Multi-Concept Customization of Diffusion Models

Mariam Gasoyan[1,3], Levon Khachatryan[1,2]
[1]Yerevan State University (YSU), [2]Picsart AI Resarch (PAIR), [3]Cognaize (PAIR)

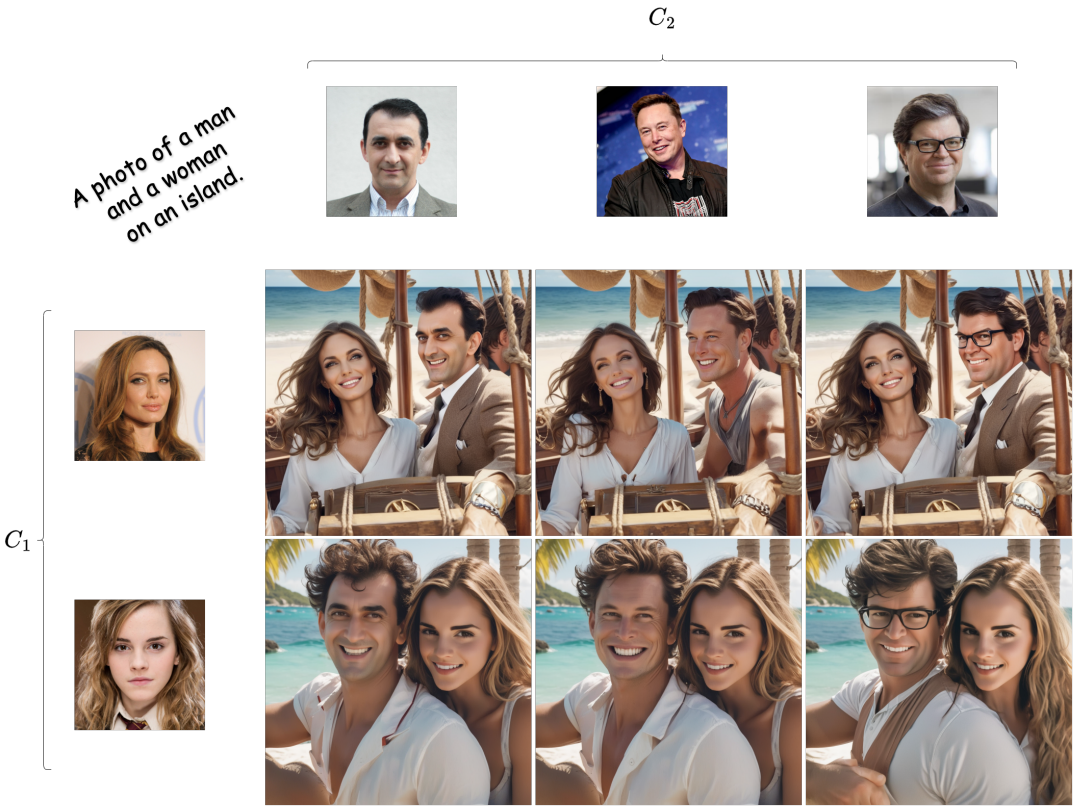

Figure 1. Dynamic Personalized Image Synthesis with Identity Preservation across Thematic Settings

Abstract

Generative diffusion models have achieved remarkable progress in high-fidelity image generation. Still, challenges persist in efficiently customizing these models for generating images that incorporate multiple distinct concepts without extensive retraining. This paper presents a novel zero-shot multi-concept customization approach for diffusion models, integrating SDXL, InstantID, and Grounding Dino to enable accurate personalization based on textual descriptions and reference images. Our method utilizes SDXL for text-guided generation, InstantID for identity preservation, and Grounding Dino for spatial mask generation. This approach allows a compelling fusion of latent representa-

tions to produce images that adhere to multiple distinct concepts, eliminating the need for extensive dataset accumulation and model retraining. Experimental results demonstrate that our method maintains high generative quality and robust adaptability to new concepts. This research advances the practical application of generative diffusion models, making sophisticated customization more accessible and efficient in real-world settings.

1. Introduction

The advancement of generative models in artificial intelligence has marked a significant milestone in the ability of machines to mimic and extend human creative capabilities. Among these generative models, diffusion models have emerged as a powerful class known for their remarkable ability to generate high-quality images. Initially popularized by their success in tasks traditionally dominated by generative adversarial networks (GANs) by [4], diffusion models have rapidly evolved, showcasing their versatility and robustness across various domains.

Diffusion models operate by gradually transforming a random noise distribution into a structured image through a process guided by a neural network by [6]. Although computationally intensive, this iterative process allows for generating images with complicated details and high fidelity. Recent research has focused on leveraging diffusion models for targeted image generation, where the goal is to produce images that stick to specific concepts or identities as defined by input conditions. However, much of the existing research relies on extensive retraining or fine-tuning of the models with large datasets for each new concept, which can be resource-intensive and impractical in many scenarios.

Our research introduces a novel approach to this challenge: zero-shot customization of diffusion models capable of generating images based on multiple identities or concepts without model retraining. This approach leverages the inherent flexibility of diffusion models and recent advances in zero-shot learning to adapt to new tasks without direct example-based training. By integrating zero-shot learning techniques, our model can interpret and synthesize complex, multi-concept images after being exposed only to textual descriptions of the desired outputs, thus circumventing the extensive data collection and training phases required.

This research is significant because it has the potential to drastically reduce the barriers to entry for the customization of generative models. Industries such as digital content creation, personalized advertising, and virtual design can benefit from the ability to rapidly prototype and customize outputs without the overhead of continuous model updates. Furthermore, this approach democratizes access to high-quality generative tools, allowing smaller entities or individuals to leverage state-of-the-art technology without significant resource expenditure.

In developing our methodology, we first examined the current state of diffusion models, focusing on their architecture and the standard procedures for conditioning these models on specific tasks. We then explored the realm of zero-shot learning, highlighting its relevance and recent advancements that make it a viable solution for model customization. We have articulated our work by developing a framework that integrates diffusion models and zero-shot learning into a cohesive system capable of simultaneously handling multiple, potentially overlapping concepts.

This paper is structured as follows: Section 2 provides a detailed review of the related work, including their theoretical foundations and typical applications. Section 3 presents the Preliminary ideas and methods used in our method. Section 4 describes our proposed methodology for multi-concept customization of diffusion models, elaborating on the architectural adaptations and operational mechanisms that facilitate zero-shot learning. Section 5 presents experimental results, showcasing the effectiveness of our approach across various domains and concept complexities. Finally, Section 6 concludes with a discussion of the implications of our findings.

By bridging the gap between advanced generative models and zero-shot learning, our research not only pushes the boundaries of what is possi-

ble in artificial intelligence but also opens up new avenues for creative and practical applications of these technologies.

In summary, our contributions are threefold:

- A novel zero-shot multi-concept diffusion customization mechanism.
- An algorithm for leveraging existing text-to-image diffusion models for high-quality, high-resolution customization.
- Superior performance in identity preservation and overall image quality, as demonstrated through extensive experiments.

2. Related Work

2.1. Text-to-Image Diffusion Models

Text-to-image diffusion models such as GLIDE [15], DALL-E 2 [19], Imagen [22], and Stable Diffusion [20] have significantly advanced the capability of AI in generating customized images from textual descriptions. These models utilize sophisticated diffusion processes that iteratively refine random noise into detailed images guided by textual prompts. Trained on vast datasets of image-caption pairs, they learn a broad array of visual representations, making them highly effective for general image synthesis.

A crucial aspect of these models is their reliance on pre-trained language models like CLIP [18] to encode text prompts into latent representations that guide the diffusion process. This integration enables a nuanced understanding of textual descriptions, translating them into complex visual outputs. For instance, Stable Diffusion operates within a latent space, significantly reducing computational demands while still producing high-resolution images [20]. Stable Diffusion XL (SDXL) introduces enhancements like a larger UNet architecture and an additional text encoder, further improving textual control over generated images [17].

However, these models often struggle to consistently replicate specific details of individual identities or subjects across different contexts. Recent developments in personalized image generation address these challenges by introducing methods like DreamBooth [21] and Textual Inversion [3], which allow for more personalized image generation by fine-tuning with limited reference images.

2.2. Fine-tuning-based Approaches

Single-concept Customization. Fine-tuning methods like DreamBooth [21] enable the customization of pre-trained text-to-image models to generate precise images of a subject in varied contexts. DreamBooth fine-tunes a diffusion model on a small dataset containing pictures of the target subject associated with unique identifier tokens, maintaining high fidelity and versatility in image generation.

Block-wise LoRA [10] improves upon standard Low-Rank Adaptation (LoRA) by fine-tuning different blocks of the U-Net architecture in Stable Diffusion. Selectively skipping certain blocks allows the model to retain its pre-trained knowledge while efficiently personalizing the concept.

CAT (Contrastive Adapter Training) [16] leverages a contrastive loss between the original model and adapter without token conditioning, helping prevent overfitting and underfitting. It achieves superior results in knowledge preservation compared to other adapter-based methods like DreamBooth and LoRA.

Multi-concept Customization. Multi-concept customization via fine-tuning is explored in Cones 2 [13], which proposes an efficient way to represent and combine multiple subjects. Cones 2 fine-tune the text encoder of a pre-trained text-to-image diffusion model using a residual embedding. Layout priors serve as spatial guidance to arrange multiple subjects, reducing interference.

Concept Weaver [9] composes multiple personalized concepts at inference time by generating a template image aligned with input prompts, followed by a concept fusion strategy. The fusion process integrates the appearance of target concepts into the template while preserving structural details.

Multi-Concept Customization of Text-to-Image Diffusion [8] presents a framework that fine-tunes key and value projection matrices in cross-attention layers to embed distinct concepts. Joint training with adaptive text prompts ensures diverse concept

combinations.

2.3. Zero-shot Approaches

Single-concept Customization. Zero-shot learning approaches like InstantID [24] and PhotoMaker [11] focus on generating personalized images without specific training on the subjects. InstantID incorporates a plug-and-play module that adjusts the model's output to maintain subject identity across different styles and scenarios.

IDAdapter [2] is a tuning-free method emphasizing diversity and identity preservation in personalized image generation from a single-face image. Pick-and-Draw [14] enhances identity consistency and generative diversity for personalization methods by combining appearance-picking guidance with layout drawing guidance.

MoMA [23] simplifies personalization by utilizing a single text prompt with a Masked Object Matching Adapter to inject subject details into the diffusion model, achieving robust personalization without explicit fine-tuning.

Multi-concept Customization. Training-free multi-concept customization methods tackle the limitations of fine-tuning-based approaches, such as computational expense, overfitting, and identity blending.

FastComposer [25] enables efficient, personalized multi-subject image generation without fine-tuning. By utilizing subject embeddings extracted by an image encoder to augment text conditioning in diffusion models, cross-attention localization supervision ensures attention maps are localized to the correct regions.

OMG [7] introduces a two-stage framework for occlusion-friendly multi-concept generation in diffusion models. The first stage generates a non-customized image layout, gathering visual comprehension information such as attention maps and concept masks. In the second stage, a concept noise blending strategy integrates multiple concepts by combining noise from different single-concept models.

MultiDiffusion [1] employs a shared denoising process that links multiple diffusion processes through constraints. While effective in producing high-quality images, multiple crops must integrate several concepts.

Mix-of-Show [5] manages occlusion in multi-concept scenes through gradient fusion and regionally controllable sampling. Although it requires merging multiple models, Mix-of-Show demonstrates the potential of spatial conditioning for enhancing layout control.

3. Preliminaries

This section explores essential methodologies used for multi-concept customization of diffusion models. We discuss Stable Diffusion and its expanded version, Stable Diffusion XL (SDXL). Additionally, we examine Grounding DINO for robust object segmentation and InstantID for image generation with high-fidelity identity preservation. This section lays the groundwork for understanding the integration of complex model architectures and their efficiency in image generation.

3.1. Stable Diffusion

Stable Diffusion employs latent diffusion models (LDMs) to synthesize high-resolution images efficiently. Instead of operating directly in pixel space, which is computationally intensive, it utilizes a latent space provided by a pretrained autoencoder. This approach significantly reduces the dimensionality of the data, thereby decreasing computational demands without sacrificing image quality.

Diffusion Process: The diffusion process involves a Markov chain in the latent space, where an image is gradually denoised over T timesteps. The training objective for the denoising autoencoder is defined as:

$$L_{\text{DM}} = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right], \quad (1)$$

where $\mathbf{x}_t$ denotes the noisy version of the image at timestep t, and $\boldsymbol{\epsilon}_\theta(\cdot)$ is the predicted noise model.

Latent Space Modeling: The transition to latent space is facilitated by an encoder E that compresses an input image $\mathbf{x}$ into a latent representation $\mathbf{z} = E(\mathbf{x})$, and a decoder D that reconstructs the image from $\mathbf{z}$. The latent space supports efficient synthesis and manipulation of images at lower

computational costs. The LDMs refine the latent representations, $\mathbf{z}_t$, iteratively enhancing the image details at each step. The effectiveness of the LDM comes from the balance between compression, preserving important semantic information, and reducing unnecessary details, which are often imperceptible.

Generative Modeling in Latent Space:

$$L_{\text{LDM}} = \mathbb{E}_{E(\mathbf{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right], \tag{2}$$

where the latent space diffusion aims to model the data distribution efficiently by focusing on essential, semantically meaningful components of the data.

By reducing the operational dimensionality and harnessing the inductive biases of U-Nets specialized for image data, LDMs offer a robust framework for various image generation tasks, including class-conditional synthesis, super-resolution, and text-to-image conversion. [20]

3.2. Stable Diffusion XL (SDXL)

SDXL represents a significant enhancement of the Stable Diffusion architecture, primarily through a threefold larger UNet backbone. This increase in model scale, incorporating additional attention blocks and expanded cross-attention context, is designed to improve high-resolution image synthesis significantly.

Enhanced Architecture: SDXL extends the latent diffusion model (LDM) framework by enlarging the U-Net architecture, incorporating a diverse distribution of transformer blocks to optimize efficiency and effectiveness at different levels of the network:

$$L_{\text{SDXL}} = \mathbb{E}_{\mathbf{x},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right], \tag{3}$$

where $\mathbf{z}_t$ represents the latent space encoding of $\mathbf{x}_t$, the noisy image at timestep t.

Refinement Model: A novel diffusion-based refinement model enhances the visual fidelity of generated samples through a subsequent image-to-image technique. This refinement process implements a noising-denoising strategy on the latent output from the main SDXL model, improving detail and minimizing artifacts.

Micro-Conditioning and Multi-Aspect Training: SDXL introduces micro-conditioning on original image sizes and aspect ratios during training, allowing for more precise control over the generated image's dimensions. Multi-aspect training enables SDXL to handle various image aspect ratios effectively, improving the model's adaptability and performance across diverse applications. [17]

3.3. Grounding DINO

Grounding DINO enhances the Transformer-based DINO detector by incorporating grounded pre-training, facilitating robust open-set object detection. This model leverages human language inputs to detect arbitrary objects, making it adaptable for open-set scenarios.

Model Architecture: Grounding DINO integrates language with vision through a three-phase process: a feature enhancer for integrating visual and textual features, a language-guided query selection for determining relevant regions based on text, and a cross-modality decoder for fusing the modalities to finalize detection:

Feature Enhancer:

$$f_{\text{enh}}(\mathbf{x}, \mathbf{t}) = \text{LayerNorm}(\text{SelfAttn}(\text{CrossAttn}(\mathbf{x}, \mathbf{t}))) \tag{4}$$

where $\mathbf{x}$ are the image features, and $\mathbf{t}$ are the text features.

Detection Strategy: The model applies a detection Transformer approach, utilizing the encoded features to predict object locations and corresponding textual descriptions. This strategy efficiently handles varying object categories and descriptions provided by users.

Loss Functions: The training process involves a combination of contrastive loss for aligning the image and text features and localization loss for accurate object detection:

$$L = \lambda_1 L_{\text{contrastive}}(\mathbf{p}, \mathbf{t}) + \lambda_2 L_{\text{localization}}(\mathbf{p}, \mathbf{b}) \tag{5}$$

where $\mathbf{p}$ are the predicted probabilities, $\mathbf{t}$ the target text tokens, and $\mathbf{b}$ the bounding box coordinates. The weights λ_1 and λ_2 balance the importance of language alignment versus spatial accuracy. [12]

3.4. InstantID

InstantID presents an innovative approach for zero-shot identity-preserving image generation. It utilizes a single facial image to achieve high fidelity in personalized image synthesis. Unlike other methods requiring extensive fine-tuning or multiple reference images, InstantID offers a plug-and-play module compatible with existing pre-trained diffusion models like SD1.5 and SDXL.

Methodology: The core of InstantID is a face encoder, termed IdentityNet, which imposes strong semantic and weak spatial conditions. This design ensures high-quality face fidelity while allowing for style variations. The process integrates facial images, landmark images, and textual prompts to guide the image synthesis effectively:

$$L = \mathbb{E}_{z_t, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, C)\|^2 \right], \quad (6)$$

where z_t is the latent representation at timestep t, C is the conditioning derived from textual prompts, and ϵ_θ is the noise predicted by the model.

Implementation: InstantID modifies the diffusion process by incorporating an identity embedding derived from the reference image to retain crucial identity features during generation. This embedding interacts with the model's latent space to preserve identity characteristics effectively:

$$z_{new} = \text{Attention}(Q, K_{id}, V_{id}), \quad (7)$$

where Q represents the queries from the diffusion model's latent space, and K_{id}, V_{id} are the keys and values generated from the identity embedding.

Advantages: InstantID requires no fine-tuning at inference time, significantly reducing computational overhead and making it feasible for real-time applications. It seamlessly integrates with various pre-trained models, maintaining adaptability and achieving state-of-the-art results in identity preservation with only a single reference image.

By leveraging InstantID in the final customization step, we harness its advanced capabilities to ensure that each concept matches its intended design and integrates harmoniously with adjacent concepts. This approach significantly enhances the practical applicability of our multi-concept customization framework, enabling precise control over each concept's aesthetics and identity characteristics within the composite image [24].

4. Method

This section presents our approach, *Multi-concept Customization*, designed to generate personalized images containing multiple concepts. This method leverages state-of-the-art diffusion models, object segmentation, and identity preservation techniques to incorporate several concepts into a single image seamlessly. Our algorithm is divided into two stages:

1. Image Generation and Object Segmentation
2. Personalization and Latent Blending

The detailed steps of the algorithm are outlined in Algorithm 1.

Stage 1: Image Generation and Object Segmentation

In the first step, we generate an initial image using the SDXL model [17]. Starting with a noisy latent variable Z'_t sampled from a Gaussian distribution $\mathbb{N}(\mu, \sigma)$, the denoising process proceeds iteratively from $t-1$ down to 0, guided by an initial prompt P_{init} containing a textual description of the scene. The output is an initial image Z'_0.

The second step involves object segmentation using the GroundingDINO [12] method to identify and separate the desired concept masks $C_1, C_2, \ldots, C_k$ from the background. Each mask M_i represents a binary mask for concept C_i. The background mask M_{BG} is calculated as the complement of the sum of all concept masks.

Stage 2: Personalization and Latent Blending

In the second stage, personalization is achieved using InstantID [24]. The latent variable Z_t is initialized as the noisy variable Z'_t from the first stage. For each concept C_i, a personalized latent variable Z^i_j is generated using the InstantID model, guided by the concept identity.

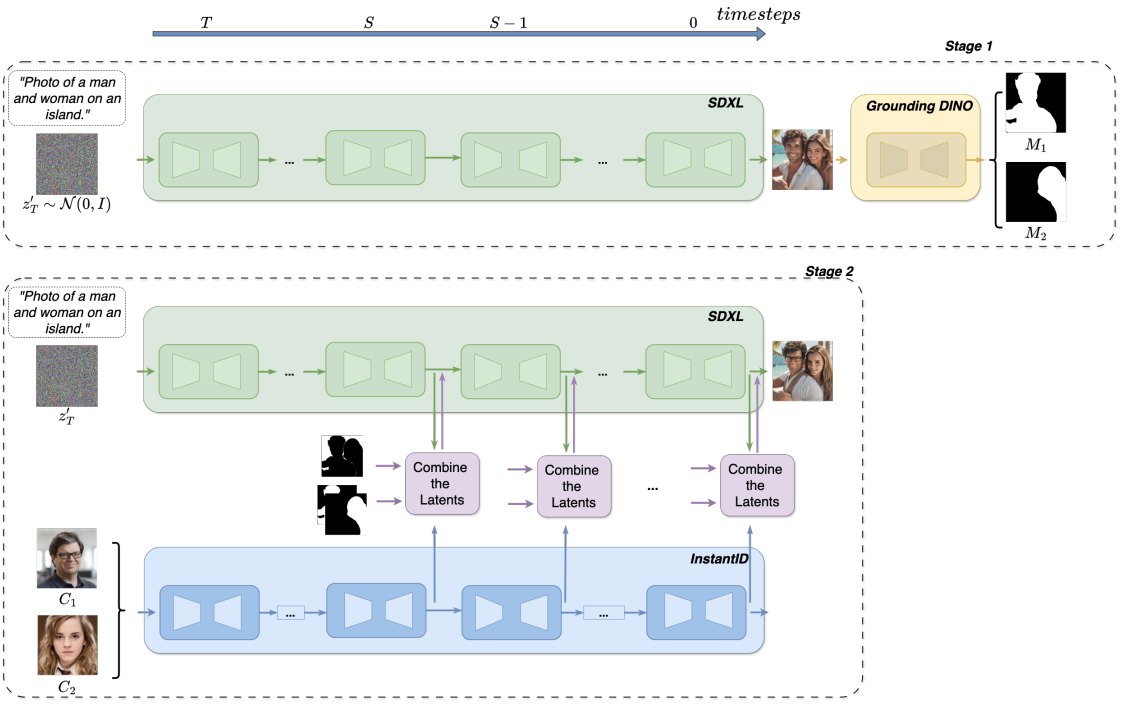

Figure 2. The overall pipeline of our proposed method

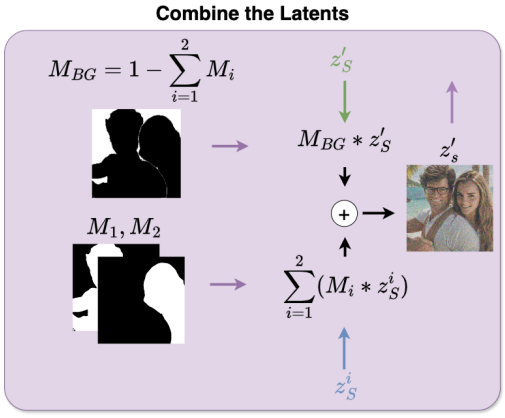

Figure 3. Integrating Latent Outputs from InstantID, SDXL, and Grounding Dino Masks

During latent blending, each personalized latent variable Z_j^i is combined with the background latent variable Z_j using the function f to generate the final image Z_0. The blending function f combines the binary masks M_i of each concept with the background mask M_{BG} to produce a consistent composite image:

$$Z_j = f(M_{BG} Z_j + \sum_{i=1}^{k} M_i Z_j^i). \tag{8}$$

5. Experiments

We conducted our experiments on an RTX A5000 instance to evaluate the performance of our proposed method compared to the baseline model, FastComposer[25]. The experiments aimed to assess the quality of generated images in terms of identity preservation, image quality, and photore-

Algorithm 1 Multi-concept Customization

Input: $P_{init}, C_1, C_2, ...C_k, Z'_t \sim \mathbb{N}(\mu, \sigma)$
Output: Final customized image Z_0
procedure MULTI-CONCEPT CUSTOMIZATION(
)
Stage 1: Image Generation and Object Segmentation
for $j \in \{t-1, ...0\}$ **do**
$Z'_j = SDXL(Z'_{j+1}, P_{init})$
end for

$\{M_1, M_2, ...M_k\} = GroundingDino(Z'_0)$
where $M_i \in [0,1]^{h*w}$ and
$M_{BG} = 1 - \sum_{i=1}^{k} M_i$

Stage 2: Personalization and Latent Blending
$Z_t = Z'_t$
for $j \in \{t-1, ...0\}$ **do**
for $i \in [1:k]$ **do**
if $j < S$ **then**
$Z'_j = SDXL(Z'_{j+1}, P_{init})$
$Z^i_j = InstantID(Z^i_{j+1}, C_i)$
else
$Z'_j = SDXL(Z_{j+1}, P_{init})$
$Z^i_j = InstantID(Z_{j+1}, C_i)$
end if
end for
$Z_j = f(M_{BG} Z_j + \sum_{i=1}^{k} M_i Z^i_j)$
end for
end procedure
Return: Final image Z_0

alism.

We generated several images of the same pairs of identities using identical textual prompts to ensure a fair comparison between the two models. The baseline model, FastComposer, served as a reference point for evaluating the effectiveness of our approach.

Our results demonstrate that our proposed method outperforms FastComposer across multiple metrics. Specifically, our approach excelled in preserving identity, producing high-quality images, and achieving photorealistic generation. Visual inspection of the generated images indicates the superiority of our method over the baseline.

Figure 4 illustrates the qualitative comparison between images generated by our method and FastComposer. It is evident from the visual comparison that our approach consistently produces images with superior identity preservation, image quality, and photorealism.

Figure 4. Qualitative comparison between images generated by our method and FastComposer. Our method (right) preserves identity better, enhances image quality, and achieves greater photorealism compared to FastComposer (left).

These findings underscore the effectiveness of our proposed method in generating high-quality, photorealistic images while preserving the identity of the subjects. Our approach holds promise for various applications, including personalized image generation and content creation.

Further quantitative evaluations and user studies are warranted to validate the robustness and generalizability of our approach across diverse datasets and scenarios.

6. Conclusions

This paper introduced a novel approach to zero-shot customization of diffusion models, enabling high-quality image generation based on multiple

identities or concepts without requiring model retraining. By integrating recent advances in zero-shot learning, our methodology allows diffusion models to interpret and synthesize complex, multi-concept images solely based on textual descriptions. This framework circumvents the need for extensive data collection and training, which typically hinder personalization.

Our experimental results demonstrated the efficacy of this approach across various domains, highlighting the significant potential to reduce barriers to entry for personalized image generation. This work bridges the gap between advanced generative models and zero-shot learning, opening new avenues for creative and practical applications in digital content creation, personalized advertising, and virtual design. Future research will focus on refining zero-shot learning techniques, expanding generalization across domains, and exploring user-friendly interfaces to make this technology more accessible.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023. 4

[2] Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models, 2024. 4

[3] 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, 2022. 3

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[5] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models, 2023. 4

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[7] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models, 2024. 4

[8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 3

[9] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models, 2024. 3

[10] Likun Li, Haoqi Zeng, Changpeng Yang, Haozhe Jia, and Di Xu. Block-wise lora: Revisiting fine-grained lora for effective personalization and stylization in text-to-image generation, 2024. 3

[11] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 4

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023. 5, 6

[13] 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, 2023. 3

[14] Henglei Lv, Jiayu Xiao, Liang Li, and Qingming Huang. Pick-and-draw: Training-free semantic guidance for text-to-image personalization, 2024. 4

[15] 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, 2022. 3

[16] Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. Cat: Contrastive adapter training for personalized image generation, 2024. 3

[17] 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, 2023. 3, 5, 6

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.

Learning transferable visual models from natural language supervision, 2021. 3

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3, 5

[21] 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, 2023. 3

[22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[23] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation, 2024. 4

[24] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. 4, 6

[25] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention, 2023. 4, 7