T2I-CONBENCH: TEXT-TO-IMAGE BENCHMARK FOR CONTINUAL POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual post-training adapts a single text-to-image diffusion model to learn new tasks without incurring the cost of separate models, but naïve post-training causes forgetting of pretrained knowledge and undermines zero-shot compositionality. We observe that the absence of a standardized evaluation protocol hampers related research for continual post-training. To address this, we introduce **T2I-ConBench**, a unified benchmark for continual post-training of text-to-image models. T2I-ConBench focuses on two practical scenarios, *item customization* and *domain enhancement*, and analyzes four dimensions: (1) retention of generality, (2) target-task performance, (3) catastrophic forgetting, and (4) cross-task generalization. It combines automated metrics, human-preference modeling, and vision-language QA for comprehensive assessment. We benchmark ten representative methods across three realistic task sequences and find that no approach excels on all fronts. Even joint "oracle" training does not succeed for every task, and cross-task generalization remains unsolved.

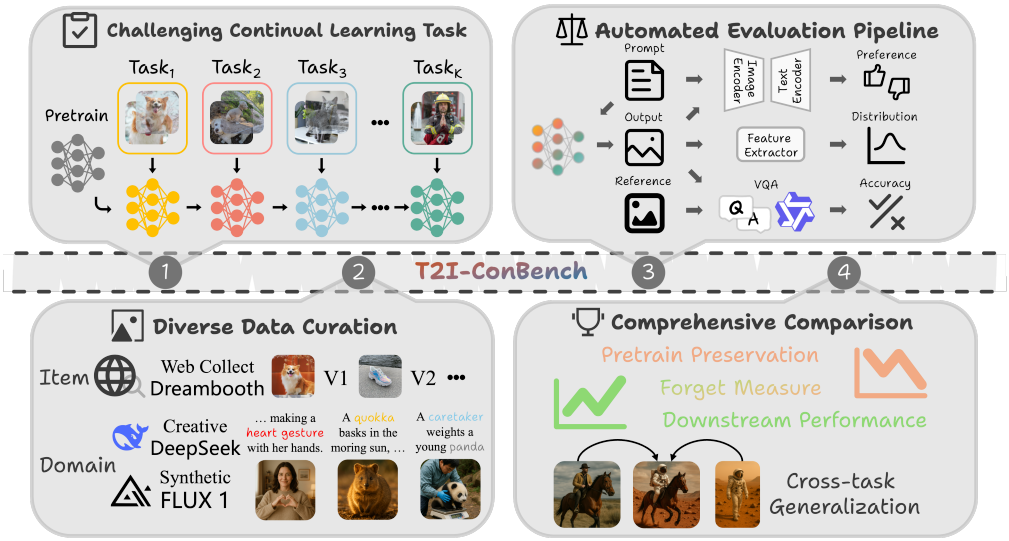

Figure 1: Overview of T2I-ConBench. Our benchmark consists of four components: (1) challenging continual post-training task sequences, (2) the curation of diverse item and domain datasets, (3) an automated evaluation pipeline, and (4) comprehensive metrics to fully assess each continual learning method's ability to update knowledge, resist forgetting, and generalize across tasks.

1 INTRODUCTION

Over the past few years, large-scale text-to-image (T2I) diffusion models (Saharia et al., 2022; Podell et al., 2024; Chen et al., 2024b) pretrained on massive image-text corpora have achieved remarkably realistic, high-resolution synthesis. However, real-world deployments (Ruiz et al., 2023; Chaichuk et al., 2025) continually require new concepts, styles, or tasks, ranging from personalized rendering of a specific object to domain-specific enhancements in medical imaging, industrial design, or cultural heritage. Training and maintaining a dedicated model for each downstream task is impractical due to prohibitive storage overhead and loss of knowledge sharing across tasks (Pilault et al., 2021; He et al., 2022). An ideal solution is to sequentially adapt a single foundation model to each new task

dataset, integrating fresh task-specific knowledge while preserving its original pretrained capabilities, commonly referred to as the continual post-training paradigm (Lu et al., 2025; Smith et al., 2024; Ke et al., 2022).

The key challenge is that, when naively post-trained on new tasks, T2I suffer *catastrophic forgetting* (French, 1993; Ratcliff, 1990): their ability to generate pretraining concepts degrades as they learn new ones. Recent work (Wang et al., 2024) has therefore adapted various continual post-training strategies to mitigate this issue, including rehearsal-based methods (Chaudhry et al., 2019), regularization-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017), and parameter-isolation methods (Hui et al., 2024; Chen et al., 2025b; Smith et al., 2024). They have shown impressive gains in specified scenarios with minimal degradation in general capability. Yet all existing methods evaluate knowledge updates within a single-granularity, sequential-task framework and overlook two critical aspects: (1) the dynamic degradation of pretrained capabilities throughout continual adaptation (Hui et al., 2024; Shi et al., 2024; Wang et al., 2023b), and (2) cross-task generalization (Okawa et al., 2023; Yin & Wang, 2025) to combine concepts across tasks. A model subjected to continual downstream learning should not only excel on each new task in isolation, but also preserve its capacity to generalize across both new and previously learned concepts. However, there is no unified benchmark to evaluate these trade-offs in continual post-training approaches.

We bridge this gap with **T2I-ConBench** (**Fig.** 1), a comprehensive benchmark for the continual post-training of text-to-image diffusion models. T2I-ConBench covers two prototypical post-training tasks of differing granularity: ❶ item customization (Zhang et al., 2025; Ruiz et al., 2023), using web-scraped real-world images to probe personalized object-level generation, and ❷ domain enhancement (Zhu et al., 2024b), using synthetic data to test improvement on generative quality and text-image alignment. For each sequence, we craft targeted prompts that challenge both general and specialized generation capabilities. We also develop an automated evaluation pipeline combining standard T2I metrics, a learned human-preference model, and visual question answering to assess ❶ preservation of pretrained generality, ❷ target-task performance, ❸ forgetting, and ❹ cross-task generalization. By unifying these dimensions within one extensible framework, T2I-ConBench enables fair comparison of continual post-training methods, illuminating their relative strengths in updating, retaining, and compositing knowledge.

Building upon T2I-ConBench, we construct three realistic continual post-training scenarios that order tasks of differing granularity, and we evaluate ten representative baseline methods on these mixed-order streams. Our experiments yield three key takeaways: ❶ *No single method excels everywhere.* ❷ *"Oracle" joint learning is not a panacea.* ❸ *Cross-task generalization remains an open challenge.*

We release all T2I-ConBench datasets, training scripts, and evaluation pipelines, providing the community with a unified, extensible platform to develop and benchmark continual post-training strategies for the next generation of T2I diffusion models.

2 Task Definition

Continual post-training (Smith et al., 2024) of large pretrained T2I diffusion models denotes the sequential adaptation of a single foundation model to a stream of small, task-specific datasets. After each adaptation task, the model must assimilate the novel concepts or domains without access to earlier data and without eroding its original generative competence. Concretely, we begin with a base model that has completed broad pretraining. We then define a sequence of downstream tasks, each associated with its own disjoint set of text–image pairs. A continual post-training algorithm produces a new model after each task so that it both adapts to the current task's data and resists degradation on all previously seen tasks. Achieving this balance requires effective mitigation of catastrophic forgetting while still integrating new knowledge. For a more formal definition of tasks, please refer to the Appendix F.

Cross-task generalization (Okawa et al., 2023; Yin & Wang, 2025) evaluates the ability to recombine knowledge acquired from different tasks into novel concepts. In addition to per-task performance metrics, our benchmark introduces a compositional generation evaluation to quantify this capability throughout continual post-training. This ability builds on the key observation that pretrained diffusion models often exhibit zero-shot generalization (Dhariwal & Nichol, 2021), e.g., after learning both "a person riding a horse" and "astronaut" in the pretraining stage, they can generate "an astronaut

riding a horse," which they have never seen during training (**Fig.** 1). We ask: if a model is first continually post-trained on the "person riding a horse" task and then on the "astronaut" task, does it still retain the ability to produce the novel combination "an astronaut riding a horse"? To answer this, we construct prompts that merge conditions from two different tasks (**Sec.** 3) and then evaluate how reliably the post-trained model generates images matching these unseen, composed prompts (**Sec.** 4). By measuring alignment of compositional generations to corresponding prompts, we can determine whether continual post-training preserves the pretrained model's generalization to blend concepts. A strong alignment indicates that the continually post-trained model not only learns each task's concepts but also preserves the representational flexibility to recombine them in novel ways, supporting long-term accumulation of knowledge.

Remark Unlike traditional T2I benchmarks (Huang et al., 2023) that compare different models, our T2I-ConBench holds both base models and task datasets fixed. We focus on the impact of the continual post-training algorithm itself, without conflating results with variations in data quality or model architecture. Such a design allows us to isolate and precisely measure the impact of continual post-training methods on knowledge retention, downstream performance, and cross-task generalization.

3 Data Curation

In real-world applications, T2I models often struggle with generating specific items and producing high-quality, domain-specific outputs. Prioritizing only one aspect would leave significant gaps in overall performance. The diverse demands of post-training for T2I models highlight the need for a systematic evaluation framework that accommodates varying data requirements. These data needs can be divided into two main categories:

- **Item Customization** focuses on data designed for the personalized generation of specific objects.
- **Domain Enhancement** involves data to improve image quality and semantic consistency within a specific domain (e.g., portrait photography, wildlife images, or natural landscapes).

Item Customization and Domain Enhancement differ in granularity and learning objectives, demanding distinct strategies for knowledge updating and retention. These differences imply that the effectiveness of continual post-training methods will depend on task types. These two scenarios form a comprehensive framework for tackling the practical challenges of post-training in T2I models.

For **Item Customization** tasks, we curate a training dataset comprising four distinct items selected from the dataset provided in(Ruiz et al., 2023). These items are: "V1 dog", "V2 dog", "V3 cat", and "V4 sneaker"[1]. The images for these subjects typically capture them under various conditions, environments, and angles to ensure diversity. We then use a large language model (LLM) to generate 10 scenarios for each customized item paired with its non-personalized class, forming the *test set for each item*.

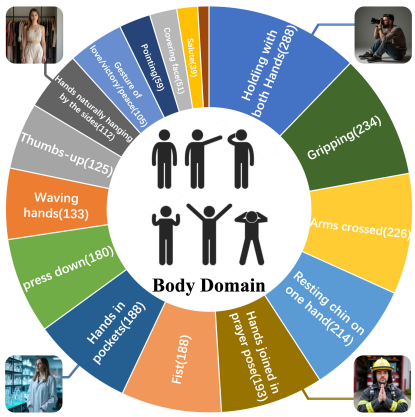

Figure 2: *Body pose* distribution.

For **Domain Enhancement** tasks, we specifically focus on two domains: natural world concepts and human portraits, which we refer to as "**Nature**" and "**Body**" domains, respectively. Domain enhancement is *not* intended to cover the entire broad categories. Instead, to improve the generation quality and semantic alignment, we first generate numerous prompts containing various concepts within each domain. We then use the base model to identify concepts concepts that exhibit high semantic or structural failure rates, such as fine-grained natural species or complex compositions. For the "**Nature**" domain, concepts requiring enhancement include: Squid, Quokka, Markhor, Gerenuk, Spix's Macaw, and Pomelo. For the "**Body**" domain, we primarily focus on improving the generation of body poses. Concepts requiring enhancement include: pointing, hands naturally hanging by the sides, arms crossed, etc. The total concepts are listed in **Fig.** 2, along with the number of training data samples for each concept.

[1] https://github.com/google/dreambooth

To acquire high-quality post-training data for these concepts, we opt for synthetic data generation. Generating synthetic data is an efficient and convenient method for obtaining large, controlled datasets. We first use LLMs to create prompts incorporating the identified concepts. These generated prompts are then sampled; most are designated for the training set, while the remainder form the *test sets for each domain*. Moreover, to enhance the model's understanding of interactive relationships between concepts across two distinct domains, we construct a training dataset for human interactions with common animals. The training prompts include one common animal concept, for which the base model shows high generation quality, and a concept from the Body domain set. We then use the Flux_dev model (Labs, 2024) to generate images for each training-set prompt. The generated data undergo meticulous manual screening to ensure that they are plausible, aesthetically pleasing, and semantically faithful to the prompts. All generated images do not involve any private data and fully comply with established safety and usage standards (Beduschi, 2024). The initial dataset size is about $80k$. After thorough manual filtering, the final dataset sizes are 2513 for the Nature domain, 2356 for body poses, and 1821 for interactions with common animals. The latter two constitute the Body domain training dataset. Comprehensive details on the dataset, its necessity, and the impact of scale are provided in Appendices G, M, and N, respectively.

Cross-Task Generalization Test Sets Considering that knowledge across distinct domains is often considered independent, we also aim to investigate the T2I model's generalization capabilities across different domains after continual training. Specifically, we explore the model's ability to synthesize concepts from different domains within a single image. Good generalization capabilities indicate that the model not only learns each task's concepts but also preserves the representational flexibility to recombine them in novel ways. We construct specialized test sets to probe this cross-dataset generalization:

- **Item+Item:** This set evaluates the model's ability to combine two different trained items in a single image, often within varying environmental contexts. We generate prompts combining pairs of the four trained items within 20 different environmental scenes.
- **Item+Domain:** These sets evaluate the model's ability to combine a trained item with concepts from either the Nature or Body domains. For the Item-Nature test set, prompts combine each of the five items with various Nature concepts. We generate 3 prompts per item for natural combinations. For the Item-Body test set, prompts combine each of the five items with specific body poses. We generate one prompt for each item-pose pair for a base set of poses, and an additional prompt per item for 11 high-frequency human pose concepts.
- **Domain+Domain:** To assess the model's ability to combine learned concepts from different domains, we create prompts that combine concepts from the Nature domain training set with concepts from the Body domain training set. This set evaluates if the knowledge learned within distinct domains can be effectively composed when prompted together. For each concept in the Nature domain, its corresponding test set comprises 20 captions, each depicting an interaction between a human and the concept.

4 Evaluation Pipeline

To comprehensively evaluate continual post-training methods, we adopt a multi-axis assessment framework for fair comparison and scalable benchmarking, spanning generation quality, semantic alignment, task-specific accuracy, backward transfer, and compositional generalization.

Pretrain Preservation To assess how well continual post-training preserves pretrained capabilities, we measure ❶ *generation quality* using Fréchet Inception Distance (**FID**) (Heusel et al., 2017) to quantify image-generation quality, where lower values indicates closer alignment to real images, computed against MS-COCO dataset (Lin et al., 2014) as our real-image reference; and ❷ *text-image alignment* using T2I-CompBench (Huang et al., 2023), which uses a visual language model (Zhu et al., 2024a) (VLM) to evaluate the T2I semantic accuracy under compositional prompts. Considering the full T2I-CompBench involves generating and scoring large, multidimensional datasets, making it costly to run after each task, we select its most representative compositional tasks as a proxy, complex generation (**Comp**). This subset serves as our metric for post-training text-image alignment.

Downstream Performance We define separate evaluation metrics for two downstream tasks with different granularity. ❶ *Item Customization*, we measure the model's accuracy at generating personalized objects. For each fine-grained concept, we prompt the post-training model to generate a test set

of images, and we use the original concept's training set of images as references. Employing a designed question prompt template, we then apply a VLM-based visual question answering (VQA) (Ma et al., 2023) pipeline to score the similarity between generated and reference images on the unique personalized concept, denoted as **Unique-Sim**. ❷ *Domain Enhancement*, we assess human aesthetic preference using the Human Preference Score (**HPS**) (Wu et al., 2023a), providing a fine-grained assessment of the aesthetic and semantic fidelity of task domain outputs from T2I models.

Forget Measure Beyond measuring degradation of pretrained capabilities relative to the base model, we also quantify forgetting in downstream performance dynamics during continual post-training. For both Item Customization and Domain Enhancement, we compute *backward transfer* (Wang et al., 2024) on their respective downstream metrics, denoted **Unique-Forget** and **Domain-Forget**. Additionally, we assess forgetting of the base class when learning personalized concepts in Item Customization. We generate images for non-personalized prompts (e.g., "a dog ...") and score their similarity to all personalized examples (e.g., "V1 dog ...") via our VQA pipeline, as **Class-Sim**. A lower Class-Sim indicates less forgetting of the broader class in favor of the specific concept.

Cross-task Generalization We generate prompts that merge concepts from different tasks and assess whether the fine-tuned model can accurately render these novel combinations. We also score cross-task performance using a VQA pipeline (**Fig.** 3). First, an LLM decomposes each compositional test-prompt into its simpler, single-object components and generates corresponding question-answer pairs that fully cover both individual object generation and their cross-task interactions. Next, we convert those Q&A pairs into VQA-style questions so that we can directly evaluate image–text alignment by comparing the VLM's answers against the ground-truth. For customized item objects or specialized fauna in the nature domain that the VLM may have never seen, we supply reference images of target objects alongside generated images when querying the VLM. Correct responses indicate successful cross-task composition. We evaluate each post-trained model on its respective cross-task test set and report the accuracy as our cross-task generalization metric, reflecting each method's effects of representational flexibility and long-term knowledge accumulation.

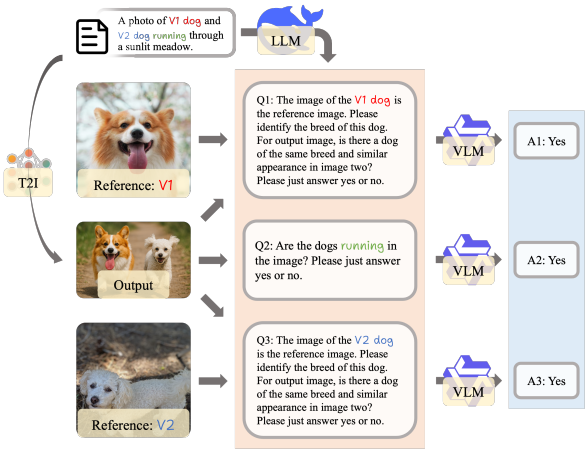

Figure 3: Evaluation pipeline of cross-task generalization.

Remark Our evaluation pipeline is fully automated, eliminating human intervention and reducing the labor cost of large-scale, multi-round assessments. The interfaces we define are model-agnostic, allowing seamless integration of advanced evaluators. Comprehensive metric definitions, formulas, and further comparative analyses are provided in Appendices H and L, respectively.

5 Continual Post-training Baselines

We refer to the pretrained model as **Base** for establishing a baseline on general generative capabilities and downstream tasks. We treat the model obtained by jointly training on all task data as the "oracle method" (Wu et al., 2023b), thereby characterizing the upper bound of performance in sequential learning, as **Joint**. For continual post-training of T2I diffusion models, we apply and adapt 10 baseline methods to mitigate catastrophic forgetting and enhance new concept learning. First, the simplest sequential fine-tuning (**SeqFT**) (Zhang et al., 2024a) updates all model parameters in task order, optimizing exclusively for the current task without preserving pretrained knowledge or retaining performance on earlier tasks. In addition, we compare the following representative baselines:

Rehearsal-based methods maintain a memory buffer that stores samples to replay prior knowledge. We store 10% of each completed task's image–text pairs in the memory buffer and mix them with new-task data during subsequent post-training. This simple **Replay** baseline (Chaudhry et al., 2019) effectively mitigates forgetting and provides a reference for more advanced rehearsal and buffer-management strategies.

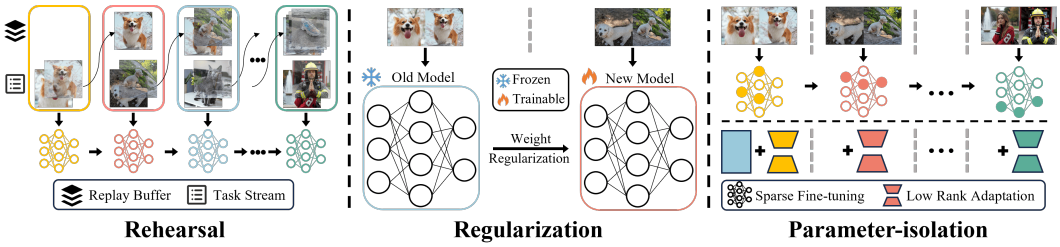

Figure 4: Overview of the continual post-training baselines evaluated in this work, encompassing rehearsal-based, regularization-based, and parameter-isolation methods (sparse fine-tuning and low-rank adaptation). These baselines are described in **Sec.** 5 and Appendix I.

Regularization-based methods add a constraint term to the training objective to balance between learning new tasks and retaining previous knowledge. We evaluate two regularization baselines:

- ℓ_2**-norm** (Zhao et al., 2024) adds an ℓ_2-norm penalty on the change from the previous task's final parameters, discouraging significant parameter updates and thus preserving earlier knowledge.
- **EWC** (Kirkpatrick et al., 2017) weights each parameter's penalty according to its estimated importance to previous tasks by Fisher information matrix (Liao et al., 2018). Parameters with higher Fisher scores incur a larger penalty for deviation, thereby more effectively preserving those weights critical to earlier tasks.

Parameter-isolation methods freeze most model parameters and update only a small subset, dramatically reducing the computation and storage costs of full-model post-training. In continual post-training for large T2I diffusion models, they split into two main categories:

❶ **Sparse fine-tuning** updates only a small, sparse subset of parameters, with all others fixed at their initial values. This reduces interference with features learned on previous tasks and mitigates forgetting. We adopt two recently proposed sparse fine-tuning baselines:

- **HFT** (Hui et al., 2024) randomly partitions parameters into two equal groups at each new task. One group (50%) is trained and the other remains frozen, thereby balancing new concept learning with preservation of prior knowledge.
- **MoFO** (Chen et al., 2025b) ranks parameters by the absolute value of their Adam momentum after each backward pass, then updates only the top subset for critical directions while freezing the rest. This momentum-driven sparse update efficiently learns new tasks and stabilizes prior performance.

❷ **Low-rank adaptation** (LoRA) assumes that the fine-tuning weight update lies in a low-dimensional subspace. Rather than updating the full weight matrix directly, LoRA factorizes weight changes into the product of two low-rank matrices, while freezing the original weights. This dramatically reduces both storage and computation costs. In the continual post-training setting, the low-rank decomposition can be extended into several variants that balance adaptation to new tasks with isolation of prior knowledge. In our experiments, we compare the following four LoRA-based baselines:

- **SeqLoRA** (Devlin et al., 2019) shares a single LoRA adapter across all tasks, updating it cumulatively each round. This approach is simple and efficient, but may suffer from accumulated interference between tasks.
- **IncLoRA** (Wang et al., 2023a) allocates a fresh, independent LoRA adapter for each new task, and sums up all adapters for final inference. By assigning each task its own low-rank subspace, it enforces strict task isolation at the cost of linearly increasing the number of parameters.
- **O-LoRA** (Wang et al., 2023a) enforces an orthogonality constraint on the up-projection matrix, making the low-rank subspaces of different tasks mutually orthogonal.
- **C-LoRA** (Smith et al., 2024) adds a self-regularization term that penalizes deviations between the LoRA update for the new task and the adapters learned for previous tasks.

Remark For detailed baseline descriptions, see Appendix I. We acknowledge that there are more advanced continual learning techniques (Ren & Sutherland, 2025) for classification or specialized continual learning methods designed for T2I diffusion models (Sun et al., 2024). However, due to the cost of their adaptation and unpredictability to our setup, we do not include them as baselines. Instead, we adopt representative and straightforward methods that capture the core properties of each category. Future work will incorporate additional approaches to provide further insights.

Table 1: Performance of continual post-training methods on the sequential item customization ("V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker") and sequential domain enhancement ("Nature" → "Body") task using PixArt-α. ↑: higher is better. ↓: lower is better. "I" and "D" denote Item and Domain, with combinations indicating cross-task generalization evaluations. Excluding *Base* and *Joint*, the best result is in **bold**, the second-best is underlined. For all metrics except Forget, red cells indicate a drop of more than 5% below *Base* for significant degradation, while green cells indicate an increase of more than 5% above *Joint* for significant outperformance of the traditional "oracle".

Order	"V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker"						"Nature" → "Body"					
Method	**Pretrain**		**Item**	**Cross**	**Forget**		**Pretrain**		**Domain**		**Cross**	**Forget**
	FID ↓	**Comp ↑**	**Unique-Sim ↑**	**I+I ↑**	**Class-Sim ↓**	**Unique-Forget ↓**	**FID ↓**	**Comp ↑**	**Body-HPS ↑**	**Nature-HPS ↑**	**D+D ↑**	**Domain-Forget ↓**
Base	26.3153	0.3378	0.0075	0.2250	0.0088	–	26.3153	0.3378	0.2966	0.2732	0.2637	–
Joint	22.9396	0.3308	0.2225	0.3694	0.0695	–	29.0167	0.3325	0.3032	0.2849	0.4577	–
SeqFT	19.7847	0.3319	0.2325	0.3222	0.0633	0.8718	29.9746	0.3382	0.2939	0.2744	0.3881	0.0392
SeqLoRA	21.9909	**0.3493**	0.0525	0.3500	**0.0263**	0.6611	28.4885	0.3433	0.2997	0.2854	0.4080	0.0083
IncLoRA	21.9657	0.3392	0.1850	0.3278	0.0863	N/A	28.2885	**0.3519**	0.3006	0.2874	0.4080	0.0007
O-LoRA	22.6171	0.3364	0.1775	0.2861	0.0968	N/A	26.5287	0.3411	0.2942	0.2880	0.4030	-0.0031
C-LoRA	23.2204	0.3411	0.1850	0.3056	0.0838	N/A	**26.1921**	0.3414	0.2920	0.2882	0.3930	-0.0031
ℓ_2-norm	20.6191	0.3417	0.1575	0.3278	0.0468	0.7962	27.1267	0.3426	0.2990	0.2863	0.3980	0.0003
EWC	19.8390	0.3399	0.2250	0.3139	0.0575	0.7017	29.7816	0.3409	0.2947	0.2746	0.3781	0.0372
HFT	20.8671	0.3357	0.1500	0.3028	0.0333	0.5833	28.8833	0.3438	**0.3010**	0.2840	0.3881	0.0104
MoFO	**19.2802**	0.3296	**0.2850**	0.3306	0.0680	0.7296	29.8326	0.3418	0.2985	0.2803	**0.4279**	0.0196
Replay	20.7805	0.3338	0.2700	**0.3694**	0.0768	**0.1428**	29.7044	0.3508	0.3007	**0.2890**	0.4179	**-0.0070**

6 Experiments

6.1 Implementation Details

Based on T2I-ConBench, we design three continual post-training scenarios for T2I diffusion models with different data granularities: (1) *Sequential item customization* with four fine-grained concepts learned in order. (2) *Sequential domain enhancement* with two broad domains trained sequentially. (3) *Sequential Item-Domain Adaptation* with a mixture of the above item and domain tasks, evaluated under two task orders. We evaluate the ten continual post-training baselines introduced in **Sec.** 5 on two diffusion architectures, PixArt-α (Chen et al., 2024b) and Stable Diffusion v1.4 (Rombach et al., 2022) (Appendix K). Detailed training protocol and hyperparameters are provided in the Appendix J.

6.2 Continue Post-training for Sequential Item Customization

The left part of **Tab.** 1 shows PixArt-α's results on the Sequential Item Customization tasks. All post-training methods achieve a substantial FID reduction versus the base model, demonstrating that targeted post-training on a small set of high-quality samples can dramatically boost image fidelity, often called *quality tuning* (Dai et al., 2023). In CompBench's text-image alignment evaluation, all methods perform roughly on par with the base model. LoRA variants struggle after learning the first task. They typically fail to acquire subsequent concepts, yielding "N/A" for forgetting metrics. This likely reflects LoRA's constrained update subspace, which cannot span widely differing concepts. Interestingly, SeqLoRA recovers item generation capability when testing on multi-item prompts, yielding an Item+Item generalization score of 0.35. This suggests that SeqLoRA has indeed internalized distinct item concepts, but they only manifest when triggered by specific prompts. Among rehearsal-free approaches, MoFO performs best, achieving a unique-item similarity of 28.5% and lower forgetting than SeqFT. Replay attains 27% unique-item similarity and a markedly lower Unique-Sim forgetting (14.28%), outperforming all rehearsal-free methods and matching Joint in cross-task generalization, benefiting from the scenario's small dataset sizes. However, despite its efficiency and strong performance, replay may pose privacy risks.

6.3 Continue Post-training for Sequential Domain Enhancement

PixArt-α's performance on the Sequential Domain Enhancement tasks is shown in the right part of **Tab.** 1. Unlike in item customization, the results of most methods get increased FID and indicate a degradation in overall image quality. The underlying reason is that fine-tuning directly on the new domain erodes the model's coverage of the general image distribution. Nonetheless, all methods achieve modest gains on CompBench, indicating improved text-image alignment with the target

Table 2: Performance of continual post-training methods for the sequential item-domain adaptation task of two orders using PixArt-α. ↑: higher is better. ↓: lower is better. "I" and "D" denote Item and Domain, respectively, with combinations indicating cross-task generalization evaluations. Excluding *Base* and *Joint*, the best result is shown in bold and the second-best is underlined. For all metrics except Forget, red cells indicate a drop of more than 5% below *Base* for significant degradation. Since the traditional "oracle" *Joint* performs poorly in this mixed adaptation scenario, it is not used as the target to surpass.

Order 1	Method	Pretrain		Item	Domain		Cross			Forget
		FID ↓	Comp ↑	Unique-Sim ↑	Body-HPS ↑	Nature-HPS ↑	I+I ↑	I+D ↑	D+D ↑	Class-Sim ↓
"V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker" → "Nature" → "Body"	*Base*	26.3154	0.3378	0.0075	0.2966	0.2732	0.2250	0.3407	0.2637	0.0088
	Joint	29.2236	0.3472	0.0725	0.3054	0.2897	0.2528	0.3898	0.4527	0.0413
	SeqFT	28.9167	0.3483	0.0225	0.3014	0.2832	0.2667	0.3796	0.3980	0.0118
	SeqLoRA	28.7234	0.3456	0.0000	0.3004	**0.2890**	0.2333	0.3571	0.4129	N/A
	IncLoRA	28.5758	0.3389	0.0000	0.2965	0.2841	0.2361	0.3919	0.3980	N/A
	O-LoRA	27.8870	0.3388	0.0600	0.2838	0.2838	0.2806	0.3530	0.3632	0.0113
	C-LoRA	**26.5394**	0.3251	0.1175	0.2908	0.2776	**0.2917**	0.3468	0.3085	0.0238
	ℓ_2-norm	27.1423	0.3425	0.0125	0.2995	0.2860	0.2306	0.3816	0.3930	**0.0000**
	EWC	28.8256	0.3461	0.0250	0.3016	0.2833	0.2639	0.3877	0.4129	0.0238
	HFT	28.8221	0.3500	0.0375	**0.3020**	0.2827	0.2444	**0.3918**	0.3930	0.0300
	MoFO	28.8221	**0.3500**	0.0350	**0.3020**	0.2827	0.2444	**0.3918**	0.3930	0.0300
	Replay	30.4569	0.3461	**0.2450**	0.3006	**0.2890**	0.2556	0.3530	**0.4527**	0.0395

Order 2	Method	Pretrain		Item	Domain		Cross			Forget
		FID ↓	Comp ↑	Unique-Sim ↑	Body-HPS ↑	Nature-HPS ↑	I+I ↑	I+D ↑	D+D ↑	Class-Sim ↓
"Nature" → "Body" → "V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker"	*Base*	26.3154	0.3378	0.0075	0.2966	0.2732	0.2250	0.3407	0.2637	0.0088
	Joint	29.2236	0.3472	0.0725	0.3054	0.2897	0.2528	0.3898	0.4527	0.0413
	SeqFT	**19.6193**	0.3359	0.2325	0.2950	0.3389	0.2833	0.4430	0.3781	0.0953
	SeqLoRA	22.2713	0.3433	0.1475	0.2921	0.2723	**0.4139**	0.4430	0.3184	0.0518
	IncLoRA	23.1411	**0.3519**	0.2300	0.2944	0.2859	0.3889	0.4470	0.3433	0.2300
	O-LoRA	22.7191	0.3411	0.0125	0.2862	0.2862	0.2361	0.3632	0.3881	N/A
	C-LoRA	23.9690	0.3414	0.0250	0.2883	0.2867	0.2583	0.3366	0.3781	N/A
	ℓ_2-norm	20.6750	0.3438	0.2150	**0.3031**	**0.2912**	0.3528	0.4245	0.3831	0.0405
	EWC	19.8055	0.3449	0.2575	0.2956	0.2775	0.3389	0.4431	0.4229	0.0750
	HFT	22.0834	0.3430	0.1450	0.3023	0.2845	0.3417	0.4368	0.4179	0.0363
	MoFO	20.5495	0.3416	**0.3950**	0.2954	0.2783	0.3583	0.4573	**0.4527**	0.1063
	Replay	29.0976	0.3471	0.0000	0.3008	0.2889	0.2389	0.3468	0.3550	**0.0213**

domain. LoRA variants perform well at domain learning. They yield strong human preference scores, even outperforming Joint on the Nature domain, and exhibit low domain forgetting. Yet they struggle to capture the more complex variations in the body domain, limiting their gains there. HFT achieves the highest HPS on the Body domain. Its strategy of reusing half the parameters and features effectively learns the detailed motions characteristic of body images. Replay remains the top performer on downstream metrics, and even achieves positive backward transfer (–0.70% Domain-Forget), implying that shared domain features can reinforce earlier knowledge. Exploring how to exploit these commonalities for more effective continual updating is a promising direction. MoFO delivers the best cross-task generalization (42.79%), though it is still behind Joint by 2.98%.

6.4 Continue Post-Training for Sequential Item-Domain Adaptation

The results for the Item-Domain Adaptation setting are reported in **Tab.** 2 for the two task orders: *Order 1* (items first, then domains) and *Order 2* (domains first, then items). The pronounced imbalance in size and quality between the item and domain datasets strongly affects continual learning.

In both task orders, *pretraining preservation* follows the second task: when domain enhancement follows item customization (Order 1), all methods see FID increase as in **Tab.** 1, mirroring the degraded image quality observed in sequential domain enhancement. Conversely, when item customization comes second (Order 2), FID decreases during that stage. Across both orders, CompBench scores improve for nearly every method, demonstrating consistent gains in text-image alignment through continual post-training. For *downstream performance*, LoRA variants split into unregularized (SeqLoRA, IncLoRA) and regularized (O-LoRA, C-LoRA) groups. The unregularized methods completely forget items in Order 1, yielding 0.0 accuracy. By contrast, the regularized methods preserve item accuracy when items are learned first but degrade significantly in Order 2, indicating that domain-task regularization can interfere with later item adaptation. Other regularization and sparse fine-tuning techniques also achieve strong results on whichever task is learned second, yet suffer severe forgetting on the first task. For example, unique-item accuracy for initially learned items nearly drops to zero in Order 1. Replay behaves differently from all others across the two orders. Its performance on the domain-enhancement task is insensitive to task order, but it only excels

when items are learned first. When items come second, Replay fails to acquire the new item-specific features. We hypothesize that, in Order 2, replaying the larger domain dataset severely interferes with learning the minority specialized item concepts. Notably, Joint also struggles in this imbalanced data-stream setup. Dominated by the larger domain dataset, Joint effectively overfits to domain enhancement and fails to learn the fine-grained personalized generation required for items.

For *cross-task generalization*, Joint also loses the benefits of separately training on items and domains in both orders. Because it underfits the item tasks, Joint performs poorly on Item+Item and Item+Domain generalization, though it remains best on Domain+Domain. The LoRA variants are primarily driven by their performance on item tasks. C-LoRA and O-LoRA achieve the highest Item+Item metrics in Order 1 but collapse in Order 2. Conversely, SeqLoRA and IncLoRA reverse that trend. All four LoRA methods exhibit weak cross-task generalization when paired with domain tasks. Regularization methods (ℓ_2-norm, EWC) and sparse fine-tuning methods (HFT, MoFO) perform poorly under Order 1 but nearly match or exceed Joint in Order 2. This indicates that task sequence not only affects knowledge updating and forgetting, but also the fusion and generalization of learned concepts. Finally, Replay fails to balance rehearsal of old data with adaptation to new data, resulting in weak cross-task generalization under both orders. Crucially, continual post-training sequences that first reinforce the coarse-grained domain and then learn fine-grained items emerge as a particularly promising direction.

6.5 Results Summary

Summarizing the experimental results across the three settings, we draw three key takeaways:

❶ *No single method excels everywhere.* Although LoRA variants indeed minimize forgetting, it severely degrades performance on item customization. Other rehearsal-free methods learn and preserve more knowledge than SeqFT, yet they still exhibit varying degrees of forgetting. Replay performs well under balanced data streams but its effectiveness becomes unstable under imbalanced streams. These results motivate the development of advanced continual post-training methods for T2I diffusion models that better reconcile the trade-off between stability and plasticity.

❷ *"Oracle" Joint learning is not a panacea.* In classical continual learning, Joint learning on all datasets is typically treated as the "oracle" upper bound. However, our study reveals that, although Joint usually outperforms baseline continual post-training methods in most scenarios, it can struggle conflicting demands of multi-task optimization, failing to reach optimal performance on specific domains, a limitation also observed in prior work (Chen et al., 2025a). Furthermore, under imbalanced tasks, Joint often overlooks few-shot concepts, such as minority items. These findings underscore both the challenge posed by our benchmark and the promising solution of continual post-training.

❸ *Cross-task generalization remains an open challenge.* In both the sequential item customization and domain enhancement scenarios, most methods fall short of Joint in cross-task generalization. Although many baselines can alleviate catastrophic forgetting, few match the oracle's ability to seamlessly recombine prior and newly acquired knowledge. This gap highlights the need for approaches that not only preserve prior representations but also actively integrate them with incoming information. For example, identifying shared parameters and features that can be reused to bootstrap new-task learning offers a promising path to enhance cross-task generalization. To accelerate this progress, we provide a standardized evaluation protocol within T2I-ConBench, empowering the continual learning community to develop and rigorously benchmark more sophisticated post-training methods.

7 Conclusions

This paper presents T2I-ConBench, a comprehensive benchmark for continual post-training of T2I diffusion models. We curate datasets spanning open-world scenarios with two levels of granularity and develop an automated evaluation pipeline that measures preservation of pretrained capabilities, downstream performance, forgetting, and cross-task generalization. We evaluate and analyze representative continual post-training methods across three sequential-task settings, establishing comparative baselines and insights to guide the development of more advanced methods. We hope that T2I-ConBench could serve as a standardized testing framework to accelerate both research and practical deployment of continual post-training techniques for T2I diffusion models.

REFERENCES

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Ana Beduschi. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data Soc.*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *ICCV*, 2021.

Mikhail Chaichuk, Sushant Gautam, Steven Hicks, and Elena Tutubalina. Prompt to polyp: Clinically-aware medical image synthesis with diffusion models. *arXiv preprint arXiv:2505.05573*, 2025.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-Σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *ECCV*, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2024b.

Yupeng Chen, Senmiao Wang, Yushun Zhang, Zhihang Lin, Haozhe Zhang, Weijian Sun, Tian Ding, and Ruoyu Sun. Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning. *arXiv preprint arXiv:2407.20999*, 2025b.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2025b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

Robert M. French. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? *NIPS*, 1993.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

Haifan Gong, Yitao Wang, Yihan Wang, Jiashun Xiao, Xiang Wan, and Haofeng Li. Diffuse-uda: Addressing unsupervised domain adaptation in medical image segmentation with appearance and structure aligned diffusion models. *arXiv preprint arXiv:2408.05985*, 2024.

Yuming Gu, You Xie, Hongyi Xu, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. *CVPR*, 2024.

Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. *arXiv preprint arXiv:2406.04295*, 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *ICLR*, 2022.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ICLR*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 2017.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 2023.

Tingfeng Hui, Zhenyu Zhang, Shuohuan Wang, Weiran Xu, Yu Sun, and Hua Wu. HFT: half fine-tuning for large language models. *arXiv preprint arXiv:2404.18466*, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *CVPR*, pp. 9307–9315, 2024.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. Continual training of language models for few-shot learning. *EMNLP*, 2022.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *ICLR*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017.

Black Forest Labs. Flux. 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ICML*, 2022.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025.

Zhibin Liao, Tom Drummond, Ian Reid, and Gustavo Carneiro. Approximate fisher information matrix to characterise the training of deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *ECCV*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

Wei Lu, Rachel K Luu, and Markus J Buehler. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *NPJ Computational Materials*, 2025.

Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *NeurIPS*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *ICLR*, 2025.

Jonathan Pilault, Amine Elhattami, and Christopher J. Pal. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. *ICLR*, 2021.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *ICCV*, pp. 2641–2649, 2015.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *ICLR*, 2024.

Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2020.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, 1990.

Yi Ren and Danica J. Sutherland. Learning dynamics of llm finetuning. *ICLR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *Comput. Res. Repos.*, 2020.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *Trans. Mach. Learn. Res.*, 2024.

Yu-Chuan Su, Kelvin C. K. Chan, Yandong Li, Yang Zhao, Han Zhang, Boqing Gong, Huisheng Wang, and Xuhui Jia. Identity encoder for personalized diffusion. *arXiv preprint arXiv:2304.07429*, 2023.

Gan Sun, Wenqi Liang, Jiahua Dong, Jun Li, Zhengming Ding, and Yang Cong. Create your world: Lifelong text-to-image diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CVPR*, 2016.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou,

Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *EMNLP*, 2023a.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023b.

Haoran Wei, Wencheng Han, Xingping Dong, and Jianbing Shen. Towards high-fidelity 3d portrait generation with rich details by cross-view prior-aware diffusion. *arXiv preprint arXiv:2411.10369*, 2024.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. *ICCV*, 2023a.

Zihao Wu, Huy Tran, Hamed Pirsiavash, and Soheil Kolouri. Is multi-task learning an upper bound for continual learning? *ICASSP*, 2023b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 2023.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv:2309.17421*, 2023.

Yutong Yin and Zhaoran Wang. Are transformers able to reason by connecting separated knowledge in training data? *arXiv preprint arXiv:2501.15857*, 2025.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *ICML*, 2017.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca++: Unleash the power of sequential fine-tuning for continual learning with pre-training. *arXiv preprint arXiv:2408.08295*, 2024a.

Haotian Zhang, Junting Zhou, Haowei Lin, Hang Ye, Jianhua Zhu, Zihao Wang, Liangcai Gao, Yizhou Wang, and Yitao Liang. Clog: Benchmarking continual learning of image generation models. *arXiv preprint arXiv:2406.04584*, 2024b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.

Xulu Zhang, Xiaoyong Wei, Wentao Hu, Jinlin Wu, Jiaxin Wu, Wengyu Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*, 2025.

Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 2024c.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. *ICML*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ICLR*, 2024a.

Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. *arXiv preprint arXiv:2306.14153*, 2024b.

APPENDIX

A ETHICS AND LEGAL CONSIDERATIONS

A.1 DATA SOURCES AND LICENSING

All item customization data in our work are derived from the publicly released *Dreambooth* dataset(Ruiz et al., 2023), which is licensed under CC-BY-4.0. The CC-BY-4.0 license explicitly permits dataset adaptation.

For domain enhancement, we rely entirely on synthetic data generated by the *FLUX.1-dev* model(Labs, 2024), with zero dependence on web-scraped content. The FLUX model itself is released under a non-commercial license permitting academic research, and we provide appropriate citations for its use.

A.2 LEGAL AND ETHICAL RESPONSIBILITY

Table A1: Harmful keyword filter results of T2I-ConBench

Keyword	hate	violence	terror	bomb	kill	racist	slur	sexual	drugs	crime
Count	0	0	0	0	0	0	0	0	0	0

Crucially, we do not propose or release any new diffusion models. Consequently, we are not subject to potential legal concerns stemming from the pre-training data sources of existing diffusion models.

During dataset construction, we are fully committed to avoiding potential legal and ethical issues for community usage. Specifically, we conducted rigorous screening of all released content and image pixels to ensure exclusion of harmful material. For instance, we implemented a keyword-based filtering mechanism (Schmidt & Wiegand, 2017) to inspect all prompts, finding no ethically harmful content, as shown in Table A1. Therefore, our dataset does not exacerbate risks of such issues arising in downstream models.

Beyond the scope of our dataset, it is important to acknowledge broader ethical and legal challenges faced by diffusion models in general. Many existing models are trained on large-scale internet datasets that may inadvertently contain copyrighted works, sensitive personal data, or materials collected without explicit consent. This raises important questions regarding licensing, consent, and fair use that continue to shape discussions around responsible dataset construction.

Another widely recognized issue is the potential for generative models to memorize portions of their training data and unintentionally reproduce sensitive or private content. This concern highlights the need for continued research on privacy-preserving training strategies, such as differential privacy or data sanitization methods, to safeguard against potential privacy leakage.

Finally, the unpredictability of continual learning and adaptation in diffusion-based systems underscores the importance of responsible deployment practices. To this end, we issue a cautionary warning: subsequent researchers should conduct independent verification, comply with relevant legal frameworks, and adopt responsible usage practices. This guidance strengthens the broader ethical discourse in the field and supports the development of diffusion models that align with societal values and legal standards.

A.3 SUPPLEMENTAL DISCLOSURES

During dataset construction, we implemented a multi-stage filtering pipeline to ensure compliance:

- Prompt-level screening using a keyword-based mechanism to remove potentially harmful or sensitive queries.
- Pixel-level inspection to exclude images with unsafe or inappropriate content.

We also recognize that, due to the inherent challenges of continual learning, diffusion models undergoing continued post-training may still produce unexpected or undesirable outputs, which poses

potential ethical risks. To mitigate this, we issue a cautionary note urging subsequent researchers to perform independent verification and to adopt responsible usage practices. This warning reinforces our discussion of ethical considerations and provides guidance for future work building upon our dataset.

B Reproducibility Statement

Due to the single-blind peer review restrictions associated with the training dataset used in this study, the data cannot be made publicly available at this time. We commit to open-sourcing the complete training dataset, training code, and evaluation code immediately upon the lifting of these restrictions, in order to support reproducible research and collectively advance the development of continual learning in the text-to-image generation field.

C Use of LLMs in this article

In this article, large language models (LLMs) are employed in two key stages of the benchmark:

1. **Prompt Generation (Construction Stage):** LLMs were used to automatically generate a large set of task-specific prompts, covering various domains such as natural scenes and human portraits. This approach avoids the limitations of manually designed prompts and ensures a diverse and challenging evaluation set.
2. **Evaluation Stage:** VQA evaluation, original text prompts are automatically decomposed into a set of questions using an LLM. These questions are then answered by a VLM-based VQA model to assess whether the generated images correctly reflect the prompts.This procedure allows systematic assessment of whether the images correctly reflect the textual prompts, leveraging the combined language understanding of LLMs and multimodal reasoning of VLMs. We strictly disclose the exact model versions.
3. **Writing Stage:** LLMs (both closed- and open-source) were used only for copyediting and grammar checking, including terminology normalization, syntactic polishing, and formatting. They were not used to generate claims, collect evidence, or construct results.

In summary, T2I-ConBench uses LLMs to construct and parse prompts, and VLMs (VQA models) to evaluate image-text alignment, forming a scalable and automated benchmark pipeline for text-to-image models.

D Related Work

D.1 Large-scale Text-to-image Generative Model

Large-scale text-to-image (T2I) diffusion models have rapidly become the backbone of generative AI. Building on latent diffusion, Stable Diffusion (Rombach et al., 2022; Podell et al., 2024) popularized an open-source U-Net (Ronneberger et al., 2015) conditioned on CLIP (Radford et al., 2021), capable of efficient generation by operating in a compressed latent space. Meanwhile, the PixArt series (Chen et al., 2024b;a) demonstrates that decomposed training stages, latent consistency modules, and weak-to-strong paradigms can reduce training cost by over 90%, while supporting 4K output and 2–4-step sampling for sub-second inference. The latest FLUX.1 models from Black Forest Labs scale diffusion transformers to 12B parameters with spatiotemporal attention and multi-stage noise scheduling, matching Midjourney and DALL E3 (Shi et al., 2020) in fidelity and prompt adherence. Crucially, pre-training on diverse, large-scale image–text data endows these models with strong zero-shot generalization, enabling them to adapt to downstream domain-specific or personalized tasks with minimal post-training.

D.2 Continual Post-training for Image Generation

Continual post-training (Lu et al., 2025; Smith et al., 2024; Ke et al., 2022; 2023) enables a single, large T2I diffusion model to absorb new, task-specific knowledge without full retraining, yielding substantial improvements on practical downstream applications. We target two key scenarios: **item**

customization (Zhang et al., 2025), where the model must learn to generate a novel object or style from only a few examples while maintaining consistency across diverse contexts, and **domain enhancement** (Guo et al., 2024), which focuses on refining overall image quality and semantic fidelity within a specialized visual domain. In item customization, methods such as C-LoRA (Smith et al., 2024) incrementally inject new concepts into cross-attention layers via low-rank adapters, while regularizing against forgetting; encoder-based adapters learn a compact network that maps reference images into embeddings fused into the diffusion process for rapid personalization (Su et al., 2023); and even zero-training approaches repurpose attention maps from exemplars at inference time to steer generation without further optimization (Hertz et al., 2022). For domain enhancement, techniques like Diffuse-UDA (Gong et al., 2024) and DiffBoost (Zhang et al., 2024c) adapt diffusion priors to medical imaging by aligning appearance and structural statistics or leveraging expert-model features, achieving high-fidelity lesion synthesis and enhanced segmentation generalization. Similarly, portrait-specific fine-tuning and 3D-aware adapter schemes improve face generation fidelity and multi-view consistency (Gu et al., 2024; Wei et al., 2024). Although these approaches deliver strong results in their respective settings, they focus on isolated, single-granularity task sequences and do not evaluate a model's capacity to recombine concepts across different domains. To address this gap, we propose a unified, sequential benchmark that integrates both item customization and domain enhancement, challenging models to preserve their pretrained versatility, master new tasks, and sustain multi-domain knowledge generalization.

D.3 Benchmarking Image Generation

Benchmarking image-generation models requires a suite of metrics that capture quality, diversity, and alignment with text prompts. Inception Score (IS) (Barratt & Sharma, 2018) evaluates sharpness and diversity by measuring the confidence and entropy of class predictions from a pretrained Inception-v3 (Szegedy et al., 2016) network on generated samples. The Fréchet Inception Distance (FID) (Heusel et al., 2017) compares the mean and covariance of deep Inception features between generated and real images, quantifying distributional similarity. To assess perceptual similarity, LPIPS (Zhang et al., 2018), CLIP-I (Radford et al., 2021), and DINO Score (Caron et al., 2021) compute distances in learned feature spaces, reflecting human judgments of visual similarity. Global text–image alignment is measured by multimodal encoders via CLIP-T, CLIPScore (Radford et al., 2021), and BLIP (Li et al., 2022), which score how well an image matches its prompt in the joint embedding space. For fine-grained semantic and logical fidelity under complex prompts, benchmarks like GenEval (Ghosh et al., 2023) using object detectors and T2I-CompBench (Huang et al., 2023) probe category- and relation-level understanding. To capture human preference, learned reward models such as HPS (Wu et al., 2023a) and ImageReward (Xu et al., 2023) encode crowd-sourced judgments into automatic scores. Recent personalization benchmarks, DreamBench (Ruiz et al., 2023) and DreamBench++ (Peng et al., 2025), leverage multimodal LLMs (Li et al., 2025; Zhu et al., 2024a) to evaluate object-level customization quality. Building on these, we introduce a vision-language-LLM-QA-based pipeline (Ma et al., 2023) to measure cross-task generalization, which is the ability to recombine old and new concepts across sequential downstream tasks. By extending static metrics into dynamic continual-learning streams, our benchmark quantifies not only per-task performance and forgetting but also knowledge transfer and synergy between tasks. CLoG (Zhang et al., 2024b) also aims at benchmarking continual learning of generative models, but unlike its continual pre-training setting starting from scratch, we focus on continual post-training, and uniquely assesses both retention of pre-trained zero-shot capabilities and knowledge generalization in mixed-task streams.

E Broader Impact and Limitations

Impact T2I-ConBench fills a critical gap in continual post-training evaluation by introducing, for the first time, a unified protocol that measures pretrained capability preservation, downstream task performance, catastrophic forgetting, and cross-task compositional generalization—laying a solid foundation for fair comparisons and reproducible research. Through two systematic tasks—"item customization" and "domain enhancement"—the benchmark not only uncovers the key trade-offs between retaining prior knowledge and adapting to new tasks, but also quantifies the dynamic degradation of old and new concept performance and the shortcomings of zero-shot composition. By releasing our datasets, prompt libraries, and evaluation pipeline, we dramatically lower the barrier

for both research and deployment, spurring innovation across diverse application domains such as industrial design, medical imaging, and cultural heritage. At the same time, we must remain vigilant about potential downsides: a standardized continual post-training toolkit could be misused to rapidly produce highly realistic deepfakes or personalized attacks, heightening privacy risks and misinformation. Moreover, the automated evaluation metrics and pretrained models underpinning our benchmark may carry social biases, risking the inadvertent perpetuation or amplification of unfairness.

Limitation Despite its unique breadth of evaluation, T2I-ConBench has several limitations. First, for precise concept-targeted generation, we rely on the FLUX model, meaning our synthetic data may inherit its biases and constraints in detail fidelity and aesthetic style, which can limit our capacity to fully assess semantic accuracy and visual consistency. Second, we focus exclusively on diffusion architectures and omit equally popular autoregressive generative models, whose differing training regimes and inductive biases could affect the relative performance of various continual post-training methods—an open question for future study. Finally, due to computational resource constraints, we evaluated on stable, mid-scale models rather than the largest and most cutting-edge networks. Nevertheless, our evaluation pipeline is model-agnostic and can readily incorporate the latest diffusion or autoregressive models going forward.

F Detailed Task Definition

Continual post-training of large pre-trained T2I diffusion models refers to the process of sequentially fine-tuning a single foundation model on a series of small, task-specific datasets. The models are expected to fine-tune on each task without revisiting earlier data to customize to new domains or concepts while preserving their original generative capabilities. Concretely, let a model M_0 with parameters θ_0 have completed a broad pre-training task $\mathcal{T}_0$. We then define a downstream task sequence $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K\}$. Each task $\mathcal{T}_i$ provides a dataset $\mathcal{D}_i = \{(x_{i,n}, y_{i,n})\}_{n=1}^{N_i}, 1 \leq i \leq K$ of N_i text-image pairs sampled from distribution P_{x_i,y_i}. The datasets are disjoint, $\mathcal{D}_i \cap \mathcal{D}_j =, i \neq j$. A continual post-training algorithm $\mathcal{A}$ produces a sequence of models $M_i = \mathcal{A}(M_{i-1}, \mathcal{D}_i)$ such that each M_i both maximizes the likelihood $p_{M_i}(\hat{y}|x_i)$ on new task $\mathcal{T}_i$ and minimizes degradation on all previous tasks $\{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{i-1}\}$. Balancing these objectives requires effective mitigation of catastrophic forgetting while still integrating new knowledge. Unlike traditional benchmarks on image generation that compare different models or training datasets, our continual post-training benchmark fixes both the base model and the task datasets. It is therefore a systematic evaluation of continual learning strategies: isolating the impact of training algorithms, without conflating results with variations in data quality or model architectures. This design enables precise measurement of how different continual post-training methods truly affect downstream performance, preservation of prior knowledge, and cross-task generalization.

Cross-task generalization evaluates the ability to recombine knowledge acquired from different tasks into novel concepts. In addition to per-task performance metrics, our benchmark introduces a compositional generation evaluation to quantify this property during continual post-training. This builds on the key observation that pre-trained diffusion models often exhibit zero-shot compositionality, e.g., after learning both "a person riding a horse" and "astronaut" in the pre-training stage, they can generate "an astronaut riding a horse," which they have never seen during the training process. We wonder whether, when a model is continually post-trained first on $\mathcal{T}_1$ ("a person riding a horse") and then on $\mathcal{T}_2$ ("astronaut"), does it retain the ability to produce the novel concept $\mathcal{T}_1 \cup \mathcal{T}_2$ ("an astronaut riding a horse")? Formally, let $g(x_i, x_j)$ be a semantic-composition function that combines two prompt conditions from tasks $\mathcal{T}_i$ and $\mathcal{T}_j$. After obtaining the model M_i via continual post-training on tasks $\{\mathcal{T}_0, \ldots, \mathcal{T}_i\}$, we measure its cross-task generalization by conditional generation likelihood $p_{M_i}(\hat{y}|g(x_i, x_j))$ for pairs (x_i, x_j) drawn from different tasks. A high generation likelihood indicates that the model not only learned each task's concepts but also preserved the representational flexibility to recombine them in novel ways. This metric thus reveals whether continual post-training sustains the emergent compositional structure of pre-trained knowledge and supports long-term accumulation of generative capabilities.

G Detailed Dataset Description

Dataset Curation Process involves Identifying Challenging Concepts, Prompt Creation, Image Generation, and Quality Filtering, details are given below:

❶ *Identifying Challenging Concepts* For the construction of our domain-specific datasets, we explicitly target concepts that are empirically difficult for T2I models. Rather than relying on purely subjective impressions, we identify "challenging concepts" via a systematic, error-based procedure consisting of three steps. **Candidate concept collection.** We first compile a pool of candidate concepts that are known or suspected to be difficult for T2I models in the `nature` and `body` domains (e.g., fine-grained anatomical structures, complex poses, or visually intricate natural phenomena). This pool is constructed based on prior empirical observations as well as common failure reports for existing models. **Prompt generation and image sampling.** For each candidate concept, we use an LLM to generate a diverse set of prompts that must explicitly contain that concept while varying other contextual factors (e.g., background, style, composition). Given these prompts, we sample multiple images from the base model for each prompt, resulting in a small image set associated with each candidate concept. **Statistical error-based labeling.** Human evaluators then inspect the generated images and count semantic and logical errors according to explicit criteria, such as incorrect object attributes, implausible spatial relationships, or anatomically inconsistent body parts (e.g., wrong number of limbs). For each candidate concept, we compute the fraction of images exhibiting such clear semantic/logical failures. If more than 50% of the images associated with a concept are labeled as erroneous, we designate that concept as a "challenging concept" and include it in the domain dataset.

❷ *Prompt Creation* Once the challenging concepts were identified, we utilized a LLM to construct a diverse set of descriptive captions featuring these specific concepts. This collection of captions was subsequently divided through random sampling to serve distinct purposes. The majority of these captions were allocated as prompts for the training dataset, while the remaining smaller portion was set aside to form the test set, intended for later evaluation of model capabilities within this domain.

❸ *Image Generation and Quality Filtering* To ensure that T2I-ConBench is built on high-quality data, we adopt a rigorous multi-stage filtering pipeline. Although this results in a relatively low acceptance rate (about 4% of the raw pool), the high rejection rate is a consequence of deliberately strict quality control rather than noisy data collection. Concretely, the filtering process consists of three stages: (i) model-based filtering, (ii) manual screening, and (iii) ethical review. **Model-based filtering.** Given the initial pool of images and prompts, we first perform an automatic model-based filtering step. For each image, we compute the mean HPS and CLIP score and remove samples whose scores fall below pre-defined thresholds. This stage eliminates approximately 60% of the lowest-quality data, primarily removing images that are visually degraded or exhibit very weak alignment with their prompts. **Manual screening.** For the remaining images, we conduct a meticulous manual screening. Human annotators examine the image–prompt alignment, overall visual quality, and semantic coherence. Particular attention is paid to semantic correctness of the depicted content, such as the correct number of limbs for humans and animals, the presence or absence of key attributes mentioned in the prompt, and obvious structural artifacts. This stage removes an additional $\sim 30\%$ of the data that pass the automatic thresholds but still fail to satisfy our standards for faithful and coherent rendering. **Ethical review.** Finally, to ensure that the benchmark adheres to ethical standards, we perform a dedicated ethical review over the remaining samples. Images that are considered ethically inappropriate (e.g., due to sensitive content, offensive stereotypes, or privacy concerns) are removed. This stage excludes roughly another 6% of the data.

The final dataset sizes are 2513 for the Nature domain, 2356 for body poses, and 1821 for interactions with common animals. The latter two constitute the Body domain training dataset. The detailed information of the domain-enhancement dataset is shown in **Tab.** A2.

Item Customization Dataset is derived from the publicly released Dreambooth dataset(Ruiz et al., 2023), which is licensed under CC-BY-4.0 and explicitly permits adaptation. In line with the official DreamBooth example implementation [2], we follow the same prompt construction strategy, using the template "A photo of V{index} {item}" as the training prompt for each image. Detailed item customization raining dataset information is provided in **Tab.** A3. To mitigate overfitting on specific

[2] `https://github.com/google/dreambooth`

items, we additionally generate 500 prior preservation images per item with the pretrained model. Although we currently experiment with only four items, the dataset design and training pipeline can be easily extended to support a larger variety of items.

H Detailed Evaluation Pipeline

Benchmarking continual learning methods requires not only the evaluation of static tasks, but also the dynamic evaluation of the performance of the text graph model to detect the performance improvement of downstream tasks and the forgetting of old task knowledge. We designed a unified indicator selection for evaluating the quality of different aspects of text graphs. In addition to the final performance and forgetting metrics commonly used in continuous learning benchmarks, we also focus on the changes in the general capabilities of large models, measure model performance from two aspects: generation quality and semantic logic, and pay attention to the evaluation of cross-task generalization capabilities.

Table A2: The detailed concepts of the training dataset.

Category	Specific Actions or Objects	Count	Total
Nature	Gerenuk	1043	2513
	Spix's Macaw	590	
	Quokka	492	
	Pomelo	363	
	Squid	25	
Body: Poses	Hands naturally hanging by the sides	112	2356
	Gestures of hearts, victory, peace	105	
	Hands joined in prayer pose	193	
	Resting chin on one hand	214	
	Holding with both hands	288	
	Hands in pockets	188	
	Covering Face	51	
	Waving hands	133	
	Arms crossed	226	
	Thumbs-up	125	
	Press down	180	
	Gripping	234	
	Pointing	59	
	Salute	39	
	Fist	188	
	Others	21	
Body: Interaction with Common Animals	Dog	247	1821
	Elephant	138	
	Panda	177	
	Tiger	135	
	Cat	192	
	Monkey	136	
	Horse	27	
	Butterfly	189	
	Lion	173	
	Giraffe	111	
	Dolphin	88	
	Kangaroo	185	
	Penguin	23	

Pretrain Preservation To assess how well continual post-training preserves pretrained capabilities, we use two metrics against the base model.

Table A3: Number of training samples per item customization task.

Item	V1 dog	V2 dog	V3 cat	V4 sneaker
Count	5	5	5	5

❶ *Generation Quality* We use Fréchet Inception Distance (**FID**) (Heusel et al., 2017) to evaluate both the quality and diversity of images generated by diffusion models. By comparing the statistical distributions of generated versus real images in the feature space of a pretrained network, FID quantifies how closely a model's output matches the true data distribution. We use fixed 30,000 captions from MS-COCO (Lin et al., 2014) to generate images for measuring the quality of T2I models. The lower the FID value, the better the quality of the generated images indicated by the trained model. We implement FID from text2image-benchmark and employ Inception V3 model (Szegedy et al., 2016) as the pretrained network with precomputed FID stats.

❷ *Text-image Alignment* T2I-CompBench (Huang et al., 2023) is a comprehensive benchmark for open-world combinatorial T2I generation, providing a large-scale dataset and semantic logic and text-image alignment evaluation metrics. The evaluation dimensions include: multi-entity relationship construction, precise attribute binding, spatial reasoning consistency, and cross-modal semantic fidelity. Through multimodal large language model evaluation, the semantic accuracy of text-to-image models under complex text prompts can be evaluated. We select the 3-in-1 evaluation for complex compositions (**Comp**) as our metrics.

Downstream Performance We define separate evaluation metrics for two downstream tasks with different granularity.

❶ *Item Customization* For each item customization task, we evaluate the model's ability to generate each fine-grained concept independently. Specifically, for each unique item, we prompt the post-trained model to produce a test image and use the original training-set image as a reference. We then convert each test prompt into a corresponding question using the template in **Tab.** A4. Finally, a VLM assesses the similarity between the generated image and its reference to produce the score $\mathrm{Sim}(i,k) = \{\text{score}\} \times 100\%$, representing the similarity score of the i-th question in the k-th unique personalized item. The pipeline is illustrated in **Fig.** A1. The **Unique-Sim** metric is calculated by the average of all unique personalized items:

$$\textbf{Unique-Sim} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}\mathrm{Sim}(i,k), \tag{A1}$$

where K is the number of item customization tasks, N_k is the number of question-image pairs corresponding to the k-th unique personalized item.

Table A4: The corresponding templates of prompt words and questions. When calculating unique accuracy, a corresponding question and reference image pair is generated for each personalized prompt. When calculating class similarity, four questions and reference image pairs are generated for each class prompt.

Class	**Unique**	**Question Template**
dog	V1 dog	What is the probability that the second image has the same *dog* as the first image? Please just answer the probability.
dog	V2 dog	What is the probability that the second image has the same *dog* as the first image? Please just answer the probability.
cat	V3 cat	What is the probability that the second image has the same *cat* as the first image? Please just answer the probability.
sneaker	V4 sneaker	What is the probability that the second image has the same *sneaker* as the first image? Please just answer the probability.

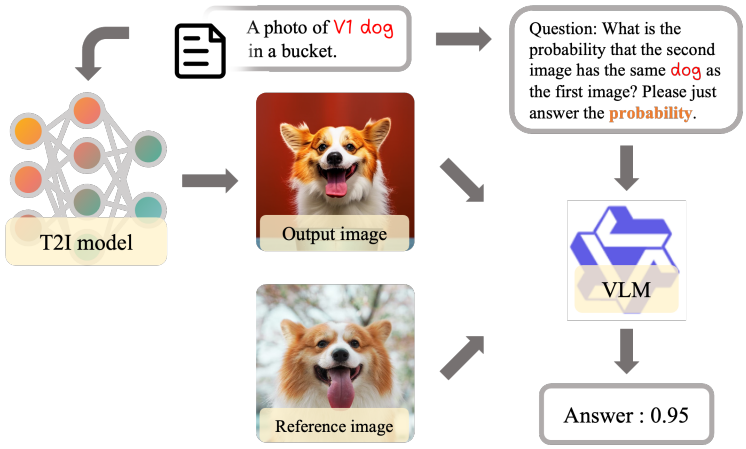

Figure A1: Evaluation pipeline of the unique personalized item similarity by VQA for Item customization tasks.

❷ *Domain Enhancement* Human Preference Score (HPS) (Wu et al., 2023a) is an automated evaluation metric trained to predict human judgments on T2I outputs by fine-tuning a CLIP model on the large-scale Human Preference Dataset. We employ HPS to evaluate the Body and the Nature domain for the alignment of the generated images with human aesthetics, respectively as **Body-HPS** and **Nature-HPS**.

Forget Measure Beyond measuring degradation of pretrained capabilities relative to the base model, we also quantify forgetting in downstream performance dynamics during continual post-training and class concept forgetting in specific item customization tasks:

❶ *Backward Transfer* For both Item Customization and Domain Enhancement, we compute *backward transfer* on their respective downstream metrics to evaluate the knowledge stability across sequential tasks. Backward transfer is the relative influence of learning the k-th task on all old tasks, defined as follows:

$$\textbf{Forget} = \frac{1}{K-1}\sum_{k=1}^{K-1}\mathrm{BWT}_k, \quad \mathrm{BWT}_k = \frac{a_{k,k}-a_{K,k}}{a_{k,k}}, \tag{A2}$$

where $a_{k,j}$ is the evaluation metric for the j-th task after the k-th round of training. Negative values indicate performance degradation on earlier tasks. By substituting the task-sequence Unique-Sim and HPS values into a, we can get the **Unique-Forget** and **Domain-Forget** metrics, respectively.

❷ *Class Concept Forgetting* When learning personalized concepts in item customization tasks, we evaluate the forgetting of their corresponding base classes. We generate images for non-personalized prompts (e.g., "a dog...") and evaluate their similarity with all learned unique personalized items (e.g., "V1 dog..."). Combined with the designed question template, each test prompt is converted into multiple corresponding personalized similarity questions. The question template is shown in **Tab.** A4. The VLM model is used to test the similarity of the generated image with the reference image in the personalized concept and obtain a score $\mathrm{Sim}(i_k, j) = \{\text{score}\} \times 100\%$ as in Unique-Sim. We use $\mathrm{Sim}(i_k, j)$ to represent the similarity score of images generated by the i-th prompt of class k with the j-th unique item. Then we need to compute **Class-Sim** as BWT by evaluating each current class's prompts with all old classes:

$$\textbf{Class-Sim} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{K}\sum_{j=1}^{K}\mathrm{Sim}(i_k, j) \tag{A3}$$

where N_k is the number of non-personalized prompts in class k.

Cross-task Generalization To evaluate the post-trained model's ability to recombine concepts from different tasks, we generate novel, compositional prompts as described in **Sec.** 3 and measure

how accurately the continually learned model renders them. This evaluation follows a three-step VQA-based pipeline (see **Fig.** 3):

❶ *Prompt Decomposition* We use an LLM to break each compositional test prompt into 2–4 simpler questions that collectively cover all relevant objects and their interactions (e.g., "Are the dogs running in the image?"). We ensure these sub-questions comprehensively probe both individual elements (vertical objects, personalized instances) and their relational actions. Example templates and generated Q&A pairs for different test sets are shown in **Fig.** A2.

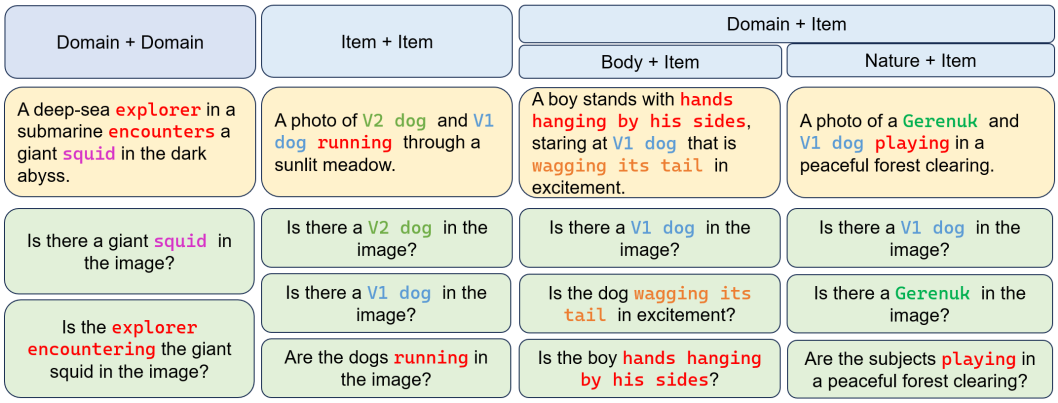

Figure A2: Example decomposition of four cross-task prompts into questions across different combination types. Colored highlights in each prompt and question indicate the key objects and actions under evaluation.

❷ *VQA Formatting* Each LLM-generated question t is formatted into a VQA-compatible query $q(t)$. For generic scenes without specialized objects, we simply append "Please answer yes or no."

$$\mathbf{t} = \text{Are the dogs running in the image?}$$
$$\mathbf{q}(\mathbf{t}) = \text{"Are the dogs running in the image?" Please just answer yes or no.}$$

For questions involving natural or personalized objects, we apply object-specific templates (**Tab.** A5).

❸ *Answer Scoring* The visual-language model (VLM) processes each image–question pair $(x, q(t))$ and returns "yes" or "no." We assign a score of 1 for "yes" and 0 otherwise. The overall cross-task score for a test set is the fraction of "yes" responses across all N image–question pairs:

$$\text{score}(x, q(t)) = \begin{cases} 1 & \text{, answer= "yes"} \\ 0 & \text{, otherwise} \end{cases} \tag{A4}$$

Finally, the test score of the cross-task test set is defined as the proportion of "yes" answers among all question-answer pairs in the test set:

$$\text{Cross} = \frac{1}{N} \sum_{i=1}^{N} \text{score}(x_i, q(t_i)) \tag{A5}$$

We denote cross-task generalization metrics for different task combinations as **Item+Item**, **Item+Domain**, and **Domain+Domain**, respectively.

Remark We use the open-source LLMs DeepSeek V3 (DeepSeek-AI, 2025b) and DeepSeek R1 (DeepSeek-AI, 2025a). Our VQA pipeline employs Qwen2.5-7B-Instruct (Qwen, 2025) as the VLM. We acknowledge that more advanced—and potentially more accurate—models like GPT-4V (Yang et al., 2023) exist for evaluation, but we provide a minimal, fully reproducible setup with open-source models better suited for benchmarking. We also retain interfaces that allow seamless integration of more advanced evaluators as the benchmark evolves.

Table A5: Example templates for VQA prompts that require reference images: each question is converted into a formatted question for the VLM, illustrating how personalized items and natural species are described and queried in a two-image comparison.

Task	Concept	Formatted Question
Item	V1 dog	"The image of the V1 dog is image 1, please identify the breed of this dog. For image 2, is there a dog of the same breed and similar appearance? Please just answer yes or no."
	V2 dog	"The image of the V2 dog is image 1, please identify the breed of this dog. For image 2, is there a dog of the same breed and similar appearance? Please just answer yes or no."
	V3 cat	"The image of the V3 cat is image 1, please identify the breed of this cat. For image 2, is there a cat of the same breed and similar appearance? Please just answer yes or no."
	V4 sneaker	"The image of the V4 sneaker is image 1, please identify the style of this sneaker. For image 2, is there a sneaker of the same style and similar appearance? Please just answer yes or no."
Nature	pomelo	"The image of the pomelo is image 1. Image 2 is a part of the pomelo. For image 3" + question t
	Spix's macaw	"The image of Spix's macaw is image 1, have over 30 percent blue feathers. For image 2," + question t
	Squid	"The image of Squid is image 1. Note that Squids have a distinct elongated body and tentacles, and should not be confused with Octopuses, which have a more rounded body and eight arms without distinct tentacles. For image 2," + question t
	Quokka	"Only animals that are many similar to the one in image 1 will be considered Quokka.For image 2," + question t
	Gerenuk	"Only animals that are many similar to the one in image 1 will be considered Gerenuk.For image 2," + question t

I Detailed Continual Post-training Baselines

- **Base** employs the pretrained model without further continual post-training, establishing a baseline on general generative capabilities and downstream tasks.
- **Joint** (Wu et al., 2023b) jointly trains the model on all task data, characterizing the upper bound of performance in sequential learning.
- **SeqFT** (Zhang et al., 2024a; Chen et al., 2025a) sequentially fine-tunes the model on each task with all parameters updated in task order. The model is optimized exclusively for the current task without preserving pretrained knowledge or retaining performance on earlier tasks.
- **Replay** (Chaudhry et al., 2019) maintains a small memory buffer that stores samples to replay prior knowledge. We store 10% of each completed task's image–text pairs in a small memory buffer and mix them with new-task data during subsequent fine-tuning. For item datasets with fewer than 10 examples, we ensure at least one sample is retained for replay.
- **ℓ_2-norm** (Zhao et al., 2024) adds an ℓ_2-norm penalty on the change from the previous task's final parameters. Concretely, when training on task i, the loss becomes $\mathcal{L}_i = \mathcal{L}_i^{\text{new}} + \lambda\Omega_i(\theta_i, \theta_{\text{old}})$, where $\mathcal{L}_i^{\text{new}}$ is the standard loss on the new task, θ_{old} are the frozen parameters from previous tasks, Ω_i is the regularization function, and λ controls its strength. Formally, the regularization term of ℓ_2-norm is $\Omega_{\ell_2} = \|\theta_i - \theta_{i-1}\|_2$. This term discourages large deviations from the starting values at the beginning of task i, thereby limiting drastic parameter shifts when learning the new task.
- **EWC** (Kirkpatrick et al., 2017) is built upon the ℓ_2-Norm baseline by weighting each parameter's penalty according to its estimated importance to previous tasks. Let F_k be the Fisher Information

Matrix (FIM) (Liao et al., 2018) computed after task k. We form a diagonal approximation on all old tasks $\hat{F}_{1:i-1} = \sum_{k=1}^{i-1} \mathrm{diag}(F_k)$. When training on task i, the regularization term is $\Omega_{\mathrm{EWC}} = (\theta_i - \theta_{i-1})^\top \hat{F}_{1:i-1} (\theta_i - \theta_{i-1})$. Parameters with higher Fisher scores incur a larger penalty for deviation, thereby more effectively preserving those weights most critical to earlier tasks.

- **HFT** (Hui et al., 2024) randomly splits parameters into two groups before each new task, i.e., $\theta_k = \{\vartheta_k, \psi_k\}$. One half (50%) ϑ_k is updated on the new task, $\vartheta_k^t \leftarrow \vartheta_k^{t-1} - \eta \nabla_{\vartheta_k} \mathcal{L}\left(\theta_k^{t-1}\right)$, while the other half ψ_k remains frozen to preserve prior knowledge, $\psi_k^t \leftarrow \psi_{k-1}^t$. During each task of continual post-training, only the active group is tuned, achieving a dynamic balance between learning new concepts and retaining old ones.
- **MoFO** (Chen et al., 2025b) leverages the momentum terms from the Adam optimizer (Kingma & Ba, 2015) to approximate parameter importance. To keep computation efficient, MoFO first groups parameters by their natural components (e.g., weight matrices versus bias vectors). After each backward pass, parameters are ranked by the absolute value of their momentum. MoFO updates only parameters with the largest $\alpha\%$ momentum in each partition for critical directions, and the rest remain frozen. By focusing updates on these "high-momentum" directions, MoFO achieves a sparse, adaptive fine-tuning that accelerates learning of new tasks without destabilizing performance on previously learned tasks.
- **SeqLoRA** (Devlin et al., 2019) shares a single LoRA adapter across all tasks. The LoRA adapter factorizes the update as $\Delta W \approx BA$, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times d}$ are low-rank matrices with $r \ll d$). During post-training on each task, the original weights remain frozen and only A and B are learned. This approach is simple and efficient, but may suffer from interference between tasks.
- **IncLoRA** (Wang et al., 2023a) allocates a fresh, independent LoRA adapter (B_i, A_i) for each new task i. The model's effective weight after i tasks for inference is $W_i = W_0 + \sum_{k=1}^{i} B_k A_k$ By assigning each task its own low-rank subspace, it enforces strict task isolation at the cost of linearly increasing the number of parameters.
- **O-LoRA** (Wang et al., 2023a) extends IncLoRA by imposing orthogonality across task adapters. When training the i-th adapter, it minimizes the task loss subject to $L_{\text{O-LoRA}} = \lambda \sum_{j=1}^{i-1} \|B_j^\top B_i\|^2$, ensuring each task's low-rank subspace remains mutually orthogonal. λ is the coefficient to control regularization strength. This further reduces parameter conflict across tasks and enhances knowledge separation.
- **C-LoRA** (Smith et al., 2024) adds a self-regularization term that penalizes deviations between the LoRA update for the new task and the adapters learned for previous tasks. The addition loss term for task i is $\mathcal{L}_{\text{C-LoRA}} = \lambda \left\| \left[\sum_{j=1}^{t-1} \boldsymbol{A}_j \boldsymbol{B}_j \right] \odot \boldsymbol{A}_i \boldsymbol{B}_i \right\|_2^2$, where λ balances adaptation to the new task with consistency to prior updates. By encouraging consistency with prior adapters, C-LoRA strikes a balance between retaining old knowledge and adapting to new tasks.

J Detailed Training Implementation

Sequential Item Customization We build on the Diffusers DreamBooth example, integrating DeepSpeed (Rajbhandari et al., 2020) Stage 2 for memory-efficient training. All methods fine-tune using the following shared settings unless noted otherwise:

- Optimizer: AdamW (Loshchilov & Hutter, 2019) with learning rate of 5×10^{-6}, weight decay 1×10^{-2}, gradient clipping at 1.0.
- Batch size $= 4$.
- Scheduler: constant learning rate.
- Training Steps: 500 for each item.

For each task, we use 500 prior class images generated by the base model to prevent overfitting on each personalized concept, with a prior regularization coefficient of 0.02.

Baseline-specific configurations:

- Joint: train 2000 steps on all item datasets.
- LoRA Variants (SeqLoRA, IncLoRA, O-LoRA, C-LoRA): rank $= 16$, LoRA $\alpha = 32$.

Table A6: Performance of continual post-training methods for the *sequential item-domain adaptation* task of *Order 2* using *SD v1.4*. ↑: higher is better. ↓: lower is better. "I" and "D" denote Item and Domain, respectively, with combinations indicating cross-task generalization evaluations. Excluding *Base* and *Joint*, the best result is shown in **bold** and the second-best is underlined.

Order 2	Method	Pretrain		Item	Domain		Cross			Forget
		FID ↓	Comp ↑	Unique-Sim ↑	Body-HPS ↑	Nature-HPS ↑	I+I ↑	I+D ↑	D+D ↑	Class-Sim ↓
"Nature" → "Body" → "V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker"	*Base*	9.9275	0.2901	0.0000	0.2118	0.2229	0.1806	0.2224	0.1493	0.0013
	Joint	22.7432	0.3097	0.1225	0.2968	0.2851	0.2028	0.3020	0.2985	0.0293
	SeqFT	19.0929	0.3043	0.3450	**0.2919**	0.2598	**0.3444**	0.2632	0.2289	0.0238
	SeqLoRA	16.5584	0.2805	0.3025	0.2519	0.2422	0.3111	0.2918	0.1940	0.0788
	IncLoRA	17.7793	0.2766	0.2675	0.2473	0.2502	0.3306	0.3061	0.1791	0.0850
	O-LoRA	14.1877	0.2727	0.2700	0.2425	0.2553	0.2778	0.2958	0.1244	0.0458
	C-LoRA	**14.1097**	0.2804	0.2975	0.2465	0.2564	0.2778	0.2754	0.1443	0.0550
	ℓ_2-norm	14.7921	0.2992	0.2300	0.2680	0.2577	0.2722	0.3081	0.2587	0.0475
	EWC	19.3321	0.2883	**0.5050**	0.2794	0.2691	0.2917	0.2959	0.1990	0.0543
	HFT	16.6841	**0.3136**	0.3525	0.2901	0.2633	0.3111	0.3121	**0.2786**	**0.0093**
	MoFO	17.0268	0.3053	0.2600	0.2907	0.2650	0.2917	0.2939	0.2239	0.0418
	Replay	17.8449	0.3084	0.3450	0.2884	**0.2796**	0.3333	**0.3122**	0.2189	0.0668

- O-LoRA: orthogonality penalty $\lambda = 1 \times 10^{-1}$.
- C-LoRA: self-regularization $\lambda = 1 \times 10^{6}$.
- ℓ_2-norm: regularization coefficient $\lambda = 1 \times 10^{-3}$.
- EWC: regularization coefficient $\lambda = 1 \times 10^{-4}$.
- HFT: freeze half of each layer's parameters (freeze ratio = 0.5).
- MoFO: partition at the parameter level, updating only the top 50% by momentum ($\alpha = 0.5$) and build upon Adam (Kingma & Ba, 2015).
- Replay: replay all previous item samples.

Sequential Domain Enhancement We build on the PixArt-α training pipeline, integrating DeepSpeed Stage 2 for efficient memory usage. All methods share these base settings unless specified otherwise:

- Optimizer: AdamW (Loshchilov & Hutter, 2019) with learning rate of 5×10^{-6}, weight decay 1×10^{-2}, gradient clipping at 1.0.
- Batch size $= 256$.
- Scheduler: constant learning rate.
- Training Steps: 3000 for each domain.

Baseline-specific configurations:

- Joint: train 48000 steps on all domain datasets.
- LoRA Variants (SeqLoRA, IncLoRA, O-LoRA, C-LoRA): rank $= 16$, LoRA $\alpha = 32$.
- O-LoRA: orthogonality penalty $\lambda = 1 \times 10^{-1}$.
- C-LoRA: self-regularization $\lambda = 1 \times 10^{6}$.
- ℓ_2-norm: regularization coefficient $\lambda = 1 \times 10^{-3}$.
- EWC: regularization coefficient $\lambda = 1 \times 10^{-4}$.
- HFT: freeze half of each layer's parameters (freeze ratio = 0.5).
- MoFO: partition at the parameter level, updating only the top 50% by momentum ($\alpha = 0.5$) and build upon Adam (Kingma & Ba, 2015).
- Replay: replay 10% domain samples.

K Additonal Experiment Results

K.1 Continue Post-training for Sequential Item-Domain Adaptation on SD v1.4

In addition to PixArt-α based on DiT (Peebles & Xie, 2023) used in **Sec.** 6, we also experiment with **Stable Diffusion v1.4** (**SD v1.4**) (Rombach et al., 2022), a U-Net–based model, as the base model. We apply the same Order 2 of Sequential Item-Domain Adaptation task (**Tab.** 2) to evaluate various continual post-training methods. The results are shown in **Tab.** A6. Overall, the findings mirror our key takeaways. A detailed analysis follows:

FID Since SD v1.4 exhibits stronger pretrained generative capabilities, all continual post-training methods show a distribution shift–induced quality drop on the new datasets.

Text–Image Alignment (Comp) Nearly every method improves alignment, except the LoRA variants and EWC.

Item Joint suffers from domain data bias and remains weak on item Unique-Sim. EWC achieves the best item metrics, and notably, Replay also successfully learns item concepts on SD v1.4 (unlike on PixArt-α), indicating that Replay's effectiveness varies across architectures.

Domain Joint continues to set the upper bound. Most methods exhibit forgetting on the first Nature domain; Replay best updates and preserves Nature domain knowledge, followed closely by EWC.

Cross-Task Generalization Joint excels only on Domain+Domain composition and underperforms on item-domain mixed generalization. All LoRA variants struggle, whereas HFT emerges as the strongest overall, underscoring the effective and efficient solution of parameter and feature reuse for knowledge fusion.

K.2 Visualization of Cross-task Generalization Results

Fig. A3,A4,A5,A6 present example cross-task generalization test images generated by the models across the three sequential task settings.

L Discussion on Alternative Evaluation Metrics

We compare the other tipical pretraining metrics(CMMD, CLIPScore, and LPIPS) with FID:

- **CLIPScore** (Hessel et al., 2021) evaluates text-image semantic alignment. It is actually a subset component of our Comp metric. The Comp metric from (Huang et al., 2023) is a 3-in-1 solution that unifies BLIP-VQA for attribute binding, UniDet for spatial relations, and CLIPScore for non-spatial relations, and has been validated to correlate strongly with human assessment. We also compare Comp and CLIPScore. Results in Table A7 indicate a larger average rank difference (avg Δ rank) of 4.0. The absolute CLIPScore differences are also minor. Given that Comp offers a significantly broader evaluation scope than CLIPScore alone, we retain Comp as our primary semantic alignment metric.
- **LPIPS** (Zhang et al., 2018) is an image similarity metric that is not typically used to assess T2I generative quality, and thus we do not consider incorporating it in T2I-ConBench.
- **CMMD** (Jayasumana et al., 2024) is a metric similar to FID and capable of measuring T2I general generation quality. In Table A7, we observe that CMMD and FID induce nearly identical method rankings (average Δ rank difference of only 1.1667). Since both capture overall quality and FID is more prevalent in the T2I literature (Chen et al., 2024a), we focus on FID in the main paper.

M Necessity of the Curated Dataset.

While general corpora such as MSCOCO or Flickr exist, they lack specialized design for T2I continual post-training. To demonstrate T2I-ConBench's unique value, we conduct comparative experiments using Flickr30K (Plummer et al., 2015):

- **Dataset Construction.**We apply multimodal retrieval to Flickr30K and extract 3,506 "Nature" and 19,625 "Body" related images and prompts. We then sort them by descending HPS score and select the top 2,500 "Nature" and 4,000 "Body" samples to mirror our T2I-ConBench dataset sizes.
- **Findings in Table A8.** We compare continual post-training performance under Joint, SeqFT, and Replay on both T2I-ConBench and filtered Flickr30K:
  - On T2I-ConBench, Joint & Replay significantly outperform SeqFT on task performance (Body-HPS, Nature-HPS) while maintaining pretrained capabilities (FID, Comp). In particular, Replay remarkably mitigates forgetting relative to SeqFT (Δ Domain-Forget –117.85%), aligning with expectations.

Table A7: Comparison of pretraining metrics (FID, CMMD, Comp and CLIPScore) on the sequential domain-enhancement task ("Nature" → "Body") using PixArt-α. Experimental settings match those in **Tab.** 1. Values in parentheses indicate the method's rank for that metric. We also report Δ ranks between FID & CMMD and Comp & CLIPScore.

Method	FID↓(rank)	CMMD↓(rank)	Δ rank	Comp↑(rank)	CLIPScore↑(rank)	Δ rank
Base	26.3153 (2)	1.1388 (4)	2	0.3378 (11)	0.3135 (9)	2
Joint	29.0167 (8)	1.1466 (9)	1	0.3325 (12)	0.3141 (4)	8
SeqFT	29.9746 (12)	1.1592 (12)	0	0.3382 (10)	0.3132 (12)	2
SeqLoRA	28.4885 (6)	1.1452 (8)	2	0.3433 (4)	0.3138 (7)	3
IncLoRA	28.2885 (5)	1.1448 (7)	2	0.3519 (1)	0.3138 (6)	5
O-LoRA	26.5287 (3)	1.1377 (3)	0	0.3411 (8)	0.3142 (3)	5
C-LoRA	26.1921 (1)	1.1233 (1)	0	0.3414 (7)	0.3147 (1)	6
ℓ_2-norm	27.1267 (4)	1.1272 (2)	2	0.3426 (5)	0.3144 (2)	3
EWC	29.7816 (10)	1.1496 (10)	0	0.3409 (9)	0.3133 (11)	2
HFT	28.8833 (7)	1.1442 (6)	1	0.3438 (3)	0.3139 (5)	2
MoFO	29.8326 (11)	1.1539 (11)	0	0.3418 (6)	0.3135 (10)	4
Replay	29.7044 (9)	1.1438 (5)	4	0.3508 (2)	0.3136 (8)	6
Avg Δ rank	–	–	1.1667	–	–	4.0000

  - On filtered Flickr, however, method rankings become inconsistent and counter-intuitive. Joint underperforms SeqFT on downstream metrics, and Replay even exhibits worse forgetting (Δ Domain-Forget +78.8%) than SeqFT, which is unexpected and confusing.
- **Analysis.** The inconsistent Flickr results stem from its inherent dataset limitations rather than experimental settings, as evidenced by T2I-ConBench yielding expected outcomes under identical conditions. We attribute Flickr's unreliability to uncontrolled diversity and ambiguous concept distribution. By contrast, T2I-ConBench eliminates these confounds through rigorous quality control and precisely scoped targets, enabling clear differentiation among continual learning approaches.
- **Cross-task Generalization.** Furthermore, T2I-ConBench also supports investigating the novel capability of cross-task generalization, which is one of our key contributions.
- **Conclusion.** While approximate benchmarks might be constructed from Flickr or other corpora through extensive retrieval, rigorous scoring, and labor-intensive filtering, we chose a synthetic, contamination-free pipeline that yields higher-quality, fully controlled data. Therefore, our carefully designed curation process is essential for studying T2I continual post-training and cannot be replicated by straightforward retrieval from existing collections.

Table A8: Comparison of methods on the sequential Domain Enhancement task ("Nature" → "Body") using PixArt-α trained on either T2I-ConBench or Flickr dataset. "Δmetrics for Method$_A$ vs. Method$_B$" is defined as $\Delta\text{metric} = \frac{\text{metric}_A - \text{metric}_B}{\text{metric}_B} \times 100$. All other settings match **Tab.** 1. *Italicized results* indicate counter-intuitive and perplexing outcomes.

Method	Data Source	Δ FID ↓	Δ Comp ↑	Δ Body-HPS ↑	Δ Nature-HPS ↑	Δ Domain-Forget ↓
Joint vs. SeqFT	T2I-ConBench	-3.19	-1.68	3.16	3.82	/
Replay vs. SeqFT	T2I-ConBench	-0.90	3.72	2.31	5.32	-117.85
Joint vs. SeqFT	Flickr	-2.02	1.25	*-1.29*	*-1.66*	/
Replay vs. SeqFT	Flickr	3.92	-0.48	*-0.19*	0.22	*78.80*

N Impact of Dataset Scale

We compare fine-tuning effects using datasets of different scales (5k, 10k, 20k) derived from Flickr30K via multimodal retrieval and HPS score ranking. Table A9 shows that expanding the filtered Flickr "Body" dataset from 5k to 20k samples does not significantly increase performance differences (all changes <1%).

This finding aligns with literature: DreamBooth (Ruiz et al., 2023) achieves effective per-item customization with fewer than 10 images per concept, and Emu (Dai et al., 2023) demonstrates

significant gains with thousands of high-quality samples. High-quality, smaller datasets are often more effective for task-specific post-training than large, diverse collections. Consequently, using several thousand high-quality samples is common practice in real-world post-training.

Our synthetic data pipeline can generate a larger initial pool (80k, as stated in Line 133 of the main text) than Flickr30K. The final curated "Body" dataset in T2I-ConBench contains 4k samples, each high-quality, high-aesthetic, and precisely aligned with text prompts. This set is sufficient to differentiate performance between methods.

Table A9: Performance comparison of PixArt-α fine-tuned on different scales of Flickr data for the "Body" task. "Data Source" labels (5k, 10k, 20k) denote datasets filtered from Flickr30K for "Body" relevance and sorted by descending HPS score. Δmetrics represents the relative percentage change: $\Delta\text{metrics} = \frac{\text{metrics}_A - \text{metrics}_B}{\text{metrics}_B} \times 100$, where A and B indicate the compared dataset scales. All other experimental settings match in **Tab.** 1.

Data Source	Δ FID ↓	Δ Comp ↑	Δ Body-HPS ↑
10k vs. 5k	0.57	-0.70	0.68
20k vs. 10k	-0.06	0.98	0.45

O Extended Scaling Experiments and Benchmark

We appreciate the concern regarding the scalability of our benchmark. A central design goal of T2I-ConBench is to serve as a *challenging yet tractable* first benchmark for continual post-training of text-to-image models. To explicitly assess how our conclusions change with a longer training horizon, we extend the item-customization track from the default 4-item sequence to an 8-item sequence by appending four additional *DreamBooth* objects. Concretely, we consider the sequence"V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker" → "V5 toy" → "V6 backpack" → "V7 sunglasses" → "V8 teapot". The overall performance of the continual post-training baselines on this 8-item sequence is reported in **Tab.** A10, while **Tab.** A11 and A12 provide per-task *Unique-Sim* and *Unique-Forget* statistics for MoFO and Replay, respectively. Across this extended sequence, Replay remains the only method that preserves a non-trivial amount of information about earlier items, indicating that it is the only approach that can still generate a reasonable number of faithful, instance-specific images for early items after all 8 items have been learned. These results confirm that replay-based training is currently the only strategy that meaningfully mitigates catastrophic forgetting in this setting, whereas all parameter-efficient or regularization-based approaches remain far from solving the problem. Importantly, the trends observed in this 8-item experiment closely mirror those already present in our default 4-item, 2-domain configuration. The shorter sequence is therefore *already sufficient* to expose severe forgetting in existing continual post-training methods. Scaling the benchmark to more items amplifies the failure modes (driving *Unique-Sim* closer to zero and *Unique-Forget* closer to one for most baselines), but does not qualitatively change our conclusions: current continual post-training techniques for text-to-image models are fundamentally inadequate in preserving previously learned item customizations.

Table A10: Performance of continual post-training methods on the *8-item* sequential item customization sequence. ("V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker" → "V5 toy" → "V6 backpack" → "V7 sunglasses" → "V8 teapot") The setting is the same as **Tab.** 1.

Method	FID↓	Comp↑	Unique-Sim↑	I+I↑	Class-Sim↓	Unique-Forget↓
SeqFT	18.5235	0.3387	0.0638	0.1623	0.0424	0.9825
SeqLoRA	24.3134	0.3519	0.0150	0.1294	0.0052	0.9752
MoFO	18.1822	0.3328	0.0950	0.1723	0.0612	0.9621
Replay	21.0281	0.3313	0.1825	0.2147	0.0915	0.7516

Table A11: Detailed Unique-Sim (similarity of generated item to the target item) and Unique-Forget on each task in the *8-item* sequential item-customization using **MoFO**. ("V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker" → "V5 toy" → "V6 backpack" → "V7 sunglasses" → "V8 teapot") The setting is the same as **Tab.** 1.

Task	**Unique-Sim measured after Task Vj is completed**								**Unique-Forget**
MoFO	V1 dog	V2 dog	V3 cat	V4 sneaker	V5 toy	V6 backpack	V7 sunglasses	V8 teapot	
V_1 dog	0.94	0.54	0.18	0	0	0	0	0	1.0000
V_2 dog	-	0.87	0.23	0.16	0.02	0	0	0	1.0000
V_3 cat	-	-	0.78	0.23	0.09	0	0	0	1.0000
V_4 sneaker	-	-	-	0.75	0.42	0.23	0.03	0	1.0000
V_5 toy	-	-	-	-	0.81	0.43	0.06	0	1.0000
V_6 backpack	-	-	-	-	-	0.65	0.24	0.03	0.9538
V_7 sunglasses	-	-	-	-	-	-	0.73	0.16	0.7808
V_8 teapot	-	-	-	-	-	-	-	0.57	-
Average	0.9400	0.7050	0.3967	0.2850	0.2680	0.2183	0.1514	0.0950	0.9621

Table A12: Detailed Unique-Sim (similarity of generated item to the target item) and Unique-Forget on each task in the *8-item* sequential item-customization using **Replay**. ("V1 dog" → "V2 dog" → "V3 cat" → "V4 sneaker" → "V5 toy" → "V6 backpack" → "V7 sunglasses" → "V8 teapot") The setting is the same as **Tab.** 1.

Task	**Unique-Sim measured after Task Vj is completed**								**Unique-Forget**
Replay	V1 dog	V2 dog	V3 cat	V4 sneaker	V5 toy	V6 backpack	V7 sunglasses	V8 teapot	
V_1 dog	0.88	0.32	0.16	0.16	0.12	0.10	0.10	0.10	0.8864
V_2 dog	-	0.73	0.42	0.15	0.23	0.23	0.20	0.18	0.7534
V_3 cat	-	-	0.54	0.24	0.12	0.23	0.15	0.15	0.7222
V_4 sneaker	-	-	-	0.53	0.32	0.30	0.25	0.22	0.5849
V_5 toy	-	-	-	-	0.63	0.23	0.18	0.13	0.7937
V_6 backpack	-	-	-	-	-	0.45	0.25	0.05	0.8889
V_7 sunglasses	-	-	-	-	-	-	0.57	0.21	0.6316
V_8 teapot	-	-	-	-	-	-	-	0.42	-
Average	0.9400	0.6330	0.3867	0.2700	0.2840	0.2567	0.2429	0.1825	0.7516

P Validation of Automated Metrics with Human Evaluation

To ensure that the automated metrics used in T2I-ConBench faithfully reflect human judgments, we conduct a series of comparative studies between model-based scores and human annotations. Specifically, we validate (i) HPS-based model ranking, (ii) the Unique-Sim and Class-Sim VQA metrics, and (iii) cross-task generalization VQA scores. In all cases, we observe that the automated metrics are highly consistent with human evaluations, supporting the reliability of our automatic evaluation pipeline.

HPS-based model ranking. We first examine whether HPS preserves the same relative ordering of methods as human annotators. We focus on the `body` and `nature` domains and compare the rankings of six baselines (Base, Joint, SeqFT, ℓ_2-norm, HFT, Replay) derived from HPS and from human preference studies. As shown in **Tab.** A13, for the `body` domain, the HPS-based ranking exactly matches the human ranking across all models. For the `nature` domain, the rankings are also nearly identical, with only a minor swap between Replay and Joint, indicating that HPS is a strong proxy for human judgments in our setting.

Validation of Unique-Sim and Class-Sim. We further validate the Unique-Sim and Class-Sim VQA metrics by comparing scores produced by an automatic VQA model (Qwen2.5-VL-7B-Instruct) with scores derived from human annotations under different evaluation orders. Concretely, we consider (i) the *Items*-only configuration and (ii) two mixed *Item-Domain* orders (Order1 and Order2), and compute Unique-Sim and Class-Sim scores for both the model and humans. As reported in **Tab.** A14, the model-based scores are very close to the human scores across all configurations, with only small deviations. This indicates that both Unique-Sim and Class-Sim reliably capture the same preference patterns as human evaluators.

Table A13: Comparison of model rankings based on HPS and human evaluation for the Body and Nature domains. Lower rank is better (1 = best).

Model	Body-HPS	Body-Human	Nature-HPS	Nature-Human
Base	6	6	6	6
Joint	1	1	2	1
SeqFT	5	5	5	5
ℓ_2-norm	4	4	3	3
HFT	2	2	4	4
Replay	3	3	1	2

Table A14: Comparison of Unique-Sim and Class-Sim VQA metrics between the automatic VQA model (Qwen2.5-VL-7B-Instruct) and human evaluation under different evaluation orders.

Order	Unique-Sim			Class-Sim		
	Items	Item-Domain Order1	Item-Domain Order2	Items	Item-Domain Order1	Item-Domain Order2
Qwen2.5-VL-7B-Instruct	0.2325	0.0225	0.2325	0.0633	0.0118	0.0953
Human	0.2414	0.0175	0.2765	0.0647	0.0085	0.0973

Cross-task generalization. Finally, we validate the automated VQA scores used for cross-task generalization. We consider multiple transfer settings between item customization (I) and domain enhancement (D), including I→I, D→D, and mixed I→D configurations, and apply two different item-domain orderings (Order1 and Order2). **Tab.** A15 compares the VQA accuracy of Qwen2.5-VL-7B-Instruct with human judgments across these settings. The distributions of model and human scores are highly similar for all item/domain combinations and orders, suggesting that the automated VQA evaluation remains well aligned with human preferences even under challenging cross-task generalization scenarios.

Discussion. Across HPS-based model ranking, Unique-Sim and Class-Sim VQA metrics, and cross-task generalization, the automated metrics consistently track human preferences, both in relative ordering and in absolute score trends. These results confirm that our automated evaluation framework provides a reliable and scalable approximation to human judgment for T2I-ConBench.

Q Seed Robustness and Statistical Stability

To assess the statistical stability of T2I-ConBench and verify that our conclusions are not artifacts of a particular random initialization, we evaluate the seed robustness of two representative continual learning methods, SeqFT and MoFO, on the item customization sequence. Concretely, we train each method with three different random seeds and report the resulting performance across all core metrics: FID, Comp, Unique-Sim, I+I, Class-Sim, and Unique-Forget. The detailed results are summarized in **Tab.** A16. For both SeqFT and MoFO, the variance across three seeds is very small on all metrics, with standard deviations typically on the order of 10^{-3} for Comp, Unique-Sim, I+I, Class-Sim, and Unique-Forget, and around 10^{-3}–10^{-2} for FID. Importantly, the relative ordering between SeqFT and MoFO is consistent across all seeds: MoFO consistently achieves lower FID and Unique-Forget and higher Unique-Sim and I+I than SeqFT, while Comp and Class-Sim remain comparable. These observations indicate that (i) training on T2I-ConBench is stable across random seeds and (ii) the comparative conclusions between continual learning methods are robust to stochasticity in optimization. Overall, these experiments demonstrate that T2I-ConBench yields statistically stable results across random seeds, and that the benchmark supports reliable comparison between different continual learning baselines.

Table A15: Comparison of cross-task generalization VQA accuracy (%) between Qwen2.5-VL-7B-Instruct and human evaluation in different item/domain transfer settings and item-domain orders. I: item customization, D: domain enhancement.

Order	Items	Domains	Item-Domain Order1			Item-Domain Order2		
	I+I	D+D	D+D	I+I	I+D	D+D	I+I	I+D
Qwen2.5-VL-7B-Instruct	32.22	38.81	39.80	26.67	37.96	37.81	28.33	44.31
Human	31.95	38.81	39.80	27.77	37.23	39.30	29.17	45.52

Table A16: Seed robustness of SeqFT and MoFO on the item customization sequence in T2I-ConBench. We report results over three random seeds as well as mean ± standard deviation. Lower is better for FID, Class-Sim, and Unique-Forget; higher is better for Comp, Unique-Sim, and I+I.

Method	FID ↓	Comp ↑	Unique-Sim ↑	I+I ↑	Class-Sim ↓	Unique-Forget ↓
SeqFT (seed1)	19.7847	0.3319	0.2325	0.3222	0.0633	0.8718
SeqFT (seed2)	19.7851	0.3301	0.2350	0.3215	0.0615	0.8623
SeqFT (seed3)	19.7723	0.3304	0.2325	0.3394	0.0694	0.8641
SeqFT (Avg $\pm\sigma$)	19.7807 ± 0.0059	0.3308 ± 0.0008	0.2333 ± 0.0012	0.3277 ± 0.0083	0.0647 ± 0.0033	0.8661 ± 0.0041
MoFO (seed1)	19.2802	0.3296	0.2850	0.3306	0.0680	0.7296
MoFO (seed2)	19.2816	0.3306	0.2875	0.3342	0.0710	0.7316
MoFO (seed3)	19.2861	0.3309	0.2875	0.3316	0.0520	0.7281
MoFO (Avg $\pm\sigma$)	19.2826 ± 0.0025	0.3304 ± 0.0006	0.2867 ± 0.0012	0.3321 ± 0.0015	0.0637 ± 0.0082	0.7298 ± 0.0014

R Robustness of the VLM-based VQA Evaluation Pipeline

T2I-ConBench relies on a VLM-based VQA pipeline to automatically assess compositional correctness and class-level consistency for both item customization and domain enhancement tasks. To ensure that this pipeline is both semantically reliable and robust to the choice of VLM, we take two complementary steps. To directly test the robustness of our VQA evaluation to the choice of VLM, we instantiate the pipeline with three different VLMs: SmolVLM2 Marafioti et al. (2025), InternVL3.5-8B Wang et al. (2025), and Qwen2.5-VL-7B-Instruct. For each VLM, we re-compute all VQA-based metrics on the SeqFT baseline, including Unique-Sim, Class-Sim, and the cross-task generalization accuracies for item customization and domain enhancement. **Tab.** A17 reports Unique-Sim and Class-Sim under three evaluation orders. Although the absolute values differ moderately across VLMs, the scores remain in a similar range and preserve consistent trends across orders, indicating that the underlying preference patterns are stable with respect to the choice of VLM. We further examine cross-task generalization by evaluating VQA accuracy for transfers between item customization and domain enhancement. **Tab.** A18 summarizes the results for SeqFT across the three VLMs. Despite using different VLM backbones, the resulting accuracies are comparable in magnitude and exhibit consistent relative behavior across transfer settings and orders. Overall, the manual verification of prompt decompositions and the multi-VLM evaluation collectively demonstrate that our VQA pipeline is robust: the automated metrics remain stable across different VLMs, and the qualitative conclusions about SeqFT and other baselines do not depend on a particular evaluator. At the same time, the pipeline is *modular* and can seamlessly incorporate future, stronger VLMs, making T2I-ConBench easily upgradable on the evaluation side.

S Computational Requirements and Resource Profile

To facilitate reproducibility and help practitioners estimate the resources required to run T2I-ConBench, we report the computational profile of both training and evaluation. Unless otherwise specified, all experiments are conducted on a single machine equipped with two RTX 4090 GPUs. Under this configuration, the full training and evaluation pipeline for all continual learning baselines and metrics is feasible within a reasonable budget on commodity research hardware. **Tab.** A19 summarizes the GPU hours required to train different continual learning baselines on the three sequences in T2I-ConBench. We report the wall-clock GPU hours summed over both GPUs under our configuration. In addition to training, T2I-ConBench reports a rich set of metrics. **Tab.** A20 reports

Table A17: Unique-Sim and Class-Sim of the SeqFT baseline evaluated with different VLMs under three evaluation orders: Items-only and two Item-Domain orders (Order1 and Order2).

VLM	Unique-Sim			Class-Sim		
	Items	Item-Domain Order1	Item-Domain Order2	Items	Item-Domain Order1	Item-Domain Order2
SmolVLM2	0.2965	0.125	0.1075	0.2875	0.110	0.0275
InternVL3.5-8B	0.3325	0.5075	0.0650	0.1945	0.135	0.0350
Qwen2.5-VL-7B-Instruct	0.2325	0.2325	0.0225	0.0633	0.0953	0.0118

Table A18: Cross-task generalization of the SeqFT baseline evaluated with different VLMs. We report VQA accuracy for item customization (I) and domain enhancement (D) under items-only, domains-only, and two Item-Domain orders (Order1 and Order2).

VLM	Items / Domains		Item-Domain Order1			Item-Domain Order2		
	I+I	D+D	D+D	I+I	I+D	D+D	I+I	I+D
SmolVLM2	0.2139	0.1791	0.2139	0.2250	0.3101	0.2289	0.2472	0.2959
InternVL3.5-8B	0.3611	0.4378	0.4229	0.3111	0.3011	0.4179	0.3306	0.4926
Qwen2.5-VL-7B-Instruct	0.3222	0.3881	0.3980	0.2667	0.3796	0.3781	0.2833	0.4431

the GPU minutes required for (i) generating images for evaluation (*Inference*) and (ii) computing each metric (*Evaluation*) under the same two-RTX-4090 configuration. This breakdown shows that, while T2I-ConBench provides a comprehensive evaluation suite, the incremental cost of computing individual metrics beyond image generation remains manageable. Overall, by reporting both training and evaluation costs, T2I-ConBench provides a transparent view of its computational requirements and offers a concrete reference for adapting the benchmark to different hardware setups.

T UNIQUE-FORGET METRIC AND "N/A" ENTRIES.

In **Tab.** 1 we report the *Unique-Forget* metric for item customization. Here we clarify its definition and explain why some entries are marked as "N/A" for IncLoRA, O-LoRA, and C-LoRA. Unique-Forget is a backward-transfer style forgetting score defined on top of the Unique-Sim metric. Let $\text{Unique-Sim}_{t,i}$ denote the Unique-Sim of item i measured on the model after finishing task t, and let T be the index of the last task in the sequence. Let $\text{Unique-Forget}_{i,i}$ be the Unique-Sim of item i right after it is customized (i.e., after finishing task $t = i$). For each item i, we define

$$\text{Unique-Forget}_{i} = \frac{\text{Unique-Sim}_{i,i} - \text{Unique-Sim}_{T,i}}{\text{Unique-Sim}_{i,i}}. \tag{A6}$$

Intuitively, Unique-Forget_{i} measures the relative drop in Unique-Sim for item i between the time it is first learned and the end of the sequence, with higher values indicating more forgetting. For some LoRA-based methods, such as C-LoRA, the strong self-regularization constraints on the LoRA updates for new tasks can effectively preserve previous knowledge, but at the cost of severely restricting the learning of later item customization tasks. In this regime, the item-specific LoRA modules for later items fail to acquire meaningful item-specific information, leading to $\text{Unique-Sim}_{i,i} = 0$ for those items. **Tab.** A21 illustrates this behavior on the sequential item customization ("V1 dog" $\rightarrow$ "V2 dog" $\rightarrow$ "V3 cat" $\rightarrow$ "V4 sneaker") using C-LoRA. For the later item (V3 cat), the diagonal Unique-Sim values of the corresponding LoRA updates are zero, and the off-diagonal entries remain low, indicating that the new LoRA modules do not successfully capture the intended item-specific customization. In this degenerate case, the normalization term in Eq. equation A6 becomes zero, and the Unique-Forget score reduces to an indeterminate form $0/0$, which is mathematically undefined. We therefore mark the corresponding entries as "N/A" in **Tab.** 1. Similar patterns are observed for IncLoRA and O-LoRA.

Table A19: GPU hours required to train different continual learning baselines in T2I-ConBench under our two-RTX-4090 setup. **Items**: item customization sequence; **Domains**: domain enhancement sequence; **Item-Domain**: mixed item-domain sequence.

	Joint	SeqFT	SeqLoRA	IncLoRA	O-LoRA	C-LoRA	ℓ_2-norm	EWC	HFT	MoFO	Replay
Items	0.7	0.7	0.8	0.8	0.8	0.8	0.7	0.8	0.6	1.8	0.9
Domains	80.4	89.2	94.0	94.8	93.6	94.8	88.4	98.0	75.6	212.8	108.8
Item-Domain	82.3	90.6	95.2	96.1	95.3	95.3	90.1	103.2	80.5	214.5	110.9

Table A20: GPU minutes for inference (image generation) and metric computation in T2I-ConBench under our two-RTX-4090 setup.

	FID	Comp	Unique-Sim	Class-Sim	Body-HPS	Nature-HPS	I+I	D+D	I+D
Inference	702.3	5.1	0.7	0.7	5.3	4.1	2.3	2.2	2.3
Evaluation	3.6	10.2	5.4	5.4	5.6	5.2	5.6	5.1	10.2

Table A21: Unique-Sim and Unique-Forget for each task in the sequential item customization dog1 $\rightarrow$ dog2 $\rightarrow$ cat $\rightarrow$ sneaker using C-LoRA. A dash (–) indicates that the corresponding item has not yet been seen at that point in the sequence. For cat and sneaker, the diagonal Unique-Sim values are zero, making the Unique-Forget ratio in Eq. equation A6 undefined ("N/A").

Order	Unique-Sim				Unique-Forget
C-LoRA	V1 dog	V2 dog	V3 cat	V4 sneaker	
V1 dog	0.89	0.52	0.18	0.18	0.7978
V2 dog	–	0.67	0.56	0.56	0.1642
V3 cat	–	–	0.00	0.00	N/A
V4 sneaker	–	–	–	0.00	–
Average	0.8900	0.5950	0.2467	0.1850	N/A

Figure A3: Results on the Item+Item cross-task test set by the models of sequential item customization in **Tab.** 1. Prompts of each column: (1)A photo of V3 cat playing with V4 sneaker in a sunlit meadow; (2)A photo of V1 dog and V2 dog relaxing near a crackling fireplace in a log cabin; (3)A photo of V1 dog and V3 cat sitting together on a cobblestone street in an old town; (4)A depiction of V1 dog sitting with V4 sneaker under a cherry blossom tree in full bloom.

Figure A4: Results on the Domain+Domain cross-task test set by the models of sequential domain enhancement in **Tab.** 1. Prompts of each column: (1)A little boy joyfully watches as a Spix's macaw mimics his whistling sounds; (2)A little girl struggles to peel a pomelo, her face lighting up as she finally separates the segments; (3)A park ranger gently feeds a quokka a small piece of fruit; (4)A girl sketches a gerenuk in her wildlife observation journal.

Figure A5: Results on the cross-task test set by the models of sequential item-domain adaptation Order 1 in **Tab.** 2. Prompts of each column: (1)Item+Item: A scene of V3 cat playing with V4 sneaker in a city park on a bright summer day; (2)Item+Body: A little boy makes a fist and shakes it playfully at a mischievous V3 cat that has just knocked over; (3)Item+Nature: A depiction of a Spix's Macaw and V2 dog relaxing on a balcony overlooking a modern cityscape; (4)Domain+Domain: An elderly man on his porch talks to his pet Spix's macaw, which responds with cheerful squawks.

Figure A6: Results on the cross-task test set by the models of sequential item-domain adaptation Order 2 in **Tab.** 2. Prompts of each column: (1)Item+Item: A scene of V3 cat playing with V4 sneaker in a city park on a bright summer day; (2)Item+Body: A little boy makes a fist and shakes it playfully at a mischievous V3 cat that has just knocked over; (3)Item+Nature: A photo of a Spix's Macaw and V3 cat perched together on a cliff overlooking the ocean; (4)Domain+Domain: An elderly man on his porch talks to his pet Spix's macaw, which responds with cheerful squawks.