of ID embeddings as an image prompt. Consequently, our method achieves a Text Alignment score of 0.692 and an Identity Alignment score of 0.500. InstantID, exhibiting superior performance with a Text Alignment score of 0.698 and an Identity Alignment score of 0.534, benefits from fine-tuning on ample facial data. It is notable that our method, in contrast, has not undergone fine-tuning on such extensive datasets.

The qualitative results for multi-concept personalization shown in Tab. 1 mainly measure the fusion ability of multiple single-concept models during multi-concept customization. It cannot reflect the generation effect of multi-concept generation. Therefore, to measure the generation effects when different methods

**Table 2:** Quantitative results for diverse concepts generation.

Method	man + woman	man + man	woman + woman	object + object	Average
	IDA	IDA	IDA	IMA	
DreamBooth [33]	0.302	0.258	0.192	<b>0.784</b>	0.384
Textual Inversion [12]	0.122	0.131	0.064	0.675	0.248
Custom Diffusion [25]	0.265	0.210	0.212	0.757	0.361
Mix-of-show [16]	0.361	0.219	0.143	0.736	0.365
OMG (Ours)	<b>0.487</b>	<b>0.377</b>	<b>0.293</b>	0.763	<b>0.480</b>

generate multiple concepts simultaneously, we propose a new calculation method. To make a fair comparison, we use the same spatial condition to generate two concepts simultaneously. We calculate the region of different concepts in the image through visual comprehension, then calculate the average Identity Alignment scores (IDA) or Image Alignment score (IMA) for two different concepts with their corresponding reference images separately. This measure approach is more effective in measuring the effects of generating multiple concepts simultaneously. The experiment results are shown in Tab. 2. OMG outperforms other methods in generating images with various concept combinations, demonstrating the effectiveness of our method in multi-concept generation.

## 4.2 Qualitative Comparison

**Single-Concept Results.** The efficacy of our method in preserving identity is demonstrated through a comparison of single-concept generation representing different identities. As previously mentioned, each concept undergoes individual fine-tuning. The experimental results are presented in Fig. 6. Each column corresponds to images sampled from the same model, representing two distinct concept identities. In both character customization and object customization, our method exhibits superior identity alignment with reference images when compared to other methods. The text prompts can be found in the supplement.

Fig. 7 illustrates the results of single-concept customization compared to InstantID [44]. Our method stands out by generating higher-quality images, underscoring its visual superiority over InstantID [44] in single-concept customization.

**Multi-Concept Results.** We take a comprehensive comparison with other methods in multi-concept customization. Owing that the Mix-of-show [16] requires additional spatial conditions, we implement identical spatial condition controls across all compared methods to make a fair comparison. The experimental results are illustrated in Fig. 8. Mix-of-show [16] generates layout conflict images, leading to object loss and identity degradation. Notably, DreamBooth [33], Textual Inversion [12], and Custom Diffusion [25] exhibit limitations in generating images with realistic identity preservation. In contrast, our proposed method demonstrates robust identity preservation for each character in the multi-concept generation, substantiating its efficacy in multi-concept customization.

Furthermore, we conduct a comparative analysis between the proposed method and InstantID [44]. To facilitate this comparison, we substitute the single-concept

model in our approach with InstantID and juxtapose the two methods. The experimental findings are visually depicted in Fig. 9. Our method produces images with enhanced realism, with more natural facial. This substantiates the superior performance of our method in multi-concept customization.

### 4.3 Ablation Study

To assess the effectiveness of various components within OMG, we conduct an ablation study encompassing the following elements: Layout Preservation, Concept Noise Blending, and Different Numbers of Concepts.

**Layout Preservation.** We present the ablation results of layout preservation in Fig. 10 (a). The left image showcases the generated image in the first stage. The other two images illustrate the generated image with and without layout preservation, respectively. By substituting the attention maps generated in the second stage, the layout of the image is well-preserved. The inclusion of layout preservation contributes to the generation of a more reasonable structure, highlighting the effectiveness of layout preservation in enhancing the overall quality.

**Concept Noise Blending.** Subsequently, we compare different sample types, specifically regionally controllable sampling [16] and the proposed concept noise blending. Given that regionally controllable sampling necessitates additional spatial conditions, we ensure a fair comparison by providing the same poses for both methods. Experimental outcomes are shown in Fig. 10 (b). In instances of regionally controllable sampling, occluded regions of two concepts may lead to missing concepts or a disorderly image layout in the generated image. In contrast, the concept noise blending is effective when multiple concepts are occluded. Furthermore, our method yields images with more harmonious illumination between the foreground and the background, resulting in a more realistic portrayal.

**Different Numbers of Concepts.** We also assess the robustness by increasing the number of concepts. As depicted in Fig. 10 (c), we showcase the generation effects when the number of concepts varies from 1 to 5. Notably, even with an escalation in the number of concepts, our method consistently preserves the identity of each concept. This substantiates the efficacy of our method in generating a diverse array of concepts while maintaining identity integrity.

## 5 Conclusion

We introduce OMG, a personalized generation framework for handling occlusion challenges in the context of generating realistic images for multiple concepts. Leveraging an image editing framework, our method specifically addresses the occlusion problem prevalent in multi-concept generation. The proposed concept noise blending further mitigates identity degradation issues. Experimental results showcase OMG’s ability to successfully generate high-quality images even when concepts experience occlusion. Additionally, our method seamlessly integrates with various single-concept customization models without additional training, enhancing its versatility and practicality.

## Acknowledgment

This work is funded in part by the National Natural Science Foundation of China (Grant No. 62372480), in part by CCF-Tencent Rhino-Bird Open Research Fund (No. CCF-Tencent RAGR20230118), in part by Theme-based Research Scheme (T45-205/21-N) from Hong Kong RGC, in part by Generative AI Research and Development Centre from InnoHK.

## References

1. Alaluf, Y., Richardson, E., Metzger, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. *ACM TOG* **42**(6), 1–10 (2023)
2. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., H. Bermano, A.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In: *SIGGRAPH Asia 2023 Conference Papers*. pp. 1–10 (2023)
3. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311* (2023)
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113* (2023)
5. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf> **2**(3) (2023)
6. Chae, D., Park, N., Kim, J., Lee, K.: Instructbooth: Instruction-following personalized text-to-image generation. *arXiv preprint arXiv:2312.03011* (2023)
7. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3558–3568 (2021)
8. Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186* (2023)
9. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481* (2023)
10. Choi, J., Choi, Y., Kim, Y., Kim, J., Yoon, S.: Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779* (2023)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *CVPR*. pp. 4690–4699 (2019)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR* (2022)
13. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228* (2023)
14. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG* **42**(4), 1–13 (2023)

15. Gong, Y., Pang, Y., Cun, X., Xia, M., Chen, H., Wang, L., Zhang, Y., Wang, X., Shan, Y., Yang, Y.: Talecrafter: Interactive story visualization with multiple characters. *Siggraph Asia* (2023)
16. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *NIPS* (2023)
17. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305* (2023)
18. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971* (2023)
19. He, X., Cao, Z., Kolkin, N., Yu, L., Rhodin, H., Kalarot, R.: A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315* (2023)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *NeurIPS* **33**, 6840–6851 (2020)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. *ICLR* (2021)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. *arXiv:2304.02643* (2023)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. *arXiv preprint arXiv:2304.02643* (2023)
25. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: *CVPR*. pp. 1931–1941 (2023)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023)
27. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461* (2023)
28. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327* (2023)
29. Ma, Y., Yang, H., Wang, W., Fu, J., Liu, J.: Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319* (2023)
30. Pang, L., Yin, J., Xie, H., Wang, Q., Li, Q., Mao, X.: Cross initialization for personalized text-to-image generation. *arXiv preprint arXiv:2312.15905* (2023)
31. Po, R., Yang, G., Aberman, K., Wetzstein, G.: Orthogonal adaptation for modular customization of diffusion models. *arXiv preprint arXiv:2312.02432* (2023)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *CVPR*. pp. 10684–10695 (2022)
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *CVPR*. pp. 22500–22510 (2023)

34. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NIPS **35**, 36479–36494 (2022)
36. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
37. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
38. Smith, J.S., Hsu, Y.C., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual diffusion: Continual customization of text-to-image diffusion with c-lora. arXiv preprint arXiv:2304.06027 (2023)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ICLR (2020)
40. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
41. Tunanyan, H., Xu, D., Navasardyan, S., Wang, Z., Shi, H.: Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. arXiv preprint arXiv:2310.07419 (2023)
42. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. ACM TOG **42**(6), 1–13 (2023)
43. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.:  $p+$ : Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
44. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
45. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
46. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
47. Yan, Y., Zhang, C., Wang, R., Zhou, Y., Zhang, G., Cheng, P., Yu, G., Fu, B.: Facestudio: Put your face everywhere in seconds. arXiv preprint arXiv:2312.02663 (2023)
48. Zhang, X.L., Wei, X.Y., Wu, J.L., Zhang, T.Y., Zhang, Z.X., Lei, Z., Li, Q.: Compositional inversion for stable diffusion models. arXiv preprint arXiv:2312.08048 (2023)
49. Zhao, R., Zhu, M., Dong, S., Wang, N., Gao, X.: Catversion: Concatenating embeddings for diffusion-based text-to-image personalization. arXiv preprint arXiv:2311.14631 (2023)
50. Zhou, Y., Zhang, R., Gu, J., Sun, T.: Customization assistant for text-to-image generation. arXiv preprint arXiv:2312.03045 (2023)
51. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv preprint arXiv:2305.13579 (2023)