DreamCatcher: Efficient Multi-Concept Customization via Representation Finetuning

Jungwon Lee Changhun Lee Eunhyeok Park
Pohang University of Science and Technology (POSTECH), Republic of Korea
{leejungwon, changhun.lee, eh.park}@postech.ac.kr

Figure 1. The comprehensive comparison of multi-concept generation methods. **DreamCatcher** overcomes the limitations (written in red) by updating partial representation using concept masking, avoiding both fusion and inference overhead. These examples are generated by the prompt "<concepts> in the simple background". The corresponding reference images for each concept are provided in the Appendix.

Abstract

Recent advances in customizing Text-to-Image models allow users to generate personalized images with just a few samples. As demand for multi-concept generation grows, methods using weight fusion and test-time optimization have emerged, integrating multiple concepts within a single image. However, these approaches inject concept knowledge into the parametric space, leading to high overhead in multi-concept generation. We introduce DreamCatcher, an efficient framework based on representation finetuning. Our key innovation embeds conceptual information into the feature space, achieving up to 5× faster multi-concept generation while reducing learnable storage per concept by 88%, all without quality loss. Besides, our method is highly versatile, enabling personalized inpainting without additional training.

1. Introduction

Text-to-Image (T2I) generative models [25, 31, 32, 35] have shown remarkable capability in generating high-quality images that accurately capture the contextual details of a given text prompt. A particularly compelling application is concept customization, where a model trained on a specific concept with just a few examples can generate contextually relevant images while preserving the integrity of the learned concept. This enables users to create and modify images based on their own photographs, fostering personalized expression and enhancing commercial appeal.

Method	Memory Efficient	Fast Fusion	Fast Multi-Generation
Mix-of-Show [5]	▲	✗	✓
Orthogonal Adaptation [28]	✗	✓	✓
LoraComposer [45]	▲	✓	✗
DreamCatcher (Ours)	✓	✓	✓

Table 1. Comparison of representative modular customization. The three symbols mean ✓(Good), ✗(Bad), ▲(Intermediate).

To customize T2I models, many methods have focused on integrating conceptual information within the parametric space by finetuning either the model's weights [14, 33] or text embeddings [4, 39]. For efficiency, Parameter-Efficient Fine-Tuning (PEFT) techniques, particularly Low-Rank Adaptation (LoRA) [10], have gained prominence, enabling the learning of new concepts with only a modest parameter increase (5–10 MB). This approach has emerged

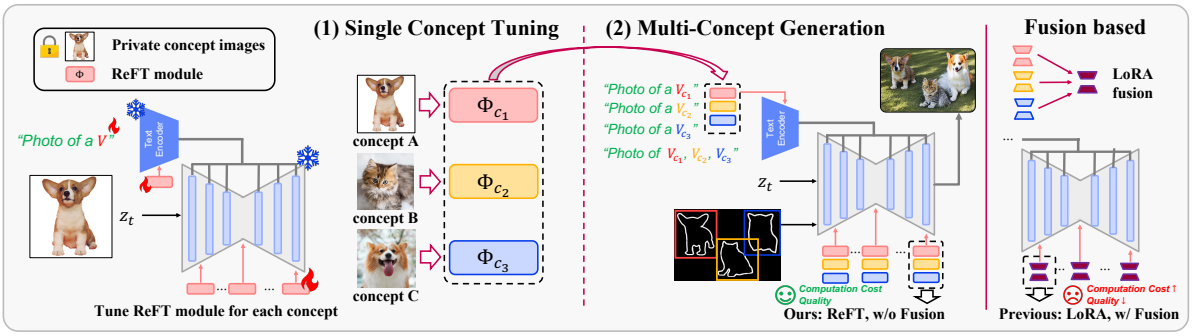

Figure 2. **Pipeline Overview.** (1) Our method independently tunes each concept using the efficient ReFT module. (2) By combining these concept modules, we generate multi-concept images that seamlessly incorporate all desired concepts. Unlike fusion-based methods using LoRA, our approach eliminates the need for fusion, thereby avoiding the associated computational overhead and performance degradation.

as the dominant paradigm in concept customization.

The growing popularity of custom concept generation has naturally expanded to include generating multiple concepts within a single image, leading to the development of multi-concept generation. However, multi-concept generation poses a greater challenge due to the difficulty of preserving the unique characteristics of each concept. Recent studies have proposed various approaches, such as fusion-based methods that combine multiple LoRA weights into a single weight (e.g., Mix-of-Show [5], Orthogonal Adaptation [28]), updating concepts through separate forward passes for each concept during inference, and test-time optimization techniques [13, 15, 45] that adjust the latent representation based on gradients during the sampling process. While these methods have enabled high-quality image generation in multi-concept scenarios, their expensive costs limit their practicality outside high-performance servers, as summarized in Figure 1, making it hard to be profitable.

We argue that these cost overheads stem from a fundamental limitation of current methods that encode concept characteristics within **parametric space**. Since these parameter update are available only during the expensive training phase, generating multiple concepts in a single image after training requires either the stage of merging weights or the complex running multiple forward-backward operations for new concept combinations, both of which demand significant computational resources [13].

In this work, we introduce DreamCatcher, a new paradigm for T2I customization. To enable efficient multi-concept generation without compromising quality, we propose Concept-Aware Intervention (CAI) directly modifying the **feature space** for each concept. By training intervention adapters tailored to specific concepts, we can inject concept-specific information into intermediate features during generation. As illustrated in Figure 2, our feature space modification enables multi-concept generation without requiring any fusion process. This approach ensures high-quality images with clearly separated concept characteristics, allowing multiple concepts to coexist within a single image through disjoint feature manipulation. Additionally, our softmax scheduling adaptively adjusts attention scores during intervention to suppress unwanted artifacts.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to demonstrate the efficiency of feature-space methods for T2I customization.
- For single-concept, DreamCatcher achieves comparable performance using up to 88% less memory than baselines.
- For multi-concept, it enables image generation up to 5× faster without additional optimization.
- It supports personalized inpainting without extra training.

2. Related Work

2.1. Concept Customization

Concept customization leverages the representational power of pretrained T2I models, such as Stable Diffusion [32] and SDXL [29], to generate images with diverse contexts for user-desired objects. In Dreambooth [33], customization is performed by finetuning the entire weights of the diffusion model using only a small number of concept images. Custom Diffusion [14], on the other hand, performs customization by training only a part of the model, specifically the cross-attention layers. However, since these methods require updating the entire or large portion of weights, the cost of finetuning and storage per concept is quite significant (e.g., over 3GB for Dreambooth and 74MB for Custom Diffusion in the case of Stable Diffusion v1.4). Recently, methods utilizing PEFT have become widely adopted, allowing for concept customization with much lower memory requirements [30]. Text Inversion (TI) [4, 23] performs customization by learning new text embeddings for the text encoder that represents the novel concept. P+ [39] extends TI by learning distinct text embeddings for each layer, thereby enhancing the model's expressive ability for customization. Additionally, other research directions [11, 12, 19, 34, 36, 41, 46, 49] focus on reducing the additional finetuning cost for new concepts by employing an image encoder that takes

a target image as input and learns supplementary attention layers alongside text prompts.

2.2. Multi-Concept Generation

Beyond single-concept customization, recent studies have focused on generating images by combining multiple concepts into a single image. In Custom Diffusion [14], multiple concepts were learned jointly, allowing for simultaneous training of several concepts. Recently, due to privacy issues restricting access to training data, modular customization methods have been explored, enabling the combination of individually learned concepts without centralized access to training data. Text Inversion [4, 23] and P+ [39] allow modular customization by combining independently learned text embeddings. Custom Diffusion achieved concept combination through the constrained customization problem. In FedAvg [22], modular customization is made possible through weight fusion, which combines separately trained LoRA [10] models into a unified set of weights. However, the fusion process in FedAvg results in an identity loss of the learned concepts. To mitigate the loss incurred during weight fusion, Mix-of-Show [5] employs gradient fusion to find weights that better preserve the identity of each concept, thereby reducing concept identity loss. Orthogonal Adaptation [28, 50] proposes adding orthogonal constraints to the LoRA weights, reducing conflicts between concepts after weight fusion. Another approach is test-time optimization [8, 13, 15, 45], which aims to ensure there is no interference between concepts during inference time.

All aforementioned approaches rely on parametric update for each concept, thereby fusing concepts in a single image is inevitably expensive. In this work, we address this inefficiency by proposing a novel idea of feature-level intervention per concept, which enables the rapid generation of high-quality images with low memory usage.

2.3. Parameter Efficient Finetuning (PEFT)

PEFT [9, 17, 18, 20, 48] was proposed to reduce the finetuning cost of Large Language Models (LLMs) and a notable example is LoRA [10], which reduces finetuning costs by adding a low-rank decomposed weight alongside with the original weight and updating only the added path. Because it significantly reduces the number of learnable parameters while LoRA-based finetuning still provides substantial quality improvements, the LoRA-based approach has been widely adopted in diffusion models [5, 21, 37].

On contrary, as an alternative PEFT approach, representation finetuning methods have been proposed in the LLM domain, which focus on finetuning representations directly. BitFit [48] and ReD [42], for instance, conduct finetuning by learning the scale or bias of hidden representations. LoFiT [47] performs finetuning by searching for the most effective attention heads and learning biases specific to those heads. Recently, ReFT [43] introduces a new PEFT framework that transforms activations in a low-rank subspace instead of weights, achieving superior memory efficiency and finetuning quality. In this work, we highlight, **for the first time**, the usefulness of this representation finetuning for model customization.

3. Method

Our DreamCatcher presents a novel framework capable of handling both single-concept and multi-concept image generation. We first introduce the core idea by demonstrating how feature-level intervention enables efficient implementation for single-concept image generation. Then, we extend this approach to multi-concept generation by explaining how individually trained intervention adapters can be modularly combined (Figure 2), allowing for flexible and scalable concept composition. In the following sections, we provide an in-depth explanation of each pipeline.

3.1. Single Concept Tuning

In diffusion models, various efforts have been made to modify representations to adjust visual characteristics. Notably, ControlNet [24] demonstrated the effectiveness of representation manipulation in image generation by adding a feature conditioned on an input image to the feature map, thereby controlling specific regions of the generated image. Similarly, in image editing, approaches have been developed to alter the properties of generated images by adjusting representations [3, 6, 16, 27, 38].

Building on these insights, we introduce the single concept tuning process of DreamCatcher, as illustrated in Figure 2. In DreamCatcher, the customization of the pretrained model follows the methods of Mix-of-Show [5], where each object is assigned a unique token, and its embedding is iteratively updated. In this work, we replace the LoRA-based module with a feature intervention module strategically interleaved after cross-attention layers to inject concept-specific information explicitly. To the best of our knowledge, this direction remains unexplored and offers distinct advantages over existing methods. To clarify our contribution, we first introduce the foundational PEFT techniques for feature intervention and explain how we leverage them.

3.1.1. Preliminary: ReFT

ReFT [43] is a representation finetuning technique that has recently gained attention for Large Language Models (LLMs). It operates by projecting the model's hidden representation $h \in \mathbb{R}^{d \times c}$ into a subspace using a low-rank projection, performing intervention within this subspace, and then restoring the representation to its original space. ReFT introduces two primary implementations for finetuning. The

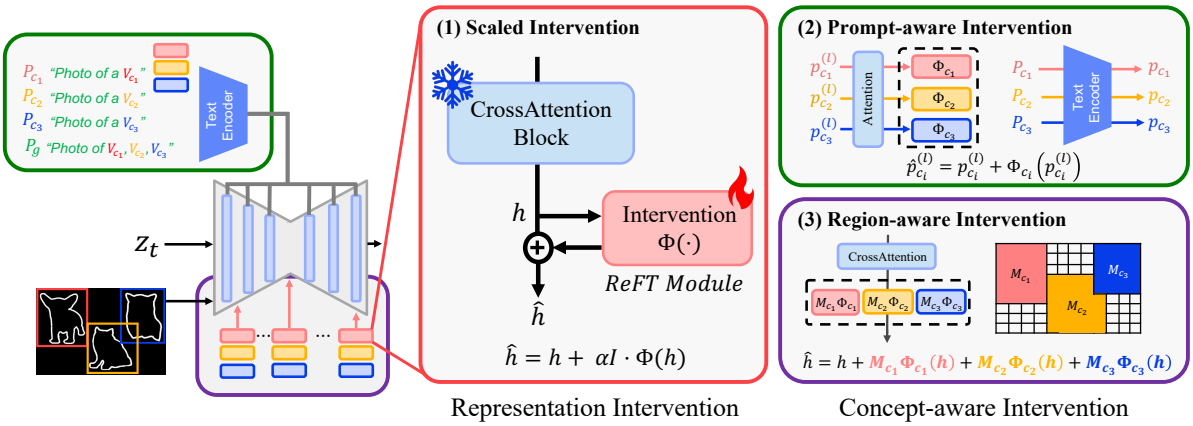

Figure 3. **Method Overview.** This is an example using three concepts, with the index of each concept indicated by subscripts. (1) Our modified ReFT module intervenes the representations of each concept region. (2) Prompt-aware Intervention. (3) Region-aware Intervention.

first one is Low-rank Linear Subspace ReFT (LoReFT):

$$\Phi_{LoReFT}(h) = R^T(Wh + b - Rh), \tag{1}$$

which consists of an orthogonal matrix $R \in \mathbb{R}^{r \times d}$ and a linear layer $W \in \mathbb{R}^{r \times d}, b \in \mathbb{R}^r$. It projects the representation into a linear subspace via the orthogonal matrix R, performs intervention in this subspace, and then restores it to the original space using R. The second approach is DiReFT:

$$\Phi_{DiReFT}(h) = W_2^\top(W_1 h + b). \tag{2}$$

DiReFT consists of low-rank projection matrices, W_1 and $W_2 \in \mathbb{R}^{r \times d}$, along with a bias term $b \in \mathbb{R}^r$. Unlike LoReFT, there are no constraints such as orthogonality on W_1 and W_2; they are fully learnable. It can be interpreted as a LoRA version for representation finetuning.

$$\hat{h} = h + \Phi(h) \tag{3}$$

In both implementations, the subspace transformation embeds essential task-related information and we can manipulate this space via intervention.

3.1.2. Scaled ReFT for Concept Customization

ReFT was initially proposed for the LLM domain, but we leverage its capability in models for the diffusion process, *where it has not been applied before.* Through empirical exploration, we find that this approach also works well for our application. However, we also discovered that a simple extension further enhances the benefits of training.

Scaled ReFT We propose the modified ReFT method optimized specifically for image processing. For both LoReFT and DiReFT, we discovered that *adding a channel-wise scaling* $\alpha \in \mathbb{R}^d$ after mapping back to the original space significantly enhances quality, expressed as:

$$\hat{h} = h + \alpha I \cdot \Phi(h). \tag{4}$$

Our intuition is that intervention in the subspace alone is insufficient to modulate high-rank information. Therefore, by adding learnable parameters for each channel, we can now amplify the channel-wise instance-specific representation, enabling better projection of each concept's characteristics. While these additional parameters take only a small proportion compared to the original learnable ReFT parameters, they play an important role during finetuning (Section 4.7), which led us to adopt this as our fundamental module.

Cross-Attention Intervention Next, we aimed to determine the optimal placement of the designed ReFT module to achieve the best quality-to-cost ratio. In LoRA-based approaches [5, 28], the LoRA adapter is applied to all linear layers in the attention block, where interventions directly take effect, requiring four LoRA modules per attention block. While adding the proposed module to all layers yields the highest image quality, it may not be the most optimal design choice, and the unique properties of ReFT might necessitate a different approach. To gain further insights, we analyzed the Singular Value Decomposition (SVD) of attention blocks at various points within the U-Net structure.

From the observations in Figure 4, we draw two main conclusions. First, the activations within attention blocks exhibit a notably low rank, suggesting that hidden representations may be projected into a low-dimensional subspace, making low-rank methods like ReFT highly suitable. Second, the rank of Cross-Attention is significantly lower than that of Self-Attention across all points, indicating that ReFT tuning with low-rank values could be particularly effective for Cross-Attention layers. Our empirical study also shows that applying ReFT solely to the output features of Cross-Attention is sufficient to achieve high-quality customization. This is intuitive as the concept semantics are embedded in the specific tokens in the text encoder, thereby the instance-wise intervention needs to be applied via cross-

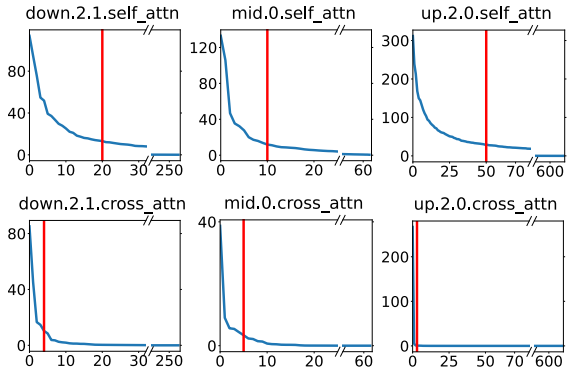

Figure 4. SVD Analysis for the Attention outputs. SVD top ranks (X-axis) and singular values (Y-axis), with the red line marking 15% of the maximum value. This indicates that cross-attention outputs can be represented with much lower rank.

attention operation. Adding ReFT to this part guarantees high-quality output with minimal cost, so we use this approach as our base implementation.

By integrating scaled ReFT into the cross-attention layers and updating the modified model using a few samples for the target concept, along with layer-wise text embedding [39], our model learns to generate images that incorporate the specific concept. This overall process is referred to as **single-concept tuning**. Note that in our study, we train adapters for each concept independently to mitigate privacy concerns.

3.2. Multi-Concept Generation

3.2.1. Concept-aware Intervention

Following the pipeline in Section 3.1, we fine-tune the scaled ReFT adapter for each concept. For multi-concept generation, we employ a plug-and-play approach called **Concept-aware Intervention**, utilizing each fine-tuned adapter. Additionally, we introduce a mechanism to prevent concept blending during generation, addressing both the Text Encoder and U-Net. Notably, this approach removes the need for a separate fusion step, making multi-concept generation in DreamCatcher training-free and enabling direct inference with the following proposed mechanisms.

Prompt-aware Intervention In this section, we describe how the text encoder separates per-concept prompt embeddings in multi-concept generation. In [5], region-aware cross-attention is employed to distinguish the influence of each concept across different regions, as defined by the region mask M. Building upon this idea, we also leverage an instance-specific mask to constrain the influence of per-concept feature intervention. In our model, the embeddings of the global prompt P_g and each concept prompt P_c are obtained through the text encoder. When extracting prompt embeddings from the text encoder, independent ReFT modules are applied to the cross-attention layers for each concept, ensuring that interventions are specific to each prompt based on the region mask:

$$p_{c_i} = \text{TextEncoder}(P_{c_i}) \tag{5}$$

$$h^{(l)}[M_{c_i}] = M_{c_i} \otimes \text{softmax}(\frac{Q(z^{(l)})K^{\top}(p_{c_i})}{\sqrt{d}}) \cdot V(p_{c_i}), \tag{6}$$

where i and l denote the concept and layer index, respectively. We refer this tailored intervention as Prompt-aware Intervention (PAI) (Figure 2 (2)). Through this method, we can obtain distinct prompt embeddings for each concept:

$$\hat{p}_{c_i}^{(l)} = p_{c_i}^{(l)} + \Phi_{c_i}^{(l)}(p_{c_i}^{(l)}). \tag{7}$$

Region-aware Intervention In addition, we propose Region-aware Intervention (RAI), which updates representations only within the masked regions on diffusion path:

$$\hat{h}^{(l)} = h^{(l)} + \Sigma_{i=1}^{N}(M_{c_i} \otimes \Phi_{c_i}^{(l)}(h^{(l)})). \tag{8}$$

In the RAI, the N ReFT modules trained for each concept exist independently, and interventions are performed only to the specific regions according to each concept mask, as shown in Figure 3 (3). Note that the per-concept embedding is obtained through PAI, and it only affects the corresponding masked region on the target image. This prevents concept mixing without the need for a specific fusion process.

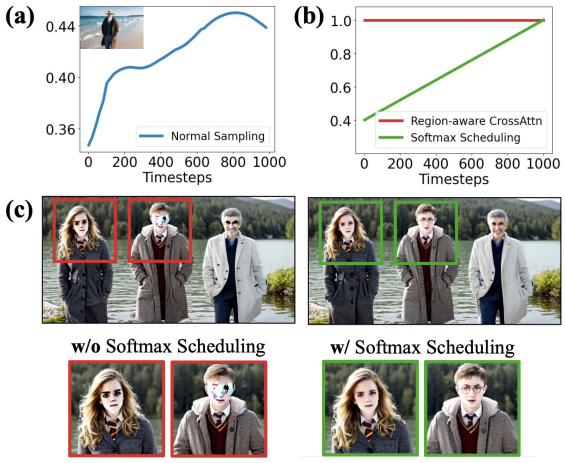

Figure 5. The concept of Softmax Scheduling and the qualitative example of using it. The changes in attention score as time steps increase for (a) normal sampling and (b) the concept region. (c) Qualitative results on the impact of Softmax Scheduling.

3.3. Softmax Scheduling

Region-aware cross-attention enables complex compositions by directing attention to specific regions, thereby minimizing interference between concepts. However, as shown

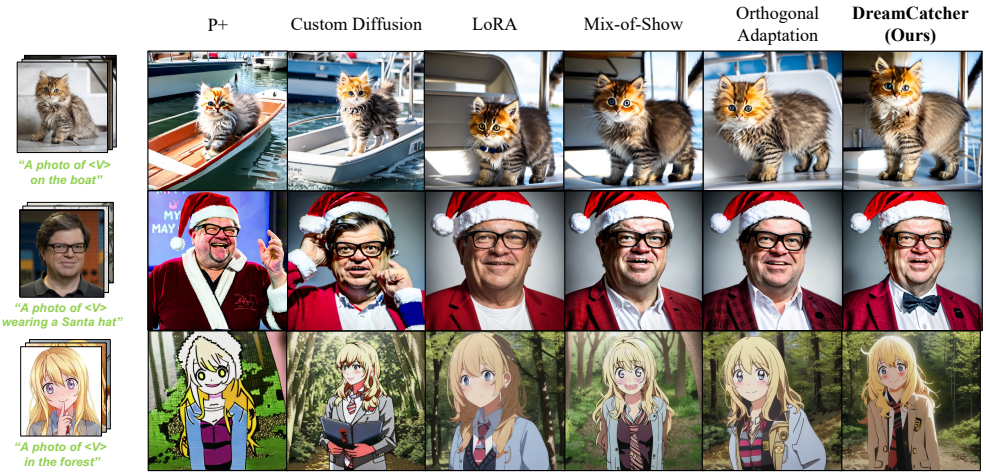

Figure 6. **Single-Concept Generation results.**

in Figure 5 (c), it has the drawback of potentially introducing artifacts due to excessive attention on small masked regions from the initial sampling timesteps. To accurately investigate this issue, we measured the attention scores in the regions where concepts are generated during the actual inference process. In Figure 5 (a), the attention score typically starts at a low value in the early stages of sampling, where noise is abundant, and gradually increases as sampling progresses. However, as seen in Figure 5 (b), region-aware cross-attention exhibits excessively high attention scores from the start of sampling, which is not matched to the real trend. To address this, we propose a Softmax Scheduling technique which modulates attention scores based on timestep t to reduce the "attention exploding" effect on localized regions caused by Softmax in the early stages.

$$AttentionScore = \lambda(t) \cdot \text{softmax}(\frac{QK^\top}{\sqrt{d_k}}) \tag{9}$$

$$\lambda(t) = \beta + (1 - \beta) \times (t/1000) \tag{10}$$

Depending on the scheduling factor β, Softmax Scheduling assigns relatively low attention scores during the initial sampling phase of the diffusion model. As noise diminishes and finer details emerge, the attention score gradually increases in the relevant regions. As shown in Figure 5 (c), this approach effectively reduces artifacts.

4. Experiments

To validate the effectiveness of the proposed method, we compared the performance of Single / Multi-Concept generation with various methods.

4.1. Experiment Setting

The baseline methods using parametric approaches apply LoRA [10] to all linear layers, with Mix-of-Show [5] using a rank of 4 and Orthogonal Adaptation [28] using a rank of 20. DreamCatcher applies scaled ReFT to the attention output of the text encoder and the cross-attention output of the U-Net, using ranks of 4 and 8, respectively. The scheduling factor β is set to 0.6.

Dataset We use the dataset from Mix-of-Show [5] for single concept tuning, which includes 6 objects, 6 real characters, and 5 anime characters, each comprising 5 to 15 images per concept. For multi-concept generation, we utilize sketches and pose conditions containing combinations of 2 to maximum 5 concepts from each category.

Metrics For evaluation, we mainly use two metrics from CLIP scores [7], which are widely used for concept customization. CLIP score measures similarity within the CLIP embedding space. CLIP-I indicates how well the generated image aligns with the concept identity of the reference image, while CLIP-T measures the alignment between the generated image and the text prompt. Usually, CLIP-T and CLIP-I have a trade-off relationship, but **CLIP-I** is crucial for estimating visual semantics. In multi-concept generation, we followed [45] to measure CLIP score: We crop each individual concept from the generated image, generating over 1000 images to evaluate performance.

4.2. Single Concept Tuning Results

Table 2 presents the results for single concept tuning. Our method demonstrates superior parameter efficiency compared to existing approaches, achieving outstanding results across all three evaluation categories. Unlike other fusion-based methods, which require time overhead for fusion and often suffer from performance degradation of post-fusion for each concept, our approach is unaffected by these issues, ensuring consistent performance. Our method with scaled ReFT modules based on DiReFT achieves a higher

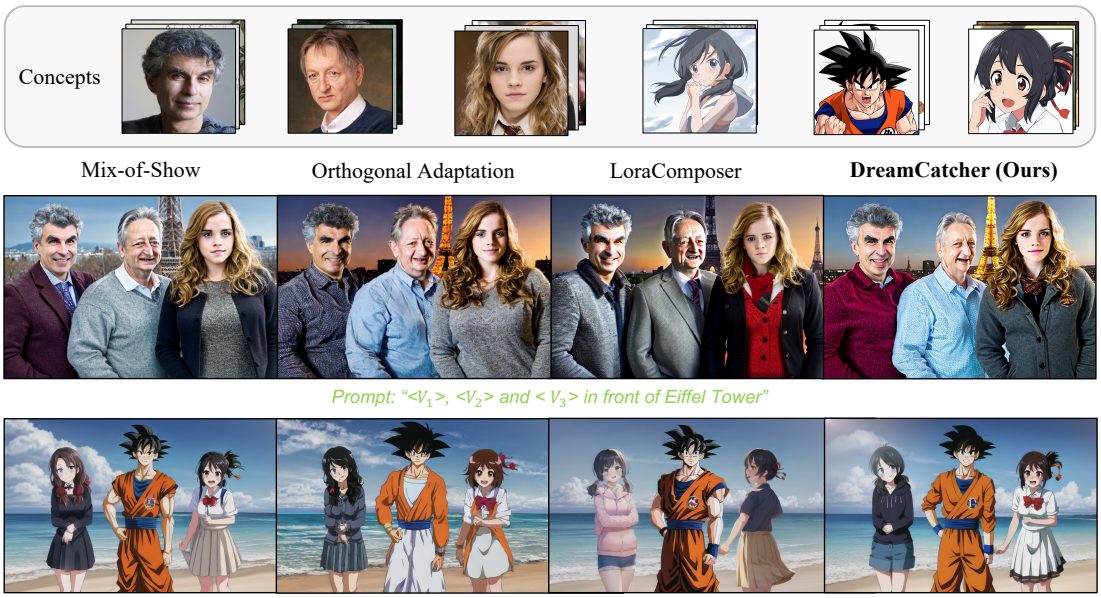

Figure 7. **Multi-Concept Generation results.**

	Methods	Real-Objects Single→Fused	Real-Characters Single→Fused	Anime-Characters Single→Fused	Storage (MB)	Fusion Time
CLIP-T	P+ [39]	0.776→0.776 **(-)**	0.684→0.684 **(-)**	0.710→0.710 **(-)**	0.05	<1s
	Custom Diffusion [14]	0.742→0.744 **(+0.02)**	0.661→0.659 **(-0.02)**	0.785→0.747 **(-0.038)**	74.0	∼2s
	LoRA [33]	0.748→0.795 **(+0.047)**	0.681→0.725 **(+0.044)**	0.756→0.760 **(+0.04)**	4.3	<1s
	Mix-of-Show [5]	0.728→0.749 **(+0.021)**	0.638→0.667 **(+0.029)**	0.701→0.712 **(+0.011)**	4.4	∼15m
	Orthogonal Adaptation [28]	0.733→0.735 **(+0.02)**	0.654→0.681 **(+0.027)**	0.707→0.709 **(+0.002)**	11.2	<1s
	Ours (LoReFT)	0.730→0.730 **(-)**	0.640→0.640 **(-)**	0.705→0.705 **(-)**	**1.3**	<1s
	Ours (DiReFT)	0.715→0.715 **(-)**	0.632→0.632 **(-)**	0.694→0.694 **(-)**	**1.3**	<1s
CLIP-I	P+ [39]	0.777→0.777 **(-)**	0.685→0.685 **(-)**	0.748→0.748 **(-)**	0.05	<1s
	Custom Diffusion [14]	0.839→0.822 **(-0.017)**	0.718→0.686 **(-0.032)**	0.790→0.750 **(-0.040)**	74.0	∼2s
	LoRA [33]	0.845→0.782 **(-0.063)**	0.735→0.617 **(-0.118)**	0.794→0.756 **(-0.038)**	4.3	<1s
	Mix-of-Show [5]	0.861→0.847 **(-0.014)**	0.789→0.747 **(-0.042)**	0.823→0.807 **(-0.016)**	4.4	∼15m
	Orthogonal Adaptation [28]	0.852→0.848 **(-0.004)**	0.771→0.752 **(-0.021)**	0.821→0.806 **(-0.015)**	11.2	<1s
	Ours (LoReFT)	0.865→0.865 **(-)**	0.792→0.792 **(-)**	0.829→0.829 **(-)**	**1.3**	<1s
	Ours (DiReFT)	0.868→0.868 **(-)**	0.791→0.791 **(-)**	0.835→0.835 **(-)**	**1.3**	<1s

Table 2. **Single-Concept Generation results.** The arrows indicate the performance change before and after fusion in fusion-based methods. While our control group methods (except for P+) experience a performance gap between the single tuned model and the fused model, our method maintains the performance since ours is not based on fusion.

Methods	CLIP-I ↑			CLIP-T ↑			Latency	Fusion Time
	Objects	Real	Anime	Objects	Real	Anime		
P+ [39]	0.789	0.727	0.687	0.679	0.575	0.564	5s	<1s
Custom Diffusion [14]	0.814	0.719	0.735	0.679	0.578	0.573	5s	∼2s
LoRA [33]	0.808	0.663	0.751	0.681	0.572	0.581	5s	<1s
Mix-of-Show [5]	0.810	0.747	0.771	0.672	0.577	0.573	5s	∼15m
Orthogonal Adaptation [28]	0.814	0.746	0.769	0.695	0.579	0.571	5s	<1s
LoraComposer [45]	0.843	0.775	0.791	0.671	0.571	0.597	25s	<1s
Ours (LoReFT)	0.836	0.755	0.784	0.669	0.579	0.585	5.4s	<1s
Ours (DiReFT)	0.837	0.764	0.795	0.667	0.576	0.581	5.4s	<1s

Table 3. **Multi-Concept Generation results.** Latency and fusion time results were measured by handling six concepts together.

CLIP-I score compared to LoReFT, while LoReFT yields a lower CLIP-T score. This suggests that the orthogonal constraint in LoReFT provides a regularization effect, leading to a well-balanced performance. Furthermore, as shown in Figure 6, ReFT effectively captures concept identity and detail across all three categories, even with fewer parameters.

4.3. Multi-Concept Generation Results

Table 3 presents the results for multi-concept generation. Our method consistently outperforms other fusion-based approaches. This is primarily due to the fact that fusion-based techniques tend to weaken concept identity during the fusion, whereas our fusion-free approach preserves concept integrity, thereby producing higher-quality images. Although LoraComposer [45], a concurrent work, achieves slightly higher scores due to gradient updates in the latent space, it requires gradient computation during inference, resulting in higher memory usage and 5× longer generation times, making it unsuitable for mobile deployment. Although our method incurs a slight increase in latency due to the intervention process, this increase is negligible, allowing for an efficient yet powerful multi-concept generation compared to other approaches. As illustrated in Figure 7, our method demonstrates superior multi-concept generation capabilities across both real and animated characters.

4.4. Evaluation on Additional Metrics

In addition to CLIP score, we provided the additional evaluation results incorporated four metrics—DINO v2 [26], FaceID [46], CLIP-IQA [40], and KID [1]—to more thoroughly measure both identity fidelity and image quality. Table 4 presents the results of our multi-concept generation experiments for real-world characters. Our method outperforms the baseline across all four metrics, demonstrating superior identity preservation as well as higher visual quality.

Method	DINO v2 ↑	FaceID ↑	CLIP-IQA ↑	KID ↓
Mix-of-Show	0.4286	0.3702	0.8259	0.0702
Orthogonal Adaptation	0.4042	0.3512	0.8274	0.0756
Ours	**0.4341**	**0.4473**	**0.8364**	**0.0656**

Table 4. Evaluation Results on Four Metrics.

4.5. User Study

To complement quantitative results, we conducted a user study with 20 test examples. For each example, 26 participants compared outputs from different methods and consistently preferred DreamCatcher, demonstrating its clear advantage over all baselines.

Method	Mix-of-Show	Orthogonal Adaptation	**Ours**
Win rate	26.5%	28.8%	44.7%

Table 5. Human Evaluation Results.

	CLIP-I ↑	CLIP-T ↑
Encoder	0.746	0.653
Decoder	0.773	0.646
Mid Blocks	0.775	0.645
Ours (All)	0.792	0.640

Methods	CLIP-I ↑	CLIP-T ↑
BitFit [48]	0.654	0.703
ReD [42]	0.707	0.736
LoReFT [43]	0.778	0.643
DiReFT [43]	0.789	0.638
Ours (LoReFT)	0.792	0.637
Ours (DiReFT)	0.791	0.634

Table 6. (Left) Comparison according to intervention position. (Right) Comparison with other intervention methods. Both experiments are conducted on real characters category in single-concept.

4.6. Personalized Image Inpainting

Another application of DreamCatcher is personalized image inpainting without additional training, enabled by region-aware intervention for partial representation updates. Personalized image inpainting [2, 44] generates customized content within a masked region. Implementation details and results are provided in the Appendix.

4.7. Ablation Study

In this section, we conduct two ablation studies. First, in Table 6 (Left), we examine the effect of intervention placement by comparing encoder, decoder, and mid blocks (the five blocks surrounding the mid block). Applying interventions only to the encoder proved difficult to train, while interventions in the decoder and mid blocks improved concept learning. Incorporating interventions across all blocks achieved the best performance, suggesting that each block contributes differently to concept learning.

The second ablation study, illustrated in Table 6 (Right), compares alternative representation finetuning methods. BitFit [48] and ReD [42], which apply scale and bias adjustments to representations, show limited effectiveness for concept customization, particularly under few-shot settings. Conversely, ReFT generates superior image quality. Furthermore, our scaled ReFT method, which incorporates an additional channel scaling factor, significantly outperforms the others, achieving markedly higher scores.

5. Conclusion

In this work, we introduced DreamCatcher, a novel and efficient approach for multi-concept customization in T2I generative models. This innovative customization framework, based on feature intervention, achieves superior image quality while maintaining efficiency. Moreover, our concept-aware intervention ensures precise control over multiple concepts, preventing unwanted blending and accelerating generation. We extend this framework to personalized image inpainting without additional training, demonstrating its adaptability. DreamCatcher has a distinct advantage for a wide range of real-world applications where lightweight and high-quality multi-concept generation is essential.

Acknowledgement

This work was supported by IITP and NRF grant funded by the Korea government(MSIT) (RS-2019-II191906, RS-2024-00457882) and Samsung Research Global AI Center.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 8

[2] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 8

[3] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*. 1, 2, 3

[5] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 6, 7

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 2022. 3

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[8] Trong-Vu Hoang, Quang-Binh Nguyen, Thanh-Toan Do, Tam V Nguyen, Minh-Triet Tran, and Trung-Nghia Le. Showflow: From robust single concept to condition-free multi-concept generation. *arXiv preprint arXiv:2506.18493*, 2025. 3

[9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[10] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. 1, 3, 6

[11] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3707–3714, 2025. 2

[12] 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

[13] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024. 2, 3

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3, 7

[15] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8880–8889, 2024. 2, 3

[16] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*. 3

[17] Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13355–13364, 2024. 3

[18] Changhun Lee, Jun-gyu Jin, Younghyun Cho, and Eunhyeok Park. Qeft: Quantization for efficient fine-tuning of llms. *arXiv preprint arXiv:2410.08661*, 2024. 3

[19] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[20] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 3

[21] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 3

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 3

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning

adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[25] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[27] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[28] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7964–7973, 2024. 1, 2, 3, 4, 6, 7

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*. 2

[30] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 2

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

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[33] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 2, 7

[34] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024. 2

[35] 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. 1

[36] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2

[37] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *Transactions on Machine Learning Research*. 3

[38] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[39] 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, 2, 3, 5, 7

[40] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 8

[41] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2

[42] Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Advancing parameter efficiency in fine-tuning via representation editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13445–13464, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3, 8

[43] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024. 3, 8

[44] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 8

[45] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024. 1, 2, 3, 6, 7, 8

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

image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 8

[47] Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *arXiv preprint arXiv:2406.01563*, 2024. 3

[48] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3, 8

[49] Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. *arXiv preprint arXiv:2505.18612*, 2025. 2

[50] Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Modular customization of diffusion models via blockwise-parameterized low-rank adaptation. *arXiv preprint arXiv:2503.08575*, 2025. 3