Locating What You Need: Towards Adapting Diffusion Models to OOD Concepts In-the-Wild

Jianan Yang[1,2] **Chenchao Gao**[4,2] **Zhiqing Xiao**[1,2] **Junbo Zhao**[1,2] **Sai Wu**[1,2]
Gang Chen[1,2], **Haobo Wang**[3,2*]
[1]College of Computer Science and Technology, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]School of Software Technology, Zhejiang University
[4]International School of Information Science and Engineering, Dalian University of Technology
{jianan0115,zhiqing.xiao,j.zhao,wusai,cg,wanghaobo}@zju.edu.cn
gccgyllyp@gmail.com

Abstract

The recent large-scale text-to-image generative models have attained unprecedented performance, while people established *adaptor* modules like LoRA and DreamBooth to extend this performance to even more unseen concept tokens. However, we empirically find that this workflow often fails to accurately depict the *out-of-distribution* concepts. This failure is highly related to the low quality of training data. To resolve this, we present a framework called **C**ontrollable **A**daptor **T**owards **O**ut-of-**D**istribution Concepts (CATOD). Our framework follows the active learning paradigm which includes high-quality data accumulation and adaptor training, enabling a finer-grained enhancement of generative results. The *aesthetics* score and *concept-matching* score are two major factors that impact the quality of synthetic results. One key component of CATOD is the weighted scoring system that automatically balances between these two scores and we also offer comprehensive theoretical analysis for this point. Then, it determines how to select data and schedule the adaptor training based on this scoring system. The extensive results show that CATOD significantly outperforms the prior approaches with an 11.10 boost on the CLIP score and a 33.08% decrease on the CMMD metric.

1 Introduction

The generative modeling for text-to-image has attained unprecedented performance most recently [37, 45, 41, 49]. Notably, by training over billions of text-image data pairs [52, 51], the family of diffusion models has allowed high-fidelity image synthesis directed by the *prompt* provided in production. Despite their massive successes, these models still fail to generate images with decent quality and matching semantics, when encountering prompts that contain unseen or *out-of-distribution* concept tokens [60, 12, 21]. Simply put, the reason causing this failure is due to that the training set is **not** unbounded with limited variations. On the side of production, this limitation could significantly impact the practicality of this technique in real-world applications.

To deal with such concepts, recent works have resorted to *adaptors* such as Textual Inversion [18], DreamBooth [47, 58, 48], and LoRA [25, 66, 76], which tunes only a small part of the text-to-image model or insert extra modules. These adaptors largely reduce the training costs, and more importantly, preserve the visual aesthetic information originally learned by the underlying model.

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 1: **Comparison of images generated before/after training adaptors over concepts with different CMMD scores.** One observation is that concepts with higher CMMD scores are notably more challenging for the underlying model to generate (the second row). Additionally, we also notice that a higher CMMD value leads to a more notable loss of visual details when training adaptors (the third row).

However, in this paper, we have found that recent works still struggle to accurately depict the visual details of *out-of-distribution* concepts (with a CMMD score above 3.5), as illustrated in Figure 1. Adapters like LoRA are able to accurately represent the shape and color of OOD concepts compared to the generative results before adaptor training, but they fall short when it comes to finer details such as texture, contours, and patterns. This issue arises from the fact that current studies predominantly focus on variations of in-distribution (ID) concepts (e.g., humans, dogs, and cats) while ignoring the out-of-distribution (OOD) ones. This failure motivates us to think about what makes the problem of distorted visual details happen when training adaptors.

In Figure 2, we observed that how an adaptor depicts OOD concepts can be significantly influenced by the quality of the training data. If the model is trained on samples containing disruptive objects, the resulting generative outputs are likely to reflect these disruptive elements. When the training data contains images with vague or very small instances of the OOD concept, the generative results may appear low-quality. In contrast, the high-quality data for adapting OOD concepts usually contains a single and clear object corresponding to the given concept, which is highly distinguishable from the background and other types of objects and helps produce accurate results with high-fidelity. However, manually picking such high-quality data requires much human labor and expertise, which may crucially limit the versatility of text-to-image models. Therefore, an effective paradigm to locate high-quality samples of OOD concepts is important for the practical shipping of this field.

To this end, we developed a framework called **C**ontrollable **A**daptor **T**owards **O**ut-of-**D**istribution Concepts (referred to as CATOD) which aims to identify high-quality samples to guide the adaptor training. This framework follows the Active Learning (AL) paradigm [54, 44], involving iteratively accumulating training data and updating the adaptor. The profound motivation of this approach is to comprehensively model the interaction between training data and the underlying text-to-image model. Specifically, CATOD includes two interconnected scores: the *aesthetic* score and the *concept-matching* score, following the observation that object clarity and uniqueness largely impact generative results, as illustrated in Figure 2. Based on this, we devised a weighted scoring system that adapts itself according to the adaptor to select high-quality data while also properly balancing the two scores. With the carefully selected high-quality data, we schedule the adaptor training based on the quality of generative results evaluated through this scoring system.

In summary, our contributions are as follows: (i)-We have identified the challenge of adapting text-to-image models to out-of-distribution (OOD) concepts, where recent studies often struggle to accurately depict them; (ii)-We have introduced a framework called CATOD that iteratively updates training data and the adaptor to generate OOD concepts precisely; (iii)-Our extensive experiments verified that CATOD achieves significant performance gain with up to 11.10 on the CLIP score and 33.08% on the CMMD metric; (iv)-We have also offered theoretical insights into the key factors: *aesthetics* and *concept-matching*, which contribute to the effectiveness of our method.

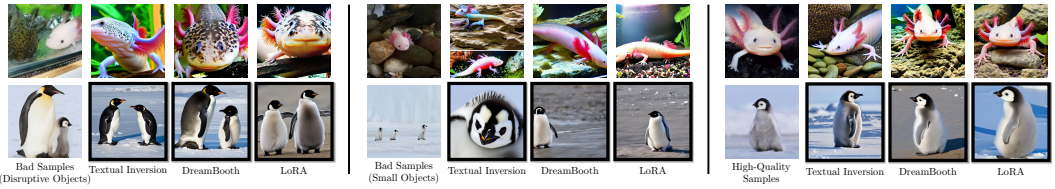


Figure 2: **Comparison of synthetic results on data with different quality.** Generated images are significantly influenced by the quality of training data. If the training data includes disruptive objects, the generative images may include disruptive visual details (*Left*). When an object within the image is too small, the results may not accurately represent the intended concepts (*Middle*). In contrast, if the image contains a high-fidelity object without disruptive elements (*Right*), the model is more likely to generate the desired result accurately.

2 OOD Problem in Latent Diffusion Models

Revisiting Latent Diffusion Models (LDMs). Latent Diffusion Models (LDMs) [45] comprise two components: a diffusion process operating the latent space and an auto-encoder which contains an encoder $\mathcal{E}$ mapping an image into the latent space and a decoder $\mathcal{D}$ that reconstruct images from latent codes. Furthermore, the diffusion process can be conditioned on the output of text embedding models, enabling the auto-encoder to integrate the information derived from texts. Let x be the image, the CLIP textual encoder c_θ that maps the corresponding text y into the latent space, the LDM loss is:

$$L_{LDM}(x, y) \coloneqq \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c_\theta(y)\right)\|_2^2\right], \tag{1}$$

where t denotes the time step, z_t denotes the latent code noised at time step t, when $\epsilon, \epsilon_\theta$ represents the noised samples and the denoising U-Net [46], respectively. Through this noising-denoising procedure applied to the latent codes, LDM enables the underlying model to integrate information derived from texts into the visual results, while also allowing more flexibility to produce images.

The OOD Concepts for LDMs. Intuitively, out-of-distribution (OOD) concepts refer to the category of data whose distribution deviates significantly from what the model has learned. This degree of drifting can be quantified by an MMD score [27], which evaluates the discrepancy between ground-truth images and the generative results in the image latent space. Formally, for two probability distributions P and Q, the MMD distance with respect to a positive definite kernel k is:

$$\text{dist}^2_{MMD}(P, Q) \coloneqq \mathbb{E}_{x_p\sim P, x'_p\sim P}\left[k(x, x')\right] + \mathbb{E}_{x_q\sim Q, x'_q\sim Q}\left[k(x_q, x'_q)\right] - 2\mathbb{E}_{x_p\sim P, x_q\sim Q}\left[k(x_p, x_q)\right], \tag{2}$$

where x_p, x'_p independently follow the distribution P while x_q, x'_q independently follow the distribution Q, with k the Gaussian RBF kernel [17]. In our implementation, we sample two sets of vectors from the distribution P and Q, then use CLIP embeddings [40] to calculate this score, which is also named as CMMD [27]. We set a concept with a high CMMD score (above 3.5) as an OOD concept.

OOD Concepts are Hard to Adapt. Recent text-to-image LDMs [45, 39] have achieved unprecedented performance on a wide range of concept tokens. However, we have found that there still exist many concepts that make LDMs fail after the adapter is fully trained. As we show in Figure 1, there are several discoveries: (i)-The concepts with higher CMMD scores are much more challenging for the underlying model to generate or adapt. The concepts with a CMMD score above 3.0 show explicit wrong visual details. For Axolotl and Frilled Lizards with CMMD above 4.0, the LDMs even generate the wrong species; (ii)-A higher CMMD score indicates a more severe loss of visual details when training adaptors. The concepts with CMMD scores above 3.5 in Axolotls and Emperor Penguin Chicks, show explicit distorted visual details, like color, texture, and delicate details. For Axolotl, the generative results show a wrong number and wrong positions of amateurs. For Emperor Penguin Chicks, the generative results show the wrong fur color of their heads and wings. To summarize, the higher the CMMD score of a concept, the more difficult it is for LDMs to adapt.

The High-Quality Matters. We further observe that the generative results are quite sensitive to training data when training adaptors meet OOD concepts. As shown in Figure 2, when the training images contain disruptive elements, the visual features of these disruptive elements will be easily introduced into the generative results. For example, when an adult emperor penguin appears in training data, then the black fur on their back can easily appear when generating their chicks, despite that their chicks have white fluffs as shown in the left part of Figure 2. If the object of the desired

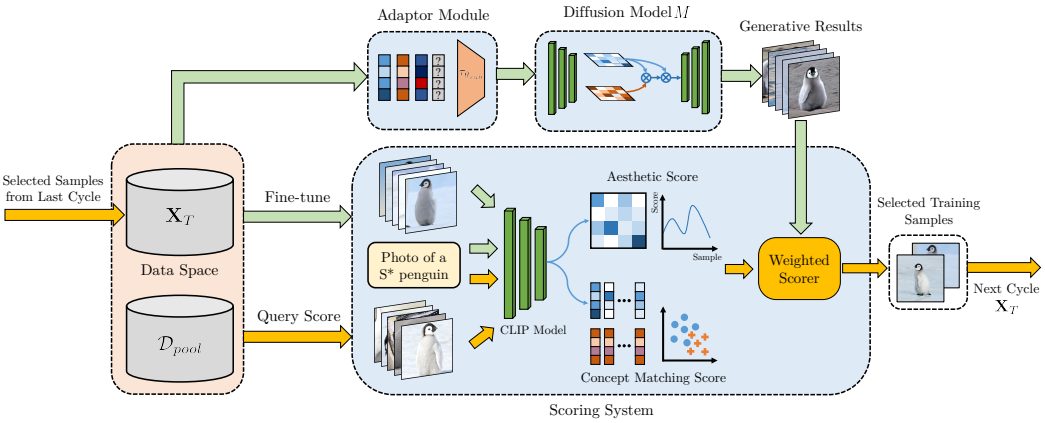


Figure 3: The overall pipeline of CATOD. In brief, CATOD alternatively performs data selection and scheduled OOD concept adaption. In each training cycle, we first generate OOD concepts according to the current adaptor and calculate the weights for the *aesthetic* score and *concept-matching* score. Then, we calculate the weighted score for each sample within the data pool $\mathcal{D}_{pool}$, select the top images accordingly, and add them to the training pool. At last, CATOD fine-tunes the scoring system and training adaptors according to the updated data pool, and proceed to the next cycle. The above three steps alternatively proceed until convergence.

concept appears to be too small within the image, then the generative results tend to be bad-looking, since the necessary visual details are not fully identified. In the middle part of Figure 2, we can see that distorted visual details like texture and shape in axolotl and emperor penguin chicks largely harm the aesthetics of the generated image. To generate images correctly and good-looking, we require enough amount of images with objects of high fidelity and do not contain disruptive elements, namely High-Quality samples as shown in the right part of Figure 2. Therefore, we aim to devise an effective data selection strategy for locating those high-quality images for these OOD concepts.

3 Method

3.1 Overall Architecture

We aim to adapt OOD concepts to LDMs correctly by an iterative data selection criterion that locates high-quality data and training adaptors accordingly as shown in Figure 3. Consequently, our method would alleviate issues introduced by disruptive elements, *e.g.*, irrelevant objects, and blurry images, while maintaining high fidelity of generative results. The core to CATOD is a scoring system, which consists of an aesthetic scorer and a concept-matching scorer, aiming to resolve the problems of incorrectly introduced visual details and distorted objects we have observed in Section 2. By properly trading off between these two scores, CATOD locates the most valuable samples for adapting underlying LDMs to OOD concepts and making the desired OOD concept depicted correctly.

3.2 The Scoring System

As described above, we need to estimate the potential impact of real-world samples over LDMs, according to which we select the most high-quality samples for training. In Section 2, we have observed two major factors that significantly impact generated image quality: object clarity and disruptive elements. Diving into the loss term $L_{LDM}(x, y)$ in Eq. (1), we may also observe that the training set $\mathbf{X}_T$ should be optimized towards both the underlying model (including the embedded adaptor) and the conditional text y, which indicates two important factors: object clarity and concept-matching achieved by preserving aesthetic information originally learned by LDMs and accurate image-text matching, respectively. Therefore, we attribute the quality of images to *aesthetic* and *concept-matching* and design two decoupled scorers accordingly.

The Aesthetic Scorer. Aesthetic evaluation is a long-standing field, which comprehensively considers whether the lighting, contrast, texture, and other photographic factors of an image are consistent with

human aesthetics. The general aesthetic scorer can be simply described as $p = S_{aes}(x)$, where S_{aes} indicates the aesthetic scoring model, p denotes the predicted score, and x represents the input image. Following the work of PA-IAA [33], we fine-tune the generic aesthetic model and make it adapt to OOD concepts with personalized preference scores. In general, we assign a high score to the samples within the training set $\mathbf{X}_T$, assign a low score to samples from irrelevant categories and samples generated by underlying LDMs without an adaptor, and use them to fine-tune the aesthetic scorer. More details of this personalization are given in the Appendix B.3.

The Concept-Matching Scorer. Intuitively, concept-matching describes whether the OOD concepts get perfectly reflected in the generated results. A similar task is image retrieval, which is designed to retrieve images containing objects describing the desired concepts. Since image retrieval also relies on feature maps and some sort of matching score to retrieve images, we adopt matching score from VLAD-related works [28, 64, 55] as follows:

$$S_{con}(x) = \frac{1}{|\mathbf{X}_T|} \sum_{k=1}^{|\mathbf{X}_T|} \phi\left(r_x, r_k\right) \exp\left(\|r_x - r_k\|\right), \tag{3}$$

where $\phi(r_x, r_k) = 1$ if r_k is the closet representation for r_x and is set to 0 otherwise. Simply put, VLAD regards the extracted representations for $\mathbf{X}_T$ as a codebook and maps each image to its nearest code. Note that samples within X_T mostly consist of clean and clear samples, this score is sufficient to quantify whether the object if exists in the given image distinguishes itself from other photographic components and matches the given concept.

3.3 Active Data Acquisition

Optimization with Active Learning (AL). Aiming to mitigate the problems caused by low-quality training samples, we proactively integrate the training data $\mathbf{X}_T$ into our objective as follows:

$$A^*, \mathbf{X}_T^* = \arg\min_{A, \mathbf{X}_T} \mathbb{E}_{x \sim \mathbf{X}_T} L_{LDM}(x, A, y). \tag{4}$$

Since the optimal set is initially unknown, a one-step optimization can easily lead to convergence to local optima. Therefore, we use an iterative paradigm to optimize adaptor A and training data $\mathbf{X}_T$, respectively. In our implementation, we adopt the paradigm of AL [44, 54] to perform the optimization of $\mathbf{X}_T^{(t)}$ by data accumulation before training adaptors, with t denotes the time step:

$$\mathcal{B}^{(t)} = \underset{\mathcal{B} \subset \mathcal{D}_{pool} - \mathbf{X}_T^{(t-1)}, |\mathcal{B}| = b}{\arg\min} \mathbb{E}_{x \sim \mathbf{X}_T^{(t-1)} \cup \mathcal{B}} L_{LDM}(x, A, y), \tag{5}$$

where b is the number of samples added to the training pool at each cycle. The main reason for using AL is its preferred sample efficiency, with better controllable data bias management [15, 50]. Instead of repeatedly selecting data from the whole real-world data pool, AL provides a more efficient procedure to optimize training data by using data selection to accumulate high-quality training data. To this stage, the learning procedure of CATOD is relatively clear: the sample pool is progressively accumulated -by Eq. (5) -and the optimization of the adaptor A is straightforward (Fig. 3).

Remember that AL involves iteratively updating the adaptor and the training data, it is important to design a training schedule for adaptors and determine how to acquire high quality based on the two scorers mentioned above. The primary objective of this design is to achieve a dynamic trade-off between the two scores. The specific details of these designs are described below.

The Active Schedule for Training Adaptors. The training schedule has been found crucial for successful adaption [47, 74, 65]. Since our training data continuously expands as the learning cycle of AL proceeds, the training schedule will be even more important. To arrange this schedule, we first calculate the aesthetics score $\gamma_{aes}(A) = \frac{1}{|g_A(\mathbf{X}_T)|} \sum_{x_g \in g_A(\mathbf{X}_T)} S_{aes}(x_g)$ and concept-matching score $\gamma_{con}(A) = \frac{1}{|g_A(\mathbf{X}_T)|} \sum_{x_g \in g_A(\mathbf{X}_T)} S_{con}(x_g)$ of the adaptor A based on its generative results $g_A(\mathbf{X}_T)$, in which both the aesthetic score $\gamma_{aes}(A)$ and $\gamma_{con}(A)$ range from 0 to 10. Then, we use a trigonometric indicator that comprehensively measures its performance:

$$\gamma(A) = 10 \sin\left(\frac{\pi}{20}\gamma_{aes}(A)\right) \sin\left(\frac{\pi}{20}\gamma_{con}(A)\right). \tag{6}$$

Notably, $\gamma(A)$ peaks when the two scores get close or around the common value of 5.0 for most samples. Meanwhile, it bottoms when the scoring exhibits a stronger signal of being biased (to either

side). The properties emphasize that adapters balancing the two factors are of better quality. In training CATOD, we use $\gamma(A)$ as a signal to reduce the learning rate and stop training in time.

Trading-off the Two Scores in Data Acquisition. After getting the adaptor A, we continue to select the most suitable samples for the next-cycle training. Notice that whether newly selected images enhance the adaptor depends on both aesthetics and concept-matching, we ought to adjust our preferences according to the adaptor. In more detail, when the adaptor has high aesthetics but does not accurately depict the OOD concept, samples with high concept-matching scores better enhance the adaptor; when the adaptor fails to exhibit photographic attributes consistent with humans, samples with high aesthetics are preferred. To implement this preference, our acquisition score is:

$$S(x) = \left(1 - \sin\left(\frac{\pi}{20}\gamma_{aes}(A)\right)\right) S_{aes}(x) + \left(1 - \sin\left(\frac{\pi}{20}\gamma_{con}(A)\right)\right) S_{con}(x). \tag{7}$$

This formulation offers some meritable advantages. On one hand, the balancing terms are not prefixed or manually tuned, but dynamically dependent on the scores marginalized over the current generations. Further, when the score of either side gets larger, the corresponding balancing coefficient, in turn, decreases, thus attaining a proper trade-off. As a result, this mechanism encourages an alternation of sample selection towards both scoring metrics by selecting Top-K samples to add to the training data $\mathbf{X}_T$ accordingly, which effectively guarantees the sample set diversity.

4 Theoretical Insights

This section presents our theoretical analyses of why aesthetic/concept-matching scores work in OOD Adaption. Specifically, we derive the distance between real-world data distributions and synthetic data distributions and then induce the important factors that affect this distance. We first introduce the *minimum mean square error* (MMSE) [22, 11, 6] to measure the discrepancy between distributions:

Definition 4.1. The *minimum mean square error (MMSE)* of estimating an input random vector $\widehat{\mathbf{X}} \in \mathbb{R}^n$ from an observation/output $\mathbf{X} \in \mathbb{R}^k$ is defined as

$$\mathrm{MMSE}(\widehat{\mathbf{X}}|\mathbf{X}) = \inf_{f \in \mathcal{M}(\mathbb{R}^n)} \mathbb{E}\left[\left\|\widehat{\mathbf{X}} - f(\mathbf{X})\right\|^2\right], \tag{8}$$

in which $\mathcal{M}(\mathbb{R}^n)$ denotes the space consisting of all measurable functions on $\mathbb{R}^n$.

Notice that we are trying to produce the best generation results which are initially unknown, our adaption task can be also regarded as estimating the optimal distribution with carefully selected data. By denoting the ideal generative results and the ideal adaptor with random variables $\widehat{\mathbf{X}}_G$ and A^*, respectively, we conclude that the LDM loss L_{LDM} is consistent with this MMSE term:

Theorem 4.2. *Let L_{LDM} be the LDM loss following Eq. 1, and the image space lies within $\mathbb{R}^n$. Then there always exists an ideal random vector $\widehat{\mathbf{X}}_G \in \mathbb{R}^n$ and an adaptor A^* for LDM, such that*

$$\arg\min_{\mathbf{X}_T \in \mathbb{R}^n} \mathrm{MMSE}(\widehat{\mathbf{X}}_G|\mathbf{X}_T) = \arg\min_{\mathbf{X}_T \in \mathbb{R}^n} \mathbb{E}_{x \sim \mathbf{X}_T}\left[L_{LDM}\left(x, A^*\right)\right]. \tag{9}$$

Based on the preceding deduction, we have confirmed that the change in MMSE can also indicate the change in LDMs. To dive deep into the MMSE term, we further decompose it as follows:

Theorem 4.3. *(Pythagorean Theorem for MMSE [13].) Following Theorem 4.2, by setting f to the generative model g_A, the MMSE term in Eq.* (8) *can be decomposed into two terms as follows:*

$$\mathbb{E}\left[\left\|\widehat{\mathbf{X}}_G - g_A(\mathbf{X}_T)\right\|\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{X}}_G - g_{A^*}(\mathbf{X}_T)\right\|\right] + \mathbb{E}\left[\|g_{A^*}(\mathbf{X}_T) - g_A(\mathbf{X}_T)\|\right], \tag{10}$$

in which A^* denotes a potential ideal adaptor containing information for text y. Notice that the generative results $\mathbf{X}_G$ relies on the training set $\mathbf{X}_T$, $g_A(\mathbf{X}_T)$ can also be written as $\mathbf{X}_G|\mathbf{X}_T$.

Upon revisiting the two terms, we can gain a deeper understanding of the relationship between MMSE and aesthetic/concept-matching score: (i)-The first term is focused on estimating the difference between the generative distribution $\mathbf{X}_G|\mathbf{X}_T$ and the ideal distribution of $\widehat{\mathbf{X}}_G$. This assessment helps

Table 1: A Comparison over the performance of CATOD, in terms of the CLIP score and CMMD score with 100 images sampled at last. This table shows the average result of 5 sub-classes within each category. The overall improvement of our proposed CATOD is provided by "Imp.". Methods with the best performance are bold-folded.

Comparison Methods	CLIP Score (↑)							CMMD [27] (↓)						
	insect	lizard	penguin	seafish	snake	Avg.	Imp.	insect	lizard	penguin	seafish	snake	Avg.	Imp.
DreamBooth[47] + RAND	65.35	67.89	68.07	65.59	72.07	67.79	⇑7.58	1.35	1.39	1.67	1.59	1.25	1.45	⇓0.50
DreamBooth[47] + CLIP	70.56	72.19	72.09	71.17	74.59	72.12	⇑3.25	1.04	1.24	1.05	1.38	1.16	1.17	⇓0.22
DreamBooth[47] + CATOD	**72.18**	**75.30**	**75.16**	**74.60**	**79.61**	**75.37**	-	**0.92**	**1.08**	**0.80**	**1.21**	**0.74**	**0.95**	-
TI[18] + RAND	58.24	59.34	63.45	61.35	62.34	60.94	⇑7.04	2.45	2.15	2.09	1.95	1.65	2.06	⇓0.63
TI[18] + CLIP	61.65	67.24	66.95	62.27	64.23	64.47	⇑3.51	2.11	1.86	1.44	1.87	1.37	1.73	⇓0.30
TI[18] + CATOD	**69.21**	**70.14**	**68.23**	**64.89**	**67.41**	**67.98**	-	**1.57**	**1.73**	**1.15**	**1.43**	**1.25**	**1.43**	-
LoRA[25] + RAND	64.39	65.02	67.49	68.87	71.09	67.37	⇑10.60	1.52	1.47	1.56	1.61	1.18	1.47	⇓0.63
LoRA[25] + CLIP	70.27	74.13	72.08	73.19	75.64	73.06	⇑4.91	1.29	1.33	1.04	1.35	0.89	1.18	⇓0.34
LoRA[25] + CATOD	**72.60**	**77.00**	**74.11**	**84.29**	**81.86**	**77.97**	-	**0.94**	**0.89**	**0.71**	**0.88**	**0.77**	**0.84**	-

us understand how the introduced samples lose visual information within the underlying LDM, thus we can connect this term to aesthetic preservation, i.e. aesthetic score; (ii)-The second term illustrates the degree to which the training set $\mathbf{X}_T$ distorts the OOD concept information within the ideal adaptor. Hence, a concept-matching score accurately portrays how newly given samples impact this term. At this stage, we have completed the theoretical support showing that both aesthetics and concept-matching are major factors that influence the performance of adaptors.

5 Experiments

In this section, we present the main experimental results both qualitatively and quantitatively. To evaluate our proposed CATOD, we combine it with several works for adaption, i.e. DreamBooth [47], Textual Inversion [18], and LoRA [25]. More experimental results can be found in the Appendix. *The source code is attached in the Supplementary.*

Datasets. We test our method on datasets with 25 OOD concepts that can hardly be generated through prompt engineering on the text-to-image model. This dataset consists of 5 categories: insect, lizard, penguin, seafish, and snake, and each category contains data from 5 OOD concepts. Each concept has 1,000 examples in total with 100 samples left out for validation. The dataset is collected from publicly available datasets including ImageNet, iNaturalist 2018 [67], IP102 [71].

Implementation Details. We conduct the active generation experiments on our proposed CATOD and three representative adaptors, i.e. DreamBooth [47], Textual Inversion [18] (termed as TI in the paper), and LoRA [25]. Since there are currently no available studies that focus on locating "high-quality" samples for training, we apply random sampling (RAND), and CLIP-score-based sampling (CLIP) for each baseline in our active learning setting. Each experiment starts with 20 randomly sampled instances, and we conducted 5 cycles of data accumulation in which we selected 20 "good" samples to add to the training pool. We train 20 epochs for all combinations of adaption techniques and sampling strategies in each active learning cycle, with a batch size of 1. Furthermore, we generate 100 images for each concept for evaluation. We use the commonly adopted Stable Diffusion 2.0 pre-trained on LAION-5B [51] following Rombach's work [45].

Evaluation Metrics. We evaluate the quality of our generated images with the widely used CLIP score [20] and the recently proposed CMMD score [27], which quantify the model performance in two aspects. Specifically, the CLIP score measures how generated images match the given text, which is expected to be as high as possible. Meanwhile, the CMMD score evaluates the discrepancy between generated images with the real ones, indicating better generative results with lower values.

5.1 Single-concept Generation Results

To evaluate the performance of CATOD on OOD concepts, we test it on all 25 target concepts one by one separately and report the average performance of 5 concepts within each category in Table 1. We show the superior results of our CATOD with respect to both qualitative and quantitative comparisons.

Qualitative Comparisons. We qualitatively compare CATOD with other sampling strategies according to their generated images based on LoRA [25]. With random sampling (RAND), we observe that the generative results only partially learned some photographic attributes like color and texture,

Table 2: A Comparison over the performance of CATOD when training with images from multiple concepts, in terms of the CLIP score and CMMD score with 100 images sampled at last. In each experiment, we sample images from all the sub-classes within each category and check whether the fine-tuned model can generate all 5 concepts. The overall improvement of our proposed CATOD is provided by "Imp.". Methods with the best performance are bold-folded.

Comparison Methods	CLIP Score (↑)							CMMD [27] (↓)						
	insect	lizard	penguin	seafish	snake	Avg.	Imp.	insect	lizard	penguin	seafish	snake	Avg.	Imp.
DreamBooth[47] + RAND	63.29	63.79	65.72	65.59	64.36	64.55	⇑8.20	1.74	2.14	2.16	2.02	1.85	1.98	⇓0.75
DreamBooth[47] + CLIP	67.57	70.35	71.10	69.34	70.38	69.75	⇑3.00	1.53	1.76	1.80	1.79	1.59	1.69	⇓0.46
DreamBooth[47] + CATOD	**70.83**	**72.28**	**74.31**	**70.90**	**75.45**	**72.75**	-	**1.39**	**1.25**	**1.34**	**1.45**	**0.73**	**1.23**	-
TI[18] + RAND	59.23	56.97	57.90	60.83	62.65	59.52	⇑5.88	2.84	2.56	2.27	2.39	2.41	2.49	⇓0.83
TI[18] + CLIP	61.74	60.72	63.79	62.71	65.71	62.93	⇑2.47	2.26	2.08	1.94	1.97	2.23	2.10	⇓0.44
TI[18] + CATOD	**64.48**	**63.93**	**65.93**	**65.44**	**67.24**	**65.40**	-	**1.76**	**1.53**	**1.65**	**1.64**	**1.73**	**1.66**	-
LoRA[25] + RAND	63.85	65.27	66.46	68.25	69.43	66.65	⇑7.64	1.79	2.05	1.93	1.85	1.55	1.83	⇓0.59
LoRA[25] + CLIP	69.25	70.84	71.39	71.91	72.78	71.23	⇑3.06	1.40	1.69	1.74	1.59	1.28	1.54	⇓0.30
LoRA[25] + CATOD	**71.19**	**74.09**	**73.68**	**75.60**	**76.90**	**74.29**	-	**1.13**	**1.37**	**1.49**	**1.26**	**0.95**	**1.24**	-

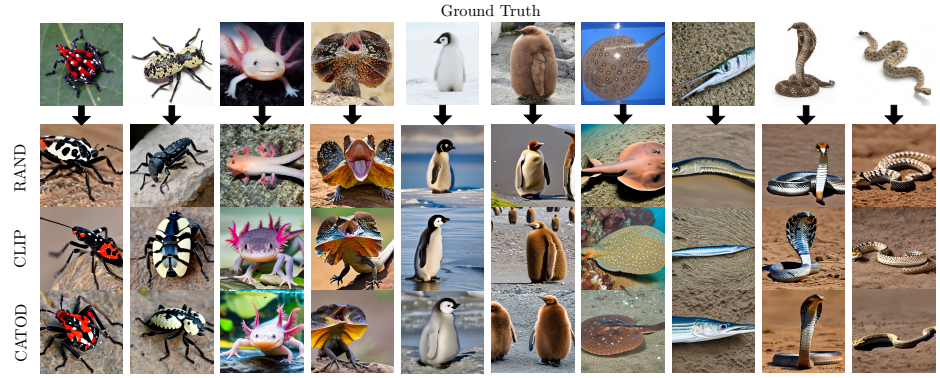


Figure 4: **A comparison of different sampling strategies with LoRA.** Specifically, we compare three lines of works: (1) RAND, in which the model is trained with 100 randomly selected samples; (2) with samples of the highest CLIP scores (100 samples); (3) 100 samples with CATOD. The model trained with randomly sampled data fails to capture the features of out-of-distribution (OOD) concepts, while the ones trained with top CLIP scores contain necessary details but also include disruptive elements.

but failed to make the objects have the correct appearance and shape. CLIP-based sampling (CLIP) somehow produces the corrected shape of the concept but still fails to capture the necessary details for describing the object. In comparison, our proposed CATOD successfully matches all the photographic attributes, while also guaranteeing the image aesthetics.

To further look at how CATOD learns the photographic attributes that precisely match the concept, we show samples generated from different cycles in Fig. 7. We can see that some attributes like color, and texture are already learned in cycle 1, but the shape of the object does not match the ground-truth ones. From cycle 1 to 3, CATOD shows a clear shape modification, making the objects more like the real ones. From cycles 3 to 5, an iterative refinement on more photographic details like light, contrast, and other minor modifications (like the beak for penguins and antenna for axolotls) is shown in generative results, making them hard to distinguish from the ground-truth ones even with a careful look. At cycle 5 and later cycles, the image quality stabilizes and we can hardly see enhancement apart from image diversity. To conclude, we can see an explicit attribute matching process from easy ones to the finer ones within CATOD, showing the importance and effectiveness of iterative training.

Quantitative Comparisons. Table 1 reports CLIP scores [20] and CMMD scores [27] from each strategy with models trained on 25 different OOD concepts and evaluated through the generative results. Specifically, we evaluate the performance of CATOD on each concept and average the results within each category in Table 1. We have the following conclusions: (1) the average performance of our strategies outperforms all compared methods by a 0.56∼11.10 CLIP score, justifying the effectiveness of locating aesthetic and clean samples over text-image matching. (2) CATOD also brings a 0.12∼0.54 CMMD decrease on various frameworks and concepts, indicating better image alignment with proper sampling guided by CATOD. To summarize, both image-matching and text-matching scores exhibit better results compared to the baselines, suggesting that our proposed CATOD significantly outperforms other strategies, which is consistent with the qualitative results in Fig. 4.

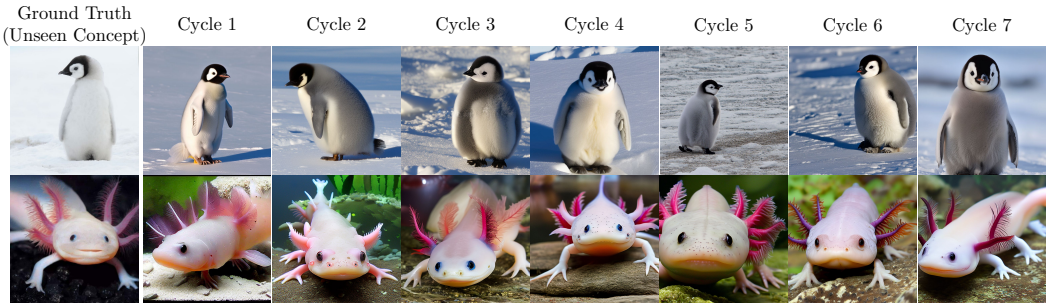


Figure 5: **Generative results as cycle proceeds**. Samples are generated with CATOD on cycles from 1 to 7. To better observe how generated images change as the cycle proceeds, we conduct another 2 cycles here. In each cycle, we select and add 20 high-quality samples. Generative samples start to converge and contain the right details within the original concept after cycle 4 or 5. We can also see that those generative results contain diverse contents within the background based on the few images given.

Table 3: **Results of Ablating Aesthetic, Concept-Matching Scorer and Weighted Scoring on CATOD.** We show the average results conducted on the categories "penguin" and "lizard" with LoRA.

Modules			CLIP(↑)		CMMD(↓)	
Aesthetic	Concept Matching	Weighted Scoring	lizard	penguin	lizard	penguin
✓			73.19	72.85	1.14	1.28
	✓		68.45	70.16	1.10	1.32
✓	✓		75.35	73.24	0.94	0.93
✓	✓	✓	**77.00**	**74.11**	**0.89**	**0.71**

Table 4: **Results of CATOD with different types of aesthetic scorers.** We show the average results conducted on the categories "penguin" and "lizard" with LoRA.

Type of Aesthetic Score	CLIP(↑)		CMMD(↓)	
	lizard	penguin	lizard	penguin
ReLIC [81]	71.35	72.39	1.29	1.59
TANet [23]	72.05	72.67	1.32	1.35
BAID [77]	73.08	72.79	1.18	1.37
Ours	**73.19**	**72.85**	**1.14**	**1.28**

5.2 Multi-concept adaption Results.

To further investigate whether CATOD could adapt multiple concepts simultaneously, we group the 25 concepts by category and train the adaptation model on each category. To be specific, we compare our baselines using a generated set consisting of 500 images (100 images per concept) and exhibit our results in Table 2. Following the setting of single-concept adaption, we conduct 10 cycles of data accumulation and get 200 samples at last, since multi-concept adaption requires more data. We still observe a notable performance gain with a 1.56∼5.07 CLIP score increase and a 0.12∼0.86 CMMD score decrease compared to baselines. These results verify that our CATOD still achieves better performance on multi-concept adaption.

5.3 Ablation Studies

We verify the efficacy of all components in our proposed CATOD in Table 3 and 4, including the aesthetical scoring module and the concept-matching mechanism within the weighted scoring system.

W/O Aesthetic Score. First, we validate the effectiveness of CATOD by removing the aesthetic scores. A significant decrease in performance (up to 8.55) on the CLIP score can be observed in Table 3. We attribute this decline to the fact that matching-based metrics prioritize image representations over the given concept, leading to a deterioration in image-text matching.

The Type of Aesthetic Scores. To further investigate the impact of aesthetic scores, we have also applied different types of aesthetics with CATOD in Table 4. We observed that recent state-of-the-art aesthetic evaluations did not improve OOD adaption and even led to minor performance loss. This could be attributed to the fact that these models were originally designed for general aesthetic assessments, whereas our aesthetic scorer is highly personalized towards specific OOD concepts.

W/O Concept-Matching Score. After removing the concept-matching scorer, we observed that the adaptor tends to perform better compared to just removing the aesthetic scorers. This might be due to the aesthetic scorer being designed based on CLIP backbones, showing some consistency with the CLIP score. However, it still demonstrates limitations based on the image-matching score CMMD. While these aesthetic qualities partly describe the clarity and accuracy of the object, they do not adequately focus on the image representation space.

The Weighted Score. We have confirmed the effectiveness of balancing two scores by simply adding them together to guide CATOD (Table 3, Line 5). We observed that this resulted in consistent performance loss for both the CLIP score (up to 1.65) and the CMMD score (up to 0.22). This suggests that both text-image matching and image-to-image matching are affected by the trade-off between aesthetics and concept matching, further emphasizing the importance of these two scores.

6 Related work

Personalized Text-to-image Synthesis with Adaptors. The task of text-to-image generation involves creating specific images based on text descriptions [3, 73, 79, 4], and has achieved impressive performance with state-of-the-art diffusion models [41, 45, 39]. Therefore, *adapting* large-scale text-to-image models to a specific concept while also preserving this amazing performance, i.e. *personalization* [7], has become another recent research interest. But this is often difficult since re-training a model with an expanded dataset for each new concept is prohibitively expensive while fine-tuning the whole model [12, 32] or transformation modules [83, 20, 60] on few examples typically leads to some rate of forgetting [30]. Therefore, *adaptors* such as Textual Inversion [18], DreamBooth [47, 48, 5, 58], LoRA [25, 66, 86, 76], along with some other works [70, 19] derived from them, have become more commonly adopted. Typically, they focus on a small but crucial part of the model or extra networks inserted into underlying models, thus more computationally-efficient, while also preserving the efficacy of the underlying models with lower computational costs. For example, *textual inversion* (TI) [18] represents the newly-given concept with pseudo word [42] and remapping it to another carefully trained embedding in the text-encoding space, guided by few images. Despite their computational efficiency, these approaches are still facing difficulties dealing with out-of-distribution concepts as we observe in Section 2.

Active Learning and Selection. Active Learning is a machine learning paradigm that involves actively selecting the most suitable data for training models from external data sources [44, 63]. The most crucial part of active learning is the strategy to locate the optimal data batch. Current studies can be roughly categorized as follows: (a) Score-based methods that prefer the samples with the highest information scores [36, 69, 10]; (b) Representation-based methods searching for the samples that are the most representative of the underlying data distribution [53, 1, 62]. The Active Selection paradigm within Active Learning serves as an efficient and powerful dataset curation tool (i.e. *adaption*), leading to numerous studies adopting this paradigm from a wide range of subjects [68, 24, 8]. Meanwhile, due to the effectiveness of Active Learning in selecting the most suitable training samples, recent studies have utilized similar sampling strategies to address challenges associated with long-tailed distributions [56, 57] and noisy data [72]. Given that Out-Of-Distribution (OOD) concepts in Latent Diffusion Models (LDMs) often involve unseen or long-tailed concept tokens [79], this motivates us to leverage Active Learning for selecting samples that are well-suited for training adaptors.

In this work, we focus on scored-based strategies in two-fold: *aesthetics* and *concept-matching*. Image Aesthetics Assessment (IAA) aims at evaluating image aesthetics computationally and automatically [9], while automatically assessing image aesthetics is useful for many applications [35, 29, 14]. To take a step further, personalized image aesthetics assessment (PIAA) [43, 85, 75] was proposed to capture unique aesthetic preferences, consistent with our goal to adjust our paradigms accordingly different concepts. At the same time, "Concept-matching" is adopted from the field of image retrieval, in which we search for relevant images in an image gallery by analyzing the visual content (e.g., objects, colors, textures, shapes *etc*.), given a query image [61, 31]. To design a paradigm that automatically meets the harsh requirements for adaptor training, we consider both two factors.

7 Conclusion

We propose CATOD, an enhanced, data-efficient, and practically useful version of the OOD concept adaptation for AIGC. This method is encapsulated in an active-learned paradigm with carefully designed acquisitional scoring mechanisms. CATOD significantly outperforms the prior approaches in many (if not most) aspects including generation quality, concept matches, technological robustness, data efficiency, etc. In the future, we hope to ship CATOD to the open-source community so as to absorb more OOD concepts that were originally uncovered.

Acknowledgements

This work is supported by the Pioneer R&D Program of Zhejiang (No. 2024C01035). This work is also partially supported by the Fundamental Research Funds for the Central Universities (226-2024-00049) and (226-2024-00145).

References

[1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[2] Song Bai, Xiang Bai, Qi Tian, and Longin Jan Latecki. Regularized diffusion process on bidirectional context for object retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1213–1226, 2018.

[3] Razan Bayoumi, Marco Alfonse, and Abdel-Badeeh M Salem. Text-to-image synthesis: A comparative study. In *Digital Transformation Technology: Proceedings of ITAF 2020*, pages 229–251. Springer, 2021.

[4] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

[5] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023.

[6] Edwin K. P. Chong. Well-conditioned linear minimum mean square error estimation. *IEEE Control. Syst. Lett.*, 6:2431–2436, 2022.

[7] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 558–577. Springer, 2022.

[8] Wupeng Deng, Quan Liu, Feifan Zhao, Duc Truong Pham, Jiwei Hu, Yongjing Wang, and Zude Zhou. Learning by doing: A dual-loop implementation architecture of deep active learning and human-machine collaboration for smart robot vision. *Robotics and Computer-Integrated Manufacturing*, 86:102673, 2024.

[9] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[10] Francesco Di Fiore, Michela Nardelli, and Laura Mainini. Active learning and bayesian optimization: a unified perspective to learn with a goal. *Archives of Computational Methods in Engineering*, pages 1–29, 2024.

[11] Mario Díaz, Peter Kairouz, and Lalitha Sankar. Lower bounds for the minimum mean-square error via neural network-based estimation. *CoRR*, abs/2108.12851, 2021.

[12] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022.

[13] Alex Dytso, H Vincent Poor, Ronit Bustin, and Shlomo Shamai. On the structure of the least favorable prior distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1081–1085. IEEE, 2018.

[14] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020.

[15] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

[16] Li Fe-Fei et al. A bayesian approach to unsupervised one-shot learning of object categories. In *proceedings ninth IEEE international conference on computer vision*, pages 1134–1141. IEEE, 2003.

[17] Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K Sriperumbudur. Characteristic kernels on groups and semigroups. *Advances in neural information processing systems*, 21, 2008.

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[19] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

[20] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[21] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[22] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Estimation of non-gaussian random variables in gaussian noise: Properties of the MMSE. In Frank R. Kschischang and En-Hui Yang, editors, *2008 IEEE International Symposium on Information Theory, ISIT 2008, Toronto, ON, Canada, July 6-11, 2008*, pages 1083–1087. IEEE, 2008.

[23] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceeding of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[24] Aral Hekimoglu, Michael Schmidt, and Alvaro Marcos-Ramiro. Monocular 3d object detection with lidar guided semi supervised active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2346–2355, 2024.

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[26] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

[27] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023.

[28] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3304–3311. IEEE Computer Society, 2010.

[29] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 662–679. Springer, 2016.

[30] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[31] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19, 2006.

[32] Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454, 2022.

[33] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910, 2020.

[34] Chundi Liu, Guangwei Yu, Maksims Volkovs, Cheng Chang, Himanshu Rai, Junwei Ma, and Satya Krishna Gorti. Guided similarity separation for image retrieval. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 386–399. Springer, 2008.

[36] Xiaoyi Mai, Salman Avestimehr, Antonio Ortega, and Mahdi Soltanolkotabi. On the effectiveness of active learning by uncertainty sampling in classification of high-dimensional gaussian mixture data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4238–4242. IEEE, 2022.

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[38] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

[39] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[42] Natalie Rathvon. *Early Reading Assessment. A Practitioner's Handbook.* ERIC, 2004.

[43] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision*, pages 638–647, 2017.

[44] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023.

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[50] Hitesh Sapkota and Qi Yu. Balancing bias and variance for active weakly supervised learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1536–1546, 2022.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[52] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[53] Silvio Savarese Sener et al. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[54] Burr Settles. Active learning literature survey. 2009.

[55] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araújo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11002–11012. IEEE, 2023.

[56] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36:75669–75687, 2023.

[57] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Proceedings of the 41st International Conference on Machine Learning*, pages 45014–45039, 2024.

[58] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

[59] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[60] Gabriel Skantze and Bram Willemsen. Collie: Continual learning of language grounding from language-image embeddings. *Journal of Artificial Intelligence Research*, 74:1201–1223, 2022.

[61] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[62] Qun Sui and Sujit K Ghosh. Similarity-based active learning methods. *Expert Systems with Applications*, 251:123849, 2024.

[63] Alaa Tharwat and Wolfram Schenck. A survey on active learning: state-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

[64] Weijie Tu, Weijian Deng, Tom Gedeon, and Liang Zheng. A bag-of-prototypes representation for dataset-level applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2881–2892. IEEE, 2023.

[65] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1921–1930. IEEE, 2023.

[66] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3266–3279, 2023.

[67] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[68] Jinsheng Wang, Guoji Xu, Peng Yuan, Yongle Li, and Ahsan Kareem. An efficient and versatile kriging-based active learning method for structural reliability analysis. *Reliability Engineering & System Safety*, 241:109670, 2024.

[69] Tianyang Wang, Xingjian Li, Pengkun Yang, Guosheng Hu, Xiangrui Zeng, Siyu Huang, Cheng-Zhong Xu, and Min Xu. Boosting active learning via improving test performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8566–8574, 2022.

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

[71] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8787–8796, 2019.

[72] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021.

[73] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[74] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. RAPHAEL: text-to-image generation via large mixture of diffusion paths. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[75] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869, 2022.

[76] Shin-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. *arXiv preprint arXiv:2309.14859*, 2023.

[77] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023.

[78] Mariia Zameshina, Olivier Teytaud, and Laurent Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023.

[79] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[81] Lin Zhao, Meimei Shang, Fei Gao, Rongsheng Li, Fei Huang, and Jun Yu. Representation learning of image composition for aesthetic prediction. *Computer Vision and Image Understanding*, 199:103024, 2020.

[82] Yan Zhao, Lei Wang, Luping Zhou, Yinghuan Shi, and Yang Gao. Modelling diffusion process by deep neural networks for image retrieval. In *BMVC*, page 161, 2018.

[83] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[84] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 52(3):1798–1811, 2020.

[85] Hancheng Zhu, Yong Zhou, Leida Li, Yaqian Li, and Yandong Guo. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia*, 2021.

[86] Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023.

A Theory

This section provides a complete derivation for the analysis given in Section 4. In brief, we first link L_{LDM} to a *minimum mean square error* (MMSE) term in Theorem A.2 (i.e. Theorem 3.2 within the main context), then decompose the MMSE term to see how we minimize MMSE in different aspects in Theorem A.3 (i.e. Theorem 3.3 within the main context) in the main context. For convenience, we reclaim some of the formulations at the beginning of this section:

Definition A.1. The *minimum mean square error (MMSE)* of estimating an input random vector $\widehat{\mathbf{X}} \in \mathbb{R}^n$ from an observation/output $\mathbf{X} \in \mathbb{R}^k$ is defined as

$$\mathrm{MMSE}(\widehat{\mathbf{X}}|\mathbf{X}) = \inf_{f \in \mathcal{M}(\mathbb{R}^n)} \mathbb{E}\left[\left\|\widehat{\mathbf{X}} - f(\mathbf{X})\right\|^2\right], \tag{11}$$

in which $\mathcal{M}(\mathbb{R}^n)$ denotes the space consisting of all measurable functions on $\mathbb{R}^n$.

Based on Eq. (1), we can adapt the LDM to an arbitrary concept y with a corresponding image set X that describes this concept. Typical adaptations are mostly based on fine-tuning all the parameters θ, which is quite costly and requires a lot of data. However, recent studies have found that training only a small part of the model [18, 47] or inserting extra networks [25, 26] could attain the same performance, which largely alleviates the computational burden with far fewer samples need for training. We call this part of parameters or networks as "*adaptors*" and denote them by A. Then, the fine-tuning process of the *adaptors*, also named *adaption*, can be formulated as follows:

$$A^* = \arg\min_A \mathbb{E}_{x\sim X, z\sim\mathcal{E}(x), \epsilon\sim\mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_{\theta,A}\left(z_t, t, c_{\theta,A}\left(y\right)\right)\|_2^2\right] \tag{12}$$

Therefore, v be a conditioning vector corresponding to some given text y, the LDM loss is:

$$L_{LDM}(x, A, y) \coloneqq \mathbb{E}_{x\sim X, z\sim\mathcal{E}(x), \epsilon\sim\mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_{\theta,A}\left(z_t, t, c_{\theta,A}\left(y\right)\right)\|_2^2\right] \tag{13}$$

Following the definition of LDM loss, we continue to link LDM loss to the MMSE term in the theorem as follows (in which we omit the parameter y since it is independent of other parameters):

Theorem A.2. *Let L_{LDM} be the LDM loss following Eq.* (13)*, and the image space lies within $\mathbb{R}^n$. Then there always exists an ideal random vector $\widehat{\mathbf{X}}_G \in \mathbb{R}^n$ and a condition vector v^* within the text embedding space for LDM, such that*

$$\arg\min_{\mathbf{X}_T \in \mathbb{R}^n} \mathrm{MMSE}(\widehat{\mathbf{X}}_G|\mathbf{X}_T) = \arg\min_{\mathbf{X}_T \in \mathbb{R}^n} \mathbb{E}_{x\sim\mathbf{X}_T}\left[L_{LDM}\left(x, A^*\right)\right].$$

Proof. By denoting an ideal generative distribution with $\widehat{\mathbf{X}}_G$, which is produced by the generative model with a minimized LDM loss, we continue our proof. Set $f_A(\mathbf{X}_T) = \mathbb{E}_{x\sim\mathbf{X}_T}[x, A]$, we can see that f_A is also a functional controlled by the adaptor A. Following the definition of MMSE and a fixed $\mathbf{X}_T$, we set A^* as:

$$A^* = \arg\min_v \mathbb{E}_{x\sim\mathbf{X}_T}\left[L_{LDM}(x, A)\right]. \tag{14}$$

With this A^*, and the Universal Approximation Theorem [38] for all deep learning models, f_{A^*} becomes the functional needed to quantify the MMSE term following its definition. Therefore, with proper $\mathbf{X}_G, A^*$, we have

$$\mathrm{MMSE}(\widehat{\mathbf{X}}_G|\mathbf{X}_T) = \mathbb{E}_{x\sim\mathbf{X}_T}\left[L_{LDM}\left(x, A^*\right)\right],$$

thus completing the proof of this theorem. □

Theorem A.3. *(Pythagorean Theorem for MMSE [13].) Following Theorem A.2, by setting f to the generative model g_{A*} with an ideal adaptor A^* containing sufficient OOD concept information, rewrite $g_A(\mathbf{X}_T)$ to $\mathbf{X}_G|\mathbf{X}_T$, and the minimum mean square error (MMSE) in Eq.* (11) *can be decomposed into two terms:*

$$\mathbb{E}\left[\left\|\widehat{\mathbf{X}}_G - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\right\|\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{X}}_G - g_{A^*}(\mathbf{X}_T)\right\|\right] + \mathbb{E}\left[\|g_{A^*}(\mathbf{X}_T) - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\|\right], \tag{15}$$

Proof. This equation can be further decomposed into two terms:

$$\begin{aligned}&\mathbb{E}\left[\left(\widehat{\mathbf{X}}_G - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\right)\phi(X_T)\right]\\&= \mathbb{E}\left[(\widehat{\mathbf{X}}_G - \mathbf{X}_G)\phi(\mathbf{X}_T)\right] + \mathbb{E}\left[\mathbf{X}_G - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\phi(\mathbf{X}_T)\right].\end{aligned}$$

Since $\widehat{\mathbf{X}}_G - \mathbf{X}_G$ represents the variance for generative results and is orthogonal to $\mathbf{X}_T$, the first term is 0, making it sufficient to consider only the second term. To prove that the second term is also 0, we first prove the orthogonality property, i,e.

$$\mathbb{E}\left[\left(\widehat{\mathbf{X}}_G - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\right)\phi(X_T)\right] = 0 \tag{16}$$

for any function ϕ.

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\phi(\mathbf{X}_T)\right]\\&= \sum_{x_t}\mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T = x_t\right]\phi(x_t)P(\mathbf{X}_T = x_t)\\&= \sum_{x_g}\left[\sum_{x_t}\frac{P(\mathbf{X}_G = x_g, \mathbf{X}_T = x_t)}{P(\mathbf{X}_T = x_t)}\right]\phi(x_t)P(\mathbf{X}_T = x_t)\\&= \sum_{x_g}\sum_{x_t}x_g\phi(x_t)P(\mathbf{X}_G = x_g, \mathbf{X}_T = x_t)\\&= \mathbb{E}\left[\mathbf{X}_G\phi(\mathbf{X}_T)\right]\end{aligned}$$

Therefore, we have the orthogonal property in Eq. (16).

Now we continue our proof for the main theorem. First, we can decompose Eq. (15) as follows:

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\mathbf{X}}_G - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\right\|^2\right]\\&= \mathbb{E}\left[\|\widehat{\mathbf{X}}_G - g_{A^*}(\mathbf{X}_T) + g_{A^*}(\mathbf{X}_T) - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\|^2\right]\\&= \mathbb{E}\left[\|\widehat{\mathbf{X}}_G - g_{A^*}(\mathbf{X}_T)\|^2\right] + \mathbb{E}\left[g_{A^*}(\mathbf{X}_T) - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\|\right]\\&+ 2\mathbb{E}\left(\widehat{\mathbf{X}}_G - g_{A^*}(\mathbf{X}_T)\right)\left(g_{A^*}(\mathbf{X}_T) - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]\right)\end{aligned}$$

Since the gap between pure generative result guided by the training set $\mathbf{X}_T$ and an ideal text embedding is purely influenced by the training set $\mathbf{X}_T$, $g_{A^*}(\mathbf{X}_T) - \mathbb{E}\left[\mathbf{X}_G|\mathbf{X}_T\right]$ can also be regarded as a functional over $\mathbf{X}_T$, making the last term equal to 0 according to the orthogonal property in Eq. (16). At this step, we obtain the result in Eq. (15). □

B More Details About CATOD.

In this section, we provide more details about our proposed Controllable Adaptor Towards Out-of-Distribution Concepts (CATOD).

B.1 An Overall Paradigm

Initially, we have a large data pool D_{pool} related to the concept given with different quality and begin with an initial randomly-sampled training dataset $\mathbf{X}_T^{(0)}$. $\mathbf{X}_T$ might contain distorted images or even mismatch the given concept, which is common in publicly available datasets. Following the typical paradigm of Active Learning, we train all the models we use on the training set $\mathbf{X}_T$. Then, we select a batch of data $\mathcal{B}$ from the data pool D_{pool} according to the acquisition functions induced from models. Finally, we move this selected batch of data $\mathcal{B}$ to $\mathbf{X}_T$, go back to the model training step to re-train or adapt the models, and repeat this cycle until $\mathbf{X}_T$ reaches its maximum capacity or the quality of models cannot be further enhanced by adding new data. An overall Algorithm for a cycle of CATOD is provided in Alg. 1.

Algorithm 1 An active selection cycle for CATOD.

Input: Text-to-image model $g_A(\cdot)$ that can be integrated with adaptor A, real-world data pool D_{pool}, training data pool $\mathbf{X}_T$, budget b, pretrained aesthetic scorer $S_{aes}(\cdot)$, concept-matching scorer $S_{con}(\cdot)$, adaptor A, learning rate group R.
Output: Updated training pool $\mathbf{X}_T$, updated adaptor A.

Fine-tune the aesthetic scorer $S_{aes}(\cdot, \theta)$ on $\mathbf{X}_T$ following Sec.B.3;
Calculate the aesthetic score $\gamma_{aes}(A) = \frac{1}{|g_A(\mathbf{X}_T)|} \sum_{x_g \in g_A(\mathbf{X}_T)} S_{aes}(x_g)$ and concept-matching score $\gamma_{con}(A) = \frac{1}{|g_A(\mathbf{X}_T)|} \sum_{x_g \in g_A(\mathbf{X}_T)} S_{con}(x_g)$ for A;
for $x \in D_{pool}$ **do**
    Calculate the integrated score $S(x)$(Eq.(7));
end for
$\mathcal{B} \leftarrow$ {Top-b samples within D_{pool} according to S};
$s \leftarrow \gamma(A)$ (Eq.(6));
$r \leftarrow$ the largest value within R;
while $r \neq \min(R)$ **do**
    $A' \leftarrow A - r\nabla_A \sum_{x \in \mathbf{X}_T \cup \mathcal{B}} L_{LDM}(x, A)$;
    Generate a small-scale $\mathbf{X}_G$ with $g(\cdot)$ conditioned on A';
    Calculate score $\gamma(A)$ for A via Eq.(6)) and $\mathbf{X}_G$;
    if $\gamma(A) > s$ **then**
        $A \leftarrow A'$;
        $s \leftarrow \gamma(A)$;
    else
        $r \leftarrow$ a smaller value than r within R
    end if
end while
$\mathbf{X}_T \leftarrow \mathbf{X}_T \cup \mathcal{B}$

We use a weighted-scoring system as the acquisition module that evaluates the quality of images and adaptors in our proposed CATOD framework. In traditional active learning or data selection, the acquisition module is only conducted on samples. However, we observed that an increase in the number of samples does not necessarily contribute to a performance boost like in a traditional active learning setting (which we will verify in Appendix C). Therefore, we propose this weighted scoring system to help schedule adaptor training, which helps locate the best version among them and select samples according to their quality.

Our Controllable Adaptor Towards Out-of-Distribution Concepts framework can be divided into 4 steps: (1) Train the scoring system on the current training set $\mathbf{X}_T$. (2) Train the adaptor on the training set $\mathbf{X}_T$ and schedule the training accordingly to the scoring system; (3) Calculate the acquisition score accordingly to the scoring system and the best adaptor for samples in dataset D_{pool}; (4) Select top-K samples from D_{pool} and move them to train set $\mathbf{X}_T$.

B.2 An Implementation of MMD score.

To give out an quantifiable and unbiased estimation of Eq.(2), we sample two sets of vectors $X = \{x_1, x_2, \ldots, x_m\}$ from P and $G = \{g_1, g_2, \ldots, g_n\}$ from Q, an estimation can be give by:

$$\text{dist}^2_{MMD}(X, G) := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(g_i, g_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, g_j). \tag{17}$$

B.3 Aesthetic Scorer

Following the description of the aesthetic scorer in Section 3.2, we explain how to personalize it here. Since our primary goal is to make the model distinguish low-quality/high-quality examples for OOD concepts, we design an automatic procedure to assign aesthetic scores to our training set $\mathbf{X}_T$ at each cycle, based on which we attain an aesthetic training set $\mathbf{X}_T^{aes}$. In this work, the aesthetics training comprises three parts: (1) The original training set $\mathbf{X}_T$; (2) Generated set D_{TG} without assistance

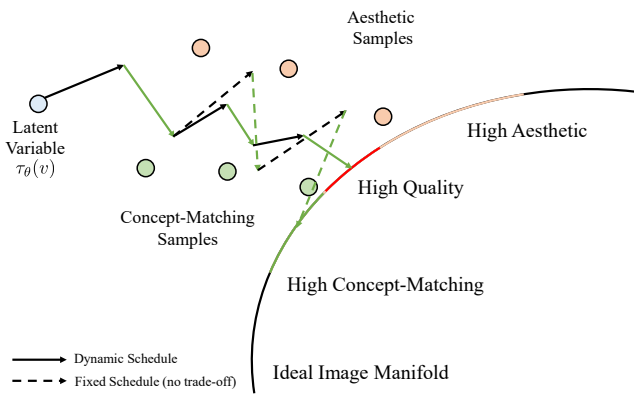


Figure 6: **An intuitive comparison for fixed/dynamic schedules.** The active learning paradigm can be viewed as guiding the iterative embedding updating through newly added samples. We can see that a fixed schedule makes the learned embedding heavily biased, which in turn leads to performance fluctuation.

from any adaptors; (3) Randomly sampled set D_{SI} from similar but irrelevant categories. Since we focus on the concepts that the text-to-image model can hardly recognize and visualize, D_{TG} contains lots of plausible, disrupted, or even irrelevant samples, which we assign a zero score. This helps the aesthetic model understand that originally generated samples are low-quality samples. In contrast, D_{SI} somehow describes some attributes for the concept but still does not match the concept, to which we assign the average score of 5.0. We employ the normal regression loss formulated as follows:

$$\mathcal{L}_{aes} = \frac{1}{|\mathbf{X}_T|} \sum_{x_t \in \mathbf{X}_T} \|p_t, \hat{p}_t\| + \frac{\lambda}{|D_{TG}| + |D_{SI}|} \sum_{x'_t \in D_{TG} \cup D_{SI}} \|p'_t - \hat{p}'_t\|, \tag{18}$$

where p denotes score predicted by the original scorer, when $\hat{p}$ denotes the score we assign, and λ is set to 1.0 in our experiments. Following this paradigm, we train an aesthetic scorer for each category of concepts, and the score personalized for each category from D_{pool} can be predicted accordingly.

After selecting top-K data according to this score based on adaptor quality, we assign a lowered aesthetic score to them:

$$\hat{p}_x = 10.0 - \frac{t_{current}}{_{total}}(10.0 - p_x), \tag{19}$$

where p_x denotes the score given by a generic aesthetic scorer, $_{total}$ represents the number of total active learning cycles, when $t_{current}$ indicates the current cycle. Note that the newly selected images typically have lower quality than those from earlier cycles, we assign progressively lowered scores to them as the learning cycle proceeds.

B.4 Adaptor Evaluation and Schedules

Following the content in the second part "The Active Schedule for Training Adaptors" in Sec. 3.3, we describe our schedule in more detail. In training CATOD, we use this indicator as a signaling proxy throughout the training process. As the number of training samples increases, the quality of selected samples at later cycles might be lower, and a fixed schedule may introduce some unnecessary details within these images (such as watermarks, borders, disturbing objects, etc.). However, these subsequent samples with relatively lower quality do contain some good photographic attributes that can help the embedding evolve. Actually, the potential problems can be alleviated by carefully-designed schedules. An intuitive illustration for this schedule is given in Fig. 6.

We first train the adaptor for 5 epochs, and save an adaptor version A_i per 5 epochs, together with its corresponding generated images, constituting a sub-cycle. When this sub-cycle is completed, we filter out the adaptor with the highest quality based on the score $\gamma(A)$ (Eq.(6)). If the selected adaptor has the most training epochs across all versions, we keep the learning rate and conduct the next sub-cycle, otherwise, we will reduce the learning rate. Note that if the adaptor quality of this cycle is not as good as that of the previous cycle (because of the quality degradation caused by introducing some unnecessary details), the initial learning rate will be reduced, which helps avoid

Table 5: The statistics of the dataset we use. The test data size $\#\mathcal{D}_{val}$ of all the concepts is fixed to 100 for a fair comparison, while the training data size $\#\mathcal{D}_{pool}$ varies since different concepts since the number of data samples across publicly available datasets is different.

Insect	CMMD (↓)	Lizard	CMMD (↓)	Penguin	CMMD (↓)	Seafish	CMMD (↓)	Snake	CMMD(↓)
Antlion	3.95	Axolotl	4.24	Emperor Penguin Chick	4.14	Crampfish	4.19	Ahaetulla nasutar	4.36
Lycorma Delicatula	3.99	Frilled Lizard	3.42	Gentoo Penguin	3.94	Dragonfish	4.16	Aipysurus laevis	3.56
Parasitic Wasps	4.61	Mediterranean House Gecko	4.45	King Penguin Chick	3.68	Garfish	3.97	Indian Cobra	3.91
Xylotrechus	4.19	Oedura	4.6	Rockhopper Penguin	4.04	Tigerfish	4.17	Pelamis Platurus	3.94
Zopherinae	4.12	Opluridae	3.88	Royal Penguin	3.98	Tuna	4.05	Sidewinder	4.11
Thrips	2.05	Whiptail	0.88	Emperor Penguin	1.19	Goldfish	1.35	Thunder snake	1.46
Flea Beetle	1.28	Alligator Lizard	1.44	King Penguin	0.75	Hammerhead	1.36	Garter snake	2.35
Aphids	1.16	Gila Monster	1.64	Little penguin	1.46	Tench	1.25	Night Snake	1.06
Red Spider	1.85	Agama	2.69	Magellanic Penguin	1.60	Tiger Shark	0.90	Rock Python	1.76
Meadow Moth	1.47	Komodo Dragon	2.00	Adelie Penguin	0.77	Killer Whale	0.70	Hognose Snake	1.94

Table 6: The statistics of the dataset we use. The test data size $\#\mathcal{D}_{val}$ of all the concepts is fixed to 100 for a fair comparison, while the training data size $\#\mathcal{D}_{pool}$ varies since different concepts since the number of data samples across publicly available datasets is different.

Category	Concept	$\#\mathcal{D}_{pool}$	$\#\mathcal{D}_{val}$
Insect	Zopherinae	1357	100
	Antlion	864	100
	Lycorma Delicatula	5108	100
	Parasitic Wasps	877	100
	Xylotrechus	1043	100
Lizard	Axolotl	1200	100
	Frilled Lizard	1008	100
	Mediterranean House Gecko	889	100
	Oedura	798	100
	Opluridae	818	100
Penguin	Emperor Penguin Chick	987	100
	Gentoo Penguin	3992	100
	King Penguin Chick	1471	100
	Rock Hopper Penguin	877	100
	Royal Penguin	754	100
Sea Fish	Crampfish	1179	100
	Dragonfish	835	100
	Garfish	1200	100
	Tigerfish	538	100
	Tuna	1236	100
Snake	Ahaetulla Nasutar	893	100
	Aipysurus Laevis	1015	100
	Indian Cobra	1200	100
	Pelamis Platurus	1091	100
	Sidewinder	1200	100

introducing disturbing elements led by some samples with insufficient quality. The sub-cycles will be continuously conducted until the minimum learning rate is reached or the adaptor quality no longer increases. Our learning rates include $5 \times 10^{-4}, 2.5 \times 10^{-4}, 7.5 \times 10^{-5}, 5 \times 10^{-5}, 2.5 \times 10^{-5}$.

C Experiments

C.1 How to Automatically Locate OOD Concepts

Since the publicly available dataset, i.e., ImageNet, iNaturalist 2018 [67], IP102 [71], which we adopt in our research contains 9,244 concepts with 14,883,455 images in total, it is infeasible to manually test and decide what concept is the out-of-distribution concept one by one. Therefore, our procedure for figuring out is two-fold: (1) automatically quantifying the distribution drift of the given concept from what the text-to-image model (Stable Diffusion 2.0, a.k.a SD 2.0) learned with

the CMMD metric (thanks to the recent work proposed by Google [27]); (2) Manually verifying whether the text-to-image model could generate the right content, according to the rank given by the CMMD metric above. Specifically, in the first step, we randomly sample 10 image data for each concept within the datasets and then generate 10 images with the model with the text prompt "A photo of S^*", in which S^* denotes the concept name. Then, we calculate the CMMD discrepancy between these two data batches and record them. The CMMD scores of some concepts (including both Out-of-Distribution concepts in the first 5 rows and In-Distribution in the last 5 rows) are listed in Table 5.

Empirically, we observed that SD 2.0 is unable to synthesize the concepts with a CMMD metric above 3.5. Therefore, we pick 25 OOD concepts accordingly divide them into 5 categories, and make our dataset as follows:

- **Insect**: Zopherinae, Antlion, Lycorma Delicatula, Parasitic Wasps, Xylotrechus;
- **Lizard**: Axolotl, Frilled Lizard, Mediterranean House Gecko, Oedura, Opluridae;
- **Penguin**: Emperor Penguin Chick, Gentoo Penguin, King Penguin Chick, Rock Hopper Penguin, Royal Penguin;
- **Sea Fish**: Crampfish, Dragonfish, Garfish, Tigerfish, Tuna;
- **Snake**: Ahaetulla Nasutar, Aipysurus Laevis, Indian Cobra, Pelamis Platurus, Sidewinder.

Actually, the dataset we collect is not class-balanced, so the number of samples we can collect varies across different concepts. The statistics of each concept are shown in Table 6. For the concepts with a $\mathcal{D}_{pool}$ less than 1,000 images, we also collect some of them from image host websites including Dreamstime, Flicker, istockphoto, pinterest, and shutterstock. The dataset will be released as soon as the code is released.

C.2 Experimental Settings

To verify the effectiveness and extensibility of our proposed CATOD framework, we conduct experiments on 25 different OOD concepts that the large-scale text-to-image (i.e. Stable Diffusion 2.0 [45]) implement different versions of CATOD. This section explains some important implementation details for the dataset, the aesthetic scorer, the concept-matching scorer we use, and how we design the learning paradigm.

The Learning Rate Schedule. As shown in Algorithm 1, we have a learning rate group R, which includes 5 learning rates: 5×10^{-4}, 2.5×10^{-4}, 7.5×10^{-5}, 5×10^{-5}, 2.5×10^{-5}. Following the indicator $\gamma(A)$ in Eq. (6), we calculate this value after every epoch. When this indicator falls below the previous evaluation, we reduce the learning rate. If the learning rate cannot be lowered further, we conclude that the model has converged and the training is complete. Note that we choose to train 20 epochs in all experiments as claimed in the main context, it is possible that the training is early stopped before reaching epoch 20.

Aesthetic Scorers. To offer the basic knowledge for aesthetic evaluation, we use pre-trained generic models, including ReLIC [81], TANet [23], SD-Chad Scorer, and VLAD [59]. Note that the scores given by the generic aesthetic scoring model tend to lie in the majority score range (5 to 6) in our dataset, we ought to personalize this model accordingly to our training set, which we refer to as PIAA [84]. Since PIAA is a typical small sample problem, we adopt similar experimental settings and evaluation criteria by referring to Few-Shot Learning [16] and a previous PIAA research work: PA-IAA [33].

Concept-Matching Scorers. Since the first step of general image retrieval is feature extraction, we use the CLIP ViT-L/14 encoder, which is a commonly used and reliable feature extractor [2, 82, 34].

C.3 About Compositional Results over Multiple-Concepts.

To further validate the effectiveness of CATOD, we also present compositional results based on the LoRA adaptors we obtained, as illustrated in Figure 7. We compose different types of concepts, including (i) background concepts such as brick wall and grassland for Emperor Penguin Chicks,

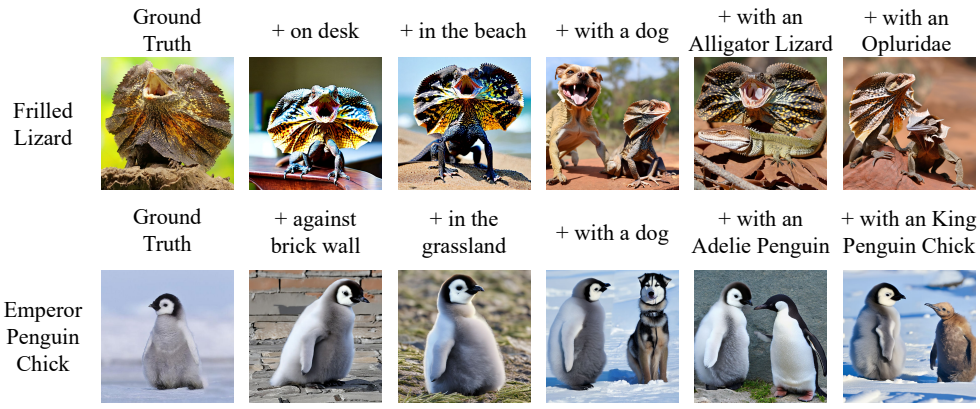
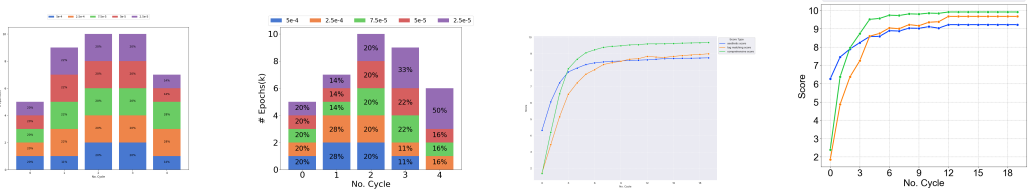


Figure 7: **Generative results with 2 concepts within one image**. Experiments are conducted based on the LoRA adaptor fully trained on concepts "Frilled Lizard" and "Emperor Penguin Chick". We try to compose these creatures with background elements, in-distribution concepts, and out-of-distribution concepts learned by other adaptors. The final results show high quality with minimal disruptive details.

and desk and beach for Frilled Lizard; (ii) in-distribution and common concepts, like the dog; (iii) in-distribution and similar concepts, like Adelie penguin for Emperor Penguin Chicks, and Alligator Lizard for Frilled Lizard; (iii) out-of-distribution concepts from other adaptors, like King Penguin Chicks for Emperor Penguin Chicks and Opluridae for Frilled Lizard. Our observations reveal that the synthetic results on out-of-distribution concepts with CATOD can seamlessly integrate different background elements, even if they were not present in real-world images (columns 1 to 2). For other in-distribution concepts like dogs, LDMs tend to represent a specific species with similar color and texture, while different concepts remain highly distinguishable within one image (columns 3 to 4). For out-of-distribution concepts, we also note that the creatures are depicted correctly, but may exhibit some confused visual details that negatively impact the aesthetics of the image (column 5).

C.4 More Analysis.



(a) The number of epochs with specified different learning rates within the schedule on RAND.

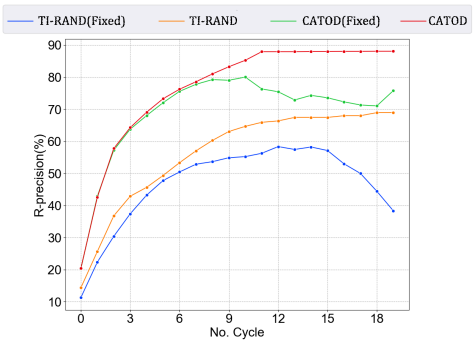
(b) The number of epochs with specified different learning rates within the schedule on CATOD.

(c) Changes on different scores as cycle proceeds on RAND.

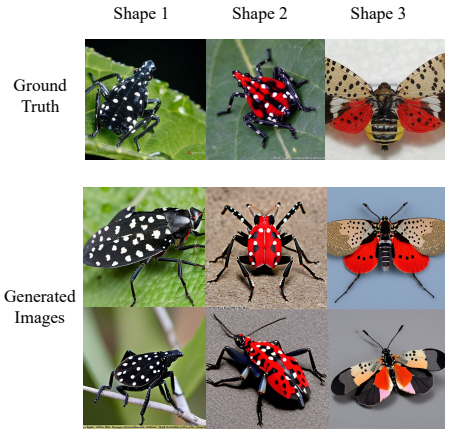
(d) Changes on different scores as the cycle proceeds on CATOD.

Figure 8: **A comparison on how the schedule and scores change on RAND(scheduled) and CATOD as cycle proceeds on concept emperor penguin(chick).** (a),(b) show how the #epochs for each learning rate in the schedule change as the cycle proceeds, when (c),(d) show how aesthetic/concept-matching/comprehensive score change on RAND (scheduled) and CATOD. The scores for CATOD stop changing at cycle 12 since more added samples do not help boost adaptor quality.

CATOD is more reliable on schedules. Figure 8(a),(b) show a comparison of how our schedules adjust as the number of cycles increases on RAND(scheduled) and CATOD. Both of them show the tendency that training epochs with larger learning rates decrease when those with small learning rates



(a) **R-precision(%) with different strategies as training pool expands.** The experiments are performed over concept "emperor penguin(chick)" and produced over RAND and CATOD. To further show the impact of the number of samples, we also compare their modified versions that do not use dynamic schedules.



(b) **A comparison on real-images with CATOD-generated ones in 3 different shapes of lycorma delicatula.** We can see that with a carefully designed selection, the adaptor helps produce all the shapes of the concept "lycorma delicatula".

Table 7: A Comparison over the performance of CATOD on different types of the initial training data pool, in terms of the CLIP score and CMMD score with 100 images sampled at last. This table shows the average result of 5 sub-classes within each category. The overall improvement of our proposed CATOD is provided by "Imp.". Methods with the best performance are bold-folded.

Comparison Methods	CLIP Score (↑)						CMMD [27] (↓)					
	insect	lizard	penguin	seafish	snake	Avg.	insect	lizard	penguin	seafish	snake	Avg.
10 HQ initial samples	73.87	79.23	75.59	83.84	82.19	78.94	0.87	0.80	0.71	0.79	0.78	0.79
10 RAND initial samples	72.26	74.06	72.89	79.75	78.08	75.41	0.92	0.84	0.73	0.80	0.93	0.84
20 HQ initial samples	75.18	77.59	76.97	82.78	83.14	79.13	0.85	0.79	0.69	0.72	0.75	0.76
20 RAND initial samples	72.08	72.82	72.97	77.74	75.06	74.13	1.01	1.03	0.92	0.97	0.83	0.95
50 HQ initial samples	75.27	77.27	76.85	82.90	83.37	79.13	0.84	0.80	0.63	0.69	0.71	0.73
50 RAND initial sample	70.34	70.53	67.49	72.10	73.74	70.84	1.13	1.29	0.94	1.06	0.97	1.08

increase, meaning that the training schedule focuses more on fine-tuning at later cycles. Comparing CATOD with RAND, we can see that the epochs with the maximal learning rate already diminish at cycle 4 on CATOD while RAND still needs large learning rates, indicating that CATOD are better at recognizing the given concept and tends to be more stable.

CATOD achieves a good score earlier than other methods. Figure 8(c),(d) shows a comparison of how the aesthetic/concept-matching/comprehensive scores change as the cycle proceeds on RAND and CATOD. We can see that CATOD already achieves a higher value on all scores and cannot be further boosted with 120 training samples on cycle 12 when the performance for RAND still fluctuates. This phenomenon illustrates that our active selection strategy makes our embedding be trained on more high-quality data.

Furthermore, we also notice that aesthetic/concept-matching scores fluctuate at later cycles when the comprehensive score is enhanced continuously, which exactly corresponds to our proposed dual scoring system that balances the importance of these two factors when ensuring the quality of the adaptor never declines.

The performance of CATOD on different initial training data pool. We also ablate CATOD over the initial training set size/quality. In detail, we retest CATOD with different numbers of initial samples (10,20,50) and different quality, i.e., high-quality (HQ) and random sampling (RAND), with 100 images sampled at last and 10 images selected per cycle. These experiments are tested with adaptor LoRA. The results are shown in Table 7:

Table 8: A Comparison over the performance of CATOD on different architectures, in terms of the CLIP score and CMMD score with 100 images sampled at last. This table shows the average result of 5 sub-classes within each category. The overall improvement of our proposed CATOD is provided by "Imp.". Methods with the best performance are bold-folded.

Comparison Methods	CLIP Score (↑)							CMMD [27] (↓)						
	insect	lizard	penguin	seafish	snake	Avg.	Imp.	insect	lizard	penguin	seafish	snake	Avg.	Imp.
SD 1.5 + RAND	63.70	64.94	65.98	66.29	69.43	66.07	⇑9.27	1.49	1.47	1.65	1.62	1.33	1.51	⇓0.61
SD 1.5 + CLIP	68.78	70.38	70.89	74.20	73.89	71.63	⇑3.71	1.19	1.45	1.28	1.45	1.21	1.32	⇓0.42
SD 1.5 + CATOD	**71.73**	**74.79**	**73.45**	**79.37**	**77.35**	**75.34**	-	**1.01**	**0.97**	**0.78**	**0.92**	**0.84**	**0.90**	-
SDXL + RAND	72.39	73.04	71.75	74.81	70.47	72.49	⇑8.05	1.39	1.45	1.49	1.58	1.01	1.38	⇓0.50
SDXL + CLIP	79.32	78.24	74.98	82.16	79.75	78.89	⇑1.65	1.16	1.27	1.09	0.93	1.07	1.10	⇓0.22
SDXL + CATOD	**80.37**	**79.58**	**77.56**	**85.20**	**79.97**	**80.54**	-	**0.95**	**0.87**	**0.96**	**0.89**	**0.75**	**0.88**	-

Table 9: A Comparison over the diversity score of CATOD, in terms of the CLIP score and CMMD score with 100 images sampled at last. This table shows the average result of 5 sub-classes within each category. The overall improvement of our proposed CATOD is provided by "Imp.". Methods with the best performance are bold-folded.

Comparison Methods	LPIPS Score(↑)						
	insect	lizard	penguin	seafish	snake	Avg.	Imp.
DreamBooth + CLIP	0.254	0.156	0.149	0.305	0.265	0.226	⇑0.052
DreamBooth + CATOD	0.362	0.198	0.197	0.349	0.286	0.278	-
TI + CLIP	0.198	0.257	0.174	0.108	0.208	0.189	⇑0.033
TI + CATOD	0.267	0.278	0.152	0.278	0.133	0.222	-
LoRA + CLIP	0.178	0.184	0.155	0.203	0.094	0.163	⇑0.054
LoRA + CATOD	0.245	0.203	0.217	0.244	0.178	0.217	-

From these results, we can draw the following conclusions:

- Initial batch size has a more significant impact on randomly initialized samples than on high-quality samples. Specifically, the performance change with high-quality samples is at most 1.40 in the CLIP score and 0.10 in the CMMD score concerning different initial numbers of training data. In contrast, the performance change on low-quality is up to 7.65 in the CLIP score and 0.45 in the CMMD score.
- The quality of initial samples does have an impact on generative results since we can see a consistent performance loss when changing HQ initial samples to randomly initialized samples. With the initial size increase, this impact tends to be even more significant.

C.5 Further Extension: Concept with multiple shapes

Recall that Stable Diffusion 2.0 fails to generate emperor penguin chicks, which accounts for the fact that adult emperor penguin chicks and their chicks have different appearances. We raise another question: can we train an adaptor with a concept that has different shapes? Inspired by this, we also test our proposed CATOD on the concept "lycorma delicatula", which is a kind of pest with 3 different forms in its life-cycle, and surprisingly find that with 300 training samples, the best version of adaptors successfully produces all these 3 shapes, as shown in Figure 9(b). These results further validate the effectiveness of our framework, even in the presence of appearance ambiguity.

C.6 About Experimental Results on other Architectures

Our experiments in the main content are all conducted on Stable Diffusion 2.0 (SD 2.0). To further validate the superiority of our proposed CATOD, we conduct additional experiments on other architectures (based on LoRA), including Stable Diffusion 1.5 (SD 1.5) and Stable Diffusion XL (SDXL). The results are shown in Table 9.

This table shows that our proposed CATOD is also compatible with different architectures with notable performance gain compared to the baselines.

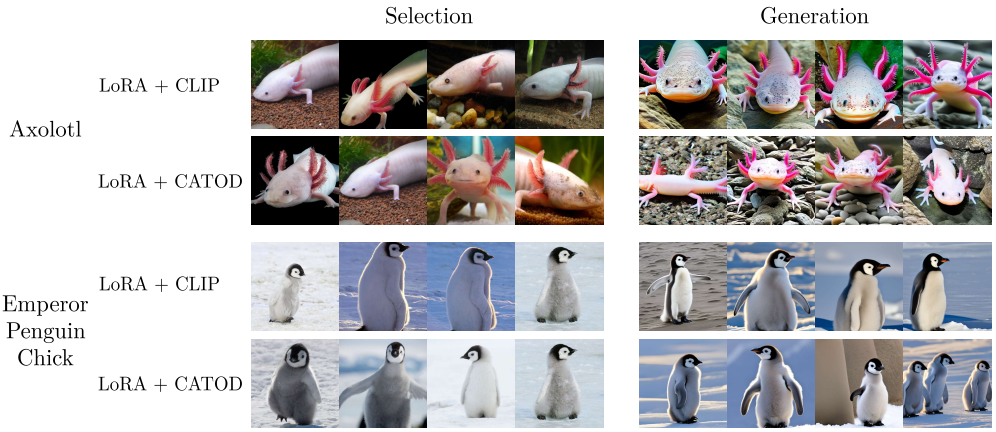


Figure 9: A comparison of selected and generate samples on different combinations of methods and concepts. We can observe that training samples with different angles selected by CATOD also lead to diverse angle in their generative results.

C.7 About the diversity

CATOD maintains the diversity to produce OOD concepts with different angles or poses in generative results. To validate the diversity of our generative results, we first provide a quantitative evaluation with LPIPS [80, 78]. The results are shown in Table 9. In this table, we can see that CATOD gives out a diversity improvement up to 0.17 in the LPIPS score compared to CLIP-based sampling, and outperforms CLIP over most categories. From this, we conclude that CATOD also preserves diversity. Since the training set contains samples with different angles in CATOD, it is also easy to produce objects with different angles, and we show the examples on concepts "Axolotl" and "Emperor Penguin Chick" in Figure 9. We can observe that training samples with different angles selected by CATOD also lead to diverse angle in their generative results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly summarized our contribution in the abstract and in the last paragraph of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: We believe that the limitation should not be regarded as an integral part of this paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided theoretical proof of our theory in Sec.4 in Appendix. A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided our code within the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This paper does not provide the dataset regarding the risk of anonymity, despite that the dataset is just a collection of publicly available datasets.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental details are provided in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiments do not have error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This is reported in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper has no risk regarding ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This has been briefly discussed in our conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: MIT License has been added already.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have a readme document for our codes.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiments were conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing experiments were conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.