Orthogonal Adaptation for Modular Customization of Diffusion Models

Supplementary Material

7. Gaussian random orthogonal matrices

Theorem 7.1. *Let* $\mathbf{v} \in \mathbb{R}^d$ *and* $\mathbf{u} \in \mathbb{R}^d$ *be two random vectors. Let* $\mathbf{v}_i \sim \mathcal{N}(0, \sigma^2 I)$ *and* $\mathbf{u}_i \sim \mathcal{N}(0, \sigma^2 I)$ *for all* $i \in [1, d]$ *independently, then* $\mathbb{E}\left[\mathbf{v}^T \mathbf{u}\right] = 0$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{v}^T \mathbf{u}\right] &= \mathbb{E}\left[\sum_{i=1}^{d} \mathbf{v}_i \mathbf{u}_i\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\mathbf{v}_i \mathbf{u}_i\right] && \text{(Linearity of expectation)} \\
&= \sum_{i=1}^{d} \mathbb{E}[\mathbf{v}_i]\mathbb{E}[\mathbf{u}_i] && \text{(Independent)} \\
&= \sum_{i=1}^{d} 0 \cdot 0 = 0.
\end{aligned}$$

□

Corollary 7.1.1. *Let* $\mathbf{A} \in \mathbb{R}^{n \times m}$ *and* $\mathbf{B} \in \mathbb{R}^{n \times m}$. *All entries of these matrices are independently sampled from* $\mathcal{N}(0, \sigma^2 I)$. *Then* $\mathbb{E}[\mathbf{A}^T \mathbf{B}] = \mathbf{0} \in \mathbb{R}^{m \times m}$.

Proof.

$$\mathbb{E}[\mathbf{A}^T \mathbf{B}]_{ij} = \mathbb{E}[\mathbf{A}_i^T \mathbf{B}_j] = 0.$$

□

8. Implementation details

Dataset. We chose to evaluate our method on human datasets due to the robustness of face recognition algorithms for evaluation purposes. While prior works [12, 13, 24, 37] have employed CLIP-based metrics as a method of evaluating identity alignment, we found that CLIP features are often poor at identifying fine details in a custom concept. In Fig. 9, we illustrate that our method works for non-human objects too.

Evaluation details. We introduce the *identity alignment* metric for measuring the ability of our method (and competing baselines) in capturing the target human identity in resulting generations. We use the ArcFace [41] facial recognition algorithm and consider a detection to be recorded when the ArcFace distance between two detected faces falls below 0.680 [41]. We choose to use detection probability as a metric rather than the raw distance metric as we found the distance metric to favor over-fitted models. Past the detection threshold, the distance metric directly measures the similarity between two faces, which is not ideal for use-cases such as re-stylization and accessorization.

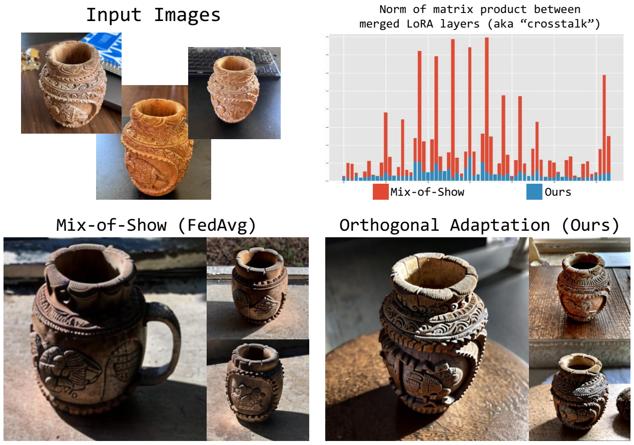


Figure 9. **Identity loss due to crosstalk.** We illustrate the effects of crosstalk by examining the effects of interfering signals between independently trained LoRAs. Measuring crosstalk through the norm of the product between two LoRA weights, our method results in lower crosstalk between independently trained LoRAs. Combined via the same method, our training regime leads to less crosstalk and therefore better identity preservation after merging.

Orthogonal adaptation details. In our method, we enforce the orthogonality constraint through the LoRA down projection matrix B. This formulation ensures orthogonality in the row-space of the resulting LoRA matrices. In theory, we can also achieve orthogonality between trained weight residuals in the column-space, in which case the orthogonality constraint would have to be enforced on the up-projection matrix A instead. We choose to enforce orthogonality in the row-space since the weight residuals interact with the layer inputs through their rows. The concept preservation formulation presented in Sec. 3 is also reliant on row-space orthogonality. In our results, we chose to use the random orthogonal basis method for enforcing orthogonality in all our results. Although the Gaussian random method results in orthogonality on expectation, the orthogonal basis method led to lower crosstalk emperically. The orthogonal basis method requires a shared orthogonal matrix to sample from. In practice, using Stable Diffusion v1.5, there are only four unique input dimensions for all layers in the diffusion model (320, 640, 768, 1280). Therefore, we only have to store four unique square matrices from which all sampled B_i's can then be sampled from. These four or-

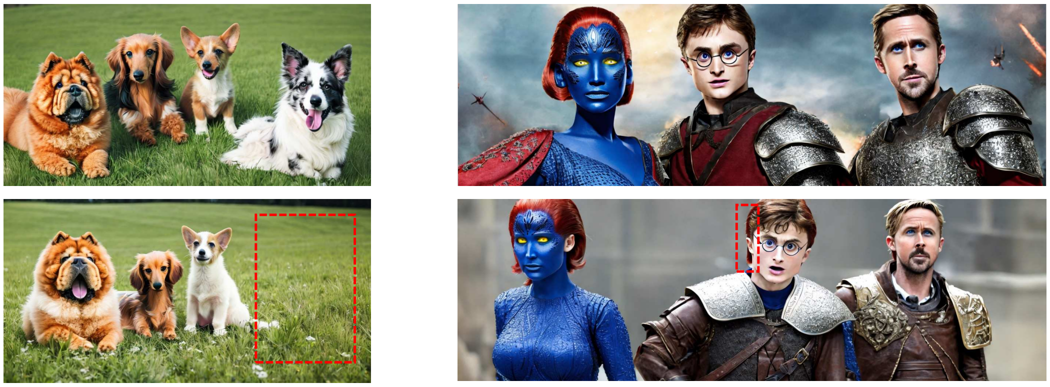


Figure 10. **Multi-concept failure cases.** Multi-concept generation remains as an open challenge. Despite employing techniques such as regionally controllable sampling from prior work [12], this method can still suffer from failure cases such as: (left) ignoring concepts, and (right) leakage of concept attributes to neighboring identities.

thogonal matrices can be downloaded along with the base model, but they can also be generated on the fly with a fix seed to ensure they are shared among all users.

FedAvg merging coefficient. Existing work considers FedAvg merging with affine coefficients. However, with a larger number of concepts, affinely combining each LoRA will lead to dilution of signal from individual LoRAs. It is also a common practice to scale individual LoRA weights post-hoc [1] for direct control over the signal strength from the fine-tuning process. We combine this scaling factor along with the FedAvg merging factor to obtain a single scale factor λ_i as shown in Eq. 1. We consider merging coefficients as a hyper-parameter that can be tuned based on user preferences.

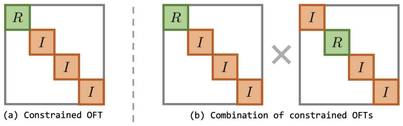
9. Additional results

Illustration of crosstalk. Fig. 9 illustrates the importance of minimizing crosstalk for identity preservation when merging LoRA weights into a single model. We measure crosstalk formally using the norm of the matrix product between individually trained LoRA weight residuals. Upper right of Fig. 9 shwos a direct comparison of the layer-wise normalized matrix product norms between two LoRAs trained with and without orthogonality constraints. Our method leads to a much lower levels of crosstalk, which translates to better identity preservation as observed from the resulting generations.

Relation to Orthogonal Fine-Tuning. Recent work by Qiu et al. [33] (OFT) proposes an alternative method for fine-tuning that leverages orthogonal rotation matrices. In this section we would like to point out some fundamental differences between our method and OFT.

Our method, fine-tunes by optimizing a weight residual $\Delta\theta$, whereas OFT uses rotation with a full-rank orthogonal matrix R. Differences between OFT and LoRA-like fine-tuning is also heavily highlighted in the OFT manuscript (Sec. 4). Our concept of orthogonality is also distinct; we aim for orthogonality between low-rank weight residuals in multi-subject generation, while OFT employs full-rank orthogonal matrices for better single-subject fine-tuning. The above sets our method apart in theory and application.

We create a modular customization baseline inspired by OFT. In the spirit of our method, we constrain OFT to a pre-determined subspace. We achieve this by picking certain blocks in OFT to be trainable, and fixing the rest to be the identity. Combining such "constrained" OFTs would then be done by sequentially applying each rotation (shown right).



For OFT, we set $r = 8$ to match the parameter count of our experiments. As shown in Tab. 3, OFT* still faces issues in identity metrics post-merging. We believe adapting OFT for modular customization is a promising new direction, and we are happy to add relevant baselines and discussion surrounding OFT to the final submission.

Method	Merge Time	Text Alignment ↑			Image Alignment ↑			Identity Alignment ↑		
		Single	Merged	Δ	Single	Merged	Δ	Single	Merged	Δ
OFT*	<1 s	.639 →	.646	+.007	.731 →	.715	-.016	.718 →	.681	-.037
Ours	<1 s	.624 →	.644	+.020	.748 →	**.741**	-.007	.740 →	**.745**	+.005

Table 3. Quantitative comparison of OFT* with our method.

Extended baseline comparisons. In Fig. 11 We show an extended version of Fig. 6 with generated images of each identity for each method before they are merged. These results aim to show that our method is capable of retaining identity alignment with the target concept before and after merging, while achieving merging of individual LoRAs instantly without any further fine-tuning or optimization stages.

Over-fitting. Since we are fine-tuning our network over a small custom dataset and we initialize our custom tokens with a user-defined class label, it may be susceptible to over-fitting. Prior works such as DreamBooth [37] and Custom Diffusion [24] alleviate this effect by adding a class preservation loss that ensures generating images from the class token still produces diverse results. In our method, we do not employ an explicit loss to prevent over-fitting, however, we found that our fine-tuned models still preserve the ability to generate diverse images for the trained class label as shown in Fig. 12

10. Limitations and future work

Our method takes an important step towards achieving modular customization. However, a few important limitations should also be addressed in future work.

Generating multiple custom concepts within the same image remains challenging. Simply prompting a merged model with multiple custom tokens usually leads to incoherent hybrids of both objects. Prior works [12] have explored spatial guidance for better disentangling concepts in a single generation, and we have also employed similar techniques to generate our results. However, these methods still lead to failure cases as illustrated in Fig. 10. Concepts are often ignored, or attributes can leak to neighboring concepts. Future work should aim to address these struggles to further enable multi-concept generations.

Storing individual LoRAs, even those trained with our method can also be expensive. Although LoRAs are already compressive due to their low-ranked nature, storing a large bank of concepts for modualr customization can still be expensive. Works such as SVDiff [13] takes steps towards further compressing LoRAs while maintaining fidelity of generated images. However, our method does not naturally fit in with the SVDiff method, implying the need for a tailored compressing methodology.

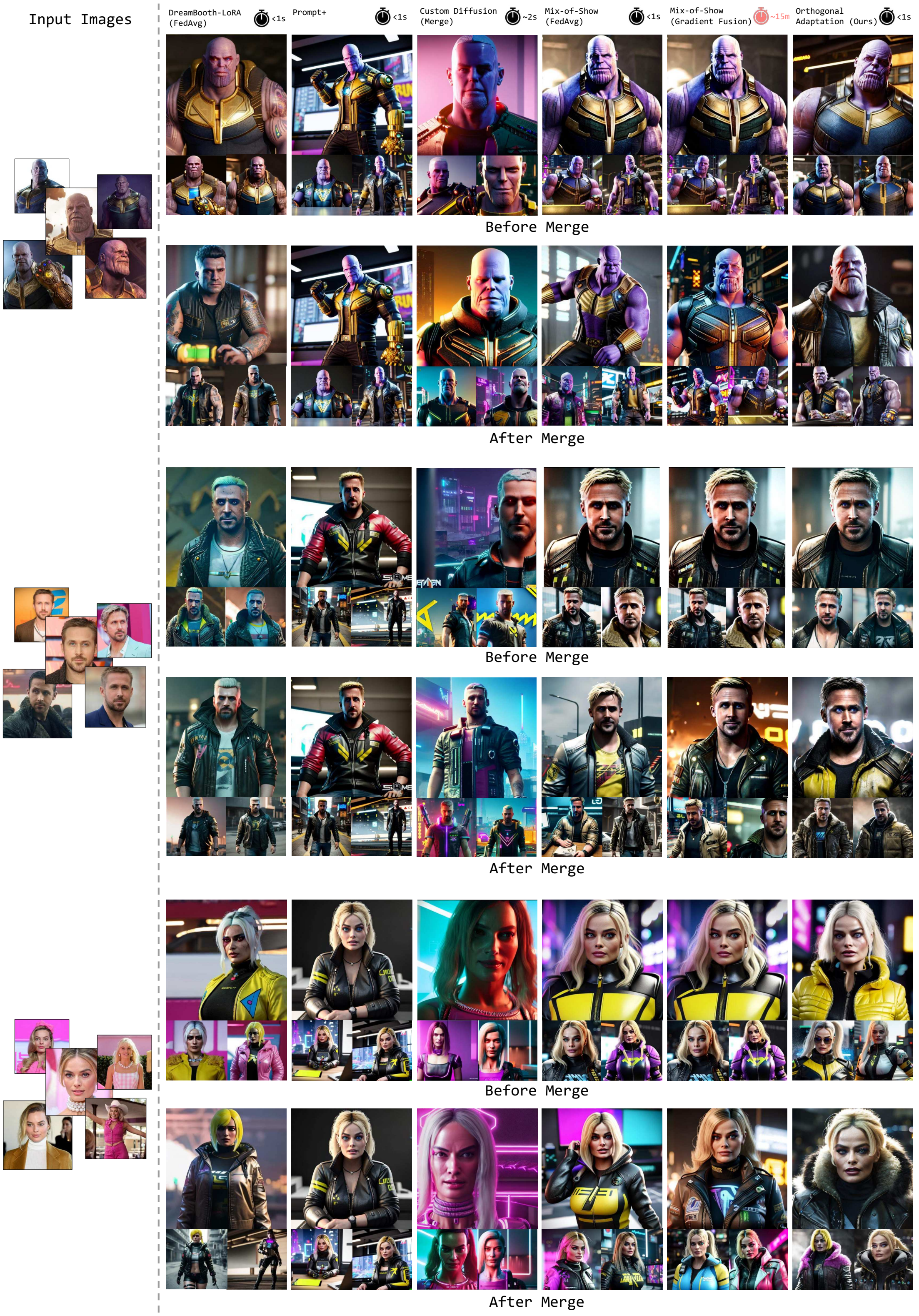


Figure 11. **Extended multi-concept results.** We show results for each method before and after merging the individually trained models into a single, merged model. Our method is able to capture the target identity with high fidelity before and after the merging process, while keeping the merging process instantaneous.

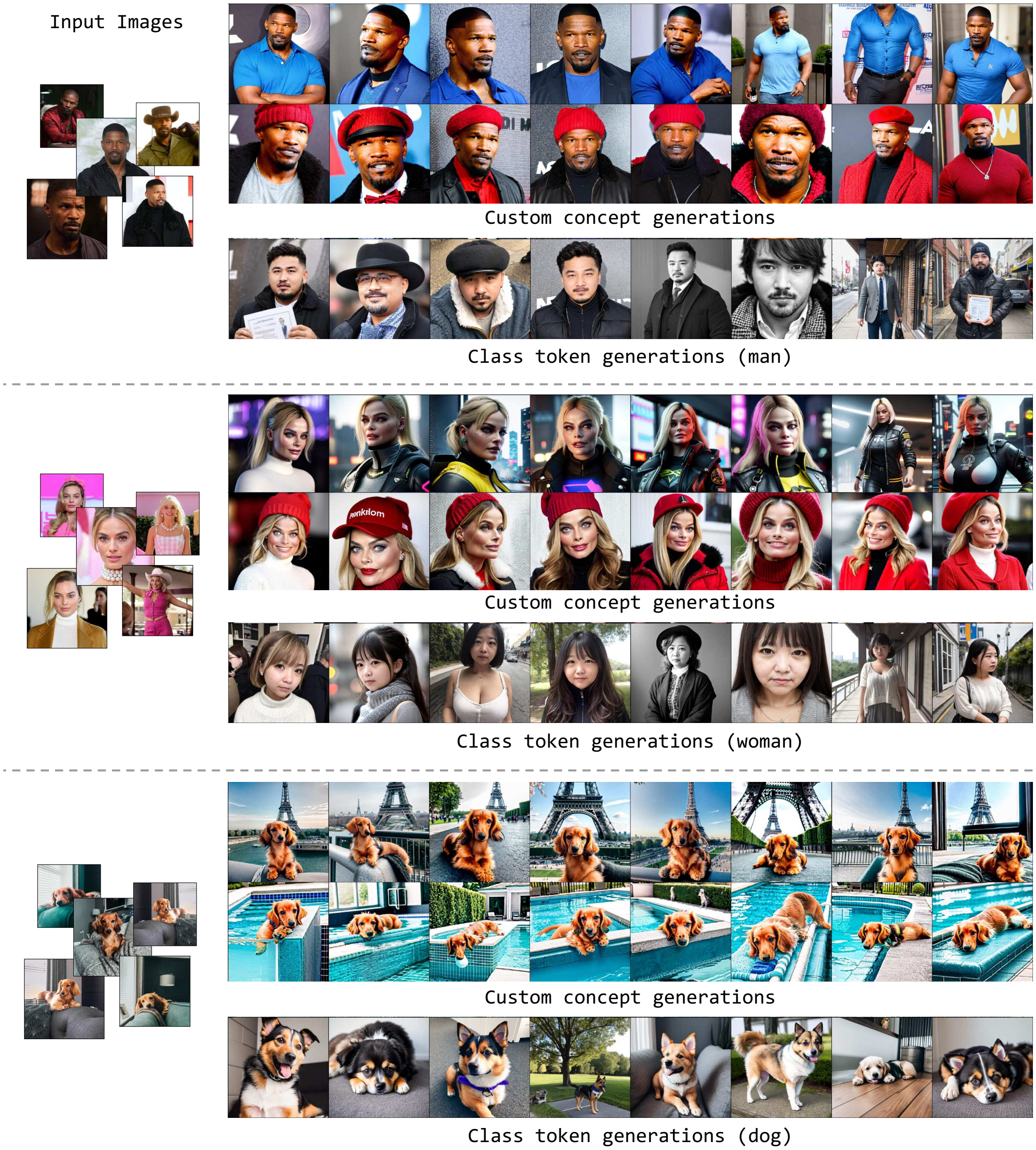

Input Images
Custom concept generations
Class token generations (man)
Custom concept generations
Class token generations (woman)
Custom concept generations
Class token generations (dog)

Figure 12. **Preservation of class label.** Although our method does not enforce an explicit class preservation loss similar to prior works [24, 37], our method is able to preserve diversity when generating images of the class label used for initialization of the custom concept token. We show this across three different classes, namely: *man*, *woman*, and *dog*.