MuseumMaker: Continual Style Customization without Catastrophic Forgetting

Chenxi Liu†, Gan Sun* *Member, IEEE*, Wenqi Liang†, Jiahua Dong, Can Qin, Yang Cong

***Abstract*—Pre-trained large text-to-image (T2I) models with an appropriate text prompt has attracted growing interests in customized images generation field. However, catastrophic forgetting issue make it hard to continually synthesize new user-provided styles while retaining the satisfying results amongst learned styles. In this paper, we propose *MuseumMaker*, a method that enables the synthesis of images by following a set of customized styles in a never-end manner, and gradually accumulate these creative artistic works as a *Museum*. When facing with a new customization style, we develop a style distillation loss module to extract and learn the styles of the training data for new image generation. It can minimize the learning biases caused by content of new training images, and address the catastrophic overfitting issue induced by few-shot images. To deal with catastrophic forgetting amongst past learned styles, we devise a dual regularization for shared-LoRA module to optimize the direction of model update, which could regularize the diffusion model from both weight and feature aspects, respectively. Meanwhile, to further preserve historical knowledge from past styles and address the limited representability of LoRA, we consider a task-wise token learning module where a unique token embedding is learned to denote a new style. As any new user-provided style come, our *MuseumMaker* can capture the nuances of the new styles while maintaining the details of learned styles. Experimental results on diverse style datasets validate the effectiveness of our proposed *MuseumMaker* method, showcasing its robustness and versatility across various scenarios.**

***Index Terms*—Text-to-Image Model, Image Generation, Style Customization, Continual Learning.**

I. Introduction

TEXT-TO-IMAGE (T2I) models have emerged in recent studies. Generative models based on diffusion model [1] [2] [3] have demonstrated remarkable efficacy and flexibility in the field of text-to-image generation. Amongst these methods, Stable Diffusion [1] stands out for its ability to generate high quality images through simple language descriptions, which advances the application of generative models to new heights. Based on Stable Diffusion, the area of image generation has attracted widespread attentions, which also promotes the research of following areas, such as image super-resolution [4] [5] and image restoration [6] [7].

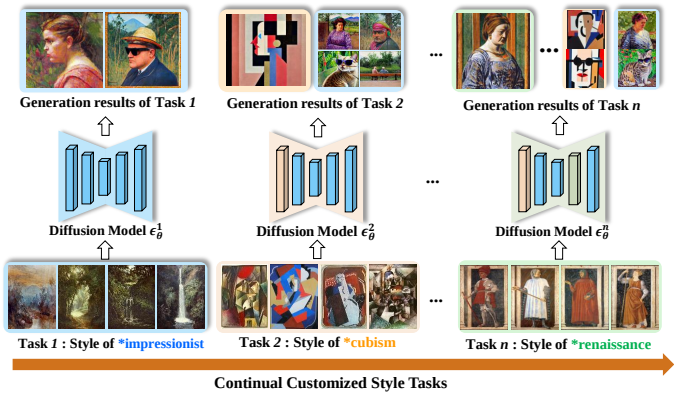

Fig. 1. Motivation of our proposed MuseumMaker model. The desired style (such as impressionist or cubism) for each generated text-to-image task can be customized by the user with a few images. Our MuseumMaker model can continually incorporate the diverse styles without catastrophic forgetting, and further accumulate these creative works as a private Museum.

Although most recent models [8] [9] focus on boosting their generative capabilities with natural languages, these models fail to control the exact texture and color palettes of generation images, when adding popular or customized styles (*e.g.*, "impressionist") to the input text prompt. A naive way is to fine-tune the T2I diffusion model using few images of the given "impressionist" style, whereas this manner cannot synthesize various images of different specific styles in a never-ending manner. For example, as shown in Fig. 1, when the user inputs a prompt such as *"a cat wearing sunglasses in the style of impressionism"* to generate an impressionist image of the cat, a naive way is to fine-tune the diffusion model with images in impressionist style. Subsequently, if the user tries to augment this diffusion model with a new style, *e.g.*, "cubism", a simple method is to fine-tune the diffusion model with both impressionistic and cubist images. As users continuously seek to incorporate new styles, the large computational requirements and ever-increasing training times swiftly become impractical. Moreover, there is a high probability for diffusion model to overfit to the new user-provided styles after continual fine-tuning operator, resulting in unsatisfactory artistic generation performance [10].

Inspired by the aforementioned practical scenarios, we assume that the pre-trained large T2I diffusion model receives data of different customized styles in a streaming manner, as shown in Fig. 1. In the scenario of continuous style stream, we aim to establish a continual style customization method with T2I diffusion model, and maintain original generative

Chenxi Liu and Wenqi Liang are with the State Key Laboratory of Robotics, the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China. (liuchenxi0101, liangwenqi0123@gmail.com.)

Gan Sun and Yang Cong are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510640, China.

Jiahua Dong is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. (dongjiahua1995@gmail.com.)

Can Qin is with the Salesforce AI Research, Palo Alto, CA, 94301, USA.

†These authors contributed equally to this work.

*The corresponding author is *Prof. Gan Sun*.

capacity even after extensive fine-tuning. To achieve this, the T2I diffusion models need to extract pure style features rather than intricate image content, and solve the following key challenges:

- "**Catastrophic Overfitting**", *i.e.*, the ability to learn pure features stylistic representation from the images provided from users, instead of influencing by the complex and intricate content of images. When learning a new style from a limited set of images, how to overcome the problem of overfitting to the specific contents of user-provided images is a common thought of style learning. For instance, if the diffusion model has learned an "impressionist" style, which contains a significant number of pictures with mountains in training data. When we input a prompt like, *"a smiling man in impressionism"*, the model may neglect or remove the concept of *"man"* and instead output a picture of mountains in impressionism.
- "**Catastrophic Forgetting**", *i.e.*, the knowledge of various styles obtained by diffusion model tends to be forgotten when incorporating with new styles. For this aspect, the personalized style knowledge forgetting learned from prior styles should be concerned. Boosting the ability to continually generate images of various styles without revisiting old data is a crucial consideration in the realm of continual style customization. For instance, the user is able to generate images of "impressionism" without accessing the training data, even after long periods of fine-tuning for new styles.

Inspired by the aforementioned considerations, we present a continual style learning approach based on T2I diffusion model in this work. Our proposed method aims to enable the continuous learning of new styles for generation purposes without accessing past datasets, which is designed to reduce memory consumption and compute efficiently. To address the two aforementioned challenges, we develop a continual style customization for diffusion model *i.e*, MuseumMaker, which could continuously adapt upcoming new styles and purely learn stylistic features from user-provided data. To tackle the catastrophic overfitting problem, we introduce a Style Distillation Loss (SDL) module, which decouples the style and content of images by extracting the representation of styles from the whole dataset. The SDL module enables the diffusion model to concentrate on acquiring the stylistic attributes of images while reducing the influence of specific content of images. To mitigate catastrophic forgetting, we propose a Dual Rregularization for shared-LoRA (DR-LoRA) module, which is designed to facilitate the smooth transfer of knowledge from old styles. This module incorporates LoRA weight regularization and stylistic feature regularization to preserve knowledge acquired from previous styles. All the customization tasks share a single set of LoRA parameters with DR-LoRA module to minimize memory consumption. Additionally, to address the limitation posed by the limited parameters of LoRA, a Task-wise Token Learning (TTL) module is develops to learn a distinct token for each style, which enables generation of different styles. The tokens corresponding to past styles are stored to further mitigate the catastrophic forgetting problem. Finally, we demonstrate the validation of our MuseumMaker through extensive experiments, and conduct ablation studies to emphasize the contribution of each module in our MuseumMaker.

To summarise, the main contributions of this paper are as follows:

- We take an earlier attempt to propose continual style customization with pre-trained large T2I diffusion model, *i.e., MuseumMaker*, which enables continual learning of various styles and effectively mitigates the catastrophic forgetting issue amongst past styles.
- To deal with the issue of catastrophic overfitting, we introduce a style distillation loss module to distill the style representation from the entire dataset into the latent representation generated from each image, which could overcome the problem of learning bias to the content of the images.
- To address the catastrophic forgetting issue, we devise a dual regularization for shared-LoRA module and a task-wise token learning module, which could maintain the style knowledge from previous learned styles. Extensive experiments have confirmed the significant performance improvements and effectiveness of our proposed MuseumMaker.

We organize the rest of the paper as follows: the first section briefly introduces some related work. The second section revisits the text-to-image diffusion model and defines the problem of continual style learning for diffusion model. Then, we describe our methods in detail in third section. To the end, we conduct various experiments to evaluate our proposed method, followed by the conclusion.

II. Related Work

A. Continual Learning

The field of continual learning has attracted significant attention in recent years. Early approaches to continual learning focus on regularization-based methods, such as EWC [11] and SI [12]. These methods introduce additional regularization terms to penalize the changes of parameters, which are important for preserving previously acquired knowledge. Subsequent works have explored extensions and variations of these techniques, including strategies for better approximating parameter importance [13] [14] and methods for considering the heterogeneous forgetting across different classes [15] [16].

Another notable direction is experience replay, where a subset of past data is stored and replayed during training on new tasks to mitigate forgetting [17]. Variations of this approach include strategies for constructing and exploiting the memory buffer [18] [19], and leveraging generated data from generative models to augment or replace the replay buffer [20] [21]. Optimization-based approaches have also been explored, such as gradient projection methods [22] [23] and meta-learning strategies [24] [25]. These techniques aim to directly manipulate the optimization process or learn inductive biases that facilitate continual learning. Architectural innovations also play a role in continual learning, with methods such as parameter isolation [26] [27] , dynamic architectures [28] [29],

and modular networks [30]. More recently, there has been a growing interest in representation-based approaches that leverage category-guided learning [31] and large-scale pre-training [32] [33] to obtain robust and transferable representation for continual learning. These methods seek to exploit the inherent advantages of self-supervised and pre-trained representation, such as improved generalization and robustness to catastrophic forgetting.

Despite significant progress, continual learning remains an active area of research. In this area, continual style learning for diffusion model is still a matter of concern, where various stylistic feature representation are hard to merge into a special diffusion model. Continual generation for different styles of images remains a challenge for diffusion model.

B. Text to Image Generation

The emergence of diffusion models has ushered in a new era of text-to-image (T2I) generation, with state-of-the-art methods achieving unprecedented levels of photorealism and caption-image alignment. GLIDE [34] pioneers the application of diffusion models to this task, which adopts classifier-free guidance by replacing class labels with text prompts. Imagen [2] further improves upon this approach by leveraging pre-trained language models as text encoders, enabling the use of rich textual representation learned from large-scale corpora.

Other advanced works explore diffusion models in latent space, rather than operating directly on pixel space. Stable Diffusion [1] is trained with a large amount of data based on latent diffusion model (LDM), which employs a VQ-GAN [35] for latent representation and adds text as conditional information to the denoising process. DALL-E 2 [36] inverts the CLIP [37] image encoder with a diffusion model, learning a prior to connect the text and the features of images in the latent space. This text-image latent prior is found to be crucial for performance.

Building upon these pioneering efforts, subsequent studies have sought to enhance T2I diffusion models in various ways. These improvements encompass model architectures [38]–[40], spatial control via sketches or spatio-textual representation [41] [42] [43], textual inversion for controlling novel concepts [44], and retrieval mechanisms [45]–[47] to handle out-of-distribution scenarios. To further tackle the issue of continual learning for diffusion, [48] [49] achieve the continuous generation of personalized concepts. However, the methods for continuous style learning and image generation are still lack of research.

C. Image Style Learning

Capturing expressive representation of image styles is pivotal for achieving high-fidelity style learning. Early techniques like texture synthesis [50] [51] and non-photorealistic rendering (NPR) [52] [53] explore artistic expression through computational methods. However, these methods still face a constraint in transfer quality, universality, and feature extraction capabilities.

Generative Adversarial Networks (GANs) have emerged as a transformative paradigm for style representation learning. [54] proposes the DCGAN model to integrate a convolutional neural network into the original GAN framework, and adopts an encoder-decoder structure to bolster stability and generation quality. [55] introduces Wasserstein GAN (WGAN), leveraging the Earth Mover (EM) distance as a loss function to measure distributional discrepancies. With the recent rapid advancement of diffusion models, extensive methods for image style learning based on diffusion models have emerged. [56] treats the style information of images as trainable textual descriptions for model learning. [57] fine-tunes the U-Net architecture to enable the model to learn the stylistic information from images, and generates images in specific styles. However, these exciting style learning methods are unable to tackle a never-ending streaming manner of styles, which constrains the further application of diffusion model to generate images of various styles.

III. Preliminaries

A. Revisiting Text-to-Image Diffusion Model

The diffusion model [58] represents a probabilistic approach utilized for image generation, which generates a image by gradually denoising a noise graph from a Gaussian distribution. As exemplified by the Stable Diffusion (SD) technique [1], the primary objective of the diffusion model is to learn a trajectory to generate the well patterned and distributed data from a random noise graph. Specifically, SD relies on a pre-trained text encoder ψ from CLIP [37], a latent encoder $\mathcal{F}$, a decoder $\mathcal{G}$ and a denoising U-Net ϵ_θ. Given a text prompt p, a timestep $t \in \text{Uniform}(1, T)$ and a random Gaussian noise graph $\varepsilon \in \mathcal{N}(0, \mathbf{I})$, the text embedding can be computed as $c = \psi(p)$,and subsequently, the image $\hat{x}$ can be generated by $\hat{x} = \mathcal{G}(\epsilon_\theta(c, t))$. Formally, the training process of SD can be defined as follow:

$$\mathcal{L}_{SD} = \mathbb{E}_{z,c,\varepsilon,t}[\|\varepsilon - \epsilon_\theta(z_t|c, t)\|_2^2], \tag{1}$$

where z_t is the image latents encoded by $\mathcal{E}$ at timestep t, and θ represents the parameters of denoising model.

B. Problem Definition

Suppose that a series of continuous tasks for Text-to-Image (T2I) diffusion model are denoted as $\mathbf{D}^K = \{\mathcal{D}^1_{n_1}, \mathcal{D}^2_{n_2}, \cdots, \mathcal{D}^K_{n_K}\}$, where K represents the number of consecutive tasks and n_k denotes the number of samples per task. Consequently, we can define all the generation tasks as $\mathcal{T} = \{\mathcal{T}^k\}_{k=1}^K$. For the k-th learning task, $\mathcal{T}^k$ corresponds to a dataset $\mathcal{D}^k_{n_k} = \{x_i^k, p_i^k\}_{i=1}^{n_k}$, where $x_i^k \in \mathcal{X}^k$ and $p_i^k \in \mathcal{P}^k$ denote the i-th image and its corresponding prompt. Different from existing works [57] [59], we consider a scenario that users provide consecutive images of different specific styles for T2I diffusion model, namely task $\mathcal{T}^k$. For each task, the number of pairs of image x_i^k and prompt p_i^k with n_k are imposed into the continual style customization diffusion model, MuseumMaker. Considering the problem of privacy and memory limitation, the dataset of past tasks is unavailable. In this scenario, the model we proposed is able to incorporate the new style while ensuring the memory of past encountered

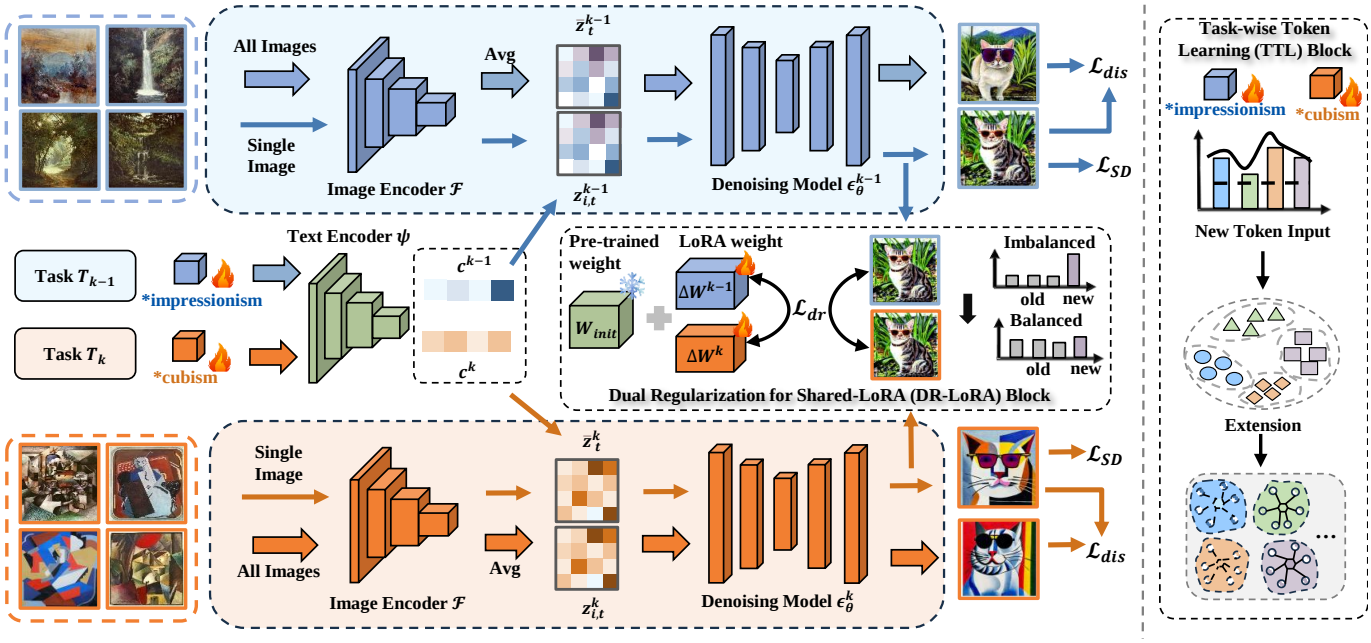

Fig. 2. Illustration of our proposed continual style customization for T2I diffusion model, *i.e.,* MuseumMaker. It mainly contains a Dual Regularization for Shared-LoRA (DR-LoRA) module to regularize the optimization of model from both weight and feature aspects, a Task-wise Token Learning (TTL) module to store the text embedding of each style learning to reduce forgetting and a Style Distillation Loss module (SDL) to make the model focus on the style of learning images.

styles. In details, the problem of continual style customization can be formulated as follows:

$$\mathcal{L}_{\mathrm{SD}}^k = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \mathbb{E}_{z_i^k, c_i^k, \varepsilon, t} \left[\| \varepsilon - \epsilon_\theta^k(z_{i,t}^k | c_i^k) \|_2^2 \right], \tag{2}$$

$$\mathcal{L}_{\mathrm{CSA}} = \sum_{k=1}^{K} (\mathcal{L}_{\mathrm{SD}}^k + \lambda \mathcal{L}_{\mathrm{c}}^k), \tag{3}$$

where ϵ denotes the denoising U-Net, θ represents its trainable parameters, and $\mathcal{B}$ is the batch size during train process. $(z_{i,t}^k, c_i^k)$ respectively represent the corresponding conditional image latents at timestep t and text embeddings on the k-th task, where $c_i^k = \psi(p_i^k)$. $\mathcal{L}_c^k$ is a loss proposed under the constraint of past prior knowledge, aiming at facilitating continuous learning of the model. Further explanation on $\mathcal{L}_c^k$ will be provided in the subsequent sections.

IV. Our Proposed Method

In this section, we firstly present the overview of our proposed MuseumMaker in Sec. IV-A, which is illustrated in Fig. 2. In Sec. IV-B, we fristly present a style distillation loss to overcome the problem of catastrophic overfitting when encountering with a new style customization task. To tackle the catastrophic forgetting issue, we introduce a dual regularization for shared-LoRA module in Sec. IV-D, and a task-wise token learning module in Sec. IV-C.

A. Overview

As illustrated in Fig. 2, we develop a progressive continual style customization method with T2I diffusion model (*i.e.,* MuseumMaker), which aims to retain knowledge of past learned styles while incorporating with new customized styles. To address the issue of catastrophic overfitting to image content, we devise a Style Distillation Loss (SDL) module for each new style, which distills the mean latent features across all images with individual image latent features to reduce the influence of image content during training. The direction of optimization is corrected to style of images by SDL, ensuring that the model focuses on style of images and alleviates overfitting to content. Considering the issue of catastrophic forgetting in continual style learning, we develop a Dual Rregularization for shared-LoRA (DR-LoRA) module and a Task-wise Token Learning (TTL) module. These two modules intend to transfer the knowledge from old model to current model and attain a unique token embedding for each specific style. Overall, MuseumMaker could offer a comprehensive solution for continual style customization.

B. Style Distillation Loss

Given a reference style customization task with $8 \sim 10$ images, continual customized style learning aims to extract its artistic style information while explicitly removing the content information, which also emerges as a key area of focus in image style transfer [60] [61]. One of feasible solutions is to explicitly defines the high-level features as content and the feature correlations (*i.e.*, Gram matrix) as style [60]. However, this method relies on additional networks, which is not feasible enough to deploy to diffusion model and not available for continual style customization. Therefore, we introduce an easily deployable style distillation loss module, which addresses catastrophic overfitting and captures the pure style representations to prevent the disruption of original concepts in diffusion model.

To capture pure style representation from user-provided images and guide the model towards learning style features, we

propose a straightforward yet highly implementable method in the continual style customization setting, *i.e.*, Style Distillation Loss (SDL) module. Specifically, given a set of input images $\mathcal{X}^k = \{x_1^k, x_2^k, ...x_{n_k}^k\}$ on the k-th style customization task. We encode all images using an VAE [62] image encoder $\mathcal{F}$ to obtain their feature representation $\mathcal{Z}^k = \{z_i^k = \mathcal{F}(x_i^k) | x_i^k \in \mathcal{X}^k\}$, and then compute the mean of all feature representation as the representation of the image style of specific dataset. The latent features of images in the k-th task can be represented as follows:

$$\bar{z}_t^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{F}(x_i^k). \tag{4}$$

By utilizing both the mean feature and individual image features to train unique styles, we then obtain a distillation loss during the denoising process:

$$\mathcal{L}_{\text{SDL}} = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \rho(\mathcal{F}(x_i^k)/\tau) \log\left(\frac{\rho(\mathcal{F}(x_i^k)/\tau)}{\rho(\bar{z}_t^k/\tau)}\right), \tag{5}$$

where τ is a temperature hyperparameter, and ρ represents the softmax function. Considering the Kullback-Leibler (KL) divergence loss between the mean latent features of all images and the latent features of individual images, SDL demonstrates its efficacy in reducing the probability of model overfitting to specific image content while learning image styles. This mechanism ensures the capacity of model to effectively concentrate on the diverse styles inherent in images.

C. DR-LoRA

To rapid learn with extracted user-provided style from a small number of images, one of the popular strategies is fine-tuning T2I diffusion model with low-rank adaptation(LoRA) [63], which has demonstrated its effectiveness in various text-to-image diffusion models [64] [48]. Inspired by LoRA, we adopt a shared-LoRA strategy for continual customized styles learning in this work, and intend to significantly reduce the computational consumption usage. However, a wide and disparate distribution existing in style data stream poses a formidable hurdle in the continual style customization setting, when the diffusion model tries to preserve the style knowledge among pre-trained model. To tackle this issue, we devise a Dual Rregularization for this LoRA module (DR-LoRA) to the balance the learning of different styles, which simultaneously considers the knowledge transfer of style distribution from both weight manifold and feature representation manifold.

Specifically, LoRA achieves fine-tuning of pre-trained large T2I models by freezing original weight while learning a pair of rank-decomposition matrices. Formally, this can be expressed as: $\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}$, where $\Delta\mathbf{W} = \mathbf{A}\mathbf{B}^T$. LoRA effectively reduces the number of parameters required during training by fine-tuning matrices $\mathbf{A}$ and $\mathbf{B}$, and makes a swift adaptation to new styles. However, without constraining the optimization of diffusion model, the model above tends to learn incoming style data and further causes the catastrophic forgetting. Therefore, we consider the weight regularization for LoRA weight, in which we compute the offset of LoRA between adjacent tasks as a loss function to optimize the direction of model updating. Mathematically, we denote the weight manifold regularization of the k-th task as follows:

$$\mathcal{L}_w = \frac{1}{L} \sum_{l=1}^{L} \left[1 - \text{Sim}\left(\text{Flatten}(\Delta\mathbf{W}_l^{k-1}), \text{Flatten}(\Delta\mathbf{W}_l^k)\right)\right], \tag{6}$$

where l represents the number of cross-attention layer of U-net, and $\text{Sim}(.)$ denotes the cosine similarity calculation. To better calculate the similarity between past LoRA weight and current one, we flatten out the LoRA weight. We slow down the updating on key weights through $\mathcal{L}_w$, which is definitively crucial for T2I diffusion model to maintain the knowledge learned from previous styles.

While $\mathcal{L}_w$ regularizes the diffusion model from the perspective of weight, it overlooks the forgetting of representation from past styles, which is an indispensable aspect to overcome catastrophic forgetting. Moreover, although shared-LoRA explores the task-shared global representation of styles, the task-specific representation for each customized style is neglected. We thus consider the unique knowledge of representation from each style, and introduce a feature representation regularization in shared-LoRA. Specifically, to transfer the style knowledge from past denoising model ϵ_θ^{k-1} to the current one ϵ_θ^k and well align feature representation, we input a set of past style prompts $\mathcal{P}^{k-1} = \{p_i^j\}_{j=1}^{k-1}$ into both past and current model. Subsequently, we can generate two noise graphs $\hat{z}_i^j$ and $\hat{z}_i^k$ from past and current denoising models, respectively. We consider noise graph $\hat{z}_i^j$ generated from ϵ_θ^{k-1} as a pseudo noise graph for past encountered styles, which implicitly provides strong semantic guidance for current model to maintain prior knowledge. Therefore, we devise a loss function to transfer the knowledge of pseudo noise graph $\hat{z}_i^{k-1}$ to current denoising model ϵ_θ^k. Formally, the loss on the k-th style generation task can be computed as the follows:

$$\mathcal{L}_f = \frac{1}{K} \frac{1}{\mathcal{B}} \sum_{j=1}^{K} \sum_{i=1}^{\mathcal{B}} \rho(\hat{z}_i^j/\tau) \log\left(\frac{\rho(\hat{z}_i^j/\tau)}{\rho(\hat{z}_i^k/\tau)}\right). \tag{7}$$

The knowledge from old styles can be transferred by $\mathcal{L}_f$, which regularizes the T2I diffusion model from the aspect of features.

Overall, the dual regularization loss for shared-LoRA is:

$$\mathcal{L}_{\text{DR}} = \lambda_1 \mathcal{L}_w + \lambda_2 \mathcal{L}_f, \tag{8}$$

where λ_1 and λ_2 are the hyperparameters to trade-off the regularization between LoRA weights and features.

D. Task-wise Token Learning

Deploying the above DR-LoRA module for continual style customization presents a promising method for sharing knowledge across diverse stylistic domains. However, when the T2I diffusion model encounters a massive continuous data stream with various styles, the DR-LoRA is limited to capture the distinct feature of each style. Additionally, as the diffusion model acquires different stylistic features, it may generate images intertwined with multiple styles, resulting in unsatisfactory generative performance. To address these challenges, we expend the model parameters of text embeddings minimally to enable the continual style learning. Furthermore,

Algorithm 1 Optimization pipeline of Our MuseumMaker.

Input: VAE image encoder $\mathcal{F}$, text encoder ψ, diffusion model ϵ_θ, dataset $\mathbf{D}^K = \{\mathcal{D}_{n_k}^1, \mathcal{D}_{n_k}^2, \cdots, \mathcal{D}_{n_k}^K\}$, past style prompts set $\mathcal{P}^{K-1} = \{p_i^j\}_{j=1}^{k-1}$, number epoch E, batch size $\mathcal{B}$, number of tasks K;

for $k = 1, 2, \cdots, K$ **do**
  if $k = 1$ **then**
    Initialize LoRA weight $\Delta\mathbf{W}^k$;
  else
    Load old LoRA weight $\Delta\mathbf{W}^{k-1}$;
  end if
  Initialize token embedding $\mathcal{V}_*^k$
  Calculate average image latent features $\bar{z}_t^k = \frac{1}{n_k}\sum_{j=1}^{n_k}\mathcal{F}(x_j^k)$, $x_j^k \in \mathcal{D}_{n_k}^k$;
  for $e = 1, 2, \cdots, E$ **do**
    Random select $\mathcal{B}$ samples from $\mathcal{D}_{n_k}^k$
    for $i = 1, 2, \cdots, \mathcal{B}$ **do**
      Calculate single image latent features $z_{i,t}^k = \mathcal{F}(x_i^k)$;
      Calculate text embedding $c_i^k = \psi(p_i^k)$;
      Generate noise graphs $\hat{x}_{i,t}^k = \epsilon_{\theta,t}(z_{i,t}^k | c_i^k, t)$, $\bar{x}_t^k = \epsilon_{\theta,t}(\bar{z}_t{}^k | c_i^k, t)$ corresponding to $z_{i,t}^K$ and $\bar{z}_t^k$;
      Calculate *style distillation loss* $\mathcal{L}_{\mathrm{SDL}}$ by Eq. (5);
      if $k > 1$ **then**
        for $j = 1, 2, ..., k-1$ **do**
          Compute pseudo noise graph $\hat{z}_i^j = \epsilon_\theta^{k-1}(p_i^j)$;
          Compute the dual regularization loss $\mathcal{L}_{\mathrm{DR}}$ by Eq. (8);
        end for
      end if
      Optimize diffusion model $\epsilon_{\theta,t}^k$ and token embedding $\mathcal{V}_*^k$ by Eq. (12);
    end for
  end for
end for
return Current token embedding $\mathcal{V}_*^K$, and current LoRA weight $\Delta\mathbf{W}^K$

we devise unique token embeddings for each style to distinguish stylistic features from different style customization tasks. To be specific, prior works [65] [66] have extensively explored token training, where a trainable token embedding can efficiently inject a special style to diffusion model for further images generation. Therefore, we develop a Task-wise Token Learning (TTL) module based on the textual inversion method. TTL module learns an independent token for each style customization task, and only stores the fine-tuned token parameters to maintain the prior knowledge, which take up extremely little memory. With a unique token corresponding to a distinctive style, the T2I diffusion model is able to capture unique features in a variety of styles. Formally, this module for the k-th task can be represented as:

$$v_*^k = \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \arg\min_{v^*} \mathbb{E}_{z_i^k, p_i^k, c_i^k, \varepsilon, t}\Big[\|\varepsilon - \epsilon_\theta(z_t^k, t, c_\theta(p_i^k))\|_2^2\Big], \tag{9}$$

where v_*^k is the learnable token embedding, and $p_i^k \in \mathcal{P}^k$ denotes the user-provided conditioned style images of the k-th task. We combine the text description v_*^k with the conditioned stylistic images by Eq. (9). The T2I diffusion model fuses the features of conditioned model, which can be triggered to generate specific style images when the corresponding v_*^k is input.

However, Textual Inversion [65] overlooks the independence of cross-attention layers at different resolutions during denoising process. Solely training the input token may not sufficiently unfold the embedding space. Inspired by this observation, we extend the input token embedding by introducing multiple independent tokens, which are learned across different cross-attention layers of the U-Net. We denote the extended v_*^k as $\mathcal{V}_*^k = \{v_1^k, v_2^k, ..., v_L^k\}$, where k represents the corresponding task index, L represents the number of layers of cross-attention. The trainable tokens set of K tasks can be written as $\boldsymbol{V}_*^K = \{\mathcal{V}_*^1, \mathcal{V}_*^2, ..., \mathcal{V}_*^K\}$. Consequently, the Eq. (9) for K tasks can be rewritten as:

$$\boldsymbol{V}_*^K = \frac{1}{\mathcal{B}}\sum_{k=1}^{K}\sum_{i=1}^{\mathcal{B}} \arg\min_{v} \mathbb{E}_{z_i^k, p_i^k, c_i^k, \varepsilon, t}\Big[\|\varepsilon - \epsilon_\theta(z_{i,t}^k, t, c_\theta(p_i^k))\|_2^2\Big], \tag{10}$$

where each task obtains a extended token. We effectively extend the dimension of the token embedding space by training an independent token embedding for each cross-attention layer, which enhances the final generation quality. After training, we store a special token embedding for each task to further mitigate catastrophic forgetting during continual style customization.

In summary, the overall pipeline optimization for our proposed MuseumMaker method is illustrated in Algorithm 1. The loss function constrained by prior knowledge to enable continual learning can be formulated as:

$$\mathcal{L}_c = \alpha\mathcal{L}_{\mathrm{SDL}} + \beta\mathcal{L}_{\mathrm{DR}}, \tag{11}$$

where α and β are manually set hyperparameters. The total optimization objective can be formulated as follows:

$$\mathcal{L}_{\mathrm{Overall}} = \sum_{k=1}^{K}(\mathcal{L}_{\mathrm{SD}}^k + \alpha\mathcal{L}_{\mathrm{SDL}} + \beta\mathcal{L}_{\mathrm{DR}}). \tag{12}$$

Regarding $\mathcal{L}_{\mathrm{Overall}}$, our MuseumMaker systematically aggregates a continual collection of styles, empowering users to draw artworks of diverse styles and build their own customized museum.

V. Experiment

A. Datasets and Evaluation

We present comparison experiments and ablation studies on Wikiart [67] to illustrate the efficacy of our MuseumMaker. Due to the particularity of imges generation, we utilize three

Fig. 3. Qualitative comparison between our method with competing methods in 10 tasks continual style adaption setting, where the first two rows represent the stylistic dataset and prompts provides users, and the rest results denote the image generated by each method with the same text prompt, *i.e., a cat, wearing a sunglasses in *style*.

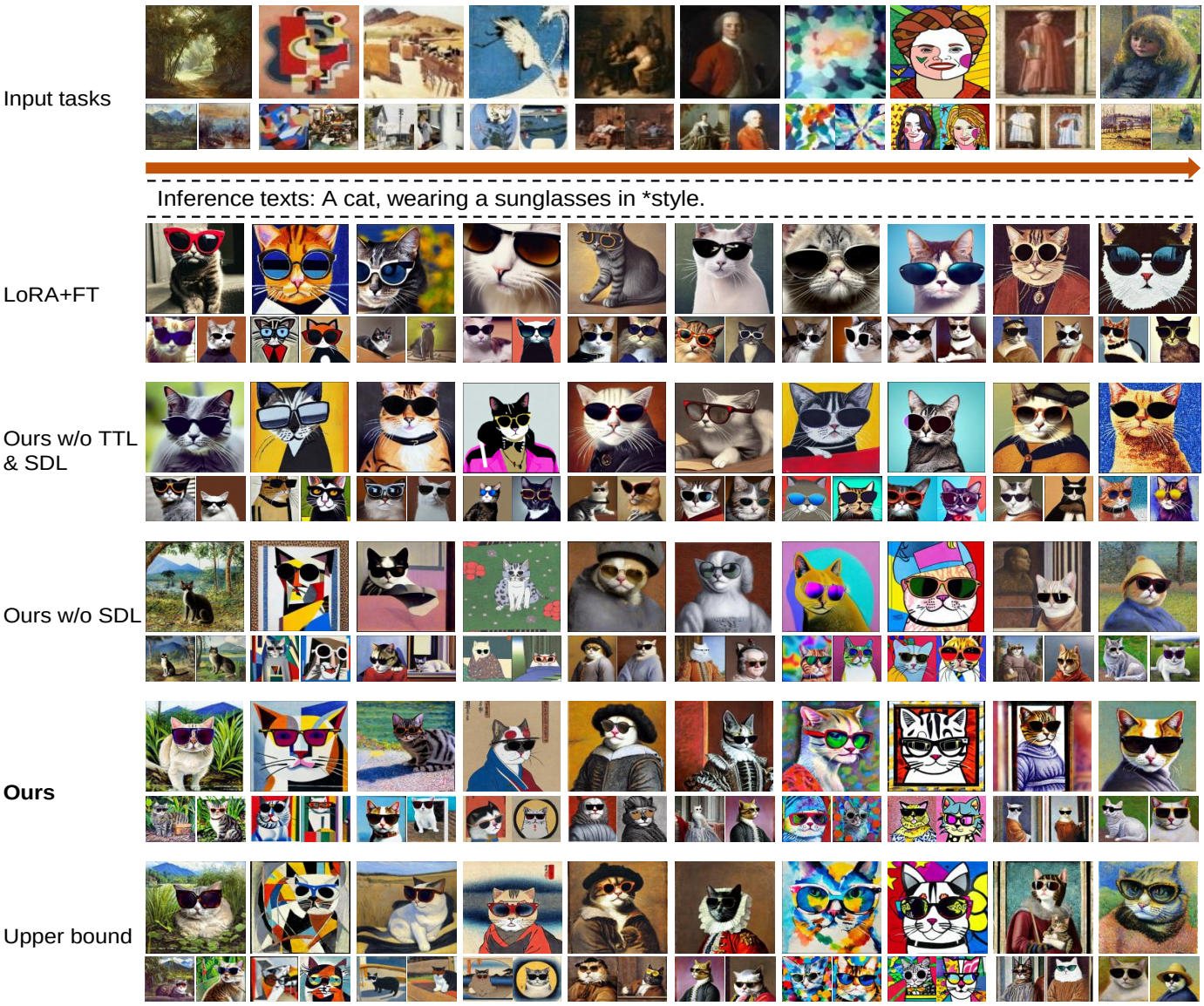

Fig. 4. Qualitative comparison of our ablation studies, where we evaluate the contribution of each module we proposed. We denotes the task-wise token learning module as TTL and style distillation loss as SDL in the fig, respectively. We input the same text prompt as we do in comparison experiments. The upper bound setting stores a learnable token embedding and LoRA weight for each style

kinds of metrics (*i.e.*, style loss [57], FID [68] and CLIP score [69]) to measure the superior performance of our proposed MuseumMaker.

Datasets: WikiArt consists of a diverse and rich artwork dataset of diferent styles, which encompasses artworks from 61 genres and localizes in 8 languages. The artworks from WikiArt is sourced from museums, universities, city halls, and other municipal buildings spanning over 100 countries. For evaluation, we select 10 datasets representing different artistic styles in continual learning manner, as follow: (1) impressionism, (2) cubism, (3) realism, (4) ukiyo, (5) baroque, (6) rococo, (7) expressionism, (8) pop art, (9) renaissance, (10) pointillism. Each dataset comprises 8-10 images, showcasing distinct distributions of artistic styles. For ease of reference, we abbreviate each style to the first three letters.

Evaluation: To fairly evaluate the generative performance of models in a continual learning scenario, we leverage hundreds of images generated from 20 prompts for quantitative evaluation, with each prompt generating 50 sets of images. In order to evaluate the similarity of style between the generated images and the original style, we introduce style loss, FID and CLIP score. The details of these three metrics are as follow:

1) **Style loss** is employed in the field of image generation to assess the stylistic similarity between two images. We take the images from the training dataset as references and calculate the style similarity between these references and all generated images, subsequently averaging the results; 2) **FID** assesses the quality of generated images by quantifying the distance between source images and generated, which considers the distribution between both source data; 3) **CLIP score** utilizes the pre-trained ViT-B/32 model from CLIP to measure the property of diffusion model. By encoding both the generated

TABLE I

CONTINUAL STYLE ADAPTATION GENERATION COMPARISON BETWEEN OURS WITH STATE-OF-THE-ARTS METHOD (*e.g.*,LORA [63]+LWF [70], LORA [63]+EWC [11], SPD [57]+LWF [70], SPD [57]+EWC [11]) IN TERMS OF FID AND CLIP SCORE. METHODS WITH THE BEST AND RUNNER-UP PERFORMANCE ARE MARKED AS BOLDED **RED** AND **BLUE**, RESPECTIVELY.

Comparison Methods	FID⇓												CLIP Score⇑											
	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**
LoRA+FT	334.7	371.8	399.2	381.2	406.8	325.0	449.4	401.6	351.2	382.7	**380.4**	⇑**66.9**	51.78	68.66	53.83	54.05	59.38	53.43	58.31	50.97	59.17	55.30	**56.49**	⇑**11.77**
LoRA+LWF	**328.1**	365.9	398.4	386.7	411.9	331.7	428.4	408.7	360.1	381.4	**380.1**	⇑**66.7**	52.09	69.72	54.05	53.70	58.17	52.03	65.77	49.36	56.69	55.10	**56.67**	⇑**11.59**
LoRA+EWC	335.7	377.1	398.3	380.6	405.1	321.9	448.4	404.9	351.1	383.9	**380.7**	⇑**67.2**	51.99	69.01	53.88	53.81	59.47	53.87	58.27	50.92	59.02	55.48	**56.57**	⇑**11.69**
SPD+FT	345.4	**262.2**	396.3	315.9	**394.6**	287.3	386.0	358.2	**314.7**	361.8	**342.2**	⇑**28.8**	58.77	**84.07**	53.71	70.15	62.23	**64.21**	70.93	56.05	68.35	56.52	**64.50**	⇑**3.76**
SPD+LWF	331.1	378.3	**395.3**	329.0	394.8	**271.6**	**264.2**	**328.4**	327.0	365.7	**338.5**	⇑**25.1**	**59.59**	82.65	53.94	**75.86**	**62.26**	**69.72**	**85.92**	**59.10**	**68.59**	55.34	**67.30**	⇑**0.96**
SPD+EWC	345.9	273.7	396.8	**298.2**	394.9	284.6	**356.3**	347.4	**313.4**	**363.5**	**337.5**	⇑**24.0**	58.72	**84.73**	**54.70**	71.22	62.10	63.92	**73.75**	57.72	**68.50**	**57.69**	**65.31**	⇑**2.96**
Ours	**250.5**	**271.2**	**374.9**	**251.8**	**348.2**	**286.3**	373.1	**312.1**	319.5	**347.2**	**313.5**	-	**68.86**	82.99	**60.18**	**79.21**	**67.15**	62.98	70.56	**59.69**	63.46	**67.53**	**68.26**	-
Upper Bound	249.5	254.6	350.0	236.5	317.3	279.8	245.0	335.6	260.4	327.5	**285.6**	-	74.35	84.20	67.35	81.84	71.94	68.40	82.32	60.37	75.46	70.71	**73.69**	-

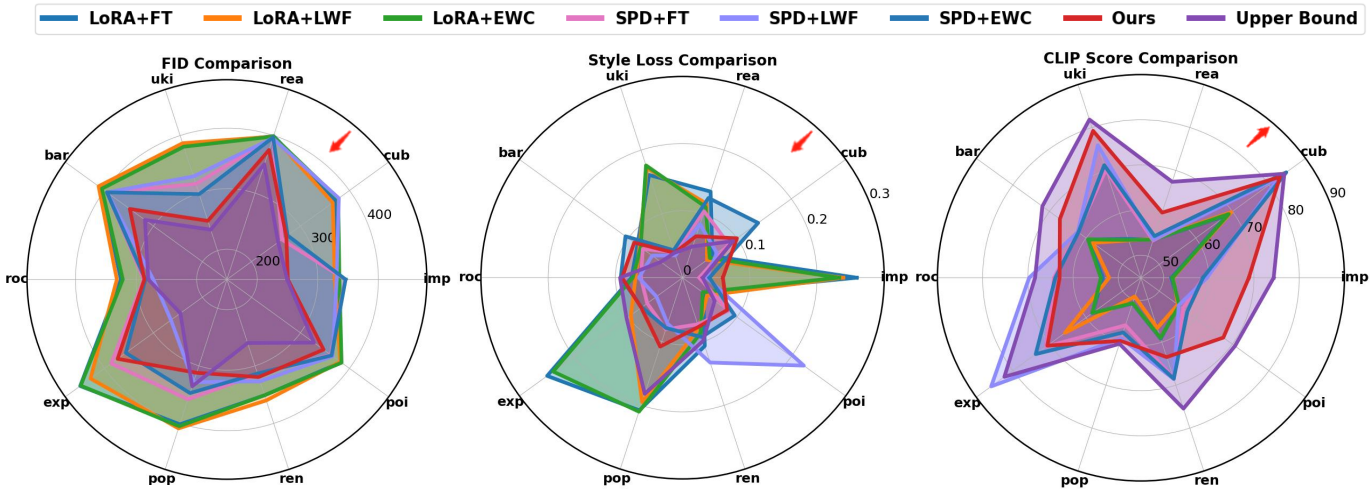

Fig. 5. Style loss, FID and CLIP score for continual style adaptation setting. The red arrow in the image indicates the direction where the score improves. Comparing with other competing method (*e.g.,* LoRA [63]+LWF, LoRA+EWC, SPD+LWF, SPD+EWC), our MuseumMaker achieves the best performance in terms of FID and CLIP score.

TABLE II

CONTINUAL STYLE ADAPTATION GENERATION COMPARISON BETWEEN OURS WITH STATE-OF-THE-ARTS METHOD (*e.g.*,LORA [63]+LWF [70], LORA [63]+EWC [11], SPD [57]+LWF [70], SPD [57]+EWC [11]) IN TERMS OF STYLE LOSS. IN ORDER TO BETTER DISPLAY THE DATA RESULTS, ALL RESULTS OF STYLE LOSS IS SCALED BY 100. METHODS WITH THE BEST AND RUNNER-UP PERFORMANCE ARE MARKED AS BOLDED **RED** AND **BLUE**, RESPECTIVELY.

Comparison Methods	Style Loss⇓											
	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**
LoRA+FT	0.259	0.055	0.135	0.161	0.083	0.076	0.249	0.208	0.107	**0.038**	**0.137**	⇑**0.058**
LoRA+LWF	0.239	**0.045**	0.116	0.172	0.079	0.070	0.095	0.195	0.079	0.045	**0.114**	⇑**0.035**
LoRA+EWC	0.233	**0.050**	0.111	0.176	0.077	0.074	0.238	0.210	0.087	**0.037**	**0.129**	⇑**0.050**
SPD+FT	**0.025**	0.085	0.104	0.040	**0.073**	**0.062**	**0.065**	0.081	**0.073**	0.084	**0.069**	⇑**0.010**
SPD+LWF	**0.036**	0.063	**0.084**	**0.035**	**0.056**	**0.065**	**0.048**	**0.072**	0.133	0.223	**0.081**	⇑**0.002**
SPD+EWC	0.040	0.139	0.125	0.043	0.105	0.092	0.075	**0.079**	0.093	0.096	**0.089**	⇑**0.010**
Ours	0.060	0.100	**0.065**	**0.039**	0.089	0.090	0.076	0.108	**0.077**	0.082	**0.079**	-
Upper Bound	0.030	0.093	0.049	0.039	0.040	0.093	0.103	0.182	0.099	0.059	**0.079**	-

images and reference images using CLIP, their similarity in the latent space can be computed. A higher similarity score indicates a better correspondence between the generated images and reference images.

B. Implementation Details and Baselines

In this work, we conduct continual style customization experiments on our proposed MuseumMaker, LoRA [63] and Specialist Diffusion (SPD) [57]. For the continual style customization experiment, we conduct the experiment on the 10 style datasets, ensuring that the previous style is not available during the training process. To assess the effectiveness of our continual style customization method in contrast to existing approaches, we deploy Fine-tuning (FT), as well as two classic continual learning methods, namely EWC [11] and LWF [70], for LoRA and SPD. These methods serve as benchmarks for comparison with our approach. Furthermore, we store a learnable token embedding and LoRA weight for each style as our upper bound. During the learning phase for each style, we employ 1000 training steps for LoRA baselines method, 100 training epoch for SPD baselines methods, 1000 training steps for ours. All the methods are implied with a batch size of 1. To achieve desirable generation performance, we set the learning rate to 1×10^{-5} for LoRA baselines method, 1×10^{-6} for SPD baselines methods and 1×10^{-5} for our method. Additionally, the hyperparameters λ_1 and λ_2 are set 0.8 and 1.0 in Eq. (8), α and β in Eq. (12) are set to 0.8 and 1.5.

TABLE III
COMPARISON EXPERIMENTS OF OUR ABLATION STUDIES, WHERE TTL AND SDL REPRESENT TASK-WISE TOKEN LEARNING MODULE AND STYLE DISTILLATION LOSS MODULE, RESPECTIVELY. METHODS EXCEPT FOR THE UPPER BOUND WITH THE BEST AND RUNNER-UP PERFORMANCE ARE MARKED AS BOLDED **RED** AND **BLUE**, RESPECTIVELY.

Comparison Methods	FID⇓											
	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**
LoRA+FT	334.7	371.8	399.2	381.2	406.8	325.0	449.4	401.6	351.2	382.7	**380.4**	⇑**66.9**
Ours w/o TTL & SDL	332.6	366.0	401.7	382.4	406.4	318.8	435.9	406.5	348.2	387.9	**378.6**	⇑**65.2**
Ours w/o SDL	**233.8**	**321.1**	**382.0**	**284.1**	**377.8**	**285.8**	**357.0**	**314.7**	**328.8**	**340.4**	**322.5**	⇑**9.1**
Ours	**250.5**	**271.2**	**374.9**	**251.8**	**348.2**	**286.3**	**373.1**	**312.1**	**319.5**	**347.2**	**313.5**	-
Upper Bound	249.5	254.6	350.0	236.5	317.3	279.8	245.0	335.6	260.4	327.5	**285.6**	-

Comparison Methods	CLIP Score⇑											
	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**
LoRA+FT	51.78	68.66	53.83	54.05	59.38	53.43	58.31	50.97	59.17	55.30	**56.49**	⇑**11.77**
Ours w/o TTL & SDL	51.68	68.93	53.87	53.35	58.43	52.10	65.87	49.45	56.23	55.11	**56.50**	⇑**11.76**
Ours w/o SDL	**74.23**	**82.26**	**65.16**	**77.13**	**66.96**	**65.00**	**73.02**	**58.92**	**68.08**	**70.25**	**70.10**	⇓**-1.84**
Ours	**68.86**	**82.99**	**60.18**	**79.21**	**67.15**	**62.98**	**70.56**	**59.69**	**63.46**	**67.53**	**68.26**	-
Upper Bound	74.35	84.20	67.35	81.84	71.94	68.40	82.32	60.37	75.46	70.71	**73.69**	-

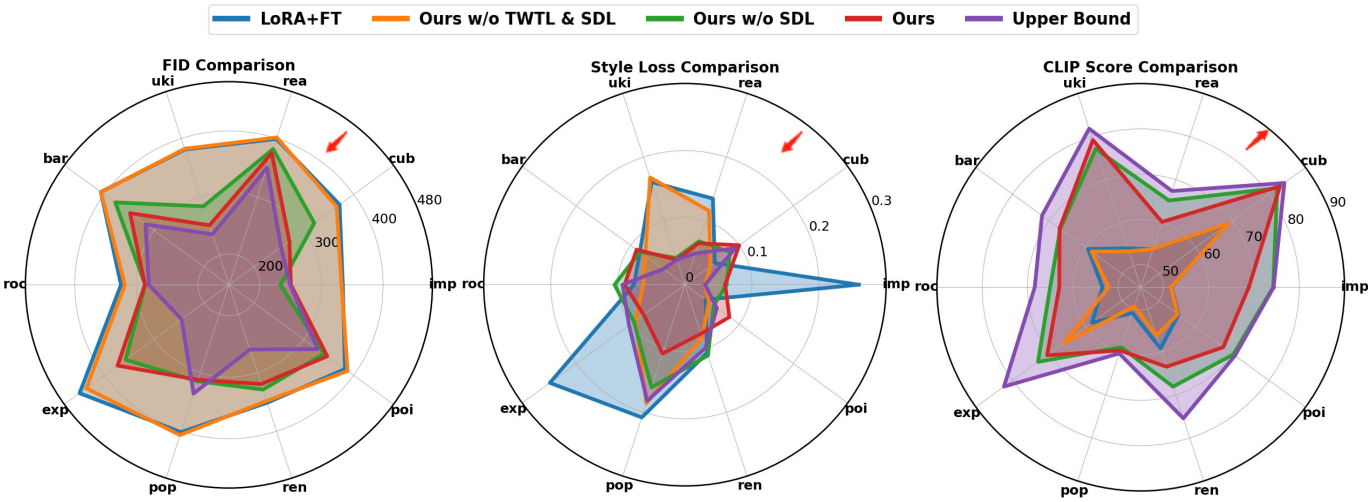

Fig. 6. Style loss, FID and CLIP score for continual style adaptation setting. We conduct ablation studies to evaluate our proposed module one by one. The red arrow in the image indicates the direction where the score improves. Our MuseumMaker achieves the best performance in terms of style loss and FID.

TABLE IV
COMPARISON EXPERIMENTS OF OUR ABLATION STUDIES, WHERE TTL AND SDL REPRESENT TASK-WISE TOKEN LEARNING MODULE AND STYLE DISTILLATION LOSS MODULE, RESPECTIVELY. IN ORDER TO BETTER DISPLAY THE DATA RESULTS, ALL RESULTS OF STYLE LOSS IS SCALED BY 100. METHODS WITH THE BEST AND RUNNER-UP PERFORMANCE ARE MARKED AS BOLDED **RED** AND **BLUE**, RESPECTIVELY.

Comparison Methods	Style Loss⇓											
	imp	cub	rea	uki	bar	roc	exp	pop	ren	poi	**Ave.**	**Imp.**
LoRA+FT	0.259	**0.055**	0.135	0.161	**0.083**	**0.076**	0.249	0.208	0.107	**0.038**	**0.137**	⇑**0.059**
Ours w/o TTL & SDL	**0.030**	**0.047**	0.116	0.168	**0.073**	**0.062**	**0.091**	0.186	**0.083**	**0.044**	**0.090**	⇑**0.044**
Ours w/o SDL	0.062	0.083	**0.068**	**0.042**	0.083	0.105	0.094	**0.161**	0.111	0.053	**0.086**	⇑**0.007**
Ours	**0.060**	0.100	**0.065**	**0.039**	0.089	0.090	**0.076**	**0.108**	**0.077**	0.082	**0.079**	-
Upper Bound	0.030	0.093	0.049	0.039	0.040	0.093	0.103	0.182	0.099	0.059	**0.079**	-

During the inference stage, we apply a DDIM sampler with 50 steps for all methods. All experiments are conducted on a single NVIDIA RTX A6000 GPU, with all generated images having a resolution of 512×512 pixels and based on Stable Diffusion v1.5.

C. Comparison Experiments

As show in Table I and Fig. 5, we present a comprehensive comparison of experiments conducted between our method and other competing approaches. To assess the efficacy of the continual style customization method proposed in this study, we first continually learn all 10 styles available in the Wikiart dataset. Subsequently, we demonstrate the results by generating each style using the same prompt. In order to comprehensively evaluate the performance of the generation, we consider the results of continual style customization from two aspects: qualitative evaluation and quantitative evaluation.

For qualitative evaluation, the results of experiments are shown in Fig. 3. The most competitive methods often suffer from catastrophic forgetting of previously encountered styles. For instance, approaches like LoRA combined with fine-tuning (LoRA+FT) and SPD combined with Fine-tuning (SPD+FT) struggle to preserve the ability to generate the textures and color schemes of previous styles. Simply applying the fine-tuning method with LoRA and SPD restricts the T2I diffusion model from retaining prior knowledge. We combine classical continual learning methods EWC and LWF with existing personalized T2I diffusion model to further compare with our method. The combinations such as LoRA+EWC, LoRA+LwF, SPD+EWC, and SPD+LwF vary degrees of forgetting regarding previously learned styles. For instance, the generated results of the pop art style and expressionism style markedly differ from the samples in the dataset. Moreover, the lack of distinct token embeddings for each style in these methods frequently results in confusion among similar styles. Specifically, combinations such as SPD+EWC and SPD+LwF tend to misclassify the baroque style as rococo style, while our method successfully generates accurate results. These outcomes successfully demonstrate that our method

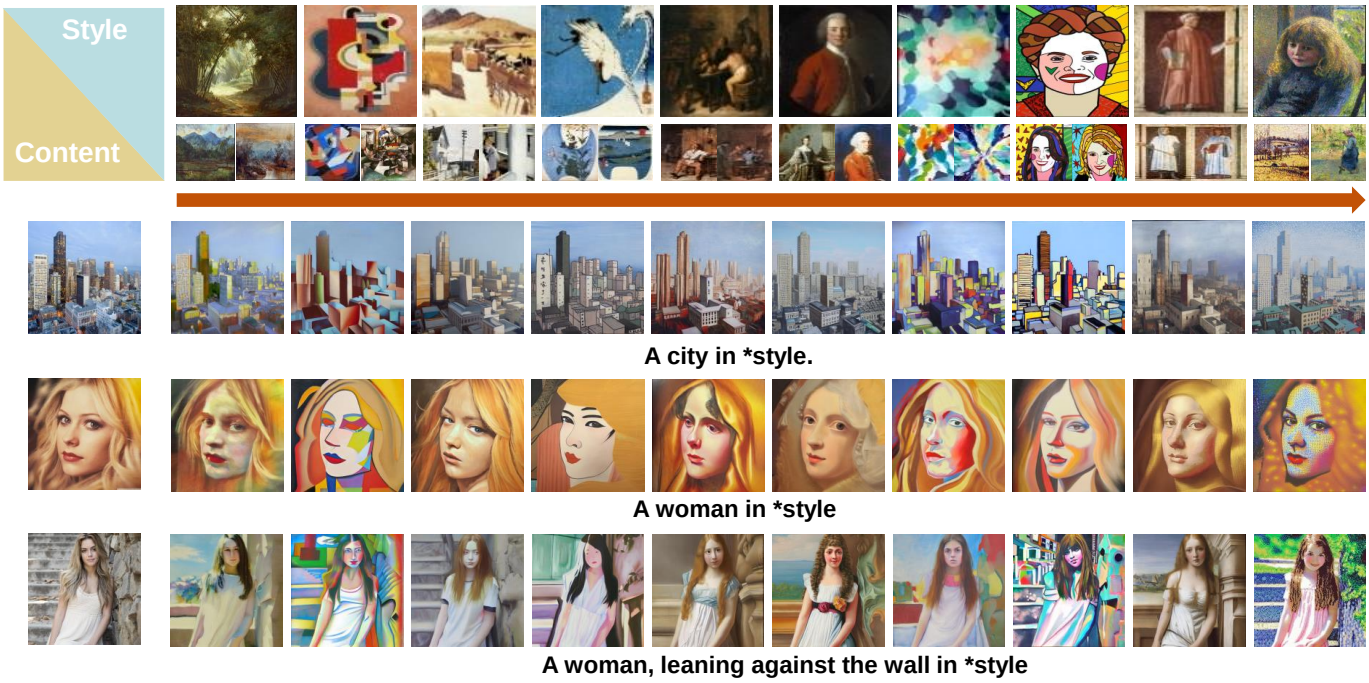

Fig. 7. We conduct the style transfer study with our proposed MuseumMaker, which demonstrates that our method successfully captures various features of 10 styles.

adeptly preserves prior knowledge, enabling the generation of a diverse range of styles while effectively distinguishing between similar styles. Our approach showcases a robust capacity to mitigate catastrophic forgetting and differentiate similar stylistic features. This demonstrates the effectiveness of our dual regularization approach, which operates across both the shared-LoRA module and the task-wise token learning module.

For quantitative evaluation, we conduct three metrics (*i.e.,* style loss, FID, CLIP score) to measure the quality of generated images. As the result shown in Table. I, Table. II and Fig. 5, we compare the performance of all the competing methods in terms of style loss, FID and CLIP score, respectively. Our method achieves the best performance, exhibiting improvements ranging from 20.7 to 67.2 in terms of FID and from 0.96 to 11.77 in terms of CLIP score. Such improvements demonstrate that our method reduce the catastrophic forgetting to old styles and catastrophic overfitting to new styles. In the realm of style loss, SPD+FT attains the highest performance, as it is directly tailored to optimize style loss. However, when it comes to FID and CLIP scores, it fails to demonstrate satisfactory performance. Our approach ranks second in terms of style loss performance. Compared to LoRA-based methods, SPD+LWF and SPD+EWC, our approach exhibits an improvement ranging from 0.002 to 0.058 in style loss. According to the above, our method attains the strongest ability to overcome the catastrophic forgetting and overfitting under the setting of continual style learning for text-to-image diffusion model.

D. Ablation Studies

In the subsection, we conduct ablation studies on the continual learning of 10 styles, which encompasses five experimental settings: LoRA+FT, Ours w/o TTL & SDL, Ours w/o SDL, our complete proposed method and upper bound. For the upper bound setting, we store a unique token embedding and a special LoRA weight for each style. To ensure a fair comparison, all five settings employ the same hyperparameters and number of training steps. Furthermore, we define the same setup as in the comparison experiments to minimize the influence of randomness. Consistent with the comparison experiments, we perform both qualitative and quantitative analyses.

For the qualitative comparison, we evaluate the contribution of each module we proposed, which is depicted in Fig 4. We propose a dual regularization for shared-LoRA (DR-LoRA) module and a task-wise token learning (TTL) module to tackle the catastrophic forgetting. We eliminate these two modules one by one to evaluate their effect, and compare the result generated by ours w/o SDL, ours w/o TTL & SDL and LoRA based fine-tuning (LoRA+FT). As the first three rows of the output section in Fig. 4, the images generated by ours w/o SDL shows the best performance on images generation of past styles, which demonstrates the effectiveness of DR-LoRA and TTL module. Additionally, we devise a style distillation loss for addressing catastrophic overfitting. As shown in the last three rows of Fig. 4, the lack of the SDL module results in overfitting to the image content compared to our proposed method, adversely affecting the generation of pre-trained concepts. The results generated by our MuseumMaker exhibit a similar performance to the upper bound. To further measure the influence of SDL module under different weights, we design a experiments with varying weights, as shown in Fig. 8. When the diffusion model has a higher weight of SDL module, the model more effectively reduces the risk of

Fig. 8. We compare the image generation performance with varying weights of the SDL module on the realism dataset.

TABLE V
THE COMPARISON OF TRAINING PARAMS AND TRAING TIME BETWEEN OUR METHOD AND OTHER COMPETITIVE METHOD.

Methods	# Params	Training Time
LoRA+FT	39.06M	105min
LoRA+LWF	39.06M	145min
LoRA+EWC	39.06M	140min
SPD+FT	859.52M	170min
SPD+LWF	859.52M	290min
SPD+EWC	859.52M	179min
Ours	39.06M	155min
Upper Bound	390.06M	105min

overfitting the content of images.

Regarding the quantitative comparison for ablation studies, we evaluate hundreds of generated images with style loss, FID and CLIP score metrics. As shown in Table. III,Table. IV and Fig. 4, we compare ours, ours w/o SDL, our w/o TTL & SDL and LoRA+FT with three metrics to evaluate the effectiveness of each proposed module. Ours method achieves the best performance on two metrics (*i.e.,* style loss and FID), which improve $0.007 \sim 0.058$ and $9.0 \sim 66.9$ in terms of styles loss and FID, respectively. As we propose DR-LoRA and TTL to tackle catastrophic forgetting, the setting of ours w/o SDL and our w/o TTL & SDL increase $0.015 \sim 0.051$ in terms of style loss, $1.8 \sim 57.9$ in terms of FID and $0.01 \sim 13.61$ in terms of CLIP score. The improvement across these three metrics highlights the effectiveness of the DR-LoRA module and TTL module in addressing catastrophic forgetting. However, the setting of ours w/o SDL achieves the best performance on CLIP score, and our method falls short by 1.84. Lower scores of ours w/o SDL in terms of style loss and FID reveal that ours w/o SDL slightly overfits to the content of images and learns impure represents of styles, which leads to a better adaptability to CLIP score. This phenomenon justifies the importance of SDL module to overcome catastrophic overfitting.

E. Model Efficiency Study

To thoroughly evaluate the efficiency and effectiveness of our proposed MuseumMaker against other competitive methods, we detailedly analyze the training parameters and training time of comparison methods, as presented in Table.V. While LoRA-based methods require a substantial number of parameters (39.06M) and training times ranging from 105 to 145 minutes, these methods show unsatisfactory generation results for continual style customization. Moreover, SPD-based methods utilize a much larger number of parameters (859.52M) and a relatively long training time, ranging from 170 to 290 minutes. Due to such a large amount of computing consumption, SPD-based methods are not efficient enough for continual style customization task. Our approach achieves comparable performance with significantly minimal parameters (39.06M) and a relatively short training time of 155 minutes. Considering the results of image generation, our approach stands out for achieving performance closest to the upper bound with minimal training parameters. These outcomes demonstrate the remarkable effectiveness of our DR-LoRA, task-wise token learning and style distillation loss techniques in the context of continuous style customization tasks, showcasing its potential for practical applications in real-world scenarios.

F. Style Transfer Study

To evaluate the efficacy of our proposed continuous style customization approach in capturing the intricate nuances and diverse characteristics of styles continually provided by users, we devise its application in style transfer studies. This expansion not only broadens the scope of our method but also demonstrates its adaptability to diverse tasks in the field of computer vision. The exploration leverages the accumulated knowledge captured from the continuous stream of style data to facilitate the transfer of style from input images. As depicted in Fig.7, we observe that MuseumMaker adeptly captures the subtle textures, color themes, and structural elements after obtaining knowledge from 10 distinct styles. These results demonstrate the remarkable effectiveness of MuseumMaker in seamlessly learning a multitude of styles for both image generation and image style transfer, showcasing its versatility beyond the realm of text-to-image tasks.

VI. CONCLUSION

This paper explores the challenge of continually learning new user-provided styles and generating a variety of stylistic images with a pre-trained diffusion model. Our proposed MuseumMaker efficiently embeds styles into diffusion model without overfitting to the content of images. Specially, we consider two critical aspects in developing our MuseumMaker, *i.e.,* "catastrophic forgetting" and "catastrophic overfitting". We devise a dual regularization for shared-LoRA module and a task-wise token learning module to address the "catasrtophic forgetting". These modules help the model retain knowledge of previously encountered styles while adapting to new ones. Furthermore, we adopt a style distillation loss to tackle the "catastrophic overfitting". This loss function ensures that the

model learns pure representation of styles without being overly influenced by the content of images. Comprehensive experiments are conducted to show the effectiveness of our method. Results show that our method outperforms existing approaches in terms of style loss, FID and CLIP score, highlighting its ability to generate diverse and high-quality stylistic images.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[3] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

[4] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 021–10 030.

[5] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, and J. Zhang, "Resdiff: Combining cnn and diffusion model for image super-resolution," pp. 8975–8983, 2024.

[6] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[7] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9935–9946.

[8] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in *International Conference on Machine Learning*, 2023, pp. 13 213–13 232.

[9] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[10] T. Hu, J. Zhang, L. Liu, R. Yi, S. Kou, H. Zhu, X. Chen, Y. Wang, C. Wang, and L. Ma, "Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2406–2415.

[11] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," vol. 114, no. 13, pp. 3521–3526, 2017.

[12] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International conference on machine learning*, 2017, pp. 3987–3995.

[13] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European conference on computer vision*, 2018, pp. 532–547.

[14] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European conference on computer vision*, 2018, pp. 139–154.

[15] J. Dong, W. Liang, Y. Cong, and G. Sun, "Heterogeneous forgetting compensation for class-incremental learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 742–11 751.

[16] J. Dong, H. Li, Y. Cong, G. Sun, Y. Zhang, and L. Van Gool, "No one left behind: Real-world federated class-incremental learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[17] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[18] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. Dokania, P. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," in *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.

[19] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," *Advances in neural information processing systems*, vol. 33, pp. 15 920–15 930, 2020.

[20] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *Advances in neural information processing systems*, vol. 30, 2017.

[21] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *International Conference on Learning Representations*, 2018.

[22] W. Cong, Y. Cong, G. Sun, Y. Liu, and J. Dong, "Self-paced weight consolidation for continual learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[23] A. Chaudhry, N. Khan, P. Dokania, and P. Torr, "Continual learning in low-rank orthogonal subspaces," *Advances in neural Information Processing Systems*, vol. 33, pp. 9900–9911, 2020.

[24] K. Javed and M. White, "Meta-learning representations for continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[25] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," *arXiv preprint arXiv:1810.11910*, 2018.

[26] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765–7773.

[27] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International conference on machine learning*, 2018, pp. 4548–4557.

[28] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," *arXiv preprint arXiv:1708.01547*, 2017.

[29] C.-Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Compacting, picking and growing for unforgetting continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[30] Y. Shen, X. Zeng, and H. Jin, "A progressive model to enable continual learning for semantic slot filling," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 1279–1284.

[31] J. Dong, Y. Cong, G. Sun, L. Wang, L. Lyu, J. Li, and E. Konukoglu, "Inor-net: Incremental 3-d object recognition network for point cloud representation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[32] S. V. Mehta, D. Patil, S. Chandar, and E. Strubell, "An empirical investigation of the role of pre-training in lifelong learning," vol. 24, no. 214, pp. 1–50, 2023.

[33] T.-Y. Wu, G. Swaminathan, Z. Li, A. Ravichandran, N. Vasconcelos, R. Bhotika, and S. Soatto, "Class-incremental learning with strong pre-trained models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9601–9610.

[34] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," pp. 16 784–16 804, 2022.

[35] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[36] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[38] W. Li, X. Xu, X. Xiao, J. Liu, H. Yang, G. Li, Z. Wang, Z. Feng, Q. She, Y. Lyu *et al.*, "Upainting: Unified text-to-image diffusion generation with cross-modal guidance," *arXiv preprint arXiv:2210.16031*, 2022.

[39] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng *et al.*, "Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 135–10 145.

[40] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[41] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual representation for controllable image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 370–18 380.

[42] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[43] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese *et al.*, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[44] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. 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*, 2023, pp. 22 500–22 510.

[45] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knn-diffusion: Image generation via large-scale retrieval," *arXiv preprint arXiv:2204.02849*, 2022.

[46] R. Rombach, A. Blattmann, and B. Ommer, "Text-guided synthesis of artistic images with retrieval-augmented diffusion models," *arXiv preprint arXiv:2207.13038*, 2022.

[47] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint arXiv:2209.14491*, 2022.

[48] J. S. Smith, Y.-C. Hsu, L. Zhang, T. Hua, Z. Kira, Y. Shen, and H. Jin, "Continual diffusion: Continual customization of text-to-image diffusion with c-lora," *arXiv preprint arXiv:2304.06027*, 2023.

[49] G. Sun, W. Liang, J. Dong, J. Li, Z. Ding, and Y. Cong, "Create your world: Lifelong text-to-image diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[50] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 571–576.

[51] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 1033–1038.

[52] B. Gooch and A. Gooch, *Non-photorealistic rendering*. AK Peters/CRC Press, 2001.

[53] T. Strothotte and S. Schlechtweg, *Non-photorealistic computer graphics: modeling, rendering, and animation*. Morgan Kaufmann, 2002.

[54] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[55] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, 2017, pp. 214–223.

[56] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 146–10 156.

[57] H. Lu, H. Tunanyan, K. Wang, S. Navasardyan, Z. Wang, and H. Shi, "Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 267–14 276.

[58] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[59] M. N. Everaert, M. Bocchio, S. Arpa, S. Süsstrunk, and R. Achanta, "Diffusion in style," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2251–2261.

[60] Z. Wang, L. Zhao, and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7677–7689.

[61] D. Kotovenko, A. Sanakoyeu, S. Lang, and B. Ommer, "Content and style disentanglement for artistic style transfer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4422–4431.

[62] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[63] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[64] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu *et al.*, "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[65] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *The Eleventh International Conference on Learning Representations*, 2022.

[66] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "$p+$: Extended textual conditioning in text-to-image generation," *arXiv preprint arXiv:2303.09522*, 2023.

[67] S. Mohammad and S. Kiritchenko, "Wikiart emotions: An annotated dataset of emotions evoked by art," in *Proceedings of the eleventh international conference on language resources and evaluation*, 2018.

[68] M. Seitzer, "pytorch-fid: FID Score for PyTorch," https://github.com/mseitzer/pytorch-fid, August 2020, version 0.3.0.

[69] S. Zhengwentai, "clip-score: CLIP Score for PyTorch," https://github.com/taited/clip-score, March 2023, version 0.1.1.

[70] Z. Li and D. Hoiem, "Learning without forgetting," vol. 40, no. 12, pp. 2935–2947, 2017.