This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Orthogonal Adaptation for Modular Customization of Diffusion Models

Ryan Po
Stanford University

Guandao Yang
Stanford University

Kfir Aberman
Snap Research

Gordon Wetzstein
Stanford University

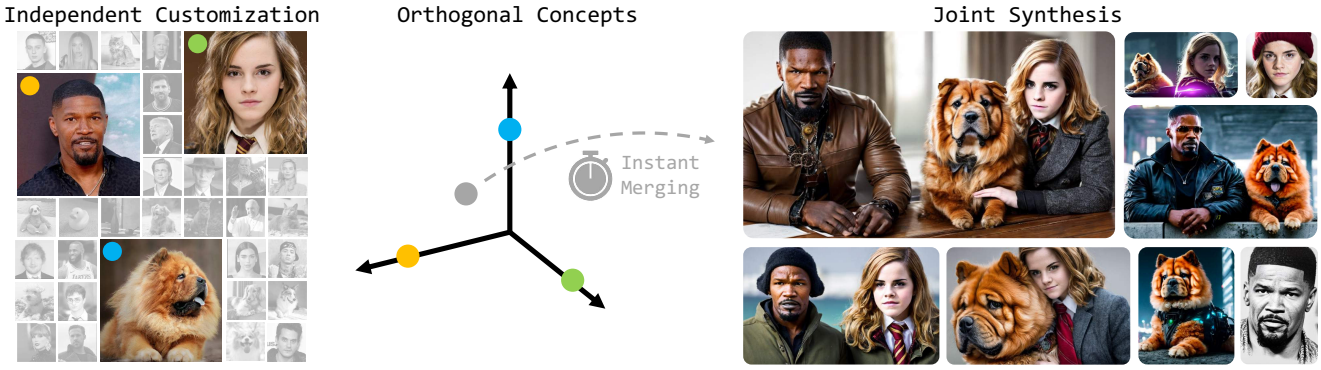

Figure 1. **Modular Customization of Diffusion Models.** Given a large set of individual concepts (left), the goal of *Modular Customization* is to enable independent customization (fine-tuning) per concept, while efficiently merging a subset of customized models during inference, so that the corresponding concepts can be jointly synthesized without compromising fidelity. To tackle this, we propose *Orthogonal Adaptation*, which encourages customized weights of one concept to be orthogonal to the customized weights of others.

Abstract

Customization techniques for text-to-image models have paved the way for a wide range of previously unattainable applications, enabling the generation of specific concepts across diverse contexts and styles. While existing methods facilitate high-fidelity customization for individual concepts or a limited, pre-defined set of them, they fall short of achieving scalability, where a single model can seamlessly render countless concepts. In this paper, we address a new problem called Modular Customization, with the goal of efficiently merging customized models that were fine-tuned independently for individual concepts. This allows the merged model to jointly synthesize concepts in one image without compromising fidelity or incurring any additional computational costs. To address this problem, we introduce Orthogonal Adaptation, a method designed to encourage the customized models, which do not have access to each other during fine-tuning, to have orthogonal residual weights. This ensures that during inference time, the customized models can be summed with minimal interference. Our proposed method is both simple and versatile, applicable to nearly all optimizable weights in the model architecture. Through an extensive set of quantitative and qualitative evaluations, our method consistently outperforms relevant baselines in terms of efficiency and identity preservation, demonstrating a significant leap toward scalable customization of diffusion models.

1. Introduction

Diffusion models (DMs) mark a paradigm shift for computer vision and beyond. DM-based foundation models for text-to-image, video, or 3D generation enable users to create and edit content with unprecedented quality and diversity using intuitive text prompts [32]. Although these foundation models are trained on a massive amount of data, in order to synthesize user-specific concepts (such as a pet, an item, or a person) with a high fidelity, they often need to be fine-tuned.

Several recent approaches to customizing DMs to individual concepts have demonstrated high-quality results [10, 18, 24, 36, 44]. A multi-concept DM strategy, however, where several pre-trained concepts are mixed in a single image, remains challenging. Existing multi-concept methods [12, 24] either show a degradation in the quality of individual concepts when merged or require access to multiple concepts during training. The latter makes the process unscalable and raises privacy concerns when the different concepts belong to different users. Furthermore, in all cases the mixing process is computationally inefficient.

We introduce orthogonal adaptation as a new approach to enabling instantaneous multi-concept customization of DMs. The primary insight of our work is that changing how the DM is fine-tuned for novel concepts can lead to very efficient mixing of these concepts. Specifically, we represent each new concept using a basis that is approx-

Figure 2. **Gallery of multi-concept generations.** Our method enables efficient merging of individually fine-tuned concepts for modular, efficient multi-concept customization of text-to-image diffusion models. Each concept shown above was fine-tuned individually using orthogonal adaptation. Fine-tuned weight residuals are then merged via summation, enabling multi-concept generation.

imately orthogonal to the basis of other concepts. These bases do not need to be know a priori and different concepts can be trained independently of each other. A key advantage of our approach is that our model does not need to be re-trained when mixing several of our orthogonal concepts together, for example to jointly synthesize different concepts that were never seen together in any training example. Importantly, our approach is **modular** in that it enables individual concepts to be learned independently and in parallel without knowledge of each other. Moreover, it is privacy aware in the sense that it never requires access to the training images of concepts to mix them.

Consider a social media platform where millions of users fine-tune a DM using their personal concepts and want to mix them with their friends' concepts on their phones. Efficiency of the customization and mixing processes as well as data privacy are key challenges in this scenario. Our method addresses precisely these issues. A core technical contribution of our work is a modular customization and scalable multi-concept merging approach that offers better quality in terms of identity preservation than baselines at similar speeds, or similar quality to state-of-the-art baselines at significantly lower processing times.

2. Related Work

Text-conditioned image synthesis. The field of text-conditioned image synthesis has experienced significant advancements, driven by developments in GANs [6, 11, 21–23] and diffusion models [8, 16, 17, 28, 30, 35, 42]. Earlier efforts focus on applying GANs to various conditional synthesis tasks, including class-conditioned image generation [6, 19, 21] and text-driven editing [2, 5, 9, 26, 29, 31, 34, 46]. More recently, the focus has shifted to large text-to-image models [33, 35, 38, 48] trained on large-scale datasets [39]. In this paper, we will utilize the open-source StableDiffusion [35] architecture and build on its pre-trained checkpoints by fine-tuning.

Method	Fidelity (Single-concept)	Efficient Merging	Fidelity (Multi-concept)
TI [10]	✗	✓	✗
DB-LoRA[1] [36]	✓	✓	✗
Custom Diffusion [24]	✗	✓	✗
Mix-of-Show [12]	✓	✗	✓
Ours	✓	✓	✓

Table 1. **Comparison of Solutions to Modular Customization.** Our customization approach excels in three key areas: (1) preserving the identity of individual concepts with high fidelity, (2) efficiently merging independently customized models, and (3) maintaining high concept fidelity for multi-concept image synthesis using the merged model.

Customization. The task of customization aims at capturing a user-defined concept, to be used for generation under various contexts. Seminal works such as Textual Inversion (TI) [10] and DreamBooth [36] tackle the problem of customization by taking a handful of images of the same concept to produce a representation of the subject to be used for controlled generation. TI captures new concepts by optimizing a text embedding to reconstruct target images using the conventional diffusion loss. Follow-up works, such as $\mathcal{P}+$ [14], extend Texture Inversion with a more expressive token representation, improving generation subject alignment/fidelity. DreamBooth [36], on the other hand, picks an uncommon word token and fine-tunes the network weights to reconstruct the target concept using diffusion loss [17]. Custom Diffusion [24] works in a similar way but only fine-tunes a subset of the diffusion model layers, namely the cross-attention layers. LoRA [18] is a low-rank matrix decomposition method that enables better parameter efficiency for fine-tuning methods, and was recently adapted to customization of text-to-image diffusion models [1] (DB-LoRA). Recent works [20, 37, 40, 41, 43, 45, 47] try to improve speed by training feed-forward networks to predict adaptation parameters from data, successfully amortize the time taken to create customize concepts.

[1] assuming DB-LoRA fine-tuned models are merged with FedAvg [25]

Multi-concept Customization. Certain existing works have taken the task of customization one step further, aiming to inject multiple novel concepts into a model at the same time. Custom Diffusion [24] achieves this through a joint optimization loss for all concepts, while Break-a-scene [3] and SVDiff [13] introduces a masked cross-attention loss to learn individual concepts in images containing multiple concepts. However, such methods require access to ground truth data of all concepts training. In this work, we are interested in the task of ***modular customization***, where concepts are learned independently, and users can then mix and match individual concepts during inference for multi-concept image synthesis (Sec. 3.1).

Prior works have provided implicit solutions to the problem of modular customization, but each existing method comes with its own set of trade-offs. TI [10, 27, 44] implicitly addresses the task by representing each concept through a unique token embedding, enabling multi-concept customization by simply querying each token. However, TI tends to suffer from low subject fidelity, as token embeddings alone provide limited expressivity. Federated Averaging (FedAvg) [25] merges fine-tuned models by simply taking a weighted average between the weights of each model, although fast and expressive, naive combination tends to lead to loss of concept identity. Custom Diffusion [24] supports merging of individually fine-tuned networks through solving a constrained customization problem. This method also struggles with expressivity, as only a small subset of the diffusion model weights are being updated. Concurrent work, Mix-of-Show (MoS) [12] expands on this method by introducing gradient fusion, enabling merging of multiple separately fine-tuned models without placing restrictions on parameter expressivity. Though expressive, gradient fusion is computationally demanding, taking ∼15-20 minutes just to combine three custom concepts into a single model, which becomes intractably expensive when deployed at scale. Table 1 summarizes the key areas in which our approach differs from previous and concurrent works.

3. Method

In this section, we first introduce the problem setting of modular customization (Sec. 3.1). We then take a look at the simple solution of FedAvg [25], and explore where and why this naive method fails to preserve identity (Sec. 3.2). Motivated by the limitations of FedAvg, we discuss the conditions to ensure concept identity preservation (Sec. 3.3), and finally introduce our solution to modular customization – ***orthogonal adaption*** (Sec. 3.4 and Sec. 3.5).

3.1. Modular Customization

In this paper, we are interested in customizing text-to-image diffusion models to generate multiple personal concepts in an efficient, scalable, and decentralized manner.

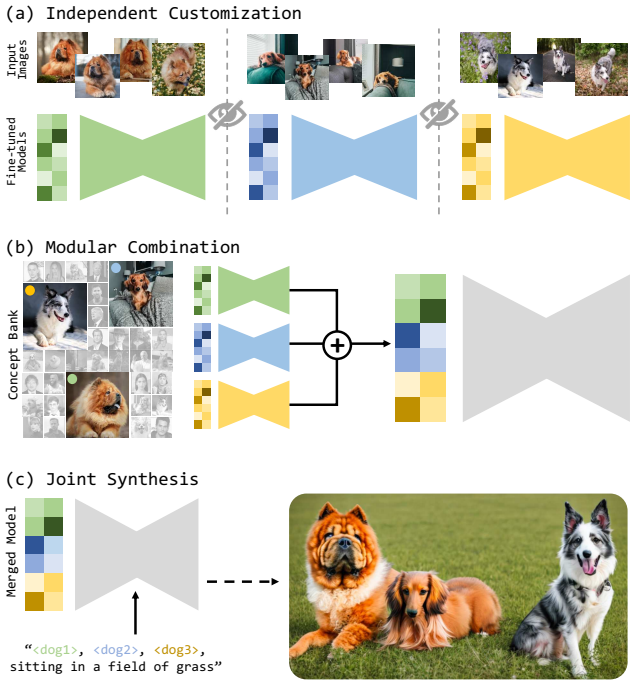

Figure 3. The three stages of **Modular Customization**: (a) Independent Customization, (b) Modular Combination, and (c) Joint Synthesis. Note that during individual fine-tuning, all processes are private, meaning each user does not have access to ground truth data for other concepts.

In addition to single-concept text-to-image customization, users are usually interested in seeing multiple concepts interacting together. This calls for a text-to-image model that is customized to a set of concepts. Being able to generate multiple personalized concepts in a single model, however, is challenging. First, the number of sets containing all possible combinations of concepts is growing exponentially with respect to the number of concepts – an intractable number even for a relatively small number of concepts. As a result, it's important for personalized concepts to be merged with interactive speed. Furthermore, users usually have limited compute at their end, which means any computation done on the users end should ideally be trivial.

These requirements motivate an efficient and scalable fine-tuning setting we call ***modular customization***, where individual fine-tuned models should act like independent modules, which can be combined with others in a plug-and-play manner without additional training. The setting of modular customization involves three stages: independent customization, modular combination and joint synthesis. Fig. 3 provides an illustration of this three stage process.

With modular customization in mind, our goal is to design a fine-tuning scheme, such that individually fine-tuned models can be trivially combined (e.g. summation) with any other fine-tuned model to enable multi-concept generation.

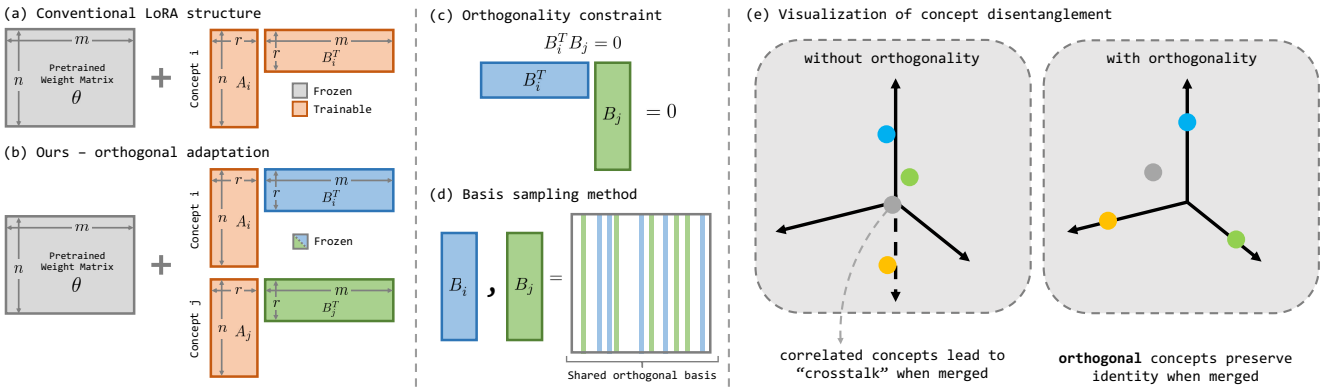

Figure 4. **Overview of Orthogonal Adaptation.** (a) LoRA [18] enables training of both low-rank decomposed matrices. (b) Orthogonal adaption constrains training only to A, leaving B fixed. (c) For two separate concepts, i and j, an orthogonality constraint is imposed between B_i and B_j. (d) When concepts i and j are trained independently, approximate orthogonality between B_i and B_j can be achieved by sampling random columns from a shared orthogonal matrix. (e) Without the orthogonality constraint, correlated concepts suffer from "crosstalk" when merged; with the orthogonality constraint, orthogonal concepts preserve their identities after merging.

3.2. Federated Averaging

Perhaps the most straight-forward technique for achieving modular customization is to take a weighted average of each individually fine-tuned model. This technique is often referred to as FedAvg [25]. Given a set of learned weight residuals $\Delta\theta_i$ optimized on concept i, the resulting merged model is simply given by

$$\theta_{\text{merged}} = \theta + \sum_i \lambda_i \Delta\theta_i, \tag{1}$$

where θ represents the pre-trained parameters of the model used for fine-tuning, and λ_i is a scalar representing the relative strength of each concept. While FedAvg is fast and places no constraints on the expressivity of each individually fine-tuned model, naively averaging these weights can lead to loss of subject fidelity due to interference between the learned weight residuals. This effect is especially severe when training multiple semantically similar concepts (e.g., human identities), as learned weight residuals tend to be very similar. We coin this undesirable phenomenon "crosstalk". Fig. 7 and Fig. 8(a) provide visualizations of the effect of crosstalk, as FedAvg causes multi-concept generations to exhibit loss of identity. Our approach is inspired by FedAvg. We adopt its computational efficiency but modify the fine-tuning process to ensure minimal interference between learned weight residuals between different concepts. We want to enable instant, multi-concept customization from individually trained models without sacrificing subject fidelity.

3.3. Preserving Concept Identity

With the goal of addressing the limitations of FedAvg, we first examine where this method fails. For simplicity, consider the case of merging two concepts i and j. After fine-tuning on each individual task, we receive a set of learned weight residuals $\Delta\theta_i$ and $\Delta\theta_j$. The output of a particular linear layer in the fine-tuned network is

$$O_i(X_i) = (\theta + \Delta\theta_i)X_i, \tag{2}$$

where X_i represents a particular input to the layer corresponding to the training data of concept i. When merging the two concepts using FedAvg with $\lambda = 1$, the resulting merged model produces

$$\hat{O}_i(X_i) = (\theta + \Delta\theta_i + \Delta\theta_j)X_i. \tag{3}$$

The goal of concept preservation is to have $\hat{O}_i(X_i) = O_i(X_i)$. Note that, without enforcing specific constraints, it is likely that $\Delta\theta_j X_i \neq 0$ and, subsequently, $\hat{O}_i \neq O_i$.

It follows that the mapping of data for concept i is preserved when $\Delta\theta_j X_i = 0$ for $j \neq i$. By symmetry, the mapping of data for concept j is preserved given $\Delta\theta_i X_j = 0$ for $i \neq j$. Intuitively, $||\Delta\theta_j X_i||$ measures the amount of crosstalk between the customized weights of concepts i and j. We would like to keep this value low to ensure subject identity is preserved even after merging. However, note that given enough data for training a certain concept i, X_i is likely to have full column rank. This makes the orthogonality condition impossible to satisfy. Instead, we propose a relaxation to this condition, choosing to minimize the crosstalk term for some projection of X_i onto a subspace S_i. This projection yields $S_i S_i^T X_i$, and our relaxed objective hopes to achieve $\hat{O}_i(S_i S_i^T X_i) = O_i(S_i S_i^T X_i)$.

3.4. Orthogonal Adaptation

Motivated by the relaxed objective above, we propose ***orthogonal adaptation***. Similar to low-rank adaptation (LoRA), we represent learned weight residuals through a low-rank decomposition of the form

$$\Delta\theta_i = A_i B_i^T, \theta_i \in \mathbb{R}^{n\times m} A_i \in \mathbb{R}^{n\times r}, B_i \in \mathbb{R}^{m\times r}, \tag{4}$$

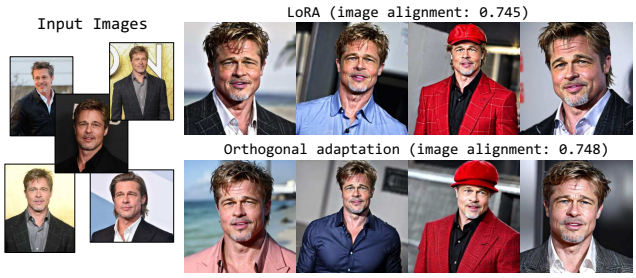

Figure 5. **Over-parameterization of text-to-image models.** Despite the added constraint on the trained weight residuals, due to the over-paramterized nature of large text-to-image diffusion models, our method is able to achieve single-concept customization results with comparable fidelity to the unconstrained setting.

where the rank $r \ll \min(n, m)$. However, contrary to conventionally fine-tuning with LoRA, we keep B_i constant, and only optimize A_i.

Consider a matrix $\bar{B}_j$, where its columns span the orthogonal complement of the column space of B_j. We show that by selecting $S_i = \bar{B}_j$, we achieve the conditions for achieving the projected preservation objective. This can be seen from the fact that,

$$\hat{O}_i(S_i S_i^T X_i) = O_i(S_i S_i^T X_i) + \Delta\theta_j S_i S_i^T X_i \tag{5}$$

$$= O_i(S_i S_i^T X_i) + A_j \cancelto{0}{B_j^T S_i} S_i^T X_i \tag{6}$$

$$= O_i(S_i S_i^T X_i). \tag{7}$$

Since $r \ll m$, the orthogonal complement of B_j covers most of $\mathbb{R}^m$. It follows that $\bar{B}_j \bar{B}_j^T X_i \approx X_i$, making $\bar{B}_j$ a reasonable candidate for S_i.

At the same time, since we expect the learned residuals for a concept to have meaningful interactions with their data, we would also like to ensure $||\Delta\theta_i X_i||$ is non-trivial. By approximating X_i with its projection onto $\bar{B}_j$, our objective changes to ensuring $||A_i B_i^T \bar{B}_j \bar{B}_j^T X_i||$ is non-trivial. Examining this term gives us the additional constraint that $B_i^T \bar{B}_j \neq 0$, meaning the columns of B_i should live in the orthogonal complement of the columns space of B_j. Therefore, to ensure meaningful fine-tuning results, we should also enforce orthogonality between the learned residuals, i.e. $B_i^T B_j = 0$.

Fig. 4 provides an overview of our orthogonal adaption method. Intuitively, as illustrated in Fig. 4(e), our method disentangles custom concepts into orthogonal directions, ensuring that there is no crosstalk between concepts. As a result, our merged model can better preserve the identity of each concept.

Expressivity of orthogonal adaption. Expressivity of our method arises as a natural concern as we are optimizing significantly fewer parameters by freezing B_i. Fortunately, text-to-image diffusion models are often over-parameterized, with millions/billion of parameters. Prior works have shown that even fine-tuning a subspace of such parameters can be expressive enough to capture a novel concept. We also show this result empirically in Fig. 5, where our method leads to results with similar fidelity, even without the need to optimize B_i during training.

3.5. Designing Orthogonal Matrices B_i's

A key challenge of the method described in previous sections is to generate a set of basis matrices B_i that are orthogonal to each other. Note that this is very difficult especially because when choosing B_i, the user is not aware of what basis the other users chose to optimize for the concepts to be combined in the future. Strictly enforcing such orthogonality might be infeasible without prior knowledge of other tasks. We instead propose a relaxation to the constraint, introducing a simple and effective method to achieve approximate orthogonality.

Randomized orthogonal basis. One method for enforcing approximate orthogonality is to determine a shared orthogonal basis. For some linear weight $\theta \in \mathbb{R}^{m \times n}$, we first generate a large orthogonal basis $O \in \mathbb{R}^{n \times n}$. This orthogonal basis is shared between all users. During training of concept i, B_i is formed from taking a random subset of k columns from O. Given $k \ll n$, the probability of two randomly chosen B_i's to share the same columns is kept low.

Randomized Gaussian. Another approach is to choose random matrix elements. Specifically, we sample each entry of B_i from a zero-mean Gaussian with standard deviation σ: $B_i[k] \sim \mathcal{N}(0, \sigma^2 I)$. When the dimensionality of B_i is high, this simple strategy creates matrices that are orthogonal in expectation: $\mathbb{E}\left[B_i^T B_j\right] = 0$ (see supplement for discussion). Naturally, this method does not require knowledge of a shared basis to sample from. In practice, however, we found randomized Gaussians lead to higher levels of crosstalk in our setting, i.e., $||B_i^T B_j||$ tends to be larger than for the randomized orthogonal basis.

4. Experiments

In this section, we show the results of our method applied to the task of modular customization. Qualitative and quantitative results indicate that our method outperforms relevant baselines [1, 12, 24] at similar speeds, and quality on par with state-of-the-art baselines that require significantly higher processing times [12].

Datasets. We perform evaluations on a custom dataset of 12 concept identities, each containing 16 unique images of the target concept in different contexts.

Implementation details. We perform fine-tuning on the Stable Diffusion [35] model, specifically the ChilloutMix checkpoint for its ability to handle high-fidelity human face

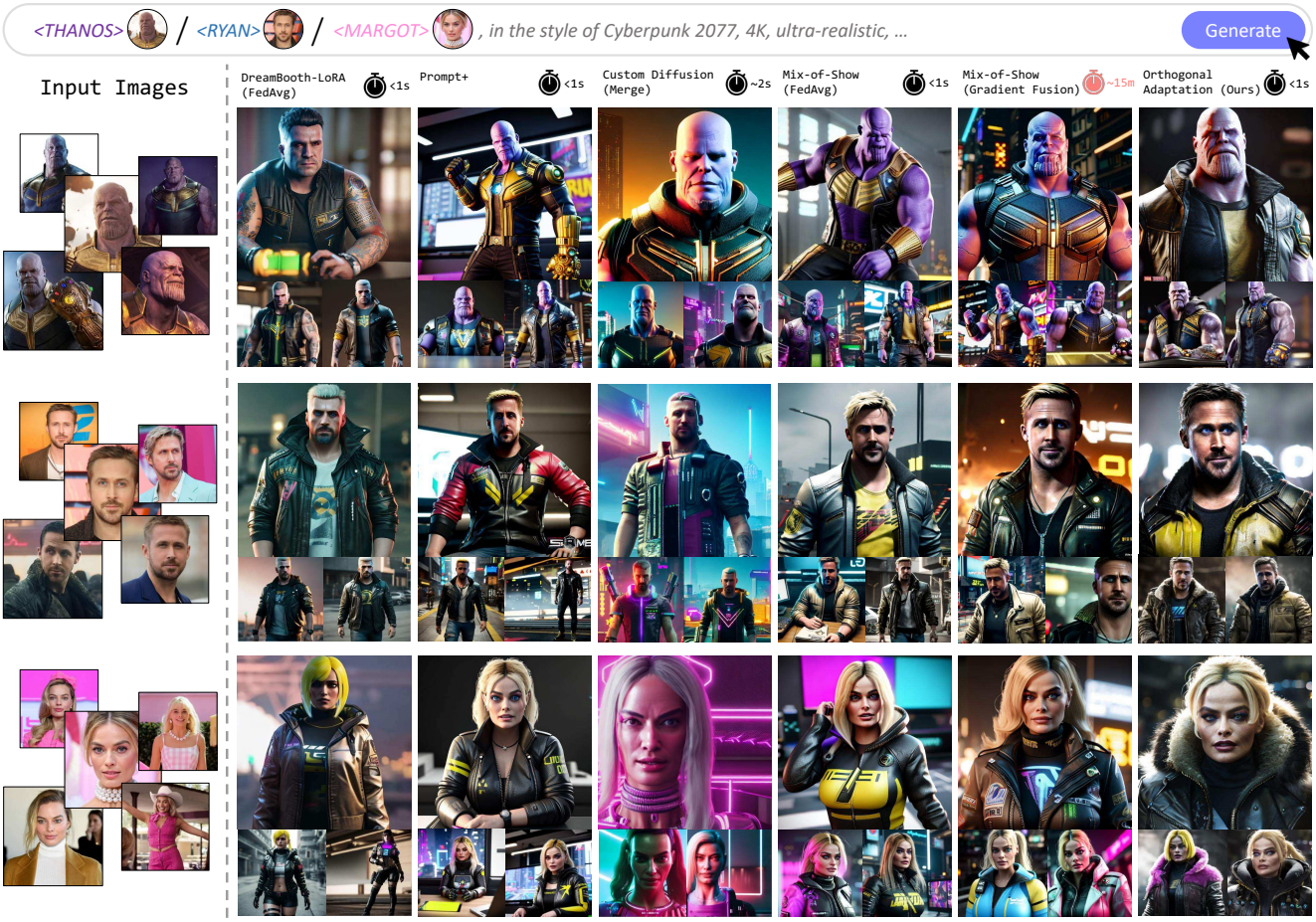

Figure 6. **Identity preservation in single-concept generations from a merged model.** We demonstrate our method's ability to maintain identity consistency across different single-concept generations. Each column showcases images from the same merged model, representing three distinct concept identities. Our approach showcases better identity alignment with the corresponding input images, offering a significant improvement over comparable merging methods. Additionally, our method's performance parallels that of Mix-of-Show (Gradient Fusion) but with the advantage of near-instantaneous merging, in contrast to the approximately 15-minute merging time required.

generation. For single-concept fine-tuning, we apply orthogonal adaptation to all linear layers in the Stable Diffusion architecture. Following prior work [12, 44], we also apply a layer-wise text embedding and represent each fine-tuned concept as two separate text tokens. We fine-tune the text embeddings with a learning rate of $1e-3$, the diffusion model parameters with a learning rate of $1e-5$ and set $r = 20$ for all experiments. Single-concept fine-tuning takes ∼10-15 minutes on two A6000 GPUs. For our method, we enforce the orthogonality constraint using the randomized orthogonal basis method for all experiments. Methods using FedAvg (including orthogonal adaption) were merged using $\lambda = 0.6$.

Baselines. We compare our method against state-of-the-art baselines on the task of modular customization, namely: DreamBooth-LoRA [1], $\mathcal{P}+$ [44], Custom Diffusion [24], and Mix-of-Show [12]. Fine-tuned models are merged differently depending on the method. DreamBooth-LoRA is merged using FedAvg, Custom Diffusion is merged using their proposed optimization-based merging method, and Mix-of-Show is merged using gradient fusion as outlined in their work. Since $\mathcal{P}+$ does not perform fine-tuning on the weights of the network, merging is done simply by querying each concept's token embedding. For completeness, we also compare against Mix-of-Show merged using FedAvg, serving as an efficient alternative to the computationally demanding gradient fusion method.

Experimental setup and metrics. First, we fine-tune each concept individually, without access to data for any other concept. Each fine-tuned model is then combined with two other concepts at random using their corresponding method for merging. Following prior work, we evaluate our method on *image alignment*, which measures the similarity of image features between generated images and the input reference image by measuring their similarity in the CLIP image feature space [10]. Similarly, we evaluate our method using *text alignment*, ensuring the output gen-

Figure 7. **Multi-concept results.** Examples of multi-concept generations, synthesized using sampling techniques from concurrent work [12]. While Mix-of-Show (FedAvg) maintains high-level features, it struggles with crosstalk, manifesting overly smooth facial features. Mix-of-Show (Gradient Fusion) exhibits good identity alignment, albeit with a computationally intensive merging process. $\mathcal{P}+$ manages to preserve identity after merging, but struggles to capture identity with high-fidelity due to limited parameter expressivity. Our method stands out by achieving high identity alignment with a significantly faster merging procedure.

Method	Merge Time	Text Alignment ↑			Image Alignment ↑			Identity Alignment ↑		
		Single	Merged	Δ	Single	Merged	Δ	Single	Merged	Δ
P+ [44]	<1 s	.643	.643	—	.683	.683	—	.515	.515	—
Custom Diffusion [24]	∼2 s	.668	.673	+.005	.648	.623	-.025	.504	.408	-.096
DB-LoRA (FedAvg) [1]	<1 s	.613	.682	+.069	.744	.531	-.213	.683	.098	-.585
MoS (FedAvg) [12]	<1 s	.625	.621	-.004	.745	.735	-.010	.728	.706	-.022
MoS (Grad Fusion) [12]	∼15 m	.625	.631	+.006	.745	.729	-.016	.728	.717	-.011
Ours	<1 s	.624	.644	-.010	.748	**.741**	-.007	.740	**.745**	+.005

Table 2. **Quantitative results.** We provide detailed qualitative comparisons for each method, evaluated both before and after the merging process. Prior to merging, our method demonstrates comparable performance in all identity-related metrics, highlighting its expressivity even with the orthogonality constraint. Post-merging, our method achieves the highest scores in image and identity alignment. Our method is also capable of maintaining text alignment scores comparable to other high-fidelity methods such as P+ and MoS.

erations still adhere to the input text-prompts by measuring the text-image similarity also using CLIP [15]. However, to further illustrate the identity preserving capabilities of our method, we also evaluate our method using the ArcFace [7] model. Using the ArcFace model, we measure the rate at which the target human identity is detected in a set of generated images, we refer to this metric as *identity alignment*.

4.1. Qualitative Comparisons

Merged single-concept results. We illustrate the identity preserving effect of our method by comparing single-concept generations of different identities from the same merged model. As mentioned above, each concept is fine-tuned individually and merged together during inference. Fig 6 shows generations for three separate concept identities, each column contains images sampled from the same model. Our method achieves better identity alignment with the input images in the merged model compared to methods with comparable merging times. We also achieve similar results to Mix-of-Show (Gradient Fusion), which requires ∼15 minutes to merge three concepts, while our method enables near instant merging.

Merged multi-concept results. We also show generated images containing all three identities in the merged model. Leveraging multi-concept sampling techniques from concurrent work [12], we show examples of multi-concept generations in Fig. 7. Once again, multi-concept models trained using our method generate images with better identity alignment than competing baselines. Due to the poor performance of DB-LoRA [1] and Custom Diffusion [24] for single-concept generations, we omit results for these methods on multi-concept generation due to space constraints.

$\mathcal{P}+$ [14] suffers from low concept fidelity due to limited expressivity in their training regime. Although Mix-of-Show [12] (FedAvg) preserves certain high-level features through the layer-wise text-embedding, it still suffers from crosstalk due to unconstrained training of weight residuals. Mix-of-Show (Gradient Fusion) shows impressive identity alignment, however, this is only enabled by a computationally demanding merging procedure. Our method achieves high identity alignment while keeping the merging process at near instant rates.

4.2. Quantitative Results

We present quantitative comparisons in Table. 2. Specifically, we show all three evaluation metrics applied to each method before and after merging. Our method achieves comparable results in all concept alignment metrics before merging, illustrating the expressivity of our method despite the orthogonality constraint. After merging, our method achieves the highest image and identity alignment scores across all methods, while maintaining comparable text alignment scores with other high-fidelity methods such as Mix-of-Show and $\mathcal{P}+$. This illustrates that our method is able to achieve high identity preservation without sacrificing the ability to generalize for different contexts.

Note that although Custom Diffusion [24] and DB-LoRA [1] achieves higher text alignment, this is at the cost of significantly lower concept alignment scores than that of competing methods.

5. Ablations

Effect of orthogonality. In Fig. 8(a), we present generated images from a model created from merging two separate fine-tuned models (concepts i and j). To illustrate the effect of orthogonality on identity preservation, we manipulate the degree of orthogonality between B_i and B_j. On the left, we have the worst case scenario, where $B_i = B_j$. On the right, we show results where perfect orthogonality is achieved, i.e. $B_i^T B_j = 0$. In between, we construct B_i and B_j from a shared orthogonal matrix, but choose half of their columns to be overlapping. Results in Fig. 8(a) show that orthogonality contributes significantly to identity preservation even in the extreme case of merging 2 concepts.

Number of merged concepts Fig. 8(b) shows results generated from models with a range of concepts merged together. With orthogonality, our model is capable of merging a high number of concepts with minimal identity loss. In contrast, without orthogonality, concept fidelity quickly degrades, even with relatively low number of concepts being combined. Running our model without orthogonality is equivalent to Mix-of-Show [12] merged using FedAvg [25].

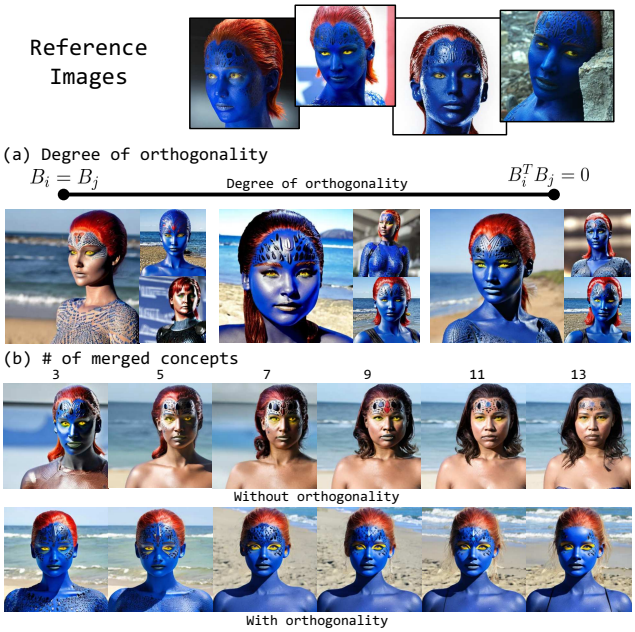

Figure 8. **Ablation studies.** (a) Images generated from a model formed by merging two separately fine-tuned models (concepts i and j), focusing on the role of orthogonality in preserving identity. (b) Image generations from models that with a varying number of merged concepts. Without orthogonality, concept identity is lost even when merging a small number of concepts.

6. Discussion

Limitations. Despite showcasing the ability to encode several custom concepts into the same text-to-image model, generating images with complex compositions/interactions between multiple custom concepts remains challenging. As concepts, such as human identities, have the tendency to either be entangled, or even completely ignored. Existing works [4, 12] have developed certain strategies for remedying this effect, but such methods are still prone to the aforementioned failure cases. Another limitation of orthogonal adaption is that it directly modifies the fine-tuning process. Therefore, existing fine-tuned networks (e.g. LoRAs [1]) can not be adapted post-hoc to ensure orthogonality.

Ethics Considerations. Generative AI could be misused for generating edited imagery of real people with the intent of spreading disinformation. Such misuse of image synthesis techniques poses a societal threat, and we do not condone using our work for such purposes. We also recognize a potential biases in the foundation model we built upon.

Conclusions. By disentangling customization concepts into orthogonal directions, orthogonal adaptation streamlines the process of integrating multiple independently fine-tuned concepts into a single model instantly and with trivial compute, while also ensuring preservation of each concept. Our work makes a significant step towards modular customization, where multi-concept customization can be achieved with individual, privately fine-tuned models.

References

[1] Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022. 2, 5, 6, 7, 8

[2] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 2

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *ArXiv*, abs/2305.16311, 2023. 3

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 8

[5] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[7] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. 7

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 2

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 3, 6

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[12] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 1, 2, 3, 5, 6, 7, 8

[13] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 3

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2, 8

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 7

[16] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 2

[17] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1, 2, 4

[19] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *European Conference on Computer Vision*, pages 91–109. Springer, 2022. 2

[20] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7, 8

[25] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. 2, 3, 4, 8

[26] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 3

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[29] Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15933–15942, 2023. 2

[30] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Trans. Mach. Learn. Res.*, 2022, 2022. 2

[31] Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, and Krishna Kumar Singh. Spatially-adaptive multilayer selection for gan inversion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11399–11409, 2022. 2

[32] Ryan Po, Wang Yifan, and Vladislav Golyanik et al. State of the art on diffusion models for visual computing. In *arxiv*, 2023. 1

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[34] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 2

[35] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 5

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 1, 2

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[40] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *ArXiv*, abs/2304.03411, 2023. 2

[41] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2

[43] Yu-Chuan Su, Kelvin C. K. Chan, Yandong Li, Yang Zhao, Han-Ying Zhang, Boqing Gong, H. Wang, and Xuhui Jia. Identity encoder for personalized diffusion. *ArXiv*, abs/2304.07429, 2023. 2

[44] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 1, 3, 6, 7

[45] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A single-pass lora-free model for stylized image generation. *ArXiv*, abs/2309.01770, 2023. 2

[46] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 2

[47] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2

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