MoA: Mixture-of-Attention for Subject-Context Disentanglement in Personalized Image Generation

KUAN-CHIEH (JACKSON) WANG, Snap Inc., USA
DANIIL OSTASHEV, Snap Inc., UK
YUWEI FANG, Snap Inc., USA
SERGEY TULYAKOV, Snap Inc., USA
KFIR ABERMAN, Snap Inc., USA

Fig. 1. **Mixture-of-Attention (MoA)** architecture enables multi-subject personalized generation with **subject-context disentanglement**. Given a multi-modal prompt that includes text and input images of human subjects, our model can generate the subjects in a fixed context and composition, without any predefined layout. MoA minimizes the intervention of the personalized part in the generation process, enabling the decoupling between the model's pre-existing capability and the personalized portion of the generation.

We introduce a new architecture for personalization of text-to-image diffusion models, coined Mixture-of-Attention (MoA). Inspired by the Mixture-of-Experts mechanism utilized in large language models (LLMs), MoA distributes the generation workload between two attention pathways: a personalized branch and a non-personalized prior branch. MoA is designed to retain the original model's prior by fixing its attention layers in the prior branch, while minimally intervening in the generation process with the personalized branch that learns to embed subjects in the layout and context generated by the prior branch. A novel routing mechanism manages the distribution of pixels in each layer across these branches to optimize the blend of personalized and generic content creation. Once trained, MoA facilitates the creation of high-quality, personalized images featuring multiple subjects with compositions and interactions as diverse as those generated by the original model. Crucially, MoA enhances the distinction between the model's pre-existing capability and the newly augmented personalized intervention, thereby offering a more disentangled subject-context control that was previously unattainable. Project page: https://snap-research.github.io/mixture-of-attention.

Additional Key Words and Phrases: Personalization, Text-to-image Generation, Diffusion Models

Authors' addresses: Kuan-Chieh (Jackson) Wang, Snap Inc., USA, jwang23@snapchat.com; Daniil Ostashev, Snap Inc., UK; Yuwei Fang, Snap Inc., USA; Sergey Tulyakov, Snap Inc., USA; Kfir Aberman, Snap Inc., USA, kaberman@snapchat.com.

1 INTRODUCTION

Recent progress in AI-generated visual content has been nothing short of revolutionary, fundamentally altering the landscape of digital media creation. Foundation models have democratized the creation of high-quality visual content, allowing even novice users to generate impressive images from simple text prompts [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022]. Among the myriad avenues of research within this field, personalization stands out as a crucial frontier. It aims at tailoring the output of a generative model to include user-specific subjects with high fidelity, thereby producing outputs that resonate more closely with individual assets or preferences [Gal et al. 2022; Ruiz et al. 2023a]. While being able to say "Create a photo of people scuba diving!" is fun, the experience becomes personal and fosters a stronger emotional connection when one can say "Create a photo of *me and my friend* scuba diving!" (see Fig. 1).

Despite the remarkable generative capabilities of these models, current personalization techniques often falter in preserving the richness of the original model. Herein, we refer to the model before personalization as the prior model. In finetuning-based personalization techniques, due to the modifications of the weights, the model tends to overfit to certain attributes in the distribution of the input images (e.g., posture and composition of subjects) or struggles to adhere adequately to the input prompt. This issue is exacerbated when it comes to multiple subjects; the personalized model struggles to generate compositions and interactions between the subjects that otherwise appear within the distribution of the non-personalized model. Even approaches that were optimized for multi-subject generation modify the original model's weights, resulting in compositions that lack diversity and naturalness [Po et al. 2023; Xiao et al. 2023]. Hence, it is advisable to purse a personalization method that is *prior preserving*. Functionally, we refer to a method as *prior preserving* if the model retains its responsiveness to changes in the text-prompt and random seed like it does in the prior model.

A good personalization method should address the aforementioned issues. In addition, it should allow the creation process to be *spontaneous*. Namely, iterating over ideas should be fast and easy. Specifically, our requirements are summarized by the following:

(1) *Prior preserving*. The personalized model should retain the ability to compose different elements, and be faithful to the *interaction* described in the text prompt like in the prior model. Also, the distribution of images should be as diverse as in the prior model.
(2) *Fast generation*. The generation should be fast to allow the users to quickly iterate over many ideas. Technically, the personalized generation process should be inference-based and should not require optimization when given a new subject.
(3) *Layout-free*. Users are not required to provide additional layout controls (e.g. segmentation mask, bounding box, or human pose) to generate images. Requiring additional layout control could hinder the creative process, and restrict the diversity of the distribution.

To achieve these goals, we introduce Mixture-of-Attention (MoA) (see Fig. 2). Inspired by the Mixture-of-Expert (MoE) layer [Jacobs et al. 1991] and its recent success in scaling language models [Roller et al. 2021], MoA extends the vanilla attention mechanism into multiple attention blocks (i.e. experts), and has a router network that softly combines the different experts. In our case, MoA distributes the generation between personalized and non-personalized attention pathways. It is designed to retain the original model's prior by fixing its attention layers in the prior (non-personalized) branch, while minimally intervening in the generation process with the personalized branch. The latter learns to embed subjects that are depicted in input images, via encoded visual tokens that is injected to the layout and context generated by the prior branch. This mechanism is enabled thanks to the router that blends the outputs of the personalized branch only at the subject pixels (i.e. foreground), by learning soft segmentation maps that dictate the distribution of the workload between the two branches. This mechanism frees us from the trade-off between identity preservation and prompt consistency.

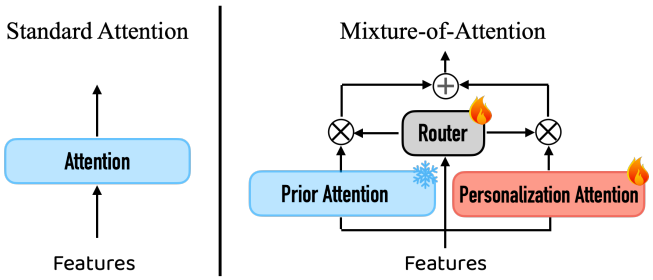

Fig. 2. **Mixture-of-Attention.** Unlike the standard attention mechanism (left), MoA is a dual attention pathways that contains a trainable personalized attention branch and a non-personalized fixed attention branch that is copied from the original model (prior attention). In addition, a routing mechanism manages the distribution of pixels in each layer across these branches to optimize the blend of personalized and generic content creation.

Since MoA distinguishes between the model's inherent capabilities and the personalized interventions, it unlocks new levels of disentangled control in personalized generative models (as demonstrated in Fig. 1). This enables us to create various applications with MoA such as subject swap, subject morphing, style transfer, etc,that was previously challenging to attain. In addition, due to the existence of the fixed prior branch, MoA is compatible with many other diffusion-based image generation and editing techniques, such as ControlNet [Zhang et al. 2023b] or inversion techniques that unlocks a novel approach to easily replace subjects in a real images (see Sec. 5).

2 RELATED WORKS

2.1 Personalized Generation

Given the rapid progress in foundation text-conditioned image synthesis with diffusion models [Dhariwal and Nichol 2021; Ho 2022; Ho et al. 2020; Nichol and Dhariwal 2021; Pandey et al. 2022; Rombach et al. 2021; Song et al. 2020a], *personalized* generation focuses on adapting and contextualizing the generation to a set of desired subject using limited input images, while retaining the powerful generative capabilities of the foundation model. Textual Inversion (TI) [Gal et al. 2022] addresses the personalization challenge by utilizing a set of images depicting the same subject to learn a special

text token that encodes the subject. Yet, using only the input text embeddings is limited in expressivity. Subsequent research, such as $\mathcal{P}+$ [Voynov et al. 2023] and NeTI [Alaluf et al. 2023], enhance TI with a more expressive token representation, thereby refining the alignment and fidelity of generated subjects. DreamBooth (DB) [Ruiz et al. 2023a] can achieve much higher subject fidelity by finetuning the model parameters. E4T [Gal et al. 2023] introduced a pretrained image encoder that jump starts the optimization with image features extracted from the subject image, and is able to substantially reduce the number of optimization steps. Other extensions include multi-subject generation [Kumari et al. 2023a], generic objects [Li et al. 2024], human-object composition [Liu et al. 2023a,b], subject editing [Tewel et al. 2023], improving efficiency [dbl 2022; Han et al. 2023; Hu et al. 2022], and using hypernetworks [Arar et al. 2023; Ruiz et al. 2023b]. These approaches fall under the *optimization-based* category where given a new subject, some parameters of the model are to be optimized. Because of the optimization which modifies the original parameters of the model, these methods are inevitably slow and prone to breaking prior preservation. In contrast, MoA falls in the *optimization-free* category. These approaches do not require optimization when given a new subject. They augment the foundation T2I model with an image encoder and finetune the augmented model to receive image inputs. Relevant methods in this category include IP-Adapter [Ye et al. 2023] and InstantID [Wang et al. 2024]. A critical difference is, in MoA, the image features are combined with a text token (e.g. 'man') and processed by the cross attention layer in the way that the cross attention layer was trained. In IP-Adapter and InstantID, the image features are combined with the output of attention layers and do not have binding to a specific text token. This design makes it hard to leverage the native text understanding and text-to-image composition of the pretrained T2I model. It also makes combining multiple image inputs challenging. For similar reasons, other optimization-free approaches that focus on the single subject setting include ELITE [Wei et al. 2023], InstantBooth [Shi et al. 2023], PhotoMaker [Li et al. 2023a], LCM-Lookahead [Gal et al. 2024]. A remedy is to introduce layout controls and mask the different image inputs in the latent space, but this led to rigid outputs and a brittle solution. In stark contrast, since MoA injects the image inputs in the text space, injecting multiple input images is trivial. In addition, by explicitly having a prior branch, MoA preserves the powerful text-to-image capabilities of the prior foundation model.

2.2 Multi-subject Generation

Extending personalized generation to the multi-subject setting is not trivial. Naive integration of multiple subjects often leads to issues like a missing subject, poor layout, or subject interference (a.k.a. identity leak) where the output subjects looks like a blended version of the inputs. Custom Diffusion [Kumari et al. 2023b] and Modular Customization [Po et al. 2023] proposed ways to combine multiple DB parameters without the subjects interfering with each other using constrained optimization techniques. Mix-of-show [Gu et al. 2024] proposed regionally controllable sampling where user specified bounding boxes are used to guide the generation process. InstantID [Wang et al. 2024] can also achieve multi-subject generation using bounding boxes as additional user control. The idea of using bounding box or segmentation mask to control the generation process has be used in other settings [Avrahami et al. 2023a; Bar-Tal et al. 2023; Hertz et al. 2023]. In addition to burdening the users to provide layout, methods that require region control naturally results in images that appear more rigid. The subjects are separated by their respective bounding boxes, and lack interaction. In contrast, while MoA can work with additional layout condition, it does not require such inputs just like the prior T2I model does not require layout guidance. Fastcomposer [Xiao et al. 2023] is the closest method as it also injects the subject image as text tokens and handles multiple subjects without layout control. Except, generated images from Fastcomposer have a characteristic layout of the subjects and lack subject-context interaction, which indicates the lack of prior preservation (see Fig. 3). Since Fastcomposer finetunes the base model's parameters, it inevitably deviates away from the prior model and has to trade-off between faithful subject injection and prior preservation. MoA is free from this trade-off thanks to the dual pathways and a learned router to combine both the frozen prior expert and learned personalization expert.

Fig. 3. **Comparing image variations.** In contrast to Fastcomposer [Xiao et al. 2023], our method (MoA) is able to generate images with diverse compositions, and foster interaction of the subject with what is described in the text prompt.

3 METHOD

In this section, we introduce the *Mixture-of-Attention* (MoA) layer (see Fig. 2) and explain how it can be integrated into text-to-image (T2I) diffusion models for subject-driven generation. In its general form, a MoA layer has multiple attention layers, each with its own projection parameters and a router network that softly combines their outputs. In this work, we use a specific instantiation suitable for personalization, which contains two branches: a fixed "prior" branch that is copied from the original network, a trainable "personalized" branch that is finetuned to handle image inputs, and a router trained to utilize the two experts for their distinct functionalities. MoA layer is used in-place of all attention layers in a pretrained diffusion U-Net (see Fig. 4). This architecture enables us to augment the T2I model with the ability to perform subject-driven generation with disentangled subject-context control, thereby preserving the diverse image distribution inherent in the prior model.

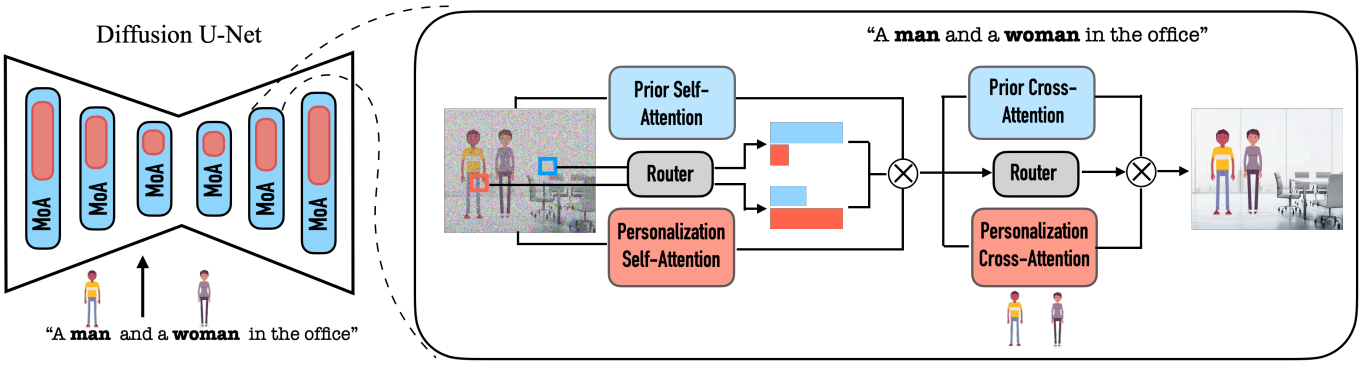

Fig. 4. **Text-to-Image Diffusion Models with MoA.** Our architecture expands the original diffusion U-Net by replacing each attention block (self and cross) with MoA. In each inference step, a MoA block receives the input image features and passes them to the router, which decides how to balance the weights between the output of the personalized attention and the output of the original attention block. Note that the images of the subjects are injected only through the personalized attention branch; hence, during training, where the router is encouraged to prioritize the prior branch, the result is that only the minimal necessary information required for generating the subjects will be transferred to the personalized attention.

3.1 Background

Attention Layer. An attention layer first computes the attention map using query, $\mathbf{Q} \in \mathbb{R}^{l_q \times d}$, and key, $\mathbf{K} \in \mathbb{R}^{l_k \times d}$ where d is the hidden dimension and l_q, l_k are the numbers of the query and key tokens, respectively. The attention map is then applied to the value, $\mathbf{V} \in \mathbb{R}^{l \times d}$. The attention operation is described as follows:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}, \tag{1}$$

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{C}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{C}\mathbf{W}_v, \tag{2}$$

where $\mathbf{W}_q \in \mathbb{R}^{d_z \times d}, \mathbf{W}_k \in \mathbb{R}^{d_c \times d}, \mathbf{W}_v \in \mathbb{R}^{d_c \times d}$ are the projection matrices of the attention operation that map the different inputs to the same hidden dimension, d. $\mathbf{Z}$ is the hidden state and $\mathbf{C}$ is the condition. In self attention layers, the condition is the hidden state, $\mathbf{C} = \mathbf{Z}$. In cross attention layers of T2I models, the condition is the text conditioning $\mathbf{C} = \mathbf{T}$.

Diffusion U-Net. At the core of a T2I diffusion model lies a U-Net which consists of a sequence of transformer blocks, each with self attention and cross attention layers. Each attention layer has its own parameters. In addition, the U-Net is conditioned on the diffusion timestep. Putting it together, the input to the specific attention layer is dependent on the U-Net layer l and the diffusion timestep t:

$$\mathbf{Q}^{t,l} = \mathbf{Z}^{t,l}\mathbf{W}_q^l, \quad \mathbf{K}^{t,l} = \mathbf{C}^{t,l}\mathbf{W}_k^l, \quad \mathbf{V}^{t,l} = \mathbf{C}^{t,l}\mathbf{W}_v^l, \tag{3}$$

where each attention layer has its own projection matrices indexed by l. The hidden state $\mathbf{Z}$ is naturally a function of both the diffusion timestep and layer. In a cross-attention layer, the text conditioning $\mathbf{C} = \mathbf{T}$ is not a function of t and l by default, but recent advances in textual inversion like NeTI [Alaluf et al. 2023] show that this spacetime conditioning can improve personalization.

Mixture-of-Expert (MoE) Layer. A MoE layer [Fedus et al. 2022; Shazeer et al. 2017] consists of N expert networks and a router network that softly combines the output of the different experts:

$$\mathbf{Z} = \sum_{n=1}^{N} \mathbf{R}_n \odot \text{Expert}_n(\mathbf{Z}), \tag{4}$$

$$\mathbf{R} = \text{Router}(\mathbf{Z}) = \text{Softmax}(f(\mathbf{Z})), \tag{5}$$

where $\odot$ denotes the Hadamard product, and $\mathbf{R} \in \mathbb{R}^{l \times N}$. The router is a learned network that outputs a soft attention map over the input dimensions (i.e., latent pixels). Functionally, the router maps each latent pixel to N logits that are then passed through a softmax. The mapping function f can be a simple linear layer or an MLP.

3.2 Mixture-of-Attention Layer

Under the general framework of MoE layers, our proposed MoA layer has two distinct features. First, each of our 'experts' is an attention layer, i.e. the attention mechanism and the learnable project layers described in Eqn. (2). Second, we have only two experts, a frozen prior expert and a learnable personalization expert. Together, our MoA layer has the following form:

$$\mathbf{Z}^{t,l} = \sum_{n=1}^{2} \mathbf{R}_n^{t,l} \odot \text{Attention}(\mathbf{Q}_n^{t,l}, \mathbf{K}_n^{t,l}, \mathbf{V}_n^{t,l}), \tag{6}$$

$$\mathbf{R}^{t,l} = \text{Router}^l(\mathbf{Z}^{t,l}). \tag{7}$$

Note, each MoA layer has its own router network, hence it is indexed by the layer l and each attention expert has its own projection layers, hence they are indexed by n. We initialize both of the experts in a MoA layer using the attention layer from a pretrained model. The prior expert is kept frozen while the personalization expert is finetuned.

3.2.1 Cross-Attention Experts. While in the MoA self attention layers, the two experts receive the same inputs, in the MoA cross attention layers, the two experts take different inputs. To fully preserve the prior, the prior expert receives the standard text-only condition. To handle image inputs, the personalization expert receives a multi-modal prompt embedding described in the following section.

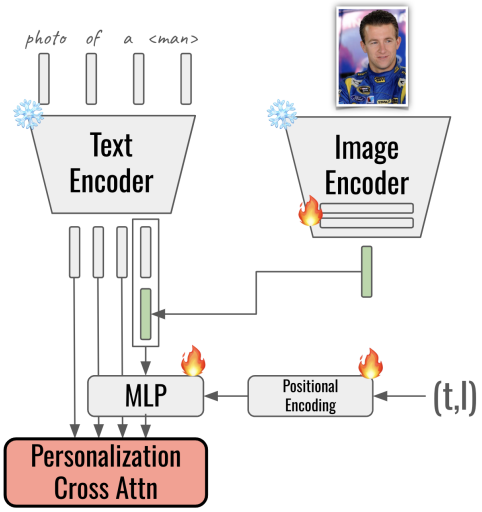

Fig. 5. **Multimodal prompts.** Our architecture enables us to inject images as visual tokens that are part of the text prompt, where each visual token is attached to a text encoding of a specific token.

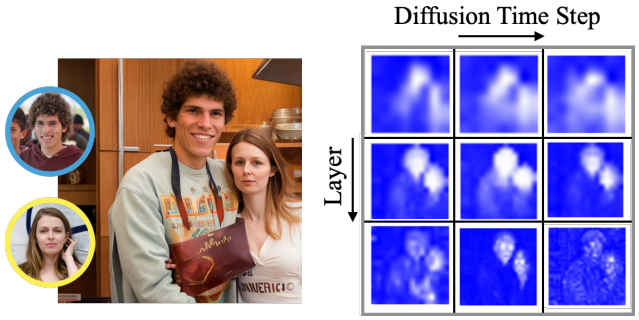

Fig. 6. **Router Visualization.** Our router learns to generate soft segmentation maps per time step in the diffusion process and per layer. Distinct parts of the subjects, in different resolutions, are highlighted across various time steps and layers.

Multimodal Prompts. Given a subject image, I, it is injected into the text prompt as shown in Fig. 5. First, image feature, $\mathbf{f}$, is extracted using a pretrained image encoder (e.g. CLIP image encoder), $\mathbf{f} = E_{\text{image}}(I)$. The image feature is concatenated with the text embedding at the corresponding token, $\mathbf{t}$, say the embedding of the 'man' token. We refer to the concatenated embedding as the multi-modal embedding, $\mathbf{m} = \text{Concat}(\mathbf{f}, \mathbf{t})$. We further condition the multi-modal embedding on two information: the diffusion timestep, t, and U-Net layer l through a learned positional encoding ($\text{PE} : \mathbb{R}^2 \mapsto \mathbb{R}^{2d_t}$) as follows

$$\bar{\mathbf{m}}_{t,l} = \text{LayerNorm}(\mathbf{m}) + \text{LayerNorm}(\text{PE}(t, l)). \tag{8}$$

From previous work on optimization-based personalization, the diffusion time and space conditioning can improve identity preservation [Alaluf et al. 2023; Voynov et al. 2023; Zhang et al. 2023a]. Finally, the embedding is passed to a learnable MLP.

3.3 Training

3.3.1 Training the Router. The router is trained with an objective that encourages the background pixels (i.e. not belonging to the image subject) to utilize the "prior" branch. The foreground pixels are not explicitly optimized w.r.t. any target. The loss is computed after accumulating the router predictions at all layers:

$$\mathcal{L}_{\text{router}} = \|(1 - \mathbf{M}) \odot (1 - \mathbf{R})\|_2^2, \tag{9}$$

$$\mathbf{R} = \frac{1}{|\mathbb{L}|} \sum_{l \in \mathbb{L}} \mathbf{R}_0^l, \tag{10}$$

where $\mathbf{R}_0^l$ is the router weight for the prior branch at U-Net layer l, and $\mathbf{M}$ is the foreground mask of the subject. $\mathbb{L}$ is the set of layers we penalize in the U-Net, and $|\mathbb{L}|$ is the number of layers. In practice, we exclude the first and last block of the U-Net (i.e. first two and last three attention layers). Empirically, they encode low-level features of the image and are less relevant to the notion of subject versus context. Across different U-Net layers and diffusion timesteps, the personalization expert focuses on regions associated with the subject, and the prior branch accounts for most of the background while still having a base level of contribution to the subjects. The routers also behave differently at different layers and different timesteps. For example, the personalization expert at one layer/timestep might attends to the face while at another layer/timestep attneds to the body, as visualized in Fig. 6. .

3.3.2 Overall Training Scheme. Typically, training or finetuning diffusion models is done by using the full (latent) image reconstruction loss, $\mathcal{L}_{\text{full}}(\mathbf{Z}, \hat{\mathbf{Z}}) = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2$. In contrast, recent personalization methods use the segmentation mask for masked (foreground only) reconstruction loss to prevent the model from confounding with the background, i.e. $\mathcal{L}_{\text{masked}}(\mathbf{Z}, \hat{\mathbf{Z}}) = \|\mathbf{M} \odot (\mathbf{Z} - \hat{\mathbf{Z}})\|_2^2$. In previous training-based methods like Fastcomposer [Xiao et al. 2023], they need to balance between preserving the prior by using the full image loss with focusing on the subject by using the masked loss, i.e. $\mathcal{L} = p\mathcal{L}_{\text{full}} + (1 - p)\mathcal{L}_{\text{masked}}$ where p is the probability of using the full loss and was set to 0.5. Because of our MoA layer, we do not need to trade-off between prior and personalization. Hence, we can use the best practice for the personalization, which is only optimizing the foreground reconstruction loss. Our frozen prior branch naturally plays the role of preserving the prior. Our finetuning consists of the masked reconstruction loss, router loss, and cross attention mask loss:

$$\mathcal{L} = \mathcal{L}_{\text{masked}} + \lambda_r \mathcal{L}_{\text{router}} + \lambda_o \mathcal{L}_{\text{object}}, \tag{11}$$

where the $\mathcal{L}_{\text{object}}$ is the balanced L1 loss proposed in Fastcomposer [Xiao et al. 2023]. We apply it to our personalization experts:

$$\mathcal{L}_{\text{object}} = \frac{1}{|\mathbb{L}||\mathbb{S}|} \sum_{l \in \mathbb{L}} \sum_{s \in \mathbb{S}} \text{mean}((1 - \mathbf{M}_s) \odot (1 - \mathbf{A}_s^l)) - \text{mean}(\mathbf{M}_s \odot \mathbf{A}_s^l), \tag{12}$$

Fig. 7. **Disentangled subject-context control with a single subject.** The top row is generetad using only the prior branch. Each column is a different random seed. MoA allows for disentangled subject-context control. Namely, injecting different subjects lead to only localized changes to the pixels pertaining to the foreground human.

where $\mathbb{S}$ denotes the set of subjects in an image, $\mathbf{M}_s$ the segmentation mask of subject s, and $\mathbf{A}_s^l$ the cross attention map at the token where subject s is injected.

4 EXPERIMENTS

In this section, we show results to highlight MoA's capability to perform disentangled subject-context control, handle occlusions through both qualitative and quantitative evaluations. We also show analysis of the router behavior as a explanation for the new capability.

4.1 Experimental Setup

Datasets. For training and quantitative evaluation, two datasets were used. For training, we used the FFHQ [Karras et al. 2019] dataset preprocessed by [Xiao et al. 2023], which also contains captions generated by BLIP-2 [Li et al. 2023b] and segmentation masks generated by MaskedFormer [Cheng et al. 2022]. For the test set of the quantitative evaluation, we used test subjects from the FFHQ dataset for qualitative results, and 15 subjects from the CelebA dataset [Liu et al. 2015] for both of the qualitative and quantitative evaluation following previous works.

Model details. For the pretrained T2I models, we use StableDiffusion v1.5 [Rombach et al. 2022]. For some qualitative results, we use the community finetuned checkpoints like `AbsoluteReality_v1.8.1`. For the image encoder, we follow previous studies and use OpenAI's `clip-vit-large-patch14` vision model. We train our models on 4 NVIDIA H100 GPUs, with a constant learning rate of 1e-5 and a batch size of 128. Following standard training for classifier-free guidance [Ho and Salimans 2022], we train the model without any conditions 10% of the time. During inference, we use the UniPC sampler [Zhao et al. 2023].

4.2 Results

All the results presented in this section are performed on the held-out test subjects of the FFHQ dataset.

Disentangled subject-context control. The first major result is the disentangled subject-context control that is enabled by our MoA architecture. We show unforeseen disentanglement between the background control using the random seed and input image subjects all in the single forward pass. In Fig. 7, we show the results of someone drinking boba in a night market and holding a bouquet of roses. Notice that as we change the input subjects while holding the seed constant, we are able to perform localized subject change without affecting the background. Moreover, in the top row, we show samples from using only the prior branch. The content is preserved after we inject different subjects. This allows for a new application where the users can quickly generate images and select them by content, and then inject the personalized information. They can then also easily swap the input subjects.

Image quality, variation, and consistency. Another unique aspect about MoA is the "localized injection in the prompt space". This feature allows for a surprising ability to handle occlusion. In Fig. 7, we can see the boba and roses occluding part of the face and body. Despite the occlusion, the face details are preserved, and the body is also consistent with the face. For example, the men holding the boba have a consistent skin tone and texture in the arms as suggested by

Fig. 8. **Images with close interactions of two subjects.** MoA can generate images with different subject layouts and different interaction types among the subjects.

their face. We show additional results of handling occlusion in Fig. 17 where we generate portraits photos with different costumes. In the generated images, a large portion of the face can be occluded. Our method is able to handle such cases while preserving the identity.

Multi-subject composition. This ability to generate full-body consistent subjects and handle occlusion unlocks the ability generate multi-subject images with close interaction between subjects. In Fig. 8, we show generated photos of couples with various prompts. Even in cases of dancing, where the subjects have substantial occlusion with each other, the generation is still globally consistent (i.e. the skin tone of the men's arms match their faces). Furthermore, in Fig. 9, we show that the *disentangled subject-context control* capability still holds in the multi-subject case. This allows the users to swap one or both of the individuals in the generated images while preserving the interaction, background, and style. Lastly, when comparing our results with Fastcomposer in the multi-subject setting, MoA is able to better preserve the context and produce more coherent images (see Fig. 10). While Fastcomposer is able to inject multiple subjects and modify the background, the subjects are not well integrated with the context. This is evident in cases where the prompt describes an activity, such as "cooking".

Analysis. For analysis, we visualize the router prediction in Fig. 6. We visualize the behavior using the same random seed, but different input subjects. The behavior of the router is consistent across the two subject pairs, and route most of the background pixel to the prior branch. We believe this explains why MoA allows for disentangled subject-context control. See the supplementary material for more visualization with different random seeds where the router behavior changes, hence leading to a different layout and background content.

5 APPLICATIONS

In this section, we demonstrate a number of applications enabled by the disentangled control of MoA and its compatibility with existing image generation/editing techniques developed for diffusion-based models. In particular, the simplicity of the design in MoA makes it compatible with using ControlNet (Sec. 5.1). MoA can create new characters by interpolating between the image features of different subjects, which we refer to as subject morphing (Sec. 5.2). Beyond generation, MoA is also compatible with real-image editing techniques based on diffusion inversion [Dhariwal and Nichol 2021; Mokady et al. 2023; Song et al. 2020b] (Sec. 5.3). We include three more applications (style swap with LoRA, time lapse, and consistent character storytelling) in Appendix E.

5.1 Controllable Personalized Generation

A key feature of MoA is its simplicity and minimal modification to the base diffusion model. This makes it naturally compatible with existing extensions like ControlNet [Zhang et al. 2023b]. Since MoA operates only within the attention mechanism, the semantics of the latent are preserved in-between U-Net blocks, where ControlNet conditioning is applied. This makes it possible to use ControlNet in exactly the same way as it would be used with the prior branch of the model. In Fig. 11, we show examples of adding pose control to MoA using ControlNet. Given the same text prompt and random seed, which specifies the context, the user can use ControlNet to change the pose of the subjects. Even in such a use case, MoA retains

Fig. 9. **Disentangled subject-context control with a multiple subjects.** MoA retains the disentangled subject-context control in the multi-subject scenario. One or both of the subjects can be swapped without substantial effect on the context.

Fig. 10. **Comparison with Fastcomposer in the multi-subject setting.**

the disentangled subject-context control and is able to swap the subjects.

5.2 Subject Morphing

By interpolating the image feature outputted by the learned image encoder in MoA, one can interpolate between two different subjects. Since MoA encodes more than the face of the subject and has a holistic understanding of the body shape and skin tone, we are able to interpolate between two very different subjects. In Fig. 12, we interpolate between the Yokozuna, who has a large body and darker skin tone, and a generated male character, who has a smaller body and a pale skin tone. The features of the interpolated subjects are preserved with different prompts like 'holding a bouquet' and 'riding a bike'.

5.3 Real Image Subject Swap

Thanks to the simplicity of MoA, and the minimal deviation from the prior model, it can be used in conjunction with DDIM Inversion [Mokady et al. 2023; Song et al. 2020b] to enable real-image editing. Fig. 13 shows results of this application. In the case of a single subject photo, we run DDIM Inversion with the prompt "a person". Starting from the inverted random noise, we run generation using MoA and inject the desired subject in the 'person' token. For swapping a subject in a couple photo, we run DDIM inversion with the prompt "a person and a person". During MoA generation, we used a crop of the subject to keep in the first 'person' token, and inject the desired subject image into the second 'person' token.

6 LIMITATIONS

Firstly, due to inherent limitations of the underlying Stable Diffusion model, our method sometimes struggles with producing high-quality small faces (see Fig. 14). This particularly affects the ability to depict multiple people in the same image, as scenes involving interactions typically demand a full-body photo from a longer distance, rather than an upper-body portrait. Secondly, generating images that depict intricate scenarios with a wide range of interactions and many individuals continues to be a challenging task. This difficulty is again largely due to the inherent limitations of Stable Diffusion and CLIP, particularly their inadequate grasp of complex compositional concepts, such as counting objects. Specifically for MoA, the current implementation has limited ability to perform text-based expression control. Since during finetuning, the expression in the input subject image and the output reconstruction target are the same, the model entangled the notion of 'identity' and 'expression'. A future direction worth exploring is to use a slightly different input image; for example, two different frames from a video explored in other topics [Kulal et al. 2023].

7 CONCLUSION

We introduce Mixture-of-Attention (MoA), a new architecture for personalized generation that augments a foundation text-to-image model with the ability to inject subject images while preserves the prior capability of the model. While images generated by existing subject-driven generation methods often lack diversity and subject-context interaction when compared to the images generated by the prior text-to-image model, MoA seamlessly unifies the two

Fig. 11. **Controllable personalized generation.** MoA is compatible with ControlNet. Given the same prompt, the user can use ControlNet for pose control. In this application, MoA still retains the disentangled subject-context control.

Fig. 12. **Subject morphing.** By interpolating between the embeddings of the image encoder in MoA, we can achieve a morphing effect between two subjects with different characteristics. On the left is the image 'Yokozuna', and on the right is an image generated by DALL-E 3.

paradigms by having two distinct experts and a router to dynamically merges the two pathways. MoA layers enables the generation of personalized context from multiple input subjects with rich interactions, akin to the original non-personalized model, within a single reverse diffusion pass and without requiring test-time fine-tuning steps, unlocking previously unattainable results. In addition, our model demonstrates previously unseen layout variation in the generated images and the capability to handle occlusion from objects or other subjects, and handle different body shapes all without explicit control. Lastly, thanks to its simplicity, MoA is naturally compatible with well-known diffusion-based generation and editing techniques like ControlNet and DDIM Inversion. As an example, the combination of MoA and DDIM Inversion unlocks the application of subject swapping in a real photo. Looking ahead, we envision further enhancements to the MoA architecture through the specialization of different experts on distinct tasks or semantic labels. Additionally, the adoption of a minimal intervention approach to personalization can be extended to various foundational models (e.g. video, and 3D/4D generation), facilitating the creation of personalized content with existing and futuristic generative models.

Fig. 13. **Real image editing with MoA.** MoA is compatible with diffusion-based image editing techniques with DDIM Inversion. Starting with the inverted noised, MoA is able to replace the subject in the reference image.

Fig. 14. **Limitation.** One key feature of MoA is enabling the generation of images with complex interaction scenarios, which result in full-body images. They inevitably contain small faces, which remains a hard task for the underlying Stable Diffusion model.

ACKNOWLEDGEMENT

The authors would like to acknowledge Colin Eles for infrastructure support, Yuval Alaluf, Or Patashnik, Rinon Gal, Daniel Cohen-Or for their feedback on the paper, and other members on the Snap Creative Vision team for valuable feedback and discussion throughout the project.

REFERENCES

2022. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. https://github.com/cloneofsimo/lora.

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–10.

Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. 2023. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023a. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *arXiv preprint arXiv:2305.16311* (2023).

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023b. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.10093* (2023).

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113* (2023).

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1290–1299.

CivitAI. 2023a. CivitAI checkpoint. https://civitai.com/models/30240/toonyou.

CivitAI. 2023b. CivitAI checkpoint. https://civitai.com/models/65203/disney-pixar-cartoon-type-a.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research* 23, 1 (2022), 5232–5270.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2024. LCM-Lookahead for Encoder-based Text-to-Image Personalization. *arXiv preprint arXiv:2404.03620* (2024).

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. 2023. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models, In NeurIPS. *NeurIPS*.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305* (2023).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-Prompt Image Editing with Cross Attention Control. *ICLR* (2023).

Jonathan Ho. 2022. Classifier-Free Diffusion Guidance. *ArXiv* abs/2207.12598 (2022). https://api.semanticscholar.org/CorpusID:249145348

Jonathan Ho, Ajay Jain, and P. Abbeel. 2020. Denoising Diffusion Probabilistic Models. *ArXiv* abs/2006.11239 (2020). https://api.semanticscholar.org/CorpusID:219955663

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17089–17099.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023a. Multi-concept customization of text-to-image diffusion. In *CVPR*. 1931–1941.

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

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems* 36 (2024).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2023a. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461* (2023).

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

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

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

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *CVPR*. 6038–6047.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*. PMLR, 8162–8171.

Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. 2022. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. *Trans. Mach. Learn. Res.* 2022 (2022). https://api.semanticscholar.org/CorpusID:245650542

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. 2023. Orthogonal adaptation for modular customization of diffusion models. *arXiv preprint arXiv:2312.02432* (2023).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. 2021. Hash layers for large sparse models. *Advances in Neural Information Processing Systems* 34 (2021), 17555–17566.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685. https://api.semanticscholar.org/CorpusID:245335280

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023a. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*. 22500–22510.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2023b. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949* (2023).

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*. 36479–36494.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1ckMDqlg

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising Diffusion Implicit Models. *ArXiv* abs/2010.02502 (2020). https://api.semanticscholar.org/CorpusID:222140788

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020b. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-Free Consistent Text-to-Image Generation. *arXiv preprint arXiv:2402.03286* (2024).

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. *P+*: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).

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

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848* (2023).

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431* (2023).

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. 2024. PeRFlow: Accelerating Diffusion models via Piecewise Rectified Flow. (2024).

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. (2023).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023a. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. 2023. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems* 36 (2023).

A ADDITIONAL EXPERIMENTAL DETAILS

Finetuning hyperaparemters. Training is done using the Accelerate library [Gugger et al. 2022] with 4 GPUs in mixed precision (`bf16`). Tab. 1 summarizes the finetuning hyperparameters.

Table 1. Prompts used for generating the qualitative results.

Name	Value
Training iteration	40k
Batch size per GPU	32
# of GPUs	4
Learning rate	5e-05
Router regularization weight (λ_r)	1e-04
Object regularization weight (λ_o)	1e-04
Prob. of removing condition	0.1
Prob. of using masked recon. loss	1
Max training diffusion timestep sampled	800

Prompts. For generating the qualitative results, we use the prompts listed in Tab. 2. The special token 'man' is replaced with 'woman' when appropriate.

Table 2. Prompts used for generating the qualitative results.

ID	Prompt
vanGogh	:a **man** and a **man** in the style of Van Gogh painting
sofa	:a **man** and a **man** sitting on a sofa
book	:a **man** and a **man** reading a book
paris	:a **man** and a **man** in Paris, shaking hands, with the Eiffel Tower in the background
mountain	:a **man** and a **man** standing on a mountain
running	:a **man** and a **man** running in a park
lavender farm	:a **man** and a **man** in a lavendar farm
lab	:a **man** in a laboratory
bike	:a **man** riding a bike
cowboy	:a portrait photo of a **man** dressed as a cowboy in the wild west, in from of a saloon, very high quality, professional photo, beautiful lighting, 8k
spaceman	:a close up portrait photo of a **man** dressed in a space suit without helmet, on Mars, in front of spacehip, beautiful starry sky, very high quality, professional photo, beautiful lighting, 8k
racecar_driver	:a portrait photo of a **man** as a race car driver, very high quality, professional photo, beautiful lighting, 8k

B ABLATION

In this section, we ablate the model by (i) removing our primary contribution, the MoA layer, (ii) remove the community checkpoint and use the vanilla SD15 (i.e.`runwayml/stable-diffusion-v1-5`) checkpoint, and (iii) removing the image feature spacetime conditioning (Eqn. (8)).

In Fig. 15, we can clearly see that removing the MoA layer produced images with substantially worse quality. While the foreground subjects are well preserved, the context is mostly lost. When comparing to images generated using the base checkpoint, the behavior is similar (i.e. both subject and context are well preserved). The primary difference is that the community checkpoint generates images with better overall texture. Similar to recent works in subject-driven generation [Gu et al. 2023; Po et al. 2023; Yan et al. 2024], we use the community checkpoint because of their better texture.

In Fig. 16, when we remove the spacetime conditioning (Eqn. (8)) of the image features, this model is more restricted than our full model. Intuitively, not having the spacetime conditioning makes the model worse at identity preservation. This indirectly affects the overall model's capability at preserving the context as well.

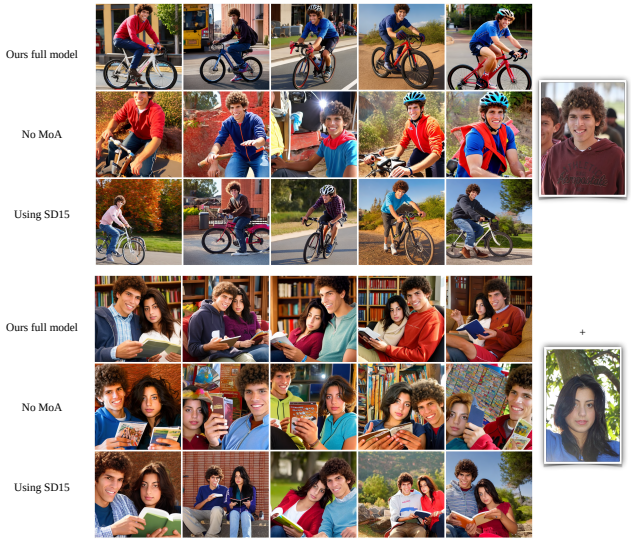

Fig. 15. Ablations: removing the MoA layer, and using the vanilla SD15 checkpoint. Constrasting with and without MoA layer, we clearly see that difference in context preservation. Without the MoA layer, the context (i.e. object, background and interaction) is lost despite the foreground being preserved well. Comparing our full model using the `AbsoluteReality` checkpoint with using the vanilla SD15 checkpoint, the behavior is similar, but overall texture differs.

C ADDITIONAL QUALITATIVE RESULTS

C.1 Handling Different Body Shapes

Given MoA's capability to preserve both the subject and context well, we found a surprising use case where subjects with different body shapes can be injected, and the body shape is naturally preserved and integrated with the context. For this section, we use an old man generated by Consistory [Tewel et al. 2024], the famous Yokozuna for a large body, and a Dalle-3 [Betker et al. 2023] generated man for skinny body type. In Fig. 18, we can see that the body types are

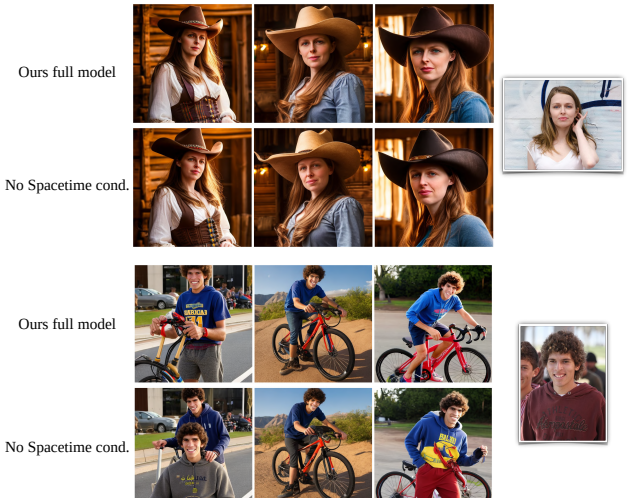

Fig. 16. Ablations: removing the spacetime conditioning in the image feature. Not having the spacetime conditing restricts the model, which results in worse identity preservation (top), and worse context preservation (bottom).

Fig. 17. **Single-subject portraits.** Our method is able to generate high-quality images of the input subjects in various imaginary scenarios and costumes.

preserved. For the second column of the men holding roses, we can see through the gap between the arm and the body for the Dalle-3 man, while Yokozuna completely blocks the background.

D QUANTITATIVE RESULTS & ANALYSIS

Evaluation metrics. The primary quantitative metrics we use are identity preservation (IP), and prompt consistency (PC). To assess IP, pairwise identity similarity is calcuated between the generated image and the input image using FaceNet [Schroff et al. 2015]. To assess PC, the average CLIP-L/14 image-text similarity is calculated following previous studies [Gal et al. 2022; Xiao et al. 2023].

We perform the same evaluation as baseline methods, and perform automated quantitative evaluation of identity preservation (IP) and prompt consistency (PC) (See Tab. 3). While we perform on par with

Fig. 18. Handling different body shapes.

Methods	*OF*	***Single-Subject***	
		IP ↑	PC ↑
ELITE	✓	0.228	0.146
Dreambooth		0.273	0.239
Custom-Diffusion		0.434	0.233
Fastcomposer	✓	0.514	0.243
Subject-Diffusion	✓	0.605	0.228
Mixture-of-Attention	✓	0.555	0.202

Table 3. Quantitative results. OF stands for "optimization-free".

baselines like FastComposer, samples from our method have more image variation (e.g. in layout) (See Fig. 20). Also, in Fig. 3, we can clearly see that our generated images have much better variations and interaction with the context. In the baseline, even when the text is "riding a bike", the bike is barely visible, and there is no clear interaction. However, in terms of the automated evaluation metric, having a small face region can lead to a lower score in the automated quantitative evaluation. Note that for a fair qualitative comparison with FastComposer, we use UniPC scheduler and our prompting strategy with their checkpoint to generate baseline results.

E ADDITIONAL APPLICATIONS

MoA is compatible with style LoRAs (Appendix E.1). By interpolating the image and text features separately, MoA is able to generate meaningful and smooth transitions (Appendix E.2). Lastly, the ability to generate multiple consistent characters allows creators to put AI generated characters in different scenarios and compose them to tell stories(Appendix E.3).

E.1 Adding Style to Personalized Generation

In addition to being compatible to ControlNet, the prior branch in MoA is also compatible with style LoRAs. By combining style LoRA with MoA, users can easily generate images in different styles. In Fig. 21, we show stylized generation using two different style LoRAs: ToonYou [CivitAI 2023a], and Pixar [CivitAI 2023b]. Preserving identity across different styles/domains is a challenging task. Identity preservation at the finest details across domain can be

Fig. 19. **Router visualization.** The 16 rows in the visualization correspond to the 16 layers in the U-Net.

Fig. 20. Samples from the quantitative evaluation.

ill-defined. Yet, from Fig. 21, we can clearly see the broad features of the subjects (e.g. hair style, face and body shape) are well preserved and easily recognizable.

E.2 Time Lapse

Similar to the subject morphing by interpolating the image features, we can achieve the 'time lapse' affect by interpolating between the text embeddings of 'person' and 'kid'. In Fig. 22, we show images of Yokozuna at different interpolated text tokens. Surprisingly, MoA is able to generate Yokozuna at different ages with only a single image of him as an adult. We hypothesize that the pretrained diffusion model has a good understanding of the visual effect of aging, and because of the strong prior preservatin of MoA, it is able to interpret the same subject at different ages.

E.3 Storytelling with Consistent Character

With the rise of AI generated content both within the research community and the artistic community, there is significant effort put in crafting visually pleasing characters. However, to tell a story with the generated characters consistently across different frames remains to be a challenge. This is another application of subject-driven generation, the task we study. In Fig. 23, we can generate consistent characters across different frames easily using our MoA model. The man is taken from The Chosen One [Avrahami et al. 2023b], and the girl from the demo of IP-adapter [Ye et al. 2023]. Compared to The Chosen One, we can easily incorporate generated

(a) Subject pair 1

(b) Subject pair 2

Fig. 21. **Stylized generation.** The three rows are: original MoA, + ToonYou LoRA, + Pixar LoRA. MoA is compatible with pretrained style LoRAs. Adding style to MoA is as simple as loading the pretrained LoRA to the prior branch of a trained MoA during generation.

Fig. 23. **Storytelling with consistent characters.** MoA makes it easy to put AI generated characters in new scenarios and combining different characters to form a story.

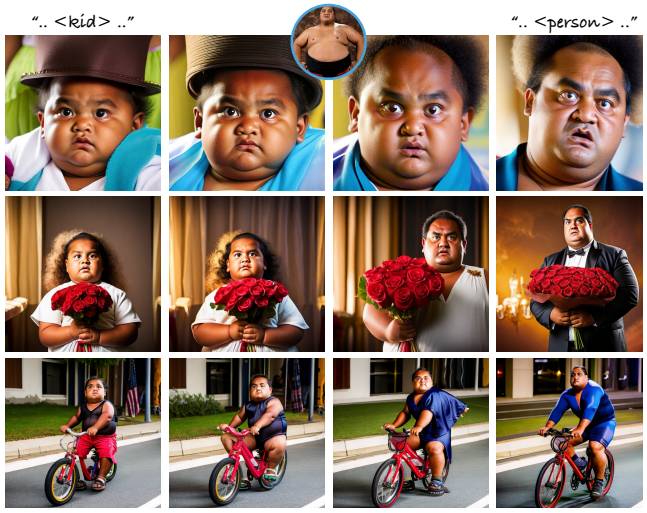

Fig. 22. **Time lapse.** By interpolating between the token 'kid' and 'person', where Yokozuna's image is injected, MoA creates this time lapse sequence between Yokozuna as a kid and an adult.

character from another method. Compare to IP-adapter, we can easily combine the two generated characters in a single frame, which IP-adapter fails to do.