Continual Conceptual Entity Learning for Text-to-Image Generative Models

Yabin Wang, Xiaopeng Hong *Senior Member, IEEE*, Zhiheng Ma *Member, IEEE*, Zhou Su *Senior Member, IEEE*, Jinpeng Zhang, Zhiwu Huang

***Abstract*—Current Text-to-Image generative models struggle to continuously learn multiple distinct entities or concepts, limiting their scalability and hindering practical deployment in dynamic environments. We formulate this task as Continual Conceptual Entity Learning (CEL) and propose a novel framework called Continual Entity Adapter Learning (CEAL). CEAL leverages a compact set of tunable parameters, termed SuperLoRA, to efficient and scalable learning of new entities. We propose a dynamic rank-increasing strategy to train the SuperLoRA, balancing computational efficiency with performance. To evaluate our method, we create three benchmarks encompassing generic objects, human faces, and artistic styles. Experimental results demonstrate that CEAL effectively learns new entities while preserving prior knowledge, outperforming existing methods in both entity fidelity and parameter efficiency.**

***Index Terms*—Continual learning, text-to-image synthesis, diffusion models.**

I. Introduction

Recently, Text-to-Image (T2I) generative models, such as GLIDE [1], DALL·E [2], and Stable Diffusion (SD) [3], have achieved remarkable success in generating diverse and complex images based on textual descriptions with high fidelity. However, they often struggle to generate styles, objects, or characters that fall outside the scope of their pre-trained datasets. For instance, while current T2I models are experts at generating images such as "a dog on the moon," they cannot create images such as "our pet dog Buddy on the moon" without being trained on Buddy's images during the pre-training phase. Thus, personalizing pre-trained T2I models for new user-provided concepts (e.g., styles [4], [5], objects [6], [7], or characters [8]) is increasingly in demand and has emerged as a research hotspot.

To bridge this gap, we introduce the task of Continual Conceptual Entity Learning (CEL). CEL involves sequentially integrating new conceptual entities (e.g., specific objects, persons, or styles) into existing T2I models in an ongoing manner over time, crucially, without extensive retraining on previous data. Unlike previous continual learning approaches for generative models [9]–[14] focused on class-level generation or single-session multi-concept customization methods (e.g., [33]) that learn multiple concepts simultaneously, CEL emphasizes entity-level sequential learning. The model must capture and reproduce individual entities' unique characteristics and details (see Fig. 1). The primary challenges of CEL stem from three core requirements: (1) sequentially learning new, distinct entities effectively from limited data, (2) mitigating catastrophic forgetting of previously acquired entities, and (3) achieving this in a parameter-efficient manner to ensure scalability.

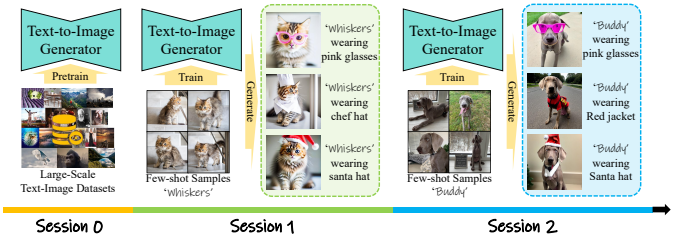

Fig. 1: In CEL, the model sequentially learns new entities (e.g., pets named 'Whiskers' or 'Buddy') from limited samples in each learning session. After learning, the T2I generator can generate images of both newly learned and previously known entities based on text prompts.

Existing approaches [6], [7] for T2I personalization are mainly designed for single-stage incremental learning and are not well-suited for the continuous, sequential learning of multiple new entities over time. When multiple new entities are continually introduced, these existing methods struggle to maintain both newly learned and previously acquired knowledge, leading to catastrophic forgetting. Recently, Low-Rank Adaptation (LoRA) [15], [16] has emerged as an efficient method for fine-tuning T2I models using low-rank adapters, allowing for scalable learning of numerous entities with small parameter expansion. However, the storage requirements for new entities remain a significant challenge, particularly in resource-constrained environments. Additionally, our findings indicate that directly merging multiple LoRA weights can result in conflicts, preventing the generation of high-quality outputs of new concepts.

To overcome these challenges, we propose a novel framework, termed **C**ontinual **E**ntity **A**dapter **L**earning (**CEAL**), for CEL. Grounded in the theoretical understanding of parameter adaptation and network function, CEAL proactively searches for a layer-wise LoRA allocation for each entity that can be integrated with minimal conflict, thereby ensuring entity fidelity, knowledge preservation, and generalization ability.

CEAL's design is grounded in the following theoretical

Y. Wang and X. Hong are with Harbin Institute of Technology, Harbin, China (e-mail: wang-yabin@outlook.com and hongxiaopeng@ieee.org). Z. Su is with Xi'an Jiaotong University, Xi'an, China (e-mail: zhousu@ieee.org). Z. Ma is with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China (e-mail: zh.ma@siat.ac.cn). J. Zhang is with Intelligent Science & Technology Academy Limited of CASIC, Beijing 100043, China. (e-mail: 2076707060@qq.com). Z. Huang is with the University of Southampton, Southampton, UK (e-mail: Zhiwu.Huang@soton.ac.uk).

principles and empirical observations: First, following the principle of Functional Localisation observed in deep learning models [17], diffusion models feature functionally distinct layers, each specializing in specific image generation aspects such as content, style, or color attributes [18], [19]. This Layer-wise Functional Specialization thus requires optimized, layer-specific LoRA configurations and provides the theoretical grounding for strategically allocating these parameters with varying ranks to mitigate interference. Second, the foundation of LoRA relies on the Intrinsic Dimension Hypothesis [20], suggesting adaptation lies in a low-dimensional subspace. Furthermore, consistent with the Lottery Ticket Hypothesis [21], which implies sparse weight updates can be sufficient, LoRA weights typically exhibit high sparsity. This means inter-entity conflicts are largely confined to a small parameter subset, thus leaving ample room to integrate new knowledge with minimal interference. Third, LoRA modules typically lack orthogonality, and their naive merging often leads to destructive interference from conflicting entity information.

Thus, CEAL allocates LoRAs with varying ranks to each layer instead of using a uniform rank. To implement this, we introduce a new structure called *Superlora*, a rank-degradable LoRA architecture viewed as a combination of multiple low-rank LoRAs, which enables weight-sharing. We train *Superlora* for each entity and search for the optimal configuration based on multiple criteria. The optimized LoRA weights are then merged into the base model, allowing for the generation of new entities without increasing the overall parameter count.

The training quality of *Superlora* directly affects the search results. Since exhaustively training all possible LoRA configurations is computationally infeasible, an efficient strategy is crucial to explore the vast configuration space effectively. We introduce a dynamic rank-increasing strategy that gradually expands the search space by introducing new rank parameters alongside the training steps. This approach allows *Superlora* to initially focus on learning essential low-rank configurations before progressively exploring more complex, higher-rank configurations. After training *Superlora* for a new entity, an optimizer explores *Superlora* using multiple criteria to determine the optimal LoRA rank configuration for that entity.

To facilitate standardized evaluation of methods addressing the unique challenges of CEL, we construct three benchmarks that encapsulate a range of different concepts, including generic objects, human faces, and distinctive artistic styles. In addition, we design evaluation metrics for CEL that evaluate the model's performance on new tasks and its susceptibility to forgetting in continuous learning. Extensive experiments on three benchmarks demonstrate the effectiveness of CEAL. Code and data will be publicly available.

The technical contributions are manifold:

- We propose a framework, CEAL, to continuously learn entities using a compact set of tunable parameters. To optimize CEAL, we define three key CEL criteria that encompass concept fidelity, parameter efficiency, and aesthetic quality.
- We propose a novel rank-degradable LoRA architecture called *Superlora*. This architecture allows for weight-sharing and dynamic allocation of LoRA ranks across different layers, enabling efficient and scalable searching. Our dynamic rank-increasing strategy further stabilizes the training process, ensuring effective learning of new entities.
- We establish three benchmarks for CEL. Each benchmark focuses on different aspects of the task, allowing for comprehensive evaluation.

II. Related Works

A. Text-to-image Personalization

Text-to-image (T2I) generation models [1], [22]–[31] have received unprecedented interest from the community in recent years. To achieve text-to-image generation, these models typically first employ a language encoder, such as CLIP [32], to encode user text input into a latent representation. This latent representation then serves as a conditional input, with the model subsequently trained on large-scale paired image-text datasets to generate corresponding images.

With the rapid advance of T2I models, their personalization (*a.k.a.* customization) is becoming increasingly crucial. Personalization methods [33]–[35] tailor a model to the specific needs of an individual or group by utilizing data unique to the intended users. DreamBooth [6] fine-tunes all parameters of the diffusion network by just giving a few images of the target entity. DreamBooth uses the pre-trained model to generate regularization data about similar concepts to relieve forgetting. However, tuning all parameters on few-shot images would increase the risk of the model forgetting previously learned knowledge and overfitting. Textural Inversion [7] proposes a concise way for customization by solely optimizing a set of word embeddings to portray a novel concept while leaving the denoising model fixed. Although it effectively preserves the original model's knowledge, it may have restricted learning capability and may encounter difficulty in comprehending intricate entities. Cones [19] identifies and manipulates concept neurons within diffusion models to enable efficient and customizable multi-subject image generation. InstantBooth [36] captures subject identity by representing the general concept as a textual embedding and feeding fine-grained visual details through lightweight adapter layers into the frozen model. HyperDreamBooth [37] uses a HyperNetwork to generate compact, personalized weights from a single image, enabling extremely fast personalization of T2I models. Custom Diffusion (CD) [38] combines these two strategies for multi-concept learning. CD learns a set of textual embeddings for a given concept and fixes the parameters of the text encoder, only fine-tuning a small subset of parameters in the cross-attention layers of the denoising model, specifically the key and value matrices. Concept Weaver [39], building upon CD, then addresses how to effectively fuse the visual features of these multiple distinct concepts into a single image at inference time. It first generates a template image aligned with the text prompt, then utilizes spatial region masks to inject the visual appearances of multiple customized concepts into their respective regions. Similarly, ED-LoRA [40] merges multiple LoRAs into a single model using Gradient Fusion and employs Regionally Controllable Sampling to accurately place the customized concepts and their attributes in images. LoRA-Composer [41] is a training-free method that

leverages the cross-attention mechanism to inject concept-specific LoRA features into image regions designated by user-provided layout conditions. MC2 [42] enables training-free composition of heterogeneous single-concept models (LoRA or TI) using inference-time optimization. Its Multi-concept Guidance (MCG) refines attention weights to spatially disentangle concepts and ensure their faithful representation while minimizing interference. However, these works primarily focus on composing multiple pre-existing customized concepts. The CEL task, in stark contrast, centers on the continual, sequential learning of new entities and the critical mitigation of catastrophic forgetting.

Recent works [16] adopt parameter-efficient fine-tuning methods, such as LoRA [15], to learn new entities without tuning the entire denoising network. Building upon the parameter efficiency of LoRA, some works [43]–[45] focus on optimizing the allocation of trainable parameters under the constraint of limited storage resources. DyLoRA [43] trains LoRA with a range of ranks, enabling dynamic adjustment of model capacity without retraining. AdaLoRA [44] adaptively allocates the parameter budget among different weight matrices based on their importance scores. IncreLoRA [45] incrementally adds trainable parameters during training. However, while these methods enhance LoRA's parameter efficiency, they are not tailored for the CEL task. They typically lack a proactive mechanism to automatically determine optimal, layer-wise LoRA configurations that minimize inter-entity conflict during sequential learning, thus often requiring users to define these specific CEL-focused setups. Moreover, *stand-alone* addition of each new concept brings substantial modifications to the model parameters, leading to an inevitable rise in storage burden and knowledge forgetting.

B. Continual Learning

Continual learning is a machine learning paradigm that learns new data over time without forgetting previously learned knowledge. Three primary strategies are used to address the challenge of *catastrophic forgetting*. *Rehearsal* methods allow a model to partially access and utilize data from previous sessions when learning new tasks. Previous data can be obtained by storing samples from past tasks [46]–[56] or synthesizing through generative models [9], [57]–[60]. *Regularization* methods [48], [61]–[68] use the knowledge from the old model to guide and constrain the learning process for new tasks, by techniques such as knowledge distillation. *Network Expansion* methods [69]–[73] augment the model's parameters to incorporate new knowledge without interfering with previous parameters, using methods like network expanding-pruning [74] and Neural Architecture Search (NAS) [75], [76]. Recently, parameter efficient tuning, such as prompt-tuning [77]–[80] and adapter [81], shows a promising way for continual learning on pre-trained models. However, most studies focus only on a class-level continual learning scenario, where the *class* serves as the learning unit. Consequently, they cannot be directly applied to the CEL scenario, where the learning unit is the *entity* rather than the *class*. Recently, C-LoRA [81], [82] introduces the use of LoRA to learn each entity for T2I diffusion models. However, C-LoRA stores all LoRA parameters in memory, which becomes problematic when dealing with a large number of entities.

III. Methods

A. Preliminary and Problem Formulation

Preliminary for T2I Diffusion Models: To achieve high-quality and creative image generation, T2I diffusion models [3], [83], [84] have become a hot topic of current research. T2I diffusion models are multimodal generators that learn precise correspondences between textual descriptions and images. Based on these learned correspondences, trained models generate images aligned with given textual prompts through a progressive denoising process. Given a diffusion model θ, the training objective can be derived as Eq. 1.

$$\mathbb{E}_{(\mathbf{x}_0,\mathcal{C},\epsilon,t)}\left[\|\epsilon-\hat{\epsilon}_\theta(\mathbf{x}_t,\mathcal{C},t)\|^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,\mathbf{I})$ is a random noise, $\mathbf{x}_0$ is the original clean image, $\mathbf{x}_t$ corresponds to the noised image at the t-th timestep, and $\mathcal{C}$ is the text embeddings as condition.

Problem Formulation: In CEL, we continually insert knowledge of new entities into a pre-trained T2I diffusion model by reducing the reconstruction loss 1. An *entity* refers to any conceptual or physical item, such as an object, character, or artistic style.

Formally, the model learns sequentially over S sessions, where each session introduces a new entity. In the session s, the T2I model θ encounters new data $D_s = \{(x_i^s, c_i^s)|i \in N^s\}$ of a specific entity v. Here, N^s represents the number of training samples. c_i^s denotes the textual description corresponding to the image x_i^s. It follows a structured template like "a photo of `id`," where `id` is the identifier for the newly introduced entity. Once trained, we can use `id` in text prompts and the model θ to create the desired images. This process is repeated until all S entities are learned. θ is expected to have the capability to generate images containing any combination of the learned S entities, as shown in Fig. 5.

B. Empirical Analysis of LoRA

LoRA [15] is an effective approach for fine-tuning the denoising model to incorporate new entities. Unlike traditional fine-tuning, LoRA introduces compact, low-rank matrices to adapt existing pre-trained weights. Only these new matrices are updated during training, keeping the original model parameters frozen. Formally, given a pre-trained T2I model θ, the weights of the l^{th} layer is $W^l \in \mathbb{R}^{n_l \times m_l}$, where n_l and m_l represent the input and output dimensions of the l^{th} layer, respectively. The calculation of low-rank adaptation applied to W^l can be expressed as Eq.2.

$$\hat{W}^l = W^l + A_r^l (B_r^l)^T, \tag{2}$$

where $A_r^l \in \mathbb{R}^{n_l \times r}$ and $B_r^l \in \mathbb{R}^{m_l \times r}$ are the low-rank matrices for the l^{th} layer and r represents the rank of LoRA.

To analyze the behavior of LoRA in personalizing T2I models, we conduct a series of experiments focusing on the entities 'dog' and 'cat' within our CEL benchmarks.

Fig. 2: **Analysis of LoRA in T2I model personalization.** (a) Performance of pruned LoRA models for 'dog' and 'cat' concepts at different ranks. (b) Layer-wise distribution of LoRA update matrix ΔW. (c) Layer-wise cosine similarity between 'dog' and 'cat' LoRA weights.

We train two independent LoRAs with a rank of $r = 32$ for each entity. After training, we prune the trained LoRA to certain ranks using SVD decomposition. Interestingly, as illustrated in Fig. 2 (a), we can find that despite removing half of the parameters, the LoRA models maintain their capability to effectively generate the intended concepts. This suggests the sparsity, where many components of the LoRA weights have minimal contributions. This observation strongly supports the Lottery Ticket Hypothesis [21] in the context of LoRA, implying that a sparse weight update is often sufficient. Consequently, sacrificing some of these weights may not significantly impair LoRA's overall functionality. Similar observations have been reported in related studies [85], [86].

Then, we conduct a layer-wise analysis of the parameter distribution in LoRA. As shown in Fig. 2 (b), we find that a significant portion of the parameters ΔW in LoRA have magnitudes very close to zero, indicating they have little impact on the output of the pre-trained models. Moreover, the weight distributions of LoRA vary across different entities, suggesting that various entities need distinct parts of the model. This observation implies that different layers within diffusion models are not uniform in function. This provides direct empirical evidence for the principle of Functional Localisation [17] within diffusion models, aligning with findings in related research [18], [19], and theoretically justifying our strategy of allocating layer-specific ranks rather than a uniform rank.

Finally, we computed the layer-wise cosine similarity distribution between the LoRA weights for 'dog' and 'cat', as depicted in Fig. 2 (c). Despite a few layers exhibiting non-zero cosine similarity, the majority of the layers demonstrate values exceedingly close to zero, implying a high degree of orthogonality. This observation suggests that directly merging the two LoRA weights together might cause the knowledge to overlap, potentially leading to less distinct representations.

C. Continual Entity Adapter Learning Framework

Based on the aforementioned theoretical principles and empirical analysis, we identify several limitations of traditional LoRA methods, including inefficient parameter usage and challenges in merging knowledge. These drawbacks make traditional LoRA unsuitable for CEL scenarios, where LoRA must sequentially incorporate new entities while maintaining parameter efficiency and ensuring seamless integration of multiple adaptations.

To address these challenges, we propose the framework, CEAL, to continuously and efficiently integrate a sequence of new entities into a pre-trained T2I diffusion model, as illustrated in Fig. 3.

The CEAL framework operates in two main stages. First, we introduce *Superlora*, a novel structure designed to effectively manage a wide spectrum of LoRA configurations within a single adapter. Through a dynamic rank-increasing strategy, *Superlora* is a flexible foundational adapter with adjustable rank complexity. This core characteristic enables the direct derivation of diverse sub-LoRA configurations, without the need for retraining. This significantly simplifies the search for optimal LoRA configurations across the extensive parameter space.

Second, leveraging the trained *Superlora*, we search for the optimal layer-wise rank allocation. This involves sampling LoRA configurations, generating images, and evaluating them against multi-objective criteria, specifically entity fidelity, parameter efficiency, and aesthetic quality, to identify compact and high-performing LoRAs.

The resulting LoRA can then be seamlessly merged into the base T2I model, providing a parameter-efficient and effective solution for the continual integration of new entities.

D. Superlora Structure and Efficient Training Strategy

This section details the *Superlora* structure and the efficient training strategies, specifically weight-sharing and dynamic rank-increasing, designed to manage its vast configuration space and enable its rank-degradable properties. To detail the *Superlora* architecture, we first establish the following definitions:

Definition 1: For a given layer l, a LoRA module attached to the original weight matrix W^l can be expressed as Eq.3.

$$\hat{W}^l = W^l + A^l_{r_l}(B^l_{r_l})^T, \tag{3}$$

where $A^l_{r_l} \in \mathbb{R}^{n_l \times r_l}$ and $B^l_{r_l} \in \mathbb{R}^{m_l \times r_l}$ are the low-rank matrices for the l^{th} layer with rank r_l.

Definition 2: The *Superlora* structure, mathematically defined as a set $\mathcal{R}$, encompasses all possible combinations of LoRA configurations across the network's layers. It is defined as:

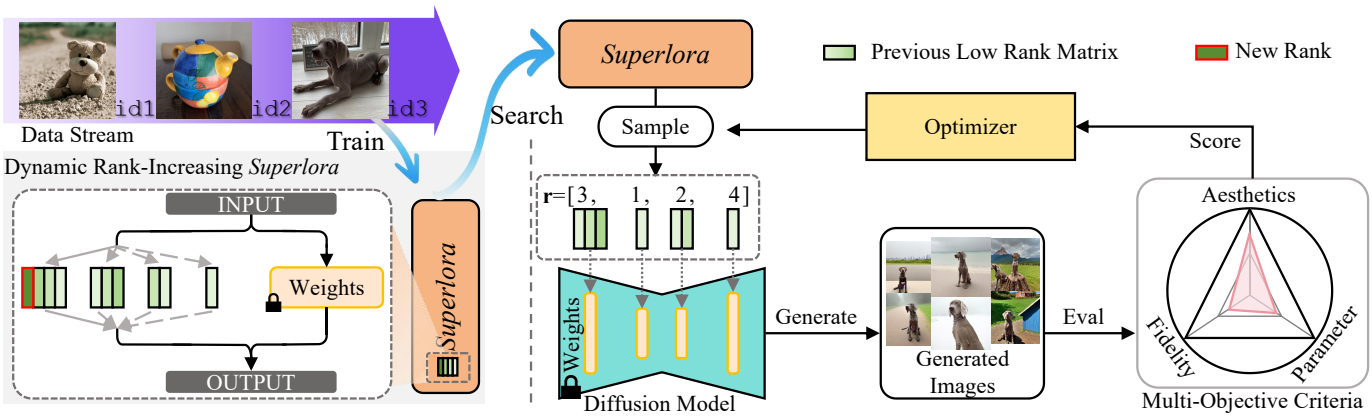

Fig. 3: **Overview of CEAL Framework.** In a training stage, CEAL continually integrates new entities into a pre-trained T2I diffusion model through a two-step optimization process, which consists of a prerequisite training process (**left**) with a rank-degradable LoRA structure, *Superlora*, and a search and optimization process (**right**) to identify the optimal LoRA configuration and optimize the parameter settings according to multiple objectives.

$\mathcal{R} = \left\{ \left(A^l_{r_l}, B^l_{r_l} \right) \mid l = 1, 2, \ldots, N \right\}$ where N is the total number of layers in the model.

Definition 3: A specific LoRA configuration within *Superlora* is represented by a rank vector $\mathbf{r} = [r_1, r_2, \ldots, r_N]$, where each r_l satisfies $0 \leq r_l \leq \hat{r}$ and $r_l \in \mathbb{Z}$. This vector $\mathbf{r}$ specifies the rank of the LoRA adaptation at each layer of the network. $\hat{r}$ is the maximum rank of *Superlora*. When $r_l = 0$, the LoRA module is not applied to the l^{th} layer.

Consequently, exhaustively training all these possible LoRA configurations is computationally intensive and time-consuming. For instance, a diffusion model with $N = 16$ layers and a maximum rank $\hat{r} = 8$ per layer would yield approximately 1×10^{14} possible rank configurations. Training such a vast number of configurations is impractical.

To address this impracticality and train *Superlora* efficiently, we employ a weight-sharing strategy. The training process does not involve training each sub-network independently. Instead, at each iteration, a specific configuration $\mathbf{r}$, is randomly sampled. Subsequently, only the parameters within the sampled rank are updated, with all other parameters kept frozen during that step. This strategy significantly reduces the computational burden of training the SuperLoRA structure itself, as it avoids the need to train all its numerous potential configurations separately, thereby enabling efficient exploration of its vast configuration space specifically during this training phase.

While weight-sharing makes *Superlora* training feasible, its vast search space can still lead to inefficient learning from uniform random sampling. To effectively train the *Superlora* in such a large search space, we introduce the dynamic rank-increasing strategy, which is similar to the curriculum learning paradigm. Initially, we set the maximum rank $\hat{r}$ to 1, focusing the training on configurations with minimal complexity. We gradually allocate more rank parameters to *Superlora* as training progresses. This rank increase operates at regular intervals, denoted by ν, throughout the training process. At each interval, we increment $\hat{r}$ by a predefined step size until it reaches a maximum rank hyperparameter, r_{final}.

By gradually increasing the rank, we allow the model to first learn low-rank adaptations, which capture the most significant features with fewer parameters. This facilitates a more stable and efficient training process. As higher ranks are introduced, the model progressively captures more complex representations without overwhelming the training dynamics. This strategy also ensures that all ranks up to r_{final} are adequately trained.

E. Searching with Multi-objective Criteria

To proactively search for the optimal LoRA for any entity and ensure high-quality image generation, we define a multi-objective function L that incorporates three essential components: aesthetic quality, parameter efficiency, and entity fidelity.

Firstly, to assess aesthetic quality, we employ an aesthetic score predictor [87], [88] trained on human aesthetic preference data. This predictor evaluates the generated images based on prompts and assigns a score, referred to as the *Human Preference Metric* (L_{hpm}), indicating the aesthetic appeal of the images. Higher scores reflect images that are more likely to be preferred by humans.

Secondly, the *Memory Cost Metric* (L_{mcm}) quantifies the number of parameters in the LoRA modules. An unconstrained number of parameters may lead to overfitting and catastrophic forgetting [67]. We calculate L_{mcm} by summing the number of parameters of the selected LoRA modules, ensuring that the LoRA has a compact storage size. This ensures that only the necessary parameter changes are performed.

Thirdly, the *Concept Similarity Metric* (L_{csm}) assesses the content fidelity between generated images and target entities. This is achieved by S-Prompts [79], a continual AI-generated detector. Specifically, S-Prompts is a binary classifier built upon the pre-trained CLIP. For each new entity, this classifier is trained to distinguish the entity's authentic training images from those generated by the model. The resulting classification accuracy of S-Prompts directly constitutes the L_{csm} score,

where higher accuracy indicates greater fidelity of the generated images to the characteristics of the target entity.

The overall function of L is a weighted sum of these three metrics as Eq. 4.

$$L = -\alpha L_{\text{csm}} + \beta L_{mcm} - \gamma L_{hpm}. \tag{4}$$

where α, β, and γ are hyperparameters that balance the magnitude of differences metrics.

Algorithm 1 Optimizer for CEAL

Input: Search space R, Initial temperature T_0, Cooling rate α, Total iterations N_I, Objective Function L
Output: Optimal solution
Initialize a solution $\mathbf{r}$ of length 16, with randomly selected values from R
Set current temperature $T = T_0$
for i in 1 to N_I **do**
    Generate a new solution $\mathbf{r}'$ by random modifying $\mathbf{r}$
    Generate images by $\mathbf{r}$ and $\mathbf{r}'$
    Evaluate $\mathbf{r}$ and $\mathbf{r}'$ using L, get fitness scores f and f'
    if $f' < f$ **then**
        Accept $\mathbf{r}'$ as the new current solution
    else
        Accept $\mathbf{r}'$ with probability $\exp(-\frac{f'-f}{T})$
    end if
    Update temperature $T = \alpha \times T$
end for
Return the final solution $\mathbf{r}$

Optimizing this multi-objective function poses significant challenges due to the discrete and high-dimensional nature of the search space. Traditional optimization methods, such as grid search or random search, are inefficient and may not effectively explore the vast combinations of rank values across different layers. Gradient-based methods are unsuitable because the rank values are discrete and non-differentiable.

To address these challenges, we use Simulated Annealing (SA) as our optimization algorithm, detailed in Algorithm 1. SA is a probabilistic technique well-suited for discrete optimization problems and can effectively explore large search spaces. It mimics the annealing process in metallurgy, where a material is heated and then slowly cooled to decrease defects and find a low-energy state.

In the context of *Superlora*, each potential combination of rank values across all block layers of the denoising U-Net in Stable Diffusion constitutes a solution, represented as a sequence $[r_1, r_2, ..., r_n]$. SA effectively explores this discrete space by initiating with a high temperature ($T_0 = 100$), allowing broad exploration of the solution landscape. At each iteration, it generates a new solution by making slight modifications to the current one, promoting diversity and avoiding premature convergence. We evaluate the fitness of both solutions using the objective function L, ensuring progress toward optimality. As the algorithm progresses, we decrease the temperature according to a cooling schedule with a rate of α. This strategy allows SA to extensively explore the search space initially and progressively focus on refining the solution, thus avoiding local minima and increasing the likelihood of finding the global optimum.

Upon obtaining a solution, we extract the corresponding LoRA modules from *Superlora* and integrate them into the base model 2. We then generate images with fixed parameters to evaluate performance L.

In CEL, we sequentially obtain a set of optimal LoRA modules for each entity and merge them into the base model. This seamless integration results in no additional computational overhead during inference and eliminates memory requirements for continual learning.

IV. Benchmark Design and Construction

A. Datasets Collection

We collect three sets as benchmarks for CEL. The benchmarks cover general objects, human faces, and artistic styles. **CEL-Objects** comprises 15 common objects, each represented by 4-6 images, sourced from [7], [38]. The dataset includes various categories such as buildings, animals, and household items. The benchmark sequence consists of 'barn', 'cat', 'cat statue', 'clock', 'colorful teapot', 'dog', 'elephant', 'mug skulls', 'physics mug', 'red teapot', 'round bird', 'teddybear', 'thin bird', 'tortoise plushy', and 'wooden pot'.

CEL-Faces is derived from the high-resolution CelebA-HQ dataset [89]. We select seven identities, each with a minimum of 20 samples, based on the highest number of available images. These celebrity identities serve as unique conceptual entities for our incremental learning task. The continual learning sequence comprises '#1 woman', '#2 man', '#3 man', '#4 man', '#5 man', '#6 woman', and '#7 woman'.

CEL-Styles features $1,207$ images representing 6 distinct artistic styles, collected from Wikiart [90]. The dataset showcases works by renowned artists representing various art movements, including Neo-impressionism, Impressionism, Post-Impressionism, Mathematical art, Russian Revivalism, and Modernism. Each image is accompanied by a descriptive caption. The dataset is split into training and validation sets using an 8:2 ratio for each artist. The sequence of continual learning comprises 'Georges Seurat', 'Konstantin Korovin', 'Maurice Prendergast', 'M.C. Escher', 'Boris Kustodiev', and 'Marc Chagall'.

B. Metrics for Evaluation

The CEL task encompasses the requirements of both continuous learning and T2I generation. Existing metrics for both of these two tasks are insufficient for comprehensive evaluation. To comprehensively assess the performance of CEL, we integrate the T2I metrics into continuous learning scenarios to form a (single-stage) performance indicator and then derive two CEL performance metrics.

The Entity Generation Performance Indicator (EGI, denoted as ρ) measures the performance during a single learning session. We define ρ as the *average* of three commonly used T2I generation metrics [3], [6], [38]: the DINO Score, the CLIP-I Score, and the CLIP-T Score.

The DINO and CLIP-I scores assess visual similarity, measuring the similarity between features of the generated

images and the target images using DINO [91] and CLIP [32], respectively. The CLIP-T Score evaluates textual similarity, defined as the average cosine similarity between the CLIP features of the generated images and their corresponding text prompts. Formally, ρ_i^j denotes the performance of the model on the i-th task after learning j sessions. Based on ρ, we define two metrics:

Incremental Concept Accuracy (ICA) measures the generation performance of new entities after the model has completed all continual learning sessions. It is defined as:

$$ICA = \frac{1}{S}\sum_{i=1}^{S}\rho_i^S, \tag{5}$$

where S is the total number of continual learning sessions.

Incremental Concept Forgetting (ICF) quantifies the decrease in performance on previously learned concepts after learning new ones. It is defined as:

$$ICF = \frac{1}{S-1}\sum_{i=1}^{S-1}\left|\rho_i^i - \rho_i^S\right|, \tag{6}$$

where ρ_i^i is the performance on the i-th task immediately after learning it, and ρ_i^S is the performance on the i-th task after all sessions have been completed.

V. Experiments

In this section, we present a comprehensive evaluation of our proposed CEAL approach for continual concept learning, encompassing quantitative and qualitative analysis.

A. Evaluation Setup

We implemented baselines of continual learning methods for CEL task: generative replay (Replay), EWC [61] and C-LoRA [82]. Additionally, we report the performance of state-of-the-art personalization methods, including Textual Inversion (TI) [7], DreamBooth (DB) [6], ED-LoRA [40] and Custom Diffusion (CD) [38]. We also compare our method with LoRA [15], using a rank of $r = 8$ for each concept. We train independent LoRA adapters for each entity and keep them in memory, which serves as a potential upper bound for CEL and offers additional context for our results. For CEAL, we set $r_{final} = 8$, which means that the maximum rank obtained by CEAL is consistent with that of LoRA with $r = 8$.

B. Implementation Details

We conducted our experiments using Stable Diffusion v1.5 [3] as the base model. The training process utilized the Mean Squared Error (MSE) loss function, and the LoRA was integrated into all attention layers 16 of the denoising U-Net architecture. *Superlora* begins with an initial rank of 1, which is progressively incremented by 1 after every $1,000$ training iteration, resulting in a final rank (r_{final}) of 8 after $8,000$ iterations. Training was carried out with a batch size of 4. For optimization, we used the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01. All input images were preprocessed to a resolution of 512×512 pixels.

For the search optimizer, we employ Simulated Annealing (SA), as presented in Algorithm 1. In SA, a solution is a sequence of rank values, where each value is selected for a corresponding layer in Superlora. The length of each solution is set to 16, corresponding to the 16 block layers in the denoising U-Net of SD. We initialize the SA parameters with an initial temperature (T_0) of 100, a cooling rate (α_{SA}) of 0.95, and a total number of iterations (N_I) of 1000. Our search space is defined as the set $\{0, 1, 2, 4, 6, 8\}$, meaning that each element in a solution can take on one of these values. The SA algorithm explores this designated space to find the optimal solution. The fitness function of SA is determined by the objective function L. Given a specific solution, we first extract the LoRA from *Superlora* with the corresponding ranks. Then we merge the extracted LoRA into the SD model and generate several images (specifically, 4 images per prompt, using prompts listed in the Supplementary Material). These images have dimensions of 512×512 pixels, a guidance scale of 7.5, and 50 sample steps. The seeds for sampling are also fixed. In our main results, we simply set the hyperparameters α, β, and γ in the objective function to 100, 0.00001, and 0.1, respectively, across all benchmarks. The purpose of this setting is to normalize the magnitudes of these objective values to the same scale, ensuring a balanced contribution from each term during the optimization process. This serves as a standardized baseline, and with minimal tuning, the search results can be adapted to meet specific requirements.

C. Main Qualitative Results

To evaluate the effectiveness of our continual learning approach, we conducted qualitative assessments for single-concept fidelity and multi-concept compositionality.

Figure 4 presents a visual comparison of different methods on the CEL-Faces dataset. The images shown are for the final three identities learned after all methods completed the full sequence of seven continual learning sessions. The results clearly highlight the limitations of existing approaches. For instance, the DB baseline suffers from complete mode collapse, failing to generate any coherent images. Other methods, while avoiding outright collapse, exhibit significant forgetting, producing faces that bear little resemblance to the target identities. In stark contrast, our method (CEAL) successfully generates high-fidelity images that are visually consistent with the target subjects, demonstrating its superior ability to mitigate catastrophic forgetting.

We further challenged the models with a multi-subject generation task on the CEL-Objects dataset, with results displayed in Figure 5. This evaluation was performed after the models had continually learned all 15 distinct entities. Although the CEL-Objects dataset primarily consists of simple, distinct items, and methods like TI and LoRA can achieve high-fidelity representations for single entities, they face significant challenges in multi-concept generation. This limitation likely stems from their training strategies, which do not explicitly optimize for the simultaneous generation of multiple entities.

In contrast, our proposed method, CEAL, excels in this complex task. It not only generates high-quality images that

Fig. 4: **Visualizations of continual generation results on CEL-Faces.** We present visualizations of generated images from various methods for newly learned entities upon completion of all 7 continual learning sessions.

accurately depict newly learned entities but also seamlessly integrates them with previously learned concepts. As shown, our approach successfully captures the complex interactions between concepts in a single frame, demonstrating remarkable performance in multi-concept compositional generation. Additional qualitative results are provided in the supplementary material.

D. Main Quantitative Results

Table I presents a comparative analysis of ICA and ICF across three benchmarks: CEL-Objects, CEL-Faces, and CEL-Styles. The primary focus of ICA is to evaluate the fidelity of new entities after continual learning, whereas ICF measures the degree of forgetting after learning new entities.

Methods such as Replay, EWC, and DB sequentially fine-tune the denoising network without increasing the number of parameters ($\#P$). TI only requires adding a single token, so its $\#P$ is almost negligible. Although these methods incorporate knowledge replay and regularization to prevent forgetting, they still exhibit inferior performance compared to parameter expansion methods, as indicated by the increase in ICF. While TI generally performs well on CEL-Faces and CEL-Styles benchmarks with no forgetting problem, its performance is suboptimal on the more challenging CEL-Objects dataset. This may indicate that learning human faces and art styles is noticeably easier for these methods than CEL-Objects, possibly due to the T2I model having a more intuitive grasp of these concepts. Our method significantly outperforms these approaches without requiring additional parameters.

For methods that expand parameters for each new entity, we report the average increase in parameters ($\#P$). LoRA increases parameters by $2.59M$ per session, CD requires an increase of $19.17M$ parameters, and ED-LoRA needs $0.81M$ per session. Although ED-LoRA and CD both employ LoRA fusion strategies, their fusion is centralized, meaning that when generating multiple concepts together, they merge LoRAs for better results but must keep all independent LoRAs in memory. C-LORA merges LoRAs together but keeps separate from the base model, resulting in a small parameter increase of $0.78M$ per entity, as reported in their paper. CEAL significantly outperforms these existing approaches without any additional parameters. This success can be attributed to our innovative multi-objective searching strategy, which strikes a remarkable balance between performance enhancement and parameter efficiency.

E. Ablation Studies

Table II presents an ablation study on the impact of the main components of CEAL in terms of ICA. We individually remove each main component from CEAL to assess its influence. Our full CEAL significantly outperforms the Naive LoRA Merging baseline (48.12), which highlights the effectiveness of CEAL's overall design.

We first evaluate CEAL's core architectural and training strategies. Replacing the entire *Superlora* structure with a standard fixed max-rank LoRA, upon which our search process is then applied (CEAL w/o Superlora), reduces performance to 58.95. This underscores Superlora's inherent advantage in facilitating the discovery of optimal layer-wise configurations. Similarly, omitting the Dynamic Rank-Increasing Strategy during Superlora training (CEAL w/o Dynamic Rank-Increasing) results in an ICA score of 59.12, confirming this strategy's

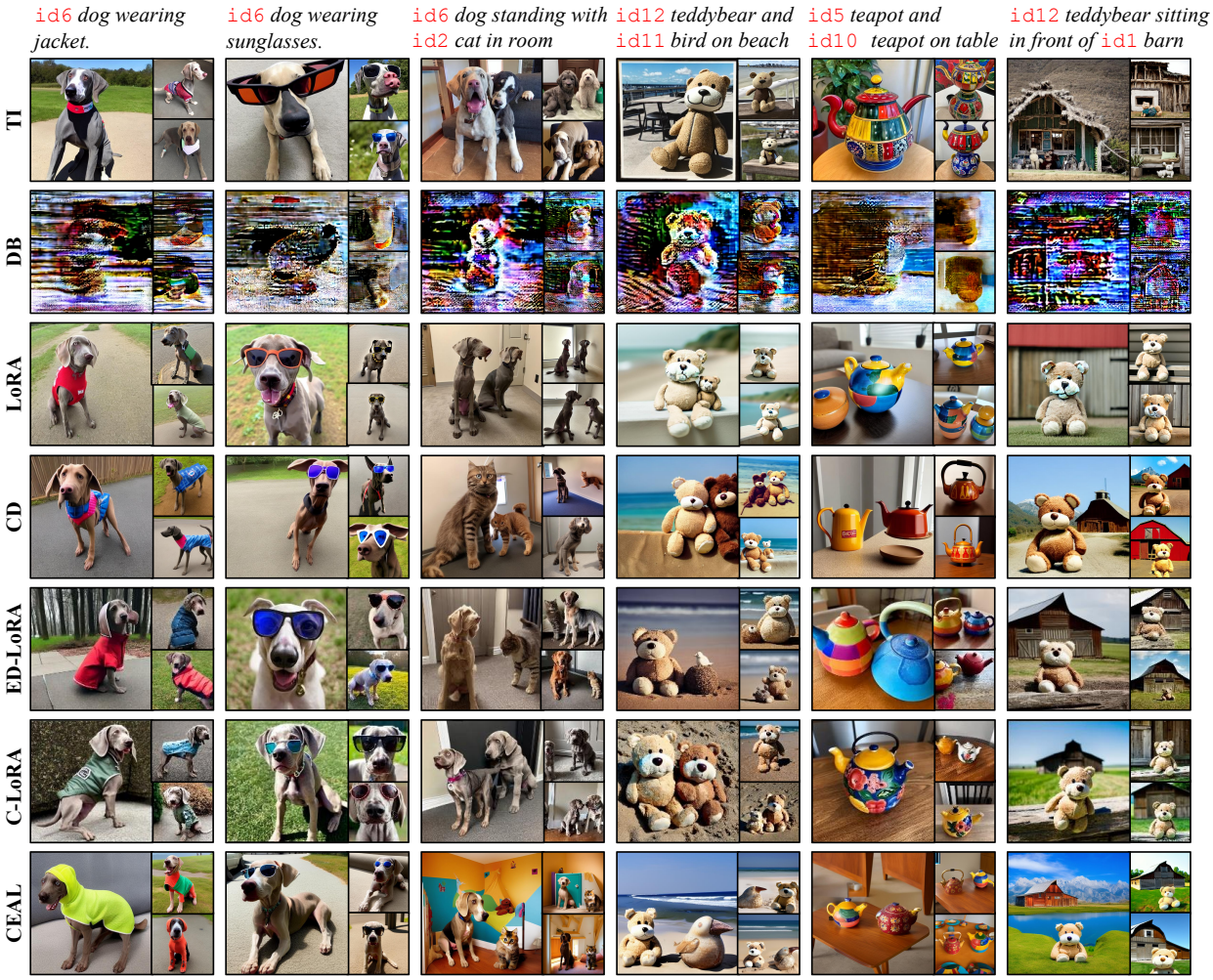

Fig. 5: **Visualization of multi-subject generation on CEL-Objects.** We show the results of multi-subject generation on CEL-Objects after all competing methods have completed 15 continual sessions. Entities are marked by their identifiers [id], with their target samples provided at the bottom and prompts at the top.

contribution to effective LoRA learning. Constraining the search to a Uniform Rank across all layers, as opposed to distinct layer-wise ranks (CEAL w/o Layer-wise Ranks), not only leads to suboptimal performance with an ICA of 60.12 but also potentially increases parameter count due to less efficient allocation. Finally, substituting the SA Optimizer with a random search mechanism (CEAL w/o SA Optimizer) causes a substantial performance drop to 53.89, accompanied by high variance. This result affirms the critical role of Simulated Annealing in effectively and consistently navigating the complex search space of LoRA configurations.

Next, we individually remove each of the three search criteria—Human Preference Metric, Memory Cost Metric, and Concept Similarity Metric—and perform the search each time. The Concept Similarity Metric proves to be the most crucial component in the CEAL method, as removing it results in a substantial accuracy drop of 12.30. This is mainly because when only two metrics remain, the search process tends to select the LoRA with the smallest number of parameters, which may not align with the objectives of the CEL.

Finally, investigating the merging process, we found that applying the optimized LoRAs independently without merging them into the base model (CEAL w/o Merging) achieves an ICA_{new} of $65.25_{\pm 0.76}$. This is slightly higher than the full CEAL with merging, suggesting a marginal trade-off between peak concept fidelity and the parameter efficiency gained by merging. However, the full CEAL's merging provides the significant benefit of zero additional parameters post-learning.

Then we study the impact of the **main** hyperparameters r_{final}, ν, α, β, and γ in Fig. 6. r_{final} constrains the maximum LoRA rank that can be searched for each entity. The results show that the performance is poor when $r_{final} = 1$, indicating that the rank of LoRA may be too limited to learn LoRA sufficiently. Interestingly, the performance is suboptimal when $r_{final} = 32$. This result supports our claim that a large search space is difficult to train, and it is essential to design mechanisms that ensure *Superlora* is adequately trained and to develop search algorithms with better evaluation metrics. ν determines the training steps required before adding a new rank to the *Superlora*. The results demonstrate that using 1000 steps as the increment is sufficient for achieving good performance, and the model's performance is not highly sensitive to the

TABLE I: **Results for all benchmarks.**

Datasets		CEL-Objects		CEL-Faces		CEL-Style	
Method	$\#\boldsymbol{P}$ (↓)	ICA (↑)	ICF (↓)	ICA (↑)	ICF (↓)	ICA (↑)	ICF (↓)
LoRA	$2.59M$	$62.38_{\pm 0.42}$	-	$57.32_{\pm 0.83}$	-	$41.55_{\pm 0.10}$	-
CD	$19.17M$	$58.86_{\pm 1.22}$	-	$55.91_{\pm 1.22}$	-	$41.47_{\pm 0.13}$	-
ED-LoRA	$0.81M$	$55.35_{\pm 0.83}$	-	$58.77_{\pm 0.44}$	-	$40.33_{\pm 0.20}$	-
C-LoRA	$0.78M$	$61.94_{\pm 0.30}$	$6.12_{\pm 0.49}$	$57.48_{\pm 0.29}$	$2.83_{\pm 0.23}$	$41.45_{\pm 0.48}$	$1.03_{\pm 0.09}$
TI	0.0	$48.61_{\pm 1.38}$	-	$59.33_{\pm 1.25}$	-	$38.30_{\pm 0.42}$	-
Replay	0.0	$27.27_{\pm 6.97}$	$25.36_{\pm 3.10}$	$14.22_{\pm 0.29}$	$8.84_{\pm 0.36}$	$20.88_{\pm 1.79}$	$5.98_{\pm 0.72}$
EWC	0.0	$49.15_{\pm 0.22}$	$12.39_{\pm 1.28}$	$56.63_{\pm 0.73}$	$4.84_{\pm 0.89}$	$37.11_{\pm 0.83}$	$1.83_{\pm 0.22}$
DB	0.0	$32.58_{\pm 5.08}$	$38.07_{\pm 2.72}$	$49.02_{\pm 0.30}$	$4.61_{\pm 0.62}$	$41.99_{\pm 0.40}$	$\mathbf{1.01_{\pm 0.29}}$
CEAL	0.0	$\mathbf{63.74_{\pm 0.31}}$	$\mathbf{9.12_{\pm 0.88}}$	$\mathbf{59.01_{\pm 0.73}}$	$\mathbf{2.71_{\pm 0.72}}$	$\mathbf{45.21_{\pm 0.53}}$	$1.04_{\pm 0.04}$

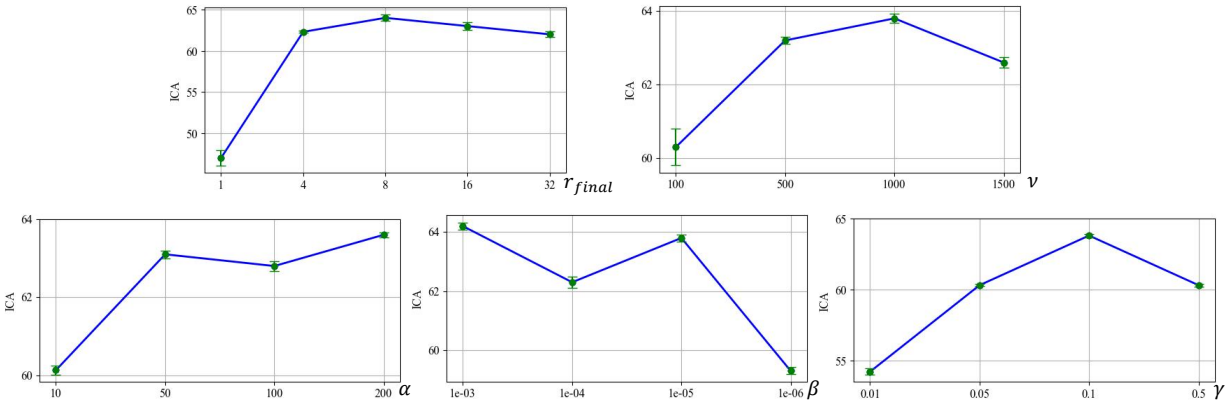

Fig. 6: Impact of r_{final}, ν, α, β, and γ on CEL-Objects.

TABLE II: Ablation study on the main components of CEAL on CEL-Objects.

Component Variant	ICA_{new}
Naive LoRA Merging	$48.12_{\pm 1.31}$
CEAL	$63.74_{\pm 0.31}$
CEAL w/o Superlora (Fixed Max-Rank LoRA + Search)	$58.95_{\pm 0.69}$
CEAL w/o Dynamic Rank-Increasing	$59.12_{\pm 0.69}$
CEAL w/o Layer-wise Ranks (Uniform Rank Search)	$60.12_{\pm 0.40}$
CEAL w/o SA Optimizer (e.g., Random Search)	$53.89_{\pm 10.34}$
CEAL w/o L_{hpm} (Human Preference Metric)	$60.48_{\pm 0.34}$
CEAL w/o L_{mcm} (Memory Cost Metric)	$62.89_{\pm 0.75}$
CEAL w/o L_{csm} (Concept Similarity Metric)	$51.44_{\pm 0.26}$
CEAL w/o Merging (Independent Optimized LoRAs)	$65.25_{\pm 0.76}$

choice of ν. Even with smaller or larger step sizes, the model still maintains relatively good accuracy. For hyperparameters α, β, and γ that balance the searching objective, we find that although there are fluctuations and impacts, the variations are not significant.

F. Computational Costs

Training and optimizing *Superlora* for each concept requires roughly 2 gpu hours on a single Nvidia RTX 4090 GPU.

The subsequent search phase for optimal rank configurations, however, is considerably more time-consuming. The dominant factor in this cost is the image generation required to evaluate each candidate solution presented by the SA optimizer. Other operations, such as extracting specific LoRA modules from the *Superlora* backbone or calculating aesthetic and CLIP scores, have a negligible computational overhead in comparison. For each prompt used in the evaluation, we generate 4 images, a process that takes approximately 4.5 seconds on an NVIDIA RTX 4090. To illustrate, consider the CEL-Objects benchmark, where each concept is typically evaluated using 20 distinct prompts. With the SA optimizer set to a maximum of $N_I = 1000$ iterations, evaluating one candidate solution (i.e., 20 prompts) takes $20 \times 4.5\text{s} = 90\text{s}$. This would lead to a maximum theoretical search duration of $90\text{s/iteration} \times 1000 \text{ iterations} = 90,000$ seconds, or approximately 25 GPU hours per concept.

However, in practice, the SA algorithm often converges to a satisfactory solution well before reaching the maximum iteration limit. Furthermore, if the same rank configuration is proposed multiple times by the SA optimizer, its previously computed fitness score can be directly reused, avoiding redundant image generation. Due to these factors, the average search time per concept is significantly lower, typically around 4 GPU hours on the same hardware.

VI. Conclusion

This paper introduces the concept of Continual Conceptual Entity Learning (CEL) for pre-trained T2I models. We introduce *Superlora*, a search space to organize potential LoRA configurations into a comprehensive set. To provide optimal customized LoRA rank configurations reasonably, we propose a novel dynamic rank-increasing strategy for training *Superlora*. We then use an efficient optimizer to continuously search for the optimal LoRA in line with multiple competing criteria. In addition, three comprehensive evaluation metrics and benchmarks for CEL are introduced. Extensive experiments on three benchmarks demonstrate the effectiveness of CEAL. However, it is worth noting that the proposed method may introduce additional computational overhead due to the search

process. On a positive note, our approach enhances parameter efficiency, potentially making AI models more accessible by reducing storage resource requirements.

Furthermore, evaluating CEAL on more diverse and challenging large-scale real-world datasets constitutes an important direction for future work to more broadly assess its generalization capabilities and practical applicability.

ACKNOWLEDGMENTS

This work was funded in part by the National Natural Science Foundation of China (62376070, 62076195) and in part by the Fundamental Research Funds for the Central Universities (AUGA5710011522).

REFERENCES

[1] 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," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NAQvF08TcyG

[8] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[9] 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.

[10] J. Pomponi, S. Scardapane, and A. Uncini, "Continual learning with invertible generative models," *Neural Networks*, vol. 164, pp. 606–616, 2023.

[11] B. Yang, X. Deng, H. Shi, C. Li, G. Zhang, H. Xu, S. Zhao, L. Lin, and X. Liang, "Continual object detection via prototypical task correlation guided gating mechanism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9255–9264.

[12] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong gan: Continual learning for conditional image generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2759–2768.

[13] Y. Zhang, X. Wang, and D. Yang, "Continual sequence generation with adaptive compositional modules," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3653–3667.

[14] F. Mi, L. Chen, M. Zhao, M. Huang, and B. Faltings, "Continual learning for natural language generation in task-oriented dialog systems," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3461–3474.

[15] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[16] Cloneofsimo, "LoRA: Using Low-rank Adaptation to Quickly Fine-tune Diffusion Models," https://github.com/cloneofsimo/lora, 2023, accessed: 2024-02-19.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[18] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "$p+$: Extended textual conditioning in text-to-image generation," *arXiv e-prints*, pp. arXiv–2303, 2023.

[19] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones: Concept neurons in diffusion models for customized generation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 21 548–21 566. [Online]. Available: https://proceedings.mlr.press/v202/liu23j.html

[20] A. Aghajanyan, S. Gupta, and L. Zettlemoyer, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7319–7328.

[21] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2018.

[22] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[23] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 214–14 223.

[24] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *International conference on machine learning*. PMLR, 2023, pp. 30 105–30 118.

[25] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *Transactions on Machine Learning Research*, 2022.

[26] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[27] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[28] C. Zhang, W. Yang, X. Li, and H. Han, "Mmginpainting: Multi-modality guided image inpainting based on diffusion models," *IEEE Transactions on Multimedia*, vol. 26, pp. 8811–8823, 2024.

[29] B. Yuan, Y. Sheng, B.-K. Bao, Y.-P. P. Chen, and C. Xu, "Semantic distance adversarial learning for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 1255–1266, 2024.

[30] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lv, C. Chen, and S. Chen, "Wavedm: Wavelet-based diffusion models for image restoration," *IEEE Transactions on Multimedia*, vol. 26, pp. 7058–7073, 2024.

[31] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "l_0-regularized intensity and gradient prior for deblurring text images and beyond," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 342–355, 2016.

[32] 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*. PMLR, 2021, pp. 8748–8763.

[33] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, and H. Zheng, "Inserting anybody in diffusion models via celeb basis," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 72 958–72 982. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/e6d37cc5723e810b793c834bcb6647cf-Paper-Conference.pdf

[34] H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang, "Freestyle layout-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 256–14 266.

[35] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[36] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8543–8552.

[37] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6527–6536.

[38] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," 2023.

[39] G. Kwon, S. Jenni, D. Li, J.-Y. Lee, J. C. Ye, and F. C. 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*, 2024, pp. 8880–8889.

[40] 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.

[41] Y. Yang, W. Wang, L. Peng, C. Song, Y. Chen, H. Li, X. Yang, Q. Lu, D. Cai, B. Wu *et al.*, "Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models," *arXiv preprint arXiv:2403.11627*, 2024.

[42] J. Jiang, Y. Zhang, K. Feng, X. Wu, W. Li, R. Pei, F. Li, and W. Zuo, "Mcˆ 2: Multi-concept guidance for customized multi-concept generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2802–2812.

[43] M. Valipour, M. Rezagholizadeh, I. Kobyzev, and A. Ghodsi, "Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 3266–3279.

[44] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *The Eleventh International Conference on Learning Representations*, 2022.

[45] F. Zhang, L. Li, J. Chen, Z. Jiang, B. Wang, and Y. Qian, "Increlora: Incremental parameter allocation method for parameter-efficient fine-tuning," *arXiv preprint arXiv:2308.12043*, 2023.

[46] Y. Ma, Z. Xie, J. Wang, K. Chen, and L. Shou, "Continual federated learning based on knowledge distillation," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, vol. 3, 2022.

[47] 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.

[48] 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.

[49] N. A. Cayco-Gajic and R. A. Silver, "Re-evaluating circuit mechanisms underlying pattern separation," *Neuron*, vol. 101, no. 4, pp. 584–602, 2019.

[50] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 831–839.

[51] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 374–382.

[52] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 208–13 217.

[53] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, and T. Moon, "Ss-il: Separated softmax for incremental learning," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 844–853.

[54] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: Pooled outputs distillation for small-tasks incremental learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 86–102.

[55] H. Cha, J. Lee, and J. Shin, "Co2l: Contrastive continual learning," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 9516–9525.

[56] F.-Y. Wang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Foster: Feature boosting and compression for class-incremental learning," in *European conference on computer vision*. Springer, 2022, pp. 398–414.

[57] C. Wu, L. Herranz, X. Liu, J. Van De Weijer, B. Raducanu *et al.*, "Memory replay gans: Learning to generate new categories without forgetting," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[58] Y. Cong, M. Zhao, J. Li, S. Wang, and L. Carin, "Gan memory with no forgetting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 481–16 494, 2020.

[59] X. Liu, C. Wu, M. Menta, L. Herranz, B. Raducanu, A. D. Bagdanov, S. Jui, and J. v. de Weijer, "Generative feature replay for class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 226–227.

[60] G. Petit, A. Popescu, H. Schindler, D. Picard, and B. Delezoide, "Fetril: Feature translation for exemplar-free class-incremental learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3911–3920.

[61] 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," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

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

[63] L. Wang, M. Zhang, Z. Jia, Q. Li, C. Bao, K. Ma, J. Zhu, and Y. Zhong, "Afec: Active forgetting of negative transfer in continual learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 379–22 391, 2021.

[64] M. M. Derakhshani, X. Zhen, L. Shao, and C. Snoek, "Kernel continual learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2621–2631.

[65] H. Ahn, S. Cha, D. Lee, and T. Moon, "Uncertainty-based continual learning with adaptive regularization," *Advances in neural information processing systems*, vol. 32, 2019.

[66] S. Jung, H. Ahn, S. Cha, and T. Moon, "Continual learning with node-importance based adaptive group sparse regularization," *Advances in neural information processing systems*, vol. 33, pp. 3647–3658, 2020.

[67] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[68] 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.

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

[70] 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.

[71] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach, "Adversarial continual learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 386–402.

[72] A. Douillard, A. Ramé, G. Couairon, and M. Cord, "Dytox: Transformers for continual learning with dynamic token expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9285–9295.

[73] Y. Wang, Z. Ma, Z. Huang, Y. Wang, Z. Su, and X. Hong, "Isolation and impartial aggregation: A paradigm of incremental learning without interference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 10 209–10 217.

[74] S. Yan, J. Xie, and X. He, "Der: Dynamically expandable representation for class incremental learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 3013–3022. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00303

[75] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *International Conference on Learning Representations*, 2018.

[76] H. Cai, L. Zhu, and S. Han, "Proxylessnas: Direct neural architecture search on target task and hardware," in *International Conference on Learning Representations*, 2018.
[77] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.
[78] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy *et al.*, "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 631–648.
[79] Y. Wang, Z. Huang, and X. Hong, "S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5682–5695, 2022.
[80] J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira, "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 909–11 919.
[81] J. S. Smith, Y.-C. Hsu, Z. Kira, Y. Shen, and H. Jin, "Continual diffusion with stamina: Stack-and-mask incremental adapters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1744–1754.
[82] 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," *Transactions on Machine Learning Research*.
[83] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[84] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.
[85] Y. Frenkel, Y. Vinker, A. Shamir, and D. Cohen-Or, "Implicit style-content separation using b-lora," in *European Conference on Computer Vision*. Springer, 2024, pp. 181–198.
[86] V. Shah, N. Ruiz, F. Cole, E. Lu, S. Lazebnik, Y. Li, and V. Jampani, "Ziplora: Any subject in any style by effectively merging loras," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–438.
[87] shunk031, "simple-aesthetics-predictor: Clip-based aesthetics predictor inspired by the interface of huggingface transformers," 2023. [Online]. Available: https://github.com/shunk031/simple-aesthetics-predictor
[88] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in neural information processing systems*, vol. 36, pp. 36 652–36 663, 2023.
[89] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.
[90] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," 05 2015.
[91] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

VII. Biography Section

Yabin Wang received the Ph.D. degree from Xi'an Jiaotong University, P. R. China, in 2025. He is an Associate Researcher with Harbin Institute of Technology (HIT). His research interests include continual learning, multi-modal learning and deepfake detection.

Xiaopeng Hong (Senior Member, IEEE) received his Ph.D. degree from Harbin Institute of Technology (HIT), P. R. China, in 2010. He is currently a professor at HIT. He has authored over 80 publications in leading journals and conferences, including IEEE T-PAMI, TIP, PIEEE, CVPR, ICCV, NeurIPS, AAAI, and ACM MM. His research on subtle facial movement analysis has been featured by international media such as MIT Technology Review. He is a co-author of the Top Paper Award at ACM Multimedia 2022 and the IEEE Finland Section Best Student Conference Paper in 2020. His current research interests include multi-modal learning, continual learning, and model interpretability.

Zhiheng Ma (Member, IEEE) received the PhD degree from Xi'an Jiaotong University, in 2021. He is a research assistant professor with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (SIAT). He has authored articles in journals and conferences, such as IEEE Transactions on Image Processing, CVPR, ICCV, and AAAI. His current research interests include incremental learning, crowd counting, and novelty detection.

Zhou Su (Senior Member, IEEE) has authored or coauthored technical papers, including top journals and top conferences, such as the IEEE Journal on Selected Areas in Communications, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Mobile Computing, IEEE/ACM Transactions on Networking, and INFOCOM. His research interests include multimedia communication, wireless communication, and network traffic. He is an associate editor for the IEEE Internet of Things Journal,IEEE Open Journal of the Computer Society, and IET Communications.

Jinpeng Zhang received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He works for the Intelligent Science & Technology Academy Limited of CASIC. His research interests include deep learning and multi-modal learning.

Zhiwu Huang received his Ph.D. in Computer Science and Technology from the University of Chinese Academy of Sciences in 2015. He then worked as a Postdoctoral and Guest Researcher in the Computer Vision Lab (CVL) at ETH Zurich from September 2015 to July 2021. From September 2021 to December 2022, he served as an Assistant Professor of Computer Science at Singapore Management University. He is currently a Lecturer (Assistant Professor) in the Vision, Learning, and Control (VLC) research group at the University of Southampton. His research interests include Computer Vision, Machine Learning, Generative Artificial Intelligence, and Geometric Deep Learning.