A Multi-Armed Bandit Approach to Online Selection and Evaluation of Generative Models

Xiaoyan Hu
The Chinese University of Hong Kong

Ho-fung Leung
Independent Researcher

Farzan Farnia
The Chinese University of Hong Kong

Abstract

Existing frameworks for evaluating and comparing generative models consider an offline setting, where the evaluator has access to large batches of data produced by the models. However, in practical scenarios, the goal is often to identify and select the best model using the fewest possible generated samples to minimize the costs of querying data from the sub-optimal models. In this work, we propose an online evaluation and selection framework to find the generative model that maximizes a standard assessment score among a group of available models. We view the task as a multi-armed bandit (MAB) and propose upper confidence bound (UCB) bandit algorithms to identify the model producing data with the best evaluation score that quantifies the quality and diversity of generated data. Specifically, we develop the MAB-based selection of generative models considering the Fréchet Distance (FD) and Inception Score (IS) metrics, resulting in the FD-UCB and IS-UCB algorithms. We prove regret bounds for these algorithms and present numerical results on standard image datasets. Our empirical results suggest the efficacy of MAB approaches for the sample-efficient evaluation and selection of deep generative models. The project code is available at `https://github.com/yannxiaoyanhu/dgm-online-eval`.

1 INTRODUCTION

Deep generative models have achieved astonishing results across a wide array of machine learning datasets. Quantitative comparisons between generative models, trained using different methods and architectures, are commonly performed by evaluating assessment metrics such as Fréchet Inception Distance (FID) (Heusel et al., 2017; Stein et al., 2023) and Inception Score (IS) (Salimans et al., 2016). Due to the growing applications of generative models to various learning tasks, the machine learning community has continuously adapted evaluation methodologies to better suit the characteristics of the newly introduced applications.

A common characteristic of standard evaluation frameworks for deep generative models is their offline assessment process, which requires a full batch of generated data for assigning scores to the models. While this offline evaluation does not incur significant costs for moderate-sized generative models, producing large batches of samples from large-scale models can be costly. In particular, generating a large batch of high-resolution image or video data could be expensive and hinder the application of existing evaluation scores for ranking generative models.

In this work, we focus on the *online evaluation and selection* of generative modeling schemes, where we consider a group of generative models and attempt to identify the model with the best score by assessing the fewest number of produced data. By limiting the number of generated samples in the assessment process, online evaluation can save on the costs associated with producing large batches of samples. Additionally, online evaluation can significantly reduce the time and computational expenses required to identify a well-performing model in assessing a large group of generative models. Figure 1 shows an example of FD-based evaluation of three pretrained models where the embeddings are extracted by InceptionNet.V3. While the offline evaluation procedure requires a large batch of generated data from each model, our online evaluation approach queries generated data from models adaptively and limits sample generation from the sub-optimal models, thus collecting samples from the best model more frequently (Figure 5).

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

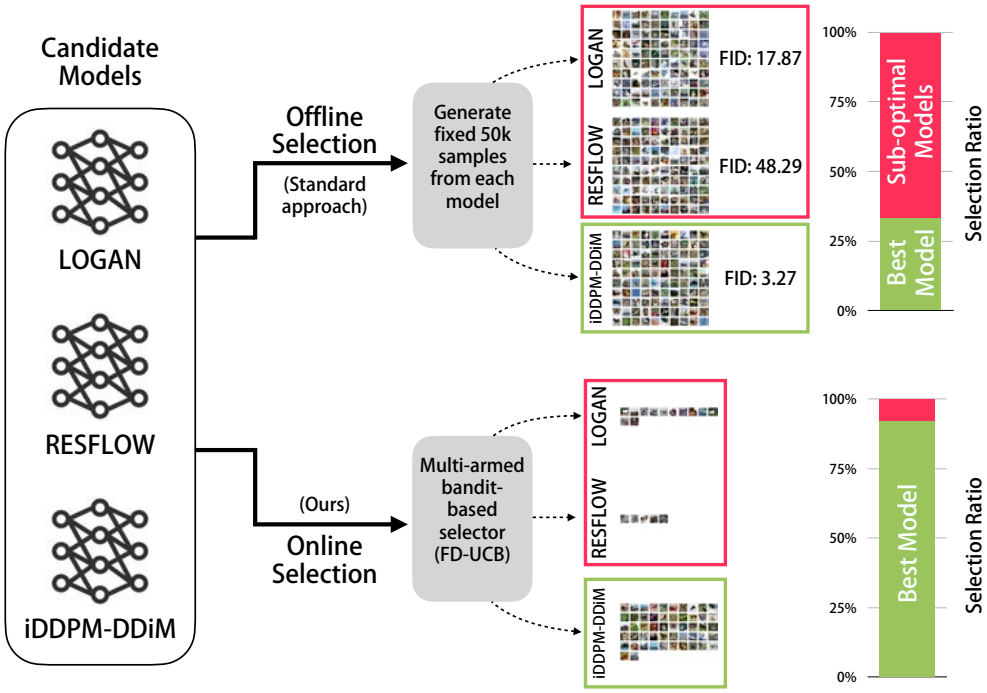

Figure 1: FID-based evaluation and selection among CIFAR10 generative models: The standard offline evaluation requires a large batch of data from every model. In contrast, our proposed online approach leverages the UCB multi-armed bandit strategy to identify the best model with fewer generations from the suboptimal models.

To measure the performance of an online evaluation algorithm, we use the *regret* notion widely used in the online learning literature (Bubeck and Cesa-Bianchi, 2012). In the online learning task, the evaluation algorithm selects a generative model and observes a mini-batch of samples generated by that model in each round. Our goal is to minimize the algorithm's regret, defined as the cumulative difference between the scores of the selected models and the best possible score from the model set. Therefore, the learner's regret quantifies the cost of generating samples from sub-optimal models. We aim to develop online learning algorithms that result in lower regret values in the assessment of generated data.

To address the described online evaluation of generative models, we utilize the multi-armed bandit (MAB) framework. In a standard MAB setup, the online learner seeks to identify the arm with the highest expected value of a random score. While this formulation has been widely considered in many learning settings, it cannot solve the online evaluation of generative models, as standard IS and FD metrics do not simplify to the expectation of a random variable and represent a non-linear function of the data distribution.

By deriving concentration bounds for the FD and IS scores, we propose the optimism-based FD-UCB and IS-UCB algorithms to address the online assessment of generative models. These algorithms apply the upper confidence bound (UCB) approach using data-dependent bounds that we establish for the FD and IS scores. Furthermore, we analytically bound the regret of the IS-UCB and FD-UCB methods, demonstrating their sub-linear regret growth assuming a full-rank covariance matrix of the embedded real data.

We discuss the results of several numerical applications of FD-UCB and IS-UCB to standard image datasets and generative modeling frameworks. We compare the performance of these algorithms with the Greedy algorithm, which selects the model with the highest estimated score, and a Naive-UCB baseline that applies the UCB algorithm with a data-independent upper confidence bound. Our numerical results show a significant improvement in our proposed algorithms compared to the Greedy and Naive-UCB baselines. Additionally, FD-UCB can attain satisfactory performance under the standard image data embeddings, including InceptionNet.V3 (Szegedy et al., 2016), DINOv2 (Oquab et al., 2024), and CLIP (Radford et al., 2021). Our work demonstrates the effectiveness of MAB-based approaches to the evaluation and selection of generative models. The following summarizes the main contributions of this work:

- Proposing a MAB-based evaluation framework for generative models that aims to minimize the regret

of misidentifying the score-maximizing model from online generated data.

- Developing the FD-UCB and IS-UCB algorithms by applying the upper-confidence-bound framework to our data-dependent estimation of the scores.
- Proving sub-linear regret bounds for the proposed FD-UCB and IS-UCB algorithms.
- Demonstrating satisfactory empirical performance of FD-UCB and IS-UCB in comparison to the Greedy and Naive-UCB baseline algorithms.

2 RELATED WORK

Assessment of deep generative models. The evaluation of generative models has been extensively studied in the literature. Several evaluation metrics are proposed, including distance-based metrics such as Wasserstein critic (Arjovsky et al., 2017), Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018), and diversity/quality-based metrics such as Precision/Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), density and coverage (Naeem et al., 2020), Vendi (Friedman and Dieng, 2023), RKE (Jalali et al., 2023), and scalable FKEA-Vendi Ospanov et al. (2024). In addition, the related works develop metrics quantifying the generalizability of the generative models, including the authenticity score (Alaa et al., 2022), the FLD score (Jiralerspong et al., 2023), the Rarity score (Han et al., 2023), and the KEN score (Zhang et al., 2024b,a) measuring the novelty of the generated samples. In this paper, we primarily focus on FID and Inception Score, which have been frequently used for evaluating generative models.

Role of embeddings in the quantitative evaluation results. Due to the high-dimensionality of images, evaluation of the generated images mostly relies on the embeddings extracted by pretrained networks on the ImageNet dataset, e.g., InceptionNet.V3 (Szegedy et al., 2016). However, Naeem et al. (2020) shows that such pretrained embeddings could exhibit unexpected behaviors. Recently, several large pretrained models have been proposed, including DINOv2 (Oquab et al., 2024) and CLIP (Cherti et al., 2023). Stein et al. (2023) shows that DINOv2-ViT-L/14 enables more interpretable evaluation of generative models. In addition, Kynkäänniemi et al. (2023) demonstrates that FID scores computed based on the embedding extracted by CLIP agree more with human-based assessments. In this paper, we provide the numerical results for the mentioned embeddings extracted by different pretrained models.

Online learning using diversity-related evaluation metrics. Online learning is a sequential decision-making framework where an agent aims to minimize a cumulative loss function revealed to her sequentially. One popular setting is the multi-armed bandit (MAB), whose study dates back to the work of Lai and Robbins (1985); Thompson (1933). At each step, the agent chooses among several arms, each associated with a reward distribution, and aims to maximize a pre-specified performance metric. The primary concern of this body of literature considers maximizing the expected return (Agrawal, 1995; Auer, 2003; Bubeck and Cesa-Bianchi, 2012). Recent works study performance metrics related to the variance or entropy of the reward distribution, including the mean-variance criterion (Sani et al., 2012; Zhu and Tan, 2020) in risk-sensitive MAB, and *informational MAB* (IMAB) where the agent maximizes the entropy rewards (Weinberger and Yemini, 2023). However, to the best of our knowledge, the evaluation of generative models has not been exclusively studied in an online learning context. In a concurrent work, Rezaei et al. (2025) aim to maximize kernel-based evaluation scores by finding a mixture of generative models. In this work, we primarily focus on identifying the best (single) model.

Online training of generative models. Several related works focus on training generative adversarial networks (GANs) (Goodfellow et al., 2014) using online learning frameworks. Grnarova et al. (2018) proposes to train *semi-shallow* GANs using the *Follow-the-Regularized-Leader* (FTRL) approach. Daskalakis et al. (2018) shows that optimistic mirror decent (OMD) can be applied to address the limit cycling problem in training Wasserstein GANs (WGANs). Also, the recent paper Park et al. (2024) studies the *no-regret* behaviors of large language model (LLM) agents. This reference proposes an unsupervised training loss, whose minimization could automatically result in known no-regret learning algorithms. On the other hand, our focus is on the evaluation of generative models which does not concern the models' training.

3 PRELIMINARIES

3.1 Inception Score

Inception score (IS) is a standard metric for evaluating generative models, defined as

$$\mathrm{IS} := \exp\left\{\mathbb{E}_{X_{\mathrm{g}}\sim p_{\mathrm{g}}}[\mathrm{KL}(p_{Y|X_{\mathrm{g}}}\|p_{Y_{\mathrm{g}}})]\right\}, \qquad (1)$$

where $X_{\mathrm{g}} \sim p_{\mathrm{g}}$ is a generated image, $p_{Y|X_{\mathrm{g}}}$ is the conditional class distribution assigned by the InceptionNet.V3 pretrained on ImageNet (Szegedy et al., 2016),

and $p_{Y_g} = \mathbb{E}_{X_g \sim p_g}[p_{Y|X_g}]$ is the marginal class distribution. Further, we have that IS $= \exp[I(X_g; Y_g)] = \exp[H(Y_g) - H(Y_g|X_g)]$, where $I(\cdot;\cdot)$ is the mutual information, and $H(\cdot)$ is the Shannon entropy. A higher IS implies that the images generated by the model have higher diversity, since p_{Y_g} would be more uniformly distributed to increase $H(Y_g)$, and possess higher fidelity, because $p_{Y|X_g}$ is closer to a one-hot vector to enforce a smaller $H(Y_g|X_g)$.

3.2 Fréchet Distance

Fréchet Distance (FD) (Dowson and Landau, 1982) is another standard metric for evaluating generative models. Let $f(x) \in \mathbb{R}^d$ denote the d-dimensional feature of an image x. The FD between the feature distributions of the generated images $p_{f(X_g)}$ and the real images $p_{f(X_r)}$, which is computed by

$$\text{FD} = \|\mu_g - \mu_r\|_2^2 + \text{Tr}[\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{\frac{1}{2}}]. \quad (2)$$

The well-known FID metric (Heusel et al., 2017) is computed as the FD where the image data feature is extracted by InceptionNet.V3 (Szegedy et al., 2016). Recently, Kynkäänniemi et al. (2023); Stein et al. (2023) propose computing the FD distance using the CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) embeddings, respectively, to boost the score's ranking consistency with human evaluations.

4 ONLINE EVALUATION OF GENERATIVE MODELS

In this section, we introduce the framework of online evaluation of generative models, which is given in Protocol 1. We denote by $\mathcal{G} := [G]$ the set of generative models. For each generator $g \in [G]$, we denote by $p_g \in \Delta(\mathcal{X})$ its generative distribution over the space $\mathcal{X}$, which can be texts or images. Given an evaluation metric, e.g., FD or IS, the corresponding *score* of the generator $g \in [G]$ is denoted by ν_g. The evaluation proceeds in T steps. At each step $t \in [T]$, the evaluating algorithm $\mathcal{A}$ picks a generator $g_t \in [G]$ and collects a batch of generated samples $\{x_t^j \sim p_g\}_{j=1}^b$, where $b \in \mathbb{N}_+$ is the (fixed) batch size. The evaluator aims to minimize the *regret*

$$\text{Regret}(T) = \sum_{t=1}^{T}(\nu^* - \nu_{g_t}), \quad (3)$$

where $\nu^* := \arg\max_{g\in[G]} \nu_g$ (if the higher the score the better).

Regarding the challenges of online evaluation of generative models, note that the empirical estimation of the score could be biased and generator-dependent.

Protocol 1 Online Evaluation and Selection of Generative Models

Require: step $T \in \mathbb{N}_+$, evaluation metric, a set $\mathcal{G} \leftarrow [G]$ of generators for evaluation, evaluator $\mathcal{A}$, batch size $b \in \mathbb{N}_+$

1: **Initialize:** estimated score $\widehat{\nu}_g$ and generated samples $\mathcal{H}_g \leftarrow \varnothing$ for each $g \in [G]$.
2: **for** step $t = 1, 2, \cdots, T$ **do**
3: Pick generator $g_t \sim \mathcal{A}$.
4: Generate a batch $\{x_t^j \sim p_{g_t}\}_{j=1}^b$ of samples from g_t.
5: Update $\mathcal{H}_{g_t} \leftarrow \mathcal{H}_{g_t} \cup \{x_t^j\}_{j=1}^b$ and estimated score $\widehat{\nu}_{g_t}$.
6: **end for**

In addition, the generative distribution typically lies in a high-dimensional space, which makes it difficult to estimate the score from limited data. Moreover, the evaluation metrics often incorporate higher-order information of the generative distribution (e.g., computing FD requires the covariance matrix). Hence, analyzing the concentration properties of the score estimation would be challenging.

Use of regret metric. The choice of regret metric 3 follows the standard online learning literature for bandit problems (Bubeck and Cesa-Bianchi, 2012). Particularly, if the evaluator can attain a *sub-linear* regret, then the overall selection converges to the best model with high probability. Additionally, in the online evaluation task, it is possible that the best model has a very similar performance to that of the second-best model. In such a case, the regret would take the difference between the models into account, and in a case where the top-k models have similar scores, generating samples from them would weighted more equally. On the other hand, if the best two models have significantly different scores, the regret would sharply grow by picking the suboptimal models.

5 FRÉCHET DISTANCE-BASED ONLINE EVALUATION AND SELECTION

In this section, we consider the online evaluation of generative models by Fréchet Distance (FD). Given $n \in \mathbb{N}_+$ generated images $x^1, \cdots, x^n \sim p_g$ queried from model g, the empirical FD is computed by

$$\widetilde{\text{FD}}_g^n = \|\widehat{\mu}_g^n - \mu_r\|_2^2 + \text{Tr}[\widehat{\Sigma}_g^n + \Sigma_r - 2(\widehat{\Sigma}_g^n\Sigma_r)^{\frac{1}{2}}], \quad (4)$$

where

$$\widehat{\mu}_g^n = \frac{1}{n}\sum_{i=1}^{n} f(x^i),\ \widehat{\Sigma}_g^n = \frac{1}{n}\sum_{i=1}^{n}(f(x^i)-\widehat{\mu}_g^n)(f(x^i)-\widehat{\mu}_g^n)^\top$$

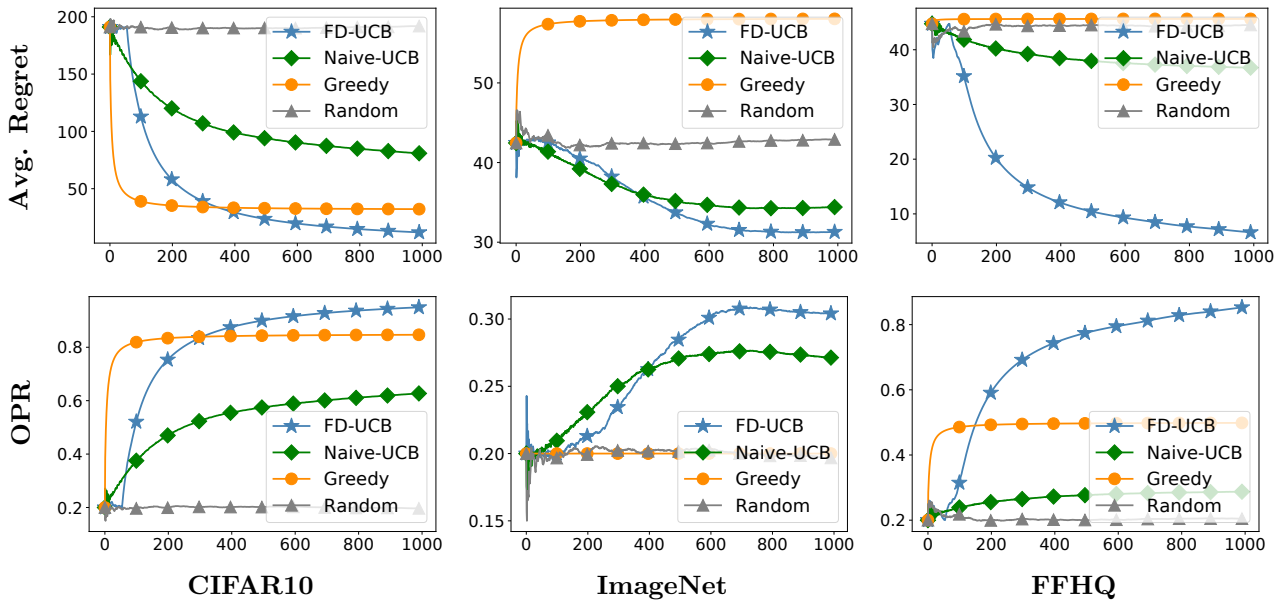

Figure 2: Online FD-based evaluation and selection among standard generative models: The x-axis is the number of online steps. At each step, the algorithm samples a batch of five generated images from the chosen model. The image data embeddings are extracted by CLIP (Cherti et al., 2023). Results are averaged over 20 trials.

are the mean vector and the empirical covariance matrix, respectively, and $f(x^i) \in \mathbb{R}^d$ is the feature of the i-th generated image which is extracted by, e.g., the InceptionNet.V3. The FD-based evaluation is typically performed on a large batch of generated samples (10-50 thousands) to reduce the estimation variance, which can be sample-inefficient and costly.

To enable sample-efficient online evaluation of generative models, we adapt the optimism-based online learning framework to the FD score. To this end, we first derive an optimistic FD score in the following theorem. We defer the theorem's proof to Appendix A.1.

Theorem 1 (Optimistic FD score)**.** *Assume for any generator g, the (random) embedding $f(X_g) \sim \mathcal{N}(\mu_g, \Sigma_g)$ follows a multivariate Gaussian, and the covariance matrix $\Sigma_{\rm r}$ of the real data is positive definite. Then, with probability at least $1-\delta$, we have*

$$\widehat{\mathrm{FD}}_g^n = \widetilde{\mathrm{FD}}_g^n - \mathcal{B}_g^n \le \mathrm{FD}_g \tag{5}$$

for any $n \ge 4\boldsymbol{r}(\Sigma_g) + \log(3/\delta)$, where $\boldsymbol{r}(\Sigma_g) := \frac{\mathrm{Tr}[\Sigma_g]}{\|\Sigma_g\|_2}$ is the effective rank of the covariance matrix Σ_g. In addition, the bonus is given by

$$\begin{aligned}\mathcal{B}_g^n :=& 2\Delta_{\mu_g}^n \cdot \left(\Delta_{\mu_g}^n + \|\widehat{\mu}_g^n - \mu_{\rm r}\|_2\right) + \mathrm{Tr}[\Sigma_{\rm r}^{\frac{1}{2}}]\sqrt{8\Delta_{\Sigma_g}^n} \\ &+ \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right),\end{aligned} \tag{6}$$

where $\Delta_{\mu_g}^n$ and $\Delta_{\Sigma_g}^n$ are defined in (23) and (24), respectively.

Remark 1 (Model-dependent parameters). The bonus term (6) involves model-dependent parameters $\mathrm{Tr}[\Sigma_g]$ and $\|\Sigma_g\|_2$ of the covariance matrix. Note that the presence of the norms of Σ_g in the concentration bound for FD is inevitable, because, without any assumptions on the norm of Σ_g, the concentration bound cannot have a finite value. To address this, two approaches can be considered: 1) when the norm of the data embeddings is bounded, e.g., the normalized CLIP embeddings where $\|X_g\|_2 \le 1$, we can substitute $\mathrm{Tr}[\Sigma_g]$ and $\|\Sigma_g\|_2$ in the above bound with the embedding upper-bound norm. In this case, we can extend our analysis beyond multivariate Gaussians (Theorem 5 in the Appendix). 2) On the other hand, when using the InceptionV3 and DINOv2 embeddings that have generally unbounded output norm, we can use the already generated data to estimate the parameters in the upper-bound. This approach avoids any offline estimation of the parameters required to have an upper-bound and preserves the online format of the FD-UCB algorithm, and conditioned that the estimated norms converge fast to the underlying value, will provide a correct solution. Our numerical results indicate that with only 50 samples, the norm terms in the bound will be relatively close to their actual values. See Part 3 in Appendix C for our experimental validation.

Based on Theorem 1, we propose FD-UCB in Algorithm 1 as an FD-based online evaluation algorithm. At the beginning, FD-UCB samples N generated data from each model, after which the algorithm can com-

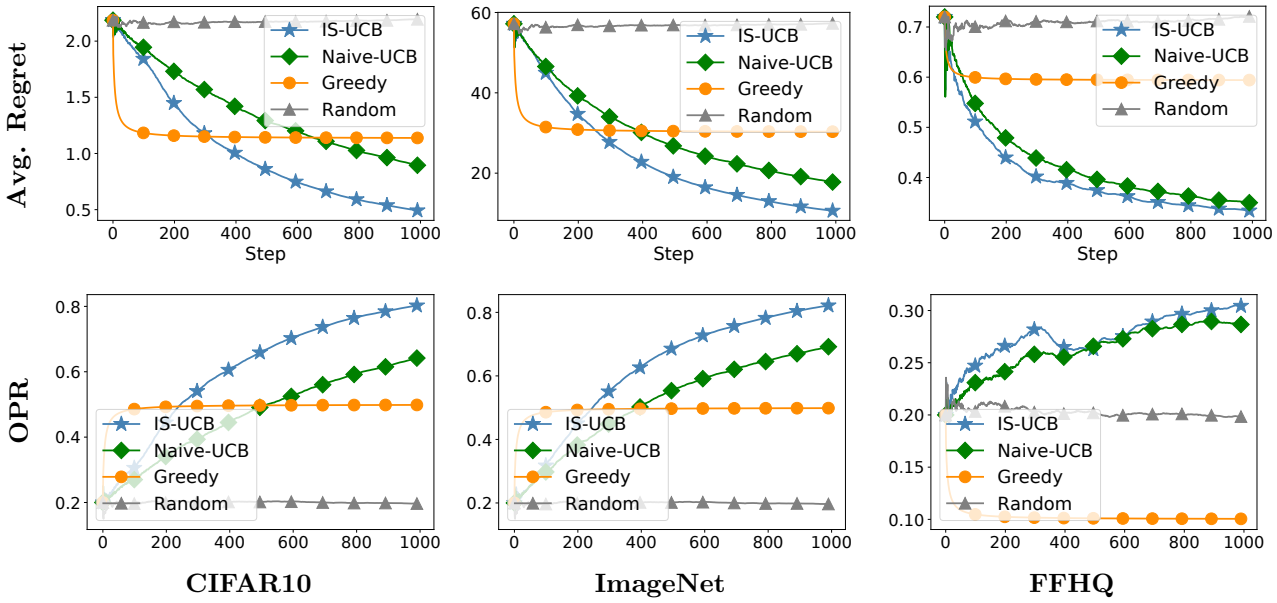

Figure 3: Online IS-based evaluation and selection among standard generative models: The x-axis is the number of steps. At each step, the algorithm samples a batch of five generated images from the chosen model. Results are averaged over 20 trials.

Algorithm 1 FD-UCB

Require: set of models $\mathcal{G}$, image embedding f, step $T \in \mathbb{N}_+$, $\mu_{\mathrm{r}} \in \mathbb{R}^d$ and $\Sigma_{\mathrm{r}} \in \mathbb{S}^d$ of real data, failure probability $\delta \leftarrow \delta/T$, batch size $b \in \mathbb{N}_+$, number of burn-in samples $N \geq \max_{g\in\mathcal{G}}\{4\boldsymbol{r}(\Sigma_g) + \log(3/\delta)\}$.

1: **Initialize:** $\widehat{\mathrm{FD}}_g \leftarrow -\infty, \mathcal{H}_g \leftarrow \varnothing, n_g \leftarrow 0$ for each $g \in \mathcal{G}$.
2: **for** model $g \in \mathcal{G}$ **do**
3: Generate M samples $\{x_0^j \sim p_g\}_{j=1}^N$.
4: Update $\mathcal{H}_g \leftarrow \mathcal{H}_g \cup \{f(x_0^j)\}_{j=1}^N$ and visitation $n_g \leftarrow N$.
5: **end for**
6: **for** step $t = 1, 2, \cdots, T$ **do**
7: Pick model $g_t \leftarrow \arg\min_{g\in\mathcal{G}} \widehat{\mathrm{FD}}_g$.
8: Model g_t generates batch samples $\{x_t^j \sim g_t\}_{j=1}^b$.
9: Update $\mathcal{H}_{g_t} \leftarrow \mathcal{H}_{g_t} \cup \{f(x_t^j)\}_{j=1}^b$ and visitation $n_{g_t} \leftarrow n_{g_t} + b$.
10: Update $\widehat{\mu}_{g_t}$ and $\widehat{\Sigma}_{g_t}$.
11: Compute the optimistic FD score

$$\widehat{\mathrm{FD}}_{g_t} \leftarrow -\mathcal{B}_{g_t}^{n_{g_t}} + \|\widehat{\mu}_{g_t} - \mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}\left[\widehat{\Sigma}_{g_t} + \Sigma_{\mathrm{r}} - 2(\widehat{\Sigma}_{g_t}\Sigma_{\mathrm{r}})^{\frac{1}{2}}\right],$$

where $\mathcal{B}_{gt}^n$ is given by Equation (6).
12: **end for**

pute an optimist FD score (5) for each model (lines 2-5). The evaluation proceeds iteratively, where at each iteration $t \in [T]$, the evaluator picks model g_t with the lowest estimated FD and queries a batch of images from the model (lines 7-8). Then, the estimated FD of generator g_t is updated (line 11). It can be shown that FD-UCB attains sub-linear regret bound, which is formalized in Theorem 3 in Appendix A.2.

6 INCEPTION SCORE-BASED ONLINE EVALUATION AND SELECTION

In this section, we focus on evaluating generative models by Inception score (IS), which is given by $\mathrm{IS} = \exp[I(X_g; Y_g)]$, where $X_g \sim p_g$ is the generated image, and Y_g is the class assigned by the InceptionNet.V3. Given $n \in \mathbb{N}_+$ generated images $x^1, \cdots, x^n \sim p_g$ queried from a generator g, we denote by

$$\widehat{H}^n(Y_g) = H\left(\frac{1}{n}\sum_{i=1}^n p_{Y|x^i}\right), \tag{7}$$

$$\widehat{H}^n(Y_g|X_g) = \frac{1}{n}\sum_{i=1}^n H(Y|x^i), \tag{8}$$

the empirical entropy of the marginal d-class distribution p_{Y_g} and the conditional d-class distribution $p_{Y|X_g}$, respectively. Then, the empirical IS is computed by

$$\widetilde{\mathrm{IS}}_g^n = \exp\left\{\widehat{H}^n(Y_g) - \widehat{H}^n(Y_g|X_g)\right\}. \tag{9}$$

For any $j \in [d]$, let $\widehat{V}^n(p_{Y|X_g}[j]) := \mathbb{V}(p_{Y|x^1}[j], \cdots, p_{Y|x^n}[j])$ denote the sample variance for the probability that the generated sample is

assigned to the j-th class. To derive an optimistic IS, we first define the *optimistic marginal class distribution* denoted by

$$\widehat{p}^n_{Y_g} := \mathrm{Clip}_{e^{-1}}(\widetilde{p}^n_{Y_g}, \boldsymbol{\epsilon}^n_g) \tag{10}$$

where $\boldsymbol{\epsilon}^n_g \in \mathbb{R}^d_+$ is a d-dimensional vector whose j-th element is given by

$$\begin{aligned}\boldsymbol{\epsilon}^n_g[j] :=& \sqrt{\frac{2\widehat{V}^n(p_{Y|X_g}[j])}{n}\log\left(\frac{4d}{\delta}\right)} \\ &+ \frac{7}{3(n-1)}\log\left(\frac{4d}{\delta}\right),\end{aligned} \tag{11}$$

and $\mathrm{Clip}_{e^{-1}}(p, \boldsymbol{\epsilon})$ is the following element-wise operator

$$\begin{cases} p[j] + \frac{e^{-1}-p[j]}{|e^{-1}-p[j]|}\boldsymbol{\epsilon}[j] & \text{, if } |e^{-1}-p[j]| \geq \boldsymbol{\epsilon}[j] \\ e^{-1} & \text{, otherwise.}\end{cases} \tag{12}$$

Particularly, the optimistic marginal class distribution ensures that $\mathcal{E}(\widehat{p}^n_{Y_g}) \geq H(Y_g)$ with high probability (see Lemma 11 in the Appendix), where we define

$$\mathcal{E}(z) := -\sum_j z[j]\cdot\log(z[j])$$

for any d-dimensional vector $z \succ \mathbf{0}$. Next, we derive a generator-dependent optimistic IS in the following theorem. We defer the theorem's proof to Appendix B.1.

Theorem 2 (Optimistic IS)**.** *Let*

$$\begin{aligned}\widehat{\mathrm{IS}}^n_g := \exp\Bigg\{&\mathcal{E}\left(\widehat{p}^n_{Y_g}\right) - \widehat{H}^n(Y_g|X_g) \\ &+ \sqrt{\frac{2\widehat{V}^n(H(Y_g|X_g))}{n}\log\left(\frac{4d}{\delta}\right)} \\ &+ \frac{7\log d}{3(n-1)}\log\left(\frac{4d}{\delta}\right)\Bigg\},\end{aligned} \tag{13}$$

where $\widehat{H}^n(Y_g|X_g)$ is defined in Equation (8), $\widehat{p}^n_{Y_g}$ the optimistic marginal class distribution (10), and $\widehat{V}^n(H(Y_g|X_g)) = \mathbb{V}(H(Y|x^1), \cdots, H(Y|x^n))$ is the empirical variance for the entropy of the conditional class distribution. Then, with probability at least $1-\delta$, we have that $\widehat{\mathrm{IS}}^n_g \geq \mathrm{IS}_g$.

Based on Theorem 2, we propose IS-UCB in Algorithm 2, an optimism-based online IS evaluation algorithm. Due to limited space in the main text, we derive the regret bound of the IS-UCB algorithm in Appendix B.2, which shows that IS-UCB attains $\widetilde{O}(\sqrt{T})$ regret.

Algorithm 2 IS-UCB

Require: steps T, set of generative models $\mathcal{G} \leftarrow [G]$, batch size $b \in \mathbb{N}_+$, failure probability $\delta \leftarrow \delta/T$.

1: **Initialize:** $\widehat{\mathrm{IS}}_g \leftarrow \infty, \widehat{H}(Y_g|X_g) \leftarrow 0, \widetilde{p}_{Y_g} \leftarrow \mathbf{0}^d, n_g \leftarrow 0$ for each $g \in [G]$.
2: **for** step $t = 1, 2, \cdots, T$ **do**
3: Pick model $g_t \leftarrow \arg\max_{g\in\mathcal{G}} \widehat{\mathrm{IS}}_g$.
4: Model g_t generates batch samples $\{x^j_t \sim p_{g_t}\}_{j=1}^b$.
5: Update

$$\widetilde{p}_{Y_{g_t}} \leftarrow \frac{N_g\cdot\widetilde{p}_{Y_g} + \sum_{j=1}^b p_{Y|x^i_t}}{n_{g_t}+b},$$

$$\widehat{H}(Y_{g_t}|X_{g_t}) \leftarrow \frac{N_g\cdot\widehat{H}(Y_{g_t}|X_{g_t}) + \sum_{j=1}^b H(Y|x^i_t)}{n_{g_t}+b}.$$

6: Update visitation $N_{g_t} \leftarrow N_{g_t} + b$.
7: Update the empirical variances $\widehat{V}(H(Y_{g_t}|X_{g_t}))$ and $\widehat{V}(p_{Y|X_{g_t}}[j])$ for any $j \in [d]$.
8: Compute the estimated IS according to Equation (13).
9: **end for**

7 EXPERIMENTAL RESULTS

In this section, we present numerical results for the FD-UCB and IS-UCB algorithms. Our extensive experiments across standard image datasets, generative modeling frameworks, and image data embeddings demonstrate the effectiveness of the proposed online learning-based approaches to model evaluation and selection. Experimental details and additional results can be found in Appendix C.

Baselines. For both FD-based and IS-based evaluation, we compare the performance of our proposed FD-UCB and IS-UCB with three baselines: Naive-UCB, Greedy algorithm, and Random algorithm. Naive-UCB is a simplification of FD-UCB and IS-UCB which replaces the generator-dependent variables in the bonus function with data-independent and dimension-based terms, which does not exploit the generated data in evaluating the confidence bound. On the other hand, the Greedy algorithm always picks the generator with the lowest empirical FD (4) or the highest empirical IS (9), and the Random algorithm selects the model randomly. For a fair comparison, we set the burn-in samples of FD-UCB to be zero. Additionally, for all these algorithms, each generator will be explored once at the beginning.

Experimental setup and performance metrics. We report two performance metrics at each step $t \in [T]$: 1) *average regret* (Avg. Regret), i.e., $(1/t)\cdot$

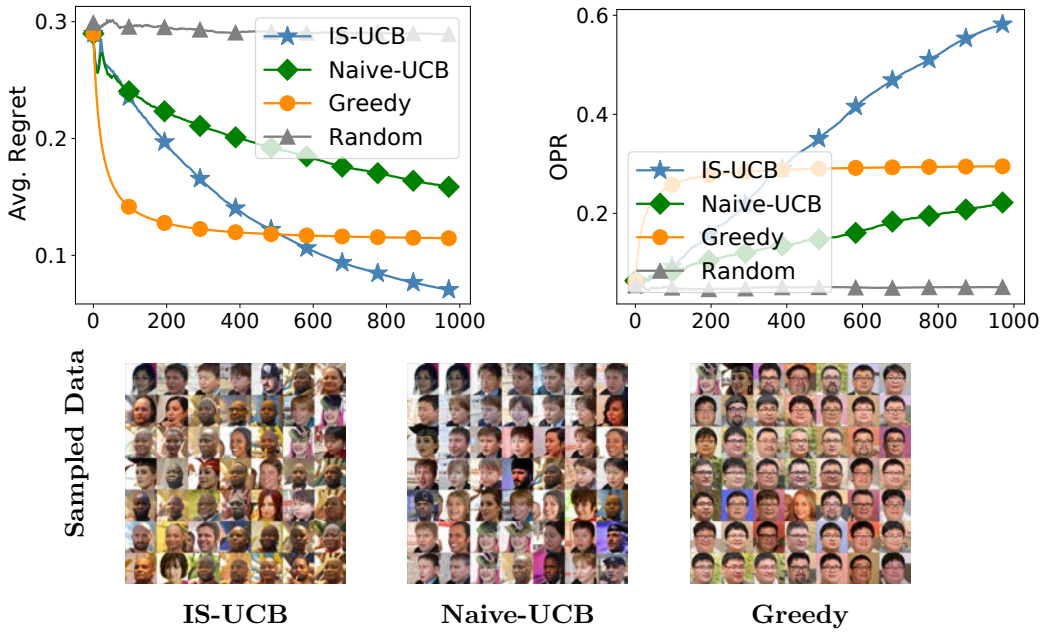

Figure 4: Online IS-based evaluation and selection among variance-controlled FFHQ models: IS-UCB can identify models that generate images with more diversity. Results are averaged over 20 trials.

Regret(t), and 2) *optimal pick ratio* (OPR), i.e., the ratio $(1/t) \cdot \sum_{i=1}^{t} \mathbf{1}(g_t = g^*)$ of picking the optimal model, which has the lowest empirical FD score or the highest empirical IS score for 50k generated images. For all the experiments, we use a batch size of 5, and the total evaluation step is $T = 1,000$. Hence the total generated samples for each trial is 5k. Results are averaged over 20 trials.

Datasets and generators. We evaluate the above algorithms on standard image datasets, including CIFAR10 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009), FFHQ (Kazemi and Sullivan, 2014), and AFHQ (Choi et al., 2020). For the first three datasets, we consider evaluation and selection among standard generative models, such as BigGAN-Deep (Brock et al., 2019), iDDPM-DDIM (Nichol and Dhariwal, 2021), and Efficient-vdVAE (Hazami et al., 2022). Details of the chosen models can be found in Appendix C. Additionally, we consider variance-controlled generators for the FFHQ and AFHQ datasets, where we apply the standard truncation technique (Kynkäänniemi et al., 2019) to the pretrained StyleGAN2-ADA model (Karras et al., 2020)[1] and synthesize $G = 20$ models. Each model generates images from truncated random noises with parameters vary from 0.01 to 0.1. A small (large) truncation parameter can lead to generated samples with low (high) diversity but high (low) quality.

[1]The repository can be found at `github.com/NVlabs/stylegan2-ada-pytorch`

7.1 Results of Online FD-based Evaluation and Selection

FID-based evaluation and selection. We first report results for the setup in Figure 1, where we select among three standard generative models on the CIFAR10 dataset, including LOGAN (Wu et al., 2020), RESFLOW (Chen et al., 2019), and iDDPM-DDiM (Nichol and Dhariwal, 2021). Our proposed FD-UCB can quickly identify the best model with fewer queries to the suboptimal models (Figure 5).

Performance on different embeddings. The results for CLIP embeddings are summarized in Figure 2, where each column and each row correspond to one dataset and one performance metric, respectively. The results show that FD-UCB outperforms Naive-UCB and the Greedy algorithm. We observe that naive-UCB converges to the best model much slower than FD-UCB, which suggests that the generator-dependent and data-driven bonus function can better exploit the properties of the generator, which is key to the sample-efficient online evaluation. We also test the InceptionNet.V3 and DINOv2 embeddings, and the results are summarized in Figures 7 and 8 in the Appendix. FD-UCB can maintain satisfactory performance over different embeddings.

Performance on variance-controlled generators. We summarize the results on the FFHQ and AFHQ-Dog datasets in Figures 9, 10, and 11 in the Appendix. FD-UCB consistently outperformed the baselines.

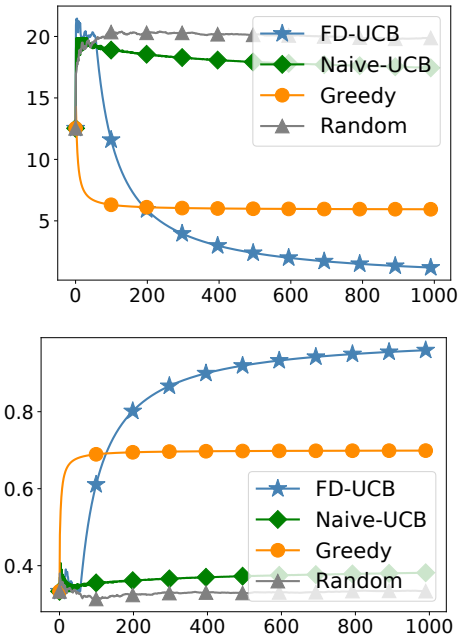

Figure 5: Online FD-based evaluation and selection among three CIFAR10 models, including LOGAN, RESFLOW, and iDDPM-DDiM (Figure 1). The image data embeddings are extracted by InceptionV3.Net. Results are averaged over 20 trials.

Performance on video and audio data. We consider the *Fréchet Video Distance* (Unterthiner et al., 2019, FVD) and *Fréchet Audio Distance* (Kilgour et al., 2019, FAD) metrics for video and audio generation, respectively. We synthesize four arms (generators) utilizing the MSR-VTT (Xu et al., 2016) and Magnatagatune (Law et al., 2009) datasets. When an arm is selected, a video/audio clip is sampled from the dataset and perturbed by Gaussian noises with an arm-specific probability. Results on the audio data are summarized in Figure 6. Additional results can be found in Figure 12 in the Appendix.

7.2 Results of Online IS-Based Evaluation and Selection

Performance on the pretrained generators. The results are summarized in Figure 3. IS-UCB (blue) attains significantly better performance than naive-UCB (orange) and the greedy algorithm (green) on CIFAR10 and FFHQ datasets.

Performance on variance-controlled generators. The results are summarized in Figure 4. The results suggest that FD-UCB consistently queries images that are diverse and of high quality. In contrast, Naive-UCB and the Greedy algorithm kept querying a proportion of images from the collapsed generators

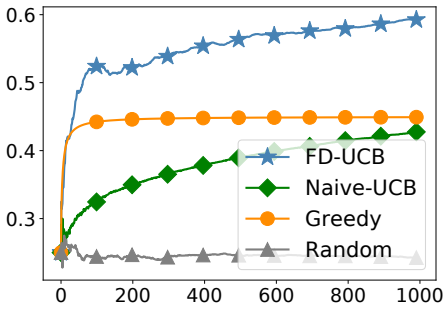

Figure 6: Online FAD-based evaluation and selection among four synthetic audio data. The embeddings are extracted by VGGish (Hershey et al., 2017). Results for the OPR metric are reported and averaged over 20 trials.

with smaller truncation parameters. The AFHQ-Dog dataset results are in the Appendix, Figure 13.

8 CONCLUSION

In this work, we studied an online learning problem aiming to identify the generative model with the best evaluation score among a set of models. We proposed the multi-armed bandit-based FD-UCB and IS-UCB algorithms to address the online learning task and showed satisfactory numerical results for their application to image-based generative models. An interesting future direction to our work is to explore the application of other online learning frameworks, such as Thompson sampling, to address the online evaluation problem. Also, while we mostly focused on standard image datasets for testing the proposed algorithms, the application of the proposed FD-UCB and IS-UCB to real text-based and video-based generative models will be interesting for future studies. Such an application only requires applying text and video-based embeddings, as in Video IS (Saito et al., 2020), FVD (Unterthiner et al., 2019), and FBD (Xiang et al., 2021). Also, the extension of our MAB framework to a contextual bandit algorithm for prompt-guided generative models, as studied by Hu et al. (2024), will remain for future studies.

Acknowledgments

The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054. The authors would also like to thank the anonymous reviewers for their constructive feedback and suggestions.

References

Agrawal, R. (1995). Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3(null):397–422.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

Bond-Taylor, S., Hessey, P., Sasaki, H., Breckon, T. P., and Willcocks, C. G. (2021). Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation.

Carreira, J. and Zisserman, A. (2018). Quo vadis, action recognition? a new model and the kinetics dataset.

Chen, R. T. Q., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. (2019). Residual flows for invertible generative modeling. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training GANs with optimism. In *International Conference on Learning Representations*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

Dowson, D. and Landau, B. (1982). The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

Friedman, D. and Dieng, A. B. (2023). The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P. W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., and Schmidt, L. (2023). Datacomp: In search of the next generation of multimodal datasets.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. (2018). An online learning approach to generative adversarial networks. In *International Conference on Learning Representations*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of

wasserstein gans. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. (2023). Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *The Eleventh International Conference on Learning Representations*.

Hazami, L., Mama, R., and Thurairatnam, R. (2022). Efficient-vdvae: Less is more.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hu, X., Leung, H.-f., and Farnia, F. (2024). An online learning approach to prompt-based selection of generative models. *arXiv preprint arXiv:2410.13287*.

Jalali, M., Li, C. T., and Farnia, F. (2023). An information-theoretic evaluation of generative models in learning multi-modal distributions. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 9931–9943. Curran Associates, Inc.

Jiralerspong, M., Bose, J., Gemp, I., Qin, C., Bachrach, Y., and Gidel, G. (2023). Feature likelihood score: Evaluating the generalization of generative models using samples. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. (2019). Fréchet audio distance: A metric for evaluating music enhancement algorithms.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. (2023). The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved precision and recall metric for assessing generative models. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. (2022). Autoregressive image generation using residual quantization.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. (2020). Reliable fidelity and diversity metrics for generative models. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7176–7185. PMLR.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G.,

Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision.

Ospanov, A., Zhang, J., Jalali, M., Cao, X., Bogdanov, A., and Farnia, F. (2024). Towards a scalable reference-free evaluation of generative models. *arXiv preprint arXiv:2407.02961*.

Park, C., Liu, X., Ozdaglar, A., and Zhang, K. (2024). Do llm agents have regret? a case study in online learning and games.

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Rezaei, P., Farnia, F., and Li, C. T. (2025). Be more diverse than the most diverse: Optimal mixtures of generative models via mixture-UCB bandit algorithms. In *The Thirteenth International Conference on Learning Representations*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Saito, M., Saito, S., Koyama, M., and Kobayashi, S. (2020). Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10–11):2586–2606.

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved techniques for training gans. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Sani, A., Lazaric, A., and Munos, R. (2012). Risk-aversion in multi-armed bandits. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. (2023). Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3732–3784. Curran Associates, Inc.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2019). FVD: A new metric for video generation.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc.

Walton, S., Hassani, A., Xu, X., Wang, Z., and Shi, H. (2023). Stylenat: Giving each head a new perspective.

Weinberger, N. and Yemini, M. (2023). Multi-armed bandits with self-information rewards. *IEEE Transactions on Information Theory*, 69(11):7160–7184.

Wu, Y., Donahue, J., Balduzzi, D., Simonyan, K., and Lillicrap, T. (2020). Logan: Latent optimisation for generative adversarial networks.

Xiang, J., Liu, Y., Cai, D., Li, H., Lian, D., and Liu, L. (2021). Assessing dialogue systems with distribution distances. *arXiv preprint arXiv:2105.02573*.

Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Zhang, J., Jalali, M., Li, C. T., and Farnia, F. (2024a). Identification of novel modes in generative models via fourier-based differential clustering. *arXiv preprint arXiv:2405.02700*.

Zhang, J., Li, C. T., and Farnia, F. (2024b). An interpretable evaluation of entropy-based novelty of generative models. In *International Conference on Machine Learning (ICML 2024)*.

Zhivotovskiy, N. (2022). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle.

Zhu, Q. and Tan, V. (2020). Thompson sampling algorithms for mean-variance bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11599–11608. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS IN SECTION 5: FD-BASED EVALUATION

A.1 Proof of Theorem 1: Optimistic FD Score

Proof. The proof of Theorem 1 is based on Theorem 4 and Lemma 1 in the Appendix. Specifically, Inequality (18) in Theorem 4 in Appendix A.3 decomposes the estimation error of the empirical FD score (4) into the errors in estimating the mean (i.e., $\|\widehat{\mu}_g^n - \mu_g\|_2$) and the covariance matrix (i.e., $\|\widehat{\Sigma}_g^n - \Sigma_g\|_2$): With probability at least $1 - \frac{\delta}{3}$, it holds that

$$\begin{aligned}\left|\widetilde{\mathrm{FD}}_g^n - \mathrm{FD}_g\right| \leq & 2\|\mu_g - \widehat{\mu}_g^n\|_2 \cdot \left(\|\mu_g - \widehat{\mu}_g^n\|_2 + \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2\right) + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\|\Sigma_g - \widehat{\Sigma}_g^n\|_2} \\ & + \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right).\end{aligned}$$

Next, Lemma 1 in Appendix A.4 derives generator-dependent concentration errors for the mean and the covariance matrix: With probability at least $1 - \frac{2\delta}{3}$, it holds that

$$\begin{aligned}\|\widehat{\mu}_g^n - \mu_g\|_2 &\leq \sqrt{\frac{1}{n}\left(\sqrt{8 \cdot \mathrm{Tr}[\Sigma_g^2]\log\left(\frac{6}{\delta}\right)} + 8\|\Sigma_g\|_2 \log\left(\frac{6}{\delta}\right)\right)} =: \Delta_{\mu_g}^n \\ \left\|\Sigma_g - \widehat{\Sigma}_g^n\right\|_2 &\leq 20\kappa^2\|\Sigma_g\|_2\sqrt{\frac{4\boldsymbol{r}(\Sigma_g) + \log(3/\delta)}{n}} + (\Delta_{\mu_g}^n)^2 =: \Delta_{\Sigma_g}^n\end{aligned}$$

Therefore, let

$$\mathcal{B}_g^n = 2\Delta_{\mu_g}^n \cdot \left(\Delta_{\mu_g}^n + \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2\right) + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\Delta_{\Sigma_g}^n} + \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right),$$

and we have that $\widetilde{\mathrm{FD}}_g^n - \mathcal{B}_g^n \leq \mathrm{FD}_g$ holds with probability at least $1 - \delta$, which concludes the proof. □

A.2 Regret of FD-UCB

Theorem 3 (Regret of FD-UCB)**.** *Under the same conditions in Theorem 1, with probability at least $1 - \delta$, the regret of the FD-UCB algorithm with a batch size $b \in \mathbb{N}_+$ is bounded by*

$$\widetilde{O}\left(\frac{D_1^{\mathcal{G}} \cdot G^{1/4}}{b^{1/4}}T^{3/4} + \frac{D_2^{\mathcal{G}} \cdot G^{1/2}}{b^{1/2}}T^{1/2} + \frac{D_3^{\mathcal{G}} \cdot G}{b}\log T + N \cdot G\right), \tag{14}$$

after running on $G := |\mathcal{G}|$ models for $T \in \mathbb{N}_+$ steps, where $D_1^{\mathcal{G}}$ and $D_2^{\mathcal{G}}$ are model-dependent parameters defined in Equations (15)-(17), and the logarithmic factors are hidden in the notation $\widetilde{O}(\cdot)$.

Proof. Recall that g_t denotes the generator picked at the t-th step. For convenience, we denote by $\widehat{\mathrm{FD}}_{g_t}$ the optimistic FD of generator g_t computed at step t. We denote by n_t the *number of images* generated by model g_t at the beginning of the t-th step. First, by Theorem 1 and a union bound over T steps, with probability at least $1 - \delta$, we have that $\mathrm{FD}_{g_t} \geq \widehat{\mathrm{FD}}_{g_t}$ for any step $t \in [T]$ and $g \in [G]$. Hence, we have that

$$\mathrm{Regret}(T) = O(N \cdot G) + \sum_{t=1}^{T}\left(\mathrm{FD}_{g_t} - \mathrm{FD}^*\right) \leq O(N \cdot G) + \sum_{t=1}^{T}\left(\mathrm{FD}_{g_t} - \widehat{\mathrm{FD}}_{g_t}\right),$$

where $\mathrm{FD}^* = \min_g \mathrm{FD}_g$, and the term $O(N \cdot G)$ corresponds to the regret incurred in the burn-in sampling phase. Further, by the definition of $\widehat{\mathrm{FD}}_{g_t}$, we further derive that

$$\mathrm{Regret}(T) \leq O(N \cdot G) + \sum_{t=1}^{T}\left(\mathrm{FD}_{g_t} - \widetilde{\mathrm{FD}}_{g_t}^{n_t} + \mathcal{B}_{g_t}^{n_t}\right)$$

$$
\begin{aligned}
\leq & O(N \cdot G) + 2\sum_{t=1}^{T} \mathcal{B}_{g_t}^{n_t} \\
\leq & \widetilde{O}\left(N \cdot G + \sum_{t=1}^{T} \left(\Delta_{\mu_{g_t}}^{n_t} \cdot \left(\Delta_{\mu_{g_t}}^{n_t} + \|\mu_{g_t} - \mu_{\mathrm{r}}\|_2 \right) + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}] \sqrt{\Delta_{\Sigma_{g_t}}^{n_t}} + \mathrm{Tr}[\Sigma_{g_t}] \sqrt{\frac{1}{n_t}} + \frac{\|\Sigma_{g_t}\|_2}{n_t} \right) \right),
\end{aligned}
$$

where we utilize $\|\widehat{\mu}_{g_t}^{n_t} - \mu_{\mathrm{r}}\|_2 \leq \Delta_{\mu_{g_t}}^{n_t} + \|\mu_{g_t} - \mu_{\mathrm{r}}\|_2$ in the last inequality. Note that

$$
\Delta_{\mu_{g_t}}^{n_t} = \sqrt{\frac{1}{n_t} \left(\sqrt{8 \cdot \mathrm{Tr}[\Sigma_{g_t}^2] \log\left(\frac{6T}{\delta}\right)} + 8\|\Sigma_{g_t}\|_2 \log\left(\frac{6T}{\delta}\right) \right)} = \widetilde{O}\left(n_t^{-1/2} \sqrt{\sqrt{\mathrm{Tr}[\Sigma_{g_t}^2]} + \|\Sigma_{g_t}\|_2} \right)
$$

and

$$
\begin{aligned}
\Delta_{\Sigma_{g_t}}^{n_t} =& 20\kappa^2 \|\Sigma_{g_t}\|_2 \sqrt{\frac{4\boldsymbol{r}(\Sigma_{g_t}) + \log(3T/\delta)}{n_t}} + (\Delta_{\mu_{g_t}}^{n_t})^2 \\
=& \widetilde{O}\left(n_t^{-1/2} \sqrt{\|\Sigma_{g_t}\|_2^2 (\boldsymbol{r}(\Sigma_{g_t}) + 1)} + n_t^{-1} \left(\sqrt{\mathrm{Tr}[\Sigma_{g_t}^2]} + \|\Sigma_{g_t}\|_2 \right) \right).
\end{aligned}
$$

Hence, the regret is further bounded by

$$
\begin{aligned}
\widetilde{O}\Bigg(N \cdot G + & \sum_{t=1}^{T} n_t^{-1/4} \cdot \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}] \left(\|\Sigma_{g_t}\|_2^2 (\boldsymbol{r}(\Sigma_{g_t}) + 1) \right)^{1/4} \\
& + n_t^{-1/2} \cdot \left((\|\mu_{g_t} - \mu_{\mathrm{r}}\|_2 + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]) \left(\sqrt{\mathrm{Tr}[\Sigma_{g_t}^2]} + \|\Sigma_{g_t}\|_2 \right)^{1/2} + \mathrm{Tr}[\Sigma_{g_t}] \right) \\
& + n_t^{-1} \cdot \left(\sqrt{\mathrm{Tr}[\Sigma_{g_t}^2]} + \|\Sigma_{g_t}\|_2 \right) \Bigg).
\end{aligned}
$$

Let

$$
D_1^{\mathcal{G}} := \max_g \left\{ \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}] \left(\|\Sigma_g\|_2^2 (\boldsymbol{r}(\Sigma_g) + 1) \right)^{1/4} \right\}, \tag{15}
$$

$$
D_2^{\mathcal{G}} := \max_g \left\{ (\|\mu_g - \mu_{\mathrm{r}}\|_2 + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]) \left(\sqrt{\mathrm{Tr}[\Sigma_g^2]} + \|\Sigma_g\|_2 \right)^{1/2} + \mathrm{Tr}[\Sigma_g] \right\}, \tag{16}
$$

$$
D_3^{\mathcal{G}} := \max_g \left\{ \sqrt{\mathrm{Tr}[\Sigma_g^2]} + \|\Sigma_g\|_2 \right\}. \tag{17}
$$

Then, we have that

$$
\mathrm{Regret}(T) \leq \widetilde{O}\left(\frac{D_1^{\mathcal{G}} \cdot G^{1/4}}{b^{1/4}} T^{3/4} + \frac{D_2^{\mathcal{G}} \cdot G^{1/2}}{b^{1/2}} T^{1/2} + \frac{D_3^{\mathcal{G}} \cdot G}{b} \log T + N \cdot G \right)
$$

where we use the fact that

$$
\sum_{t=1}^{T} \frac{1}{n_t^{\alpha}} \leq \frac{2^{\alpha}}{b^{\alpha}(1-\alpha)} \sum_{g=1}^{G} (N_g(T))^{1-\alpha} \leq \frac{2^{\alpha}}{b^{\alpha}(1-\alpha)} G^{\alpha} T^{1-\alpha}, \text{ where } 0 \leq \alpha < 1,
$$

$$
\sum_{t=1}^{T} \frac{1}{n_t} \leq \frac{2}{b} \sum_{g=1}^{G} \log(N_g(T)) \leq \frac{2G}{b} \log T,
$$

where b is the batch size, and $N_g(T)$ is the number of picks of model g at the last step T. Therefore, we conclude the proof. □

A.3 Concentration of Empirical FD (4)

The following theorem expresses the estimation error of the empirical FD by the concentration of the sample mean and covariance matrix, which facilitates the derivation of the bonus function.

Theorem 4 (Concentration of empirical FD (4))**.** *Assume the covariance matrix $\Sigma_{\mathrm{r}} \succ \mathbf{0}$ is positive strictly definite. Then, with probability at least $1-\frac{\delta}{3}$, we have that*

$$\begin{aligned}\left|\widetilde{\mathrm{FD}}_g^n - \mathrm{FD}_g\right| \leq & 2\|\mu_g - \widehat{\mu}_g^n\|_2 \cdot \left(\|\mu_g - \widehat{\mu}_g^n\|_2 + \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2\right) + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\|\Sigma_g - \widehat{\Sigma}_g^n\|_2} \\ & + \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right).\end{aligned} \tag{18}$$

Proof. For any generator g and $n \in \mathbb{N}_+$, we have that

$$\begin{aligned}\mathrm{FD}_g &= \|\mu_g - \mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}[\Sigma_g + \Sigma_{\mathrm{r}} - 2(\Sigma_g\Sigma_{\mathrm{r}})^{\frac{1}{2}}], \\ \widetilde{\mathrm{FD}}_g^n &= \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}[\widehat{\Sigma}_g^n + \Sigma_{\mathrm{r}} - 2(\widehat{\Sigma}_g^n\Sigma_{\mathrm{r}})^{\frac{1}{2}}].\end{aligned}$$

(1) Bound $\|\mu_g - \mu_{\mathbf{r}}\|_2^2 - \|\widehat{\mu}_g^n - \mu_{\mathbf{r}}\|_2^2$. We derive

$$\begin{aligned}&\left|\|\mu_g - \mu_{\mathrm{r}}\|_2^2 - \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2^2\right| \\ =&(\|\mu_g - \mu_{\mathrm{r}}\|_2 + \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2) \cdot \left|\|\mu_g - \mu_{\mathrm{r}}\|_2 - \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2\right| \\ \leq&(\|\mu_g - \widehat{\mu}_g^n\|_2 + 2\|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2) \cdot \|\widehat{\mu}_g^n - \mu_g\|_2.\end{aligned} \tag{19}$$

(2) Bound $\mathbf{Tr}[(\Sigma_g\Sigma_{\mathbf{r}})^{\frac{1}{2}} - (\widehat{\Sigma}_g^n\Sigma_{\mathbf{r}})^{\frac{1}{2}}]$. By Lemma 2, if the covariance matrix Σ_{r} is positive strictly definite, then it holds that

$$\left|\mathrm{Tr}[(\Sigma_g\Sigma_{\mathrm{r}})^{\frac{1}{2}}] - \mathrm{Tr}\left[(\widehat{\Sigma}_g^n\Sigma_{\mathrm{r}})^{\frac{1}{2}}\right]\right| \leq \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{2\|\Sigma_g - \widehat{\Sigma}_g^n\|_2}. \tag{20}$$

(3) Bound $\mathbf{Tr}[\Sigma_g - \widehat{\Sigma}_g^n]$. Note that

$$\begin{aligned}\mathrm{Tr}[\widehat{\Sigma}_g^n] =&\mathrm{Tr}\left[\frac{1}{n}\sum_{i=1}^n (f(x^i) - \widehat{\mu}_g^n)(f(x^i) - \widehat{\mu}_g^n)^\top\right] \\ =&\mathrm{Tr}\left[\frac{1}{n}\sum_{i=1}^n (f(x^i) - \mu_g + \mu_g - \widehat{\mu}_g^n)(f(x^i) - \mu_g + \mu_g - \widehat{\mu}_g^n)^\top\right] \\ =&\mathrm{Tr}\left[\frac{1}{n}\sum_{i=1}^n (f(x^i) - \mu_g)(f(x^i) - \mu_g)^\top - (\mu_g - \widehat{\mu}_g^n)(\mu_g - \widehat{\mu}_g^n)^\top\right] \\ =&\frac{1}{n}\sum_{i=1}^n \|f(x^i) - \mu_g\|_2^2 - \|\mu_g - \widehat{\mu}_g^n\|_2^2\end{aligned} \tag{21}$$

Hence, we obtain

$$\left|\mathrm{Tr}[\Sigma_g] - \mathrm{Tr}[\widehat{\Sigma}_g^n]\right| \leq \left|\mathbb{E}[\|f(X_g) - \mu_g\|_2^2] - \frac{1}{n}\sum_{i=1}^n \|f(x^i) - \mu_g\|_2^2\right| + \|\mu_g - \widehat{\mu}_g^n\|_2^2 \tag{22}$$

Note that $f(X_g) - \mu_g \sim \mathcal{N}(0, \Sigma_g)$. By Lemma 4, with probability at least $1-\frac{\delta}{3}$, it holds that

$$\left|\frac{1}{n}\sum_{i=1}^n \|f(X^i) - \mu_g\|_2^2 - \mathbb{E}[\|f(X_g) - \mu_g\|_2^2]\right| \leq \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right).$$

Putting all together. Therefore, combining Inequalities (19), (20), and (22), with probability at least $1-\frac{\delta}{3}$, it holds that

$$\begin{aligned}
&\left|\widetilde{\mathrm{FD}}_g^n - \mathrm{FD}_g\right| \\
\le& \underbrace{(\|\mu_g - \widehat{\mu}_g^n\|_2 + 2\|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2)\cdot\|\widehat{\mu}_g^n - \mu_g\|_2}_{\text{(1) error of } \|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2^2} + \underbrace{\mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\|\Sigma_g - \widehat{\Sigma}_g^n\|_2}}_{\text{(2) error of } 2\mathrm{Tr}[(\widehat{\Sigma}_g^n\Sigma_{\mathrm{r}})^{\frac{1}{2}}]} \\
&+ \underbrace{\mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right) + \|\mu_g - \widehat{\mu}_g^n\|_2^2}_{\text{(3) error of } \mathrm{Tr}[\widehat{\Sigma}_g^n]} \\
=& 2\|\mu_g - \widehat{\mu}_g^n\|_2\cdot\left(\|\mu_g - \widehat{\mu}_g^n\|_2 + \|\widehat{\mu}_g^n - \mu_{\mathrm{r}}\|_2\right) + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\|\Sigma_g - \widehat{\Sigma}_g^n\|_2} \\
&+ \mathrm{Tr}[\Sigma_g]\sqrt{\frac{8}{n}\log\left(\frac{6}{\delta}\right)} + \frac{8\|\Sigma_g\|_2}{n}\log\left(\frac{6}{\delta}\right),
\end{aligned}$$

which concludes the proof. □

A.4 Concentration of Mean Vector and Covariance Matrix

Lemma 1 ($L2$-norm error for mean and covariance matrix)**.** *Under the same conditions in Theorem 4, with probability at least* $1-\frac{2\delta}{3}$*, we have that*

$$\|\widehat{\mu}_g^n - \mu_g\|_2 \le \sqrt{\frac{1}{n}\left(\sqrt{8\cdot\mathrm{Tr}[\Sigma_g^2]\log\left(\frac{6}{\delta}\right)} + 8\|\Sigma_g\|_2\log\left(\frac{6}{\delta}\right)\right)} =: \Delta_{\mu_g}^n \tag{23}$$

and

$$\left\|\Sigma_g - \widehat{\Sigma}_g^n\right\|_2 \le 20\kappa^2\|\Sigma_g\|_2\sqrt{\frac{4\boldsymbol{r}(\Sigma_g)+\log(3/\delta)}{n}} + (\Delta_{\mu_g}^n)^2 =: \Delta_{\Sigma_g}^n \tag{24}$$

for $n \ge 4\boldsymbol{r}(\Sigma_g)+\log(3/\delta)$*, where* $\boldsymbol{r}(\Sigma_g) = \frac{\mathrm{Tr}[\Sigma_g]}{\|\Sigma_g\|_2}$ *is the effective rank of* Σ_g*.*

Proof. **1. Concentration of squared $L2$-norm error.** Note that

$$\|\widehat{\mu}_g^n - \mu_g\|_2^2 = \underbrace{\|\widehat{\mu}_g^n - \mu_g\|_2^2 - \mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2]}_{\text{(*): concentration}} + \underbrace{\mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2]}_{\text{expected L2-norm error}},$$

where $\widehat{\mu}_g^n - \mu_g \sim \mathcal{N}(0, \frac{\Sigma_g}{n})$. By Lemma 4, we have $\mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2] = \frac{\mathrm{Tr}[\Sigma_g]}{n}$, and with probability at least $1-\frac{\delta}{3}$, it holds that

$$\left|\|\widehat{\mu}_g^n - \mu_g\|_2^2 - \mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2]\right| \le \frac{1}{n}\left(\sqrt{8\cdot\mathrm{Tr}[\Sigma_g^2]\log\left(\frac{6}{\delta}\right)} + 8\|\Sigma_g\|_2\log\left(\frac{6}{\delta}\right)\right).$$

2. Concentration of sample covariance matrix. We have that

$$\begin{aligned}
&\left\|\Sigma_g - \widehat{\Sigma}_g^n\right\|_2 \\
=& \left\|\Sigma_g - \frac{1}{n}\sum_{i=1}^n (f(x^i) - \widehat{\mu}_g^n)(f(x^i) - \widehat{\mu}_g^n)^\top\right\|_2 \\
=& \left\|\Sigma_g - \frac{1}{n}\sum_{i=1}^n (f(x^i) - \mu_g + \mu_g - \widehat{\mu}_g^n)(f(x^i) - \mu_g + \mu_g - \widehat{\mu}_g^n)^\top\right\|_2 \\
=& \left\|\Sigma_g - \frac{1}{n}\sum_{i=1}^n \left((f(x^i) - \mu_g)(f(x^i) - \mu_g)^\top - 2(f(x^i) - \mu_g)(\mu_g - \widehat{\mu}_g^n)^\top + (\mu_g - \widehat{\mu}_g^n)(\mu_g - \widehat{\mu}_g^n)^\top\right)\right\|_2
\end{aligned}$$

$$\leq \left\| \Sigma_g - \frac{1}{n}\sum_{i=1}^{n}(f(x^i)-\mu_g)(f(x^i)-\mu_g)^\top \right\|_2 + \left\|(\mu_g-\widehat{\mu}_g^n)(\mu_g-\widehat{\mu}_g^n)^\top\right\|_2. \tag{25}$$

For the first term, note that $f(X_g)-\mu_g \sim \mathcal{N}(0,\Sigma_g)$. Then, by Lemma 6, with probability at least $1-\frac{\delta}{3}$, it holds that

$$\left\| \Sigma_g - \frac{1}{n}\sum_{i=1}^{n}(f(x^i)-\mu_g)(f(x^i)-\mu_g)^\top \right\|_2 \leq 20\kappa^2\|\Sigma_g\|_2\sqrt{\frac{4\boldsymbol{r}(\Sigma_g)+\log(3/\delta)}{n}} \tag{26}$$

for $n \geq 4\boldsymbol{r}(\Sigma_g)+\log(3/\delta)$, where $\boldsymbol{r}(\Sigma_g)=\frac{\mathrm{Tr}[\Sigma_g]}{\|\Sigma_g\|_2}$ and κ is a constant defined therein. For the second term, we have that

$$\left\|(\mu_g-\widehat{\mu}_g^n)(\mu_g-\widehat{\mu}_g^n)^\top\right\|_2 = \|\mu_g-\widehat{\mu}_g^n\|_2^2 \leq (\Delta_{\mu_g}^n)^2,$$

which concludes the proof. □

A.5 Concentration Bound for Norm-Bounded Embeddings

We extend the concentration bound for norm-bounded embeddings.

Theorem 5. *Assume for any generator g, the (random) embedding $f(X_g)$ is L2-norm bounded, i.e., $\|f(X_g)\|_2 \leq C$ for some positive number $C>0$, and the covariance matrix Σ_{r} of the real data is positive definite. Then, with probability at least $1-\delta$, we have*

$$\widetilde{\mathrm{FD}}_g^n - \mathcal{C}_g^n \leq \mathrm{FD}_g,$$

where

$$\mathcal{C}_g^n := 2(\Delta_1^n + \|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2)\cdot\Delta_1^n + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\Delta_2^n} + 4C^2\sqrt{\frac{1}{2n}\log\left(\frac{3}{\delta}\right)},$$

and Δ_1^n and Δ_2^n are defined in (28) and (29), respectively.

Proof. The proof is similar to Theorem 1. Recall that for any generator g, we have that

$$\mathrm{FD}_g = \|\mu_g-\mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}[\Sigma_g+\Sigma_{\mathrm{r}}-2(\Sigma_g\Sigma_{\mathrm{r}})^{\frac{1}{2}}],$$
$$\widetilde{\mathrm{FD}}_g^n = \|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}[\widehat{\Sigma}_g^n+\Sigma_{\mathrm{r}}-2(\widehat{\Sigma}_g^n\Sigma_{\mathrm{r}})^{\frac{1}{2}}].$$

(1) Bound $\|\mu_g-\mu_{\mathbf{r}}\|_2^2 - \|\widehat{\mu}_g^n-\mu_{\mathbf{r}}\|_2^2$. We derive

$$\left|\|\mu_g-\mu_{\mathrm{r}}\|_2^2 - \|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2^2\right| \leq (\|\mu_g-\widehat{\mu}_g^n\|_2 + 2\|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2)\cdot\|\widehat{\mu}_g^n-\mu_g\|_2.$$

(2) Bound $\mathbf{Tr}[(\Sigma_g\Sigma_{\mathbf{r}})^{\frac{1}{2}} - (\widehat{\Sigma}_g^n\Sigma_{\mathbf{r}})^{\frac{1}{2}}]$. By Lemma 2, if the covariance matrix Σ_{r} is positive strictly definite, then it holds that

$$\left|\mathrm{Tr}[(\Sigma_g\Sigma_{\mathrm{r}})^{\frac{1}{2}}] - \mathrm{Tr}\left[(\widehat{\Sigma}_g^n\Sigma_{\mathrm{r}})^{\frac{1}{2}}\right]\right| \leq \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{2\|\Sigma_g-\widehat{\Sigma}_g^n\|_2}.$$

(3) Bound $\mathbf{Tr}[\Sigma_g-\widehat{\Sigma}_g^n]$. Note that

$$\left|\mathrm{Tr}[\Sigma_g] - \mathrm{Tr}[\widehat{\Sigma}_g^n]\right| \leq \left|\mathbb{E}[\|f(X_g)-\mu_g\|_2^2] - \frac{1}{n}\sum_{i=1}^{n}\|f(x^i)-\mu_g\|_2^2\right| + \|\mu_g-\widehat{\mu}_g^n\|_2^2.$$

As $\|f(X_g)\|_2 \leq C$, we have $\|f(X_g)-\mu_g\|_2^2 \leq (\|f(X_g)\|_2+\|\mu_g\|_2)^2 \leq (2C)^2$. By Hoeffding's inequality, we have that with probability at least $1-\frac{\delta}{3}$, it holds that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|f(X^i)-\mu_g\|_2^2 - \mathbb{E}[\|f(X_g)-\mu_g\|_2^2]\right| \leq 4C^2\sqrt{\frac{1}{2n}\log\left(\frac{6}{\delta}\right)}$$

Therefore, we have

$$\left|\mathrm{FD}_g - \widetilde{\mathrm{FD}}_g^n\right| \leq 2(\|\mu_g-\widehat{\mu}_g^n\|_2 + \|\widehat{\mu}_g^n-\mu_{\mathrm{r}}\|_2)\cdot\|\widehat{\mu}_g^n-\mu_g\|_2 + \mathrm{Tr}[\Sigma_{\mathrm{r}}^{\frac{1}{2}}]\sqrt{8\|\Sigma_g-\widehat{\Sigma}_g^n\|_2} + 4C^2\sqrt{\frac{1}{2n}\log\left(\frac{3}{\delta}\right)}. \tag{27}$$

Next, we analyze the concentration error of the $L2$-norm error. Since

$$\|\widehat{\mu}_g^n - \mu_g\|_2^2 = \underbrace{\|\widehat{\mu}_g^n - \mu_g\|_2^2 - \mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2]}_{\text{concentration}} + \underbrace{\mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2]}_{\text{expected } L2\text{-norm error}},$$

the expected $L2$-norm error is given by

$$\mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2] = \frac{\mathrm{Tr}[\Sigma_g]}{n}.$$

Additionally, we have that $|\widehat{\mu}_g^n[i] - \mu_g[i]| \le 2C$ for the i-th entry where $i \in [d]$. Hence, by Hoeffding's inequality and a union bound over all dimensions, with probability at least $1 - \frac{\delta}{3}$, it holds that

$$\|\widehat{\mu}_g^n - \mu_g\|_2^2 - \mathbb{E}[\|\widehat{\mu}_g^n - \mu_g\|_2^2] \le \frac{2dC^2}{n} \log\left(\frac{6d}{\delta}\right).$$

Hence, we have

$$\|\widehat{\mu}_g^n - \mu_g\|_2 \le \sqrt{\frac{2dC^2}{n} \log\left(\frac{6d}{\delta}\right) + \frac{\mathrm{Tr}[\Sigma_g]}{n}} := \Delta_1^n \tag{28}$$

Further, we have that with probability at least $1 - \frac{\delta}{3}$, it holds that

$$\left\|\Sigma_g - \widehat{\Sigma}_g^n\right\|_2 \le 20\kappa^2 \|\Sigma_g\|_2 \sqrt{\frac{4\boldsymbol{r}(\Sigma_g) + \log(3/\delta)}{n}} + (\Delta_1^n)^2 := \Delta_2^n \tag{29}$$

for $n \ge 4\boldsymbol{r}(\Sigma_g) + \log(3/\delta)$, which concludes the proof.

□

A.6 Auxiliary Definitions and Lemmas

Lemma 2 (Error decomposition for $\mathrm{Tr}[(\widehat{\Sigma}_g^n \Sigma_\mathrm{r})^{\frac{1}{2}}]$)**.** *Under the same conditions in Theorem 4, we have that*

$$\left|\mathrm{Tr}[(\Sigma_g \Sigma_\mathrm{r})^{\frac{1}{2}}] - \mathrm{Tr}[(\widehat{\Sigma}_g^n \Sigma_\mathrm{r})^{\frac{1}{2}}]\right| \le \mathrm{Tr}[\Sigma_\mathrm{r}^{\frac{1}{2}}] \sqrt{2\|\Sigma_g - \widehat{\Sigma}_g^n\|_2}. \tag{30}$$

Proof. Let $\eta := \|\Sigma_g - \widehat{\Sigma}_g^n\|_2$ and $\Sigma_\mathrm{r} := BB^\top$ denote the Cholesky decomposition of Σ_r, where B is invertible by the assumption $\Sigma_\mathrm{r} \succ \mathbf{0}$. Note that matrix $A\Sigma_\mathrm{r} = ABB^\top$ has the same set of eigenvalues with $B^\top AB$. We have that

$$\begin{aligned}
&\mathrm{Tr}[(\Sigma_g \Sigma_\mathrm{r})^{\frac{1}{2}}] - \mathrm{Tr}[(\widehat{\Sigma}_g^n \Sigma_\mathrm{r})^{\frac{1}{2}}] \\
\le &\mathrm{Tr}\left[\sqrt{B^\top(\Sigma_g - \widehat{\Sigma}_g^n + \widehat{\Sigma}_g^n + \eta I)B}\right] - \mathrm{Tr}\left[\sqrt{B^\top \widehat{\Sigma}_g^n B}\right] && \text{(Lemma 7)} \\
\le &\mathrm{Tr}\left[\sqrt{B^\top(\Sigma_g - \widehat{\Sigma}_g^n + \eta I)B} + \sqrt{B^\top \widehat{\Sigma}_g^n B}\right] - \mathrm{Tr}\left[\sqrt{B^\top \widehat{\Sigma}_g^n B}\right] \\
\le &\mathrm{Tr}\left[\sqrt{2\eta B^\top B}\right] = \mathrm{Tr}[\Sigma_\mathrm{r}^{\frac{1}{2}}]\sqrt{2\eta},
\end{aligned}$$

where the second inequality holds by Lemma 8 and the fact that $B^\top(\Sigma_g - \widehat{\Sigma}_g^n + \eta I)B, B^\top \widehat{\Sigma}_g^n B \succeq \mathbf{0}$ are PSD, and the last inequality holds by $B^\top(2\eta I)B \succeq B^\top(\Sigma_g - \widehat{\Sigma}_g^n + \eta I)B$. By the same analysis, we can also derive that

$$\mathrm{Tr}[(\widehat{\Sigma}_g^n \Sigma_\mathrm{r})^{\frac{1}{2}}] - \mathrm{Tr}[(\Sigma_g \Sigma_\mathrm{r})^{\frac{1}{2}}] \le \mathrm{Tr}\left[\sqrt{2\eta B^\top B}\right] = \mathrm{Tr}[\Sigma_\mathrm{r}^{\frac{1}{2}}]\sqrt{2\eta},$$

which concludes the proof. □

Definition 1 (Sub-Gaussian random variable)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian with (positive) parameter σ, which is denoted by $X \in \mathrm{SG}(\sigma)$, if

$$\mathbb{E}[\exp(s(X - \mu))] \le \exp\left(\frac{s^2\sigma^2}{2}\right)$$

for all $s \in \mathbb{R}$.

Definition 2 (Sub-Gaussian random vector)**.** A random vector $X \in \mathbb{R}^d$ is a sub-Gaussian random vector with (positive) parameter σ if $y^\top X \in \mathrm{SG}(\sigma)$. Particularly, the multivariate Gaussian $X \sim \mathcal{N}(0, \Sigma) \in \mathrm{SG}(\sqrt{\|\Sigma\|_2})$.

Definition 3 (Sub-exponential random variable)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential with (non-negative) parameters (ν, α), which denoted by $X \in \mathrm{SE}(\nu, \alpha)$, if

$$\mathbb{E}[\exp(s(X-\mu))] \le \exp\left(\frac{s^2\nu^2}{2}\right)$$

for all $|s| < \frac{1}{\alpha}$. Moreover, we have

$$\mathbb{P}[X - \mathbb{E}[X] \ge t] \le \begin{cases} \exp(-\frac{t^2}{2\nu^2}) & \text{if } 0 \le t \le \frac{\nu^2}{\alpha}, \\ \exp(-\frac{t}{2\alpha}) & \text{for } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Equivalently, with probability at least $1-\delta$, it holds that

$$|X - \mathbb{E}[X]| \le \max\left\{\sqrt{2\nu^2 \log\left(\frac{2}{\delta}\right)}, 2\alpha \log\left(\frac{2}{\delta}\right)\right\}.$$

Lemma 3 (Weighted sum of sub-exponential r.v.'s)**.** *Let $X_1, \cdots, X_n$ be n independent random variables, where $X_i \in \mathrm{SE}(\nu_i, \alpha_i)$ is sub-exponential with mean $\mathbb{E}[X_i] = \mu_i$ and parameters (ν_i, α_i). Then, for any $W = (w_1, \cdots, w_n)^\top \in \mathbb{R}^n_+$, the (non-negative) weighted sum $\sum_{i=1}^n w_i X_i$ is sub-exponential with parameters $(\sqrt{\sum_{i=1}^n w_i^2 \nu_i^2}, \max_i\{w_i \cdot \alpha_i\})$.*

Proof. By Definition 3, we have

$$\mathbb{E}[\exp(s \cdot w_i(X_i - \mu_i))] \le \exp\left(\frac{s^2 w_i^2 \nu_i^2}{2}\right)$$

for any $i \in [n]$ and $|s| \cdot w_i < \frac{1}{\alpha_i}$. Further, since $X_1, \cdots, X_n$ are independent, it holds that

$$\mathbb{E}[\exp(s \cdot \sum_{i=1}^n w_i(X_i - \mu_i))] \le \prod_{i=1}^n \exp\left(\frac{s^2 w_i^2 \nu_i^2}{2}\right) = \exp\left(\frac{s^2\left(\sqrt{\sum_{i=1}^n w_i^2 \nu_i^2}\right)^2}{2}\right)$$

for any $|s| < \min_i \frac{1}{w_i \cdot \alpha_i}$, which concludes the proof. □

Lemma 4 (Squared L2-norm of multivariate Gaussian)**.** *Let $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$ be any d-dimensional multivariate Gaussian with zero mean. Then, the r.v. $\|X\|_2^2 \sim \sum_{i=1}^d \lambda_i \chi_1^2$ is a weighted sum of i.i.d. chi-squared random variables with one degree of freedom, where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of the covariance matrix Σ. Further, let $X_1, \cdots, X_n \sim \mathcal{N}(\mathbf{0}, \Sigma)$ denote i.i.d. samples. Then, with probability at least $1-\delta$, it holds that*

$$\left|\frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 - \mathbb{E}[\|X\|_2^2]\right| \le \max\left\{\sqrt{\frac{8 \cdot \mathrm{Tr}[\Sigma^2]}{n} \log\left(\frac{2}{\delta}\right)}, \frac{8\|\Sigma\|_2}{n} \log\left(\frac{2}{\delta}\right)\right\}, \tag{31}$$

where $\mathbb{E}[\|X\|_2^2] := \sum_{i=1}^d \lambda_i = \mathrm{Tr}[\Sigma]$.

Proof. Let $\Sigma = Q\Lambda Q^\top$ be the eigendecomposition of the covariance matrix, where we define $\Lambda := \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$. Hence, we have $X \sim QZ$, where $Z \sim \mathcal{N}(\mathbf{0}, \Lambda)$ is *isotropic* multivariate Gaussian. Hence, we derive

$$\|X\|_2^2 = (QZ)^\top QZ = Z^\top Q^\top QZ = Z^\top Z = \sum_{i=1}^d \lambda_i \chi_1^2.$$

Since $\chi_1^2 \in \mathrm{SE}(2, 4)$ is sub-exponential with parameters $(2, 4)$, by Lemma 3 (with $w_i = \lambda_i$), we have $\|X\|_2^2 \in \mathrm{SE}(2\sqrt{\mathrm{Tr}[\Sigma^2]}, 4\|\Sigma\|_2)$ and has mean $\mathbb{E}[\|X\|_2^2] = \mathrm{Tr}[\Sigma]$. We conclude the proof by Lemma 5. □

Lemma 5 (Bernstein-type bound for sub-exponential r.v.). *Let $X_1, \cdots, X_n$ be n i.i.d. random variables, where each $X_i \in \mathrm{SE}(\nu, \alpha)$. Then, with probability at least $1-\delta$, it holds that*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \le \max\left\{\sqrt{\frac{2\nu^2}{n}\log\left(\frac{2}{\delta}\right)}, \frac{2\alpha}{n}\log\left(\frac{2}{\delta}\right)\right\}$$

Proof. By Lemma 3, we have $\frac{1}{n}\sum_{i=1}^{n} X_i \in \mathrm{SE}(\frac{\nu}{\sqrt{n}}, \frac{\alpha}{n})$, which concludes the proof. □

Lemma 6 (Dimension-free concentration of sample covariance matrix (Lounici, 2014; Koltchinskii and Lounici, 2017; Zhivotovskiy, 2022)). *Let $X_1, \cdots, X_n$ are i.i.d sub-Gaussian random vectors with zero mean and covariance matrix Σ. Let*

$$\boldsymbol{r}(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_2} \tag{32}$$

denote the effective rank. *Then, with probability at least $1-\delta$, the sample covariance matrix satisfies that*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top - \Sigma\right\|_2 \le 20\kappa^2\|\Sigma\|_2\left(\sqrt{\frac{4\boldsymbol{r}(\Sigma)+\log(1/\delta)}{n}}\right) \tag{33}$$

for $n \ge 4\boldsymbol{r}(\Sigma)+\log(1/\delta)$. Here, the parameter κ is a constant such that $\|y^\top X\|_{\psi_2} \le \kappa\sqrt{y^\top \Sigma y}$, where $\|Z\|_{\psi_2} := \inf\{t>0 : \mathbb{E}[\exp(Z^2/t^2)] \le 2\}$ is the sub-Gaussian norm.

Lemma 7. *Let $A, B \in \mathbb{S}^d$ be positive semi-definite (PSD) matrices. Then, we have that*

$$\mathrm{Tr}(\sqrt{A+B}) \ge \max\{\mathrm{Tr}(\sqrt{A}), \mathrm{Tr}(\sqrt{B})\} \tag{34}$$

Proof. Since $A+B \succeq A$ and $A+B \succeq B$, by Courant-Fischer min-max theorem, it holds that $\lambda_i(A+B) \ge \lambda_i(A)$ and $\lambda_i(A+B) \ge \lambda_i(B)$ for their i-th largest eigenvalues, which concludes the proof. □

Lemma 8. *Let $A, B \in \mathbb{S}^d$ be positive semi-definite (PSD) matrices. Then, we have that*

$$\mathrm{Tr}[\sqrt{A+B}] \le \mathrm{Tr}[\sqrt{A}] + \mathrm{Tr}[\sqrt{B}]. \tag{35}$$

Proof. First, we assume that both matrices A and B are positive strictly definite. Then,

$$\begin{aligned}
\mathrm{Tr}(\sqrt{A+B}) =& \mathrm{Tr}[(A+B)^{-\frac{1}{4}}(A+B)(A+B)^{-\frac{1}{4}}] \\
=& \mathrm{Tr}[(A+B)^{-\frac{1}{4}}A(A+B)^{-\frac{1}{4}}] + \mathrm{Tr}[(A+B)^{-\frac{1}{4}}B(A+B)^{-\frac{1}{4}}] \\
=& \mathrm{Tr}[\sqrt{A}(A+B)^{-\frac{1}{2}}\sqrt{A}] + \mathrm{Tr}[\sqrt{B}(A+B)^{-\frac{1}{2}}\sqrt{B}] \\
\le& \mathrm{Tr}[\sqrt{A}] + \mathrm{Tr}[\sqrt{B}],
\end{aligned}$$

where the last inequality holds by $A^{-\frac{1}{2}} \succeq (A+B)^{-\frac{1}{2}}$ and $B^{-\frac{1}{2}} \succeq (A+B)^{-\frac{1}{2}}$ (see Lemma 9). For general PSD matrices A and B, note that

$$\mathrm{Tr}(\sqrt{A+B+\epsilon I}) \le \mathrm{Tr}(\sqrt{A+\epsilon I}) + \mathrm{Tr}(\sqrt{B+\epsilon I})$$

holds for any $\epsilon > 0$ by our previous analysis. Since the trace map is continuous, □

Lemma 9. *Let A, B denote two PSD matrices such that $A \succeq B$. Then, their inverse $B^{-1} \succeq A^{-1}$.*

Proof. It suffices to show that $\sqrt{A}B^{-1}\sqrt{A} \succeq I$. Since $\sqrt{A}B^{-1}\sqrt{A} = (\sqrt{A}B^{-\frac{1}{2}})(\sqrt{A}B^{-\frac{1}{2}})^\top$ shares the same eigenvalues with $(\sqrt{A}B^{-\frac{1}{2}})^\top(\sqrt{A}B^{-\frac{1}{2}}) = B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \succeq I$, which concludes the proof. □

B PROOFS IN SECTION 6: IS-BASED EVALUATION

B.1 Proof of Theorem 2: Generator-Dependent Optimistic IS

Proof. By Lemma 10, we have that with probability at least $1-\frac{\delta}{2}$, $\mathcal{E}(\widehat{p}^n_{Y_g}) \geq H(Y_g)$. To show Theorem 2, it suffices to show that

$$H_g(Y_g|X_g) \geq \widehat{H}^n(Y_g|X_g) - \sqrt{\frac{2\widehat{V}^n_g(H(Y_g|X_g))}{n}\log\left(\frac{4d}{\delta}\right)} - \frac{7\log d}{3(n-1)}\log\left(\frac{4d}{\delta}\right)$$

with probability at least $1-\frac{\delta}{2}$, which holds by Theorem 12 and the fact that $H_g(Y_g|X_g) \leq \log d$. Therefore, we conclude the proof. □

B.2 Regret of IS-UCB

Theorem 6 (Regret of IS-UCB)**.** *With probability at least $1-\delta$, the regret of the IS-UCB algorithm after T steps is bounded by*

$$\mathrm{Regret}(T) \leq \widetilde{O}\left(G\Delta^{-2}\log d + e^C \cdot \left((dm+\log d)\cdot\sqrt{T}+\log T\right)\right), \tag{36}$$

where $C = \widetilde{O}(d)$, $m := \max_{g\in[G],j\in[d]:p_{Y_g}[j]>0}\{|u'(p_{Y_g}[j])|\}$, $\Delta := \min_{g\in[G],j\in[d]:p_{Y_g}[j]\neq e^{-1}}\{|p_{Y_g}[j]-e^{-1}|\}$, and $u(x) = -x\log x$.

Proof. By Theorem 2 and union bound over T steps, with probability at least $1-\delta$, it holds that $\widehat{\mathrm{IS}}_{g_t} \geq \mathrm{IS}_{g_t}$ for all steps $t\in[T]$ and $g\in[G]$, where $\widehat{\mathrm{IS}}_g$ is given by Equation (13). For convenience, we denote by $\widehat{\mathrm{IS}}_{g_t}$ the optimistic IS of generator g_t computed at the t-th step. Hence, we have that

$$\mathrm{IS}^* - \mathrm{IS}_{g_t} \leq \widehat{\mathrm{IS}}_{g_t} - \mathrm{IS}_{g_t} = e^{\mathcal{E}\left(\widehat{p}_{Y_{g_t}}\right)-\widehat{H}(Y_{g_t}|X_{g_t})+\mathcal{B}_{g_t}} - \mathrm{IS}_{g_t},$$

where we denote

$$\mathcal{B}_{g_t} := \sqrt{\frac{2\widehat{V}^{n_t}_g(H(Y_g|X_g))}{n_t}\log\left(\frac{4Td}{\delta}\right)} + \frac{7\log d}{3(n_t-1)}\log\left(\frac{4Td}{\delta}\right)$$

for convenience. Let $\mathcal{T}_1, \mathcal{T}_2 \subset [T]$ be two disjoint set of steps such that $\mathcal{T}_2 = [T]\backslash\mathcal{T}_1$. Specifically, $\mathcal{T}_1$ contains steps where no element of the empirical marginal class distribution $\widetilde{p}_{Y_{g_t}}$ is clipped to e^{-1}, i.e., $|e^{-1} - \widetilde{p}_{Y_{g_t}}[j]| \geq \boldsymbol{\epsilon}_{g_t}[j]$ for all $[j]\in d$. Let $\mathrm{IS}^* = \arg\max_{g\in[G]}\mathrm{IS}_g$ denote the optimal Inception score. Hence, we have that

$$\mathrm{Regret}(T) = \sum_{t\in\mathcal{T}_1}(\mathrm{IS}^* - \mathrm{IS}_{g_t}) + \sum_{t\in\mathcal{T}_2}(\mathrm{IS}^* - \mathrm{IS}_{g_t}).$$

Recall $u(x) = -x\log x$ for $x\in\mathbb{R}_+$. For the first part, we further derive that

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_1}(\mathrm{IS}^* - \mathrm{IS}_{g_t}) \\
\leq &\sum_{t\in\mathcal{T}_1}\left(\exp\{\mathcal{E}\left(\widehat{p}_{Y_{g_t}}\right) - \widehat{H}(Y_{g_t}|X_{g_t}) + \mathcal{B}_{g_t}\} - \exp\{H(Y_g) - H(Y_g|X_g)\}\right) \\
\leq &e^C \cdot \sum_{t\in\mathcal{T}_1}\left(\mathcal{E}\left(\widehat{p}_{Y_{g_t}}\right) - \widehat{H}(Y_{g_t}|X_{g_t}) + \mathcal{B}_{g_t} - H(Y_g) + H(Y_g|X_g)\right) \\
\leq &e^C \cdot \sum_{t\in\mathcal{T}_1}\left(\sum_{j=1}^d\left(u(\widehat{p}_{Y_{g_t}}[j]) - u(p_{Y_{g_t}}[j])\right) + 2\mathcal{B}_{g_t}\right) \\
\leq &O\left(e^C \cdot \sum_{t\in\mathcal{T}_1}\left(m\cdot\sum_{j=1}^d|\widehat{p}_{Y_{g_t}}[j] - p_{Y_{g_t}}[j]| + \mathcal{B}_{g_t}\right)\right)
\end{aligned}$$

$$\leq O\left(e^C \cdot \sum_{t\in\mathcal{T}_1}\left(m\cdot\sum_{j=1}^{d}\boldsymbol{\epsilon}_{g_t}[j]+\mathcal{B}_{g_t}\right)\right)$$
$$\leq \widetilde{O}\left(e^C\cdot\sum_{t=1}^{T}\left((dm+\log d)\sqrt{\frac{1}{N_{g_t}}}+\frac{\log d}{N_{g_t}}\right)\right), \tag{37}$$

where $m=\max_{g\in[G],j\in[d]:p_{Y_g}[j]>0}\{|u'(p_{Y_g}[j])|\}$, and $C=\widetilde{O}(d)$ is the upper bound of $\mathcal{E}(\widehat{p}_{Y_{g_t}})-\widehat{H}(Y_{g_t}|X_{g_t})+\mathcal{B}_{g_t}$. For the second part, note that the number of steps where at least one element of $\widetilde{p}_{Y_{g_t}}$ is clipped to at most $\widetilde{O}(G\Delta^{-2})$, where $\Delta=\min_{g\in[G],j\in[d]:p_{Y_g}[j]\neq e^{-1}}\{|p_{Y_g}[j]-e^{-1}|\}$. Therefore, we have that

$$\text{Regret}(T)\leq\widetilde{O}\left(G\Delta^{-2}\log d+e^C\cdot\left((dm+\log d)\cdot\sqrt{T}+\log T\right)\right),$$

which concludes the proof. □

B.3 Data-dependent optimistic marginal class distribution

Lemma 10 (Optimistic marginal class distribution)**.** *Let $x^1,\cdots,x^n\sim p_g$ be n generated images from generator g. Define $\boldsymbol{\epsilon}_g^n\in\mathbb{R}_+^d$ denote the error vector whose j-th element is given by*

$$\boldsymbol{\epsilon}_g^n[j]=\sqrt{\frac{2\widehat{V}^n(p_{Y|X_g}[j])}{n}\log\left(\frac{4d}{\delta}\right)}+\frac{7}{3(n-1)}\log\left(\frac{4d}{\delta}\right),$$

where $\widehat{V}^n(p_{Y|X_g}[j])=\mathbb{V}(p_{Y|x^1}[j],\cdots,p_{Y|x^n}[j])$ is the empirical variance for the j-th class density. Then, with probability at least $1-\frac{\delta}{2}$, we have that

$$\widehat{p}_{Y_g}^n=\text{Clip}_{e^{-1}}\left(\widetilde{p}_{Y_g}^n,\boldsymbol{\epsilon}_g^n\right)$$

satisfies that $\mathcal{E}(\widehat{p}_{Y_g}^n)\geq H(Y_g)$.

Proof. It suffices that show that $\boldsymbol{\epsilon}_g^n[j]\geq|\widehat{p}_{Y_g}^n[j]-p_{Y_g}[j]|$ for all $j\in[d]$ with probability at least $1-\frac{\delta}{2}$. We evoke Lemma 12 and conclude the proof. □

B.4 Auxiliary Definitions and Lemmas

Lemma 11 (Optimistic marginal class distribution)**.** *Let $\widetilde{p}_{Y_g}^n:=\frac{1}{n}\sum_{i=1}^n p_{Y_g|x^i}$ denote the empirical marginal class distribution. Let*

$$\widehat{p}_{Y_g}^n=\text{Clip}_{e^{-1}}\left(\widetilde{p}_{Y_g}^n,|\widetilde{p}_{Y_g}^n-p_{Y_g}|\right),$$

where $\text{Clip}_{e^{-1}}(p,\boldsymbol{\epsilon})$ is the following element-wise operator

$$\begin{cases}p[j]+\frac{e^{-1}-p[j]}{|e^{-1}-p[j]|}\boldsymbol{\epsilon}[j] & \text{, if } |e^{-1}-p[j]|\geq\boldsymbol{\epsilon}[j]\\ e^{-1} & \text{, otherwise.}\end{cases}$$

for any vectors $p,\boldsymbol{\epsilon}\in\mathbb{R}_+^d$. Then, we have that $\mathcal{E}(\widehat{p}_{Y_g}^n)\geq H(Y_g)$.

Proof. Note that the function $u(z)=-z\log z$ is concave and attains maximum at $z=e^{-1}$. It suffices to show that $u(\widehat{p}_{Y_g}[j])\geq u(p_{Y_g}[j])$ for all $j\in[d]$. If $|e^{-1}-\widehat{p}_{Y_g}^n[j]|<|\widetilde{p}_{Y_g}^n[j]-p_{Y_g}[j]|$, then $\widehat{p}_{Y_g}[j]=e^{-1}$, which ensures that $u(\widehat{p}_{Y_g}[j])\geq u(p_{Y_g}[j])$. In addition, if $|e^{-1}-\widehat{p}_{Y_g}^n[j]|\geq|\widetilde{p}_{Y_g}^n[j]-p_{Y_g}[j]|$, then

$$\widehat{p}_{Y_g}^n[j]=\widetilde{p}_{Y_g}^n[j]+\frac{e^{-1}-\widetilde{p}_{Y_g}^n[j]}{|e^{-1}-\widetilde{p}_{Y_g}^n[j]|}|\widetilde{p}_{Y_g}^n[j]-p_{Y_g}[j]|=\begin{cases}\widetilde{p}_{Y_g}^n[j]+|\widetilde{p}_{Y_g}^n[j]-p_{Y_g}[j]|, & \text{if } p_{Y_g}[j]<e^{-1}\\ \widetilde{p}_{Y_g}^n[j]-|\widetilde{p}_{Y_g}^n[j]-p_{Y_g}[j]|, & \text{otherwise.}\end{cases}.$$

The first case ensures that $p_{Y_g}[j]<\widehat{p}_{Y_g}^n[j]\leq e^{-1}$, and the second case ensures that $p_{Y_g}[j]>\widehat{p}_{Y_g}^n[j]\geq e^{-1}$. Both cases satisfy that $u(\widehat{p}_{Y_g}^n[j])\geq u(p_{Y_g}[j])$. Therefore, we have that $H(Y_g)=\mathcal{E}(p_{Y_g})=\sum_{j=1}^d u(p_{Y_g}[j])\leq\sum_{j=1}^d u(\widehat{p}_{Y_g}^n[j])=\mathcal{E}(\widehat{p}_{Y_g}^n)$, which concludes the proof. □

Lemma 12 (Restatement of (Maurer and Pontil, 2009, Theorem 4)). *Let $Z, Z_1, \cdots, Z_n$ be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. we have that with probability at least $1-\delta$ in the i.i.d vector $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$ that*

$$\mathbb{E}Z - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{2V_n(\boldsymbol{Z})}{n}\log\left(\frac{2}{\delta}\right)} + \frac{7}{3(n-1)}\log\left(\frac{2}{\delta}\right),$$

where $V_n(\boldsymbol{Z}) := \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}(Z_i - Z_j)^2$ is the sample variance.

C EXPERIMENTAL DETAILS

1. List of pretrained generative models. For the CIFAR10 dataset, we compare pretrained generative models including iDDPM-DDIM (Nichol and Dhariwal, 2021), LOGAN (Wu et al., 2020), WGAN-GP (Gulrajani et al., 2017), NVAE (Vahdat and Kautz, 2020), and RESFLOW (Chen et al., 2019). For ImageNet, we compare pretrained models including DiT-XL-2 (Peebles and Xie, 2023), ADMG (Dhariwal and Nichol, 2021), BigGAN (Brock et al., 2019), RQ-Transformer (Lee et al., 2022), and ADM (Dhariwal and Nichol, 2021). For the FFHQ dataset, we compare StyleNAT (Walton et al., 2023), StyleGAN2-ADA (Karras et al., 2020), LDM (Rombach et al., 2022), Unleashing-Transformers (Bond-Taylor et al., 2021), and Efficient-vdVAE (Hazami et al., 2022). We utilize the generated image datasets downloaded from the dgm-eval repository (Stein et al., 2023).

2. List of embeddings for FD-based evaluation and selection. We consider three standard encoders for image data: InceptionV3.Net (Szegedy et al., 2016), DINOv2 (Oquab et al., 2024), and CLIP (Radford et al., 2021). Following (Stein et al., 2023), we utilize the DINOv2 ViT-L/14 model and the OpenCLIP ViT-L/14 trained on DataComp-1B (Gadre et al., 2023). Embeddings extracted by InceptionV3.Net, DINOv2, and CLIP have 2048, 1024, and 1024 dimensions, respectively. For video data, we utilize the I3D model (Carreira and Zisserman, 2018) pretrained on the Kinetics-400 (Carreira and Zisserman, 2018) dataset following (Unterthiner et al., 2019), where the logits layer with 400 dimensions is used as the embedding. For audio data, we utilize VGGish (Hershey et al., 2017) following (Kilgour et al., 2019), where the activations from the 128 dimensional layer prior to the final classification layer are used as the embedding.

3. Implementation details of FD-UCB. As we consider InceptionV3, DINOv2, and (unnormalized) CLIP embeddings, which are generally unbounded, we utilize the collected data to estimate the model-dependent parameters in the bonus (6). This approach preserves the online format of the FD-UCB algorithm. Our numerical results indicate that with only 50 samples, the norm terms in the bound would be close to their underlying values (Tables 1, 2, and 3). To have a better estimation of the covariance matrix, we also adopt the thresholding method in (Cai and Liu, 2011). We treat the batch size, parameter κ in the bonus (6), and parameter M for thresholding as hyperparameters. We conduct ablation study on these hyperparameters summarize the results in Figure 14.

Parameters	$\mathrm{Tr}(\Sigma)$	$\mathrm{Tr}(\widehat{\Sigma})$	$\sqrt{\mathrm{Tr}[\Sigma^2]}$	$\sqrt{\mathrm{Tr}[\widehat{\Sigma}^2]}$	$\|\Sigma\|_2$	$\|\widehat{\Sigma}\|_2$	$\boldsymbol{r}(\Sigma)$	$\boldsymbol{r}(\widehat{\Sigma})$	Range of $\|f(X_g)\|_2$
DiT-XL-2	166.4	160.0	12.5	20.9	9.6	11.2	17.3	14.3	$[9.7, 33.6]$
ADMG	157.7	150.8	12.1	19.4	9.2	10.5	17.2	14.4	$[10.8, 34.5]$
BigGAN	154.5	161.8	13.5	22.3	11.0	13.2	14.0	12.2	$[9.7, 31.8]$
RQ-Transformer	166.9	164.5	13.8	22.2	10.7	12.7	15.5	13.0	$[11.0, 29.7]$
ADM	176.5	182.5	15.7	25.6	12.9	15.2	13.7	12.0	$[0.5, 34.8]$
ImageNet	181.2	141.2	12.4	20.2	9.2	8.7	19.7	16.2	$[10.9, 33.0]$

Table 1: Data-dependent parameters on the ImageNet dataset and standard generative models: We present both the estimated and reference values, which are computed from 50 and $5,000$ images, respectively. The image data embeddings are extracted by InceptionV3.Net.

4. Implementation details of Naive-UCB. Naive-UCB is a simplification of FD-UCB and IS-UCB which replaces the generator-dependent variables in the bonus function with data-independent and dimension-based

Parameters	$\mathrm{Tr}(\Sigma)$	$\mathrm{Tr}(\widehat{\Sigma})$	$\sqrt{\mathrm{Tr}[\Sigma^2]}$	$\sqrt{\mathrm{Tr}[\widehat{\Sigma}^2]}$	$\|\Sigma\|_2$	$\|\widehat{\Sigma}\|_2$	$\boldsymbol{r}(\Sigma)$	$\boldsymbol{r}(\widehat{\Sigma})$	Range of $\|f(X_g)\|_2$
DiT-XL-2	591.2	601.9	59.1	101.4	38.2	48.0	15.5	12.5	$[34.6, 36.9]$
ADMG	579.9	566.1	62.1	101.3	41.2	50.4	14.1	11.2	$[34.9, 37.0]$
BigGAN	514.9	486.8	62.7	90.8	43.8	54.0	11.8	9.0	$[34.9, 36.8]$
RQ-Transformer	545.5	551.7	58.8	99.1	39.3	51.0	13.9	10.8	$[35.0, 36.8]$
ADM	560.2	561.8	59.0	100.7	38.3	52.5	14.6	10.7	$[34.9, 36.8]$
ImageNet	636.8	573.1	53.5	118.5	32.1	74.1	19.9	7.7	$[35.1, 36.7]$

Table 2: Data-dependent parameters on the ImageNet dataset and standard generative models: We present both the estimated and reference values, which are computed from 50 and 5,000 images, respectively. The image data embeddings are extracted by CLIP.

Parameters	$\mathrm{Tr}(\Sigma)$	$\mathrm{Tr}(\widehat{\Sigma})$	$\sqrt{\mathrm{Tr}[\Sigma^2]}$	$\sqrt{\mathrm{Tr}[\widehat{\Sigma}^2]}$	$\|\Sigma\|_2$	$\|\widehat{\Sigma}\|_2$	$\boldsymbol{r}(\Sigma)$	$\boldsymbol{r}(\widehat{\Sigma})$	Range of $\|f(X_g)\|_2$
DiT-XL-2	2079.5	2085.6	99.9	309.4	36.5	81.4	57.0	25.6	$[38.5, 48.7]$
ADMG	2068.2	2102.1	103.0	311.4	42.5	79.1	48.7	25.6	$[38.6, 49.0]$
BigGAN	1885.5	1799.6	120.3	285.4	68.8	110.4	27.4	16.3	$[36.5, 48.6]$
RQ-Transformer	1955.3	1968.0	111.4	302.5	54.4	91.1	36.0	21.6	$[38.4, 49.3]$
ADM	2013.3	1988.1	106.7	306.0	50.5	101.0	39.9	19.7	$[39.1, 49.2]$
ImageNet	2112.7	2095.4	85.3	347.3	21.0	104.3	100.7	20.1	$[42.2, 48.3]$

Table 3: Data-dependent parameters on the ImageNet dataset and standard generative models: We present both the estimated and reference values, which are computed from 50 and 5,000 images, respectively. The image data embeddings are extracted by DINOv2.

terms. For FD-based evaluation, Naive-UCB sets $\mathrm{Tr}[\Sigma_g] = O(d), \mathrm{Tr}[\Sigma_g^2] = O(d)$, and $\|\Sigma_g\|_2 = O(1)$ in FD-UCB. For IS-based evaluation, the Naive-UCB method sets $\widehat{V}^n(Y_g|X_g) = (\log d)^2$ and $\widehat{V}^n(p_{Y|X_g}[j]) = 1$ for any $j \in [d]$.

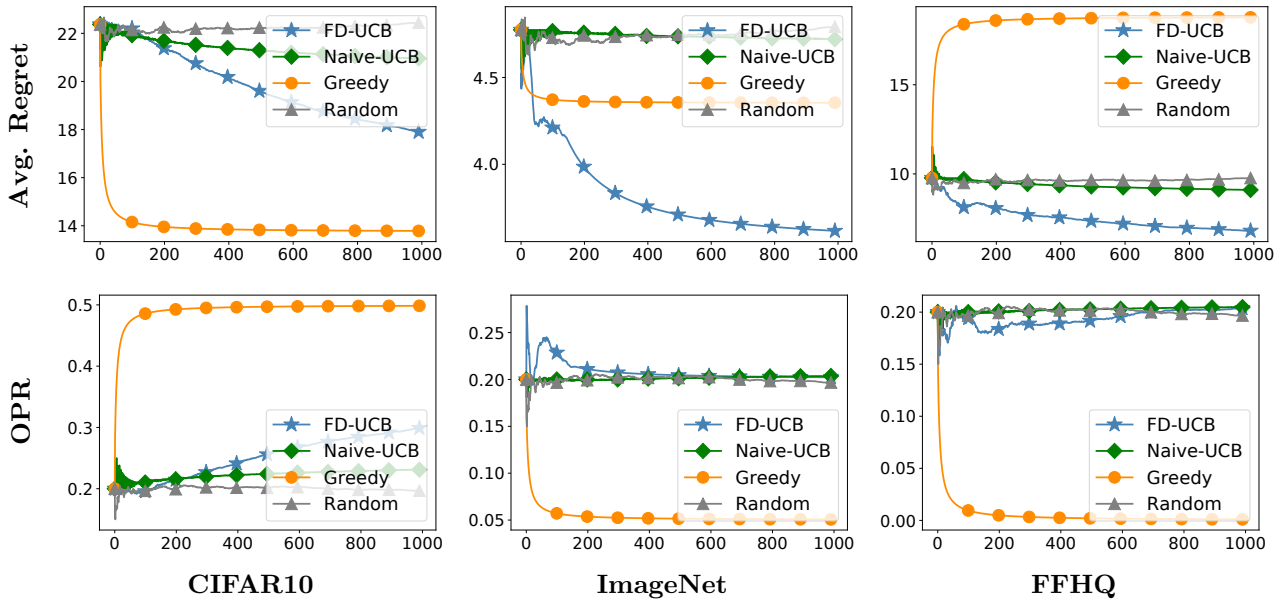

Figure 7: Online FD-based evaluation and selection among standard generative models: The image data embeddings are extracted by InceptionV3.Net. Results are averaged over 20 trials.

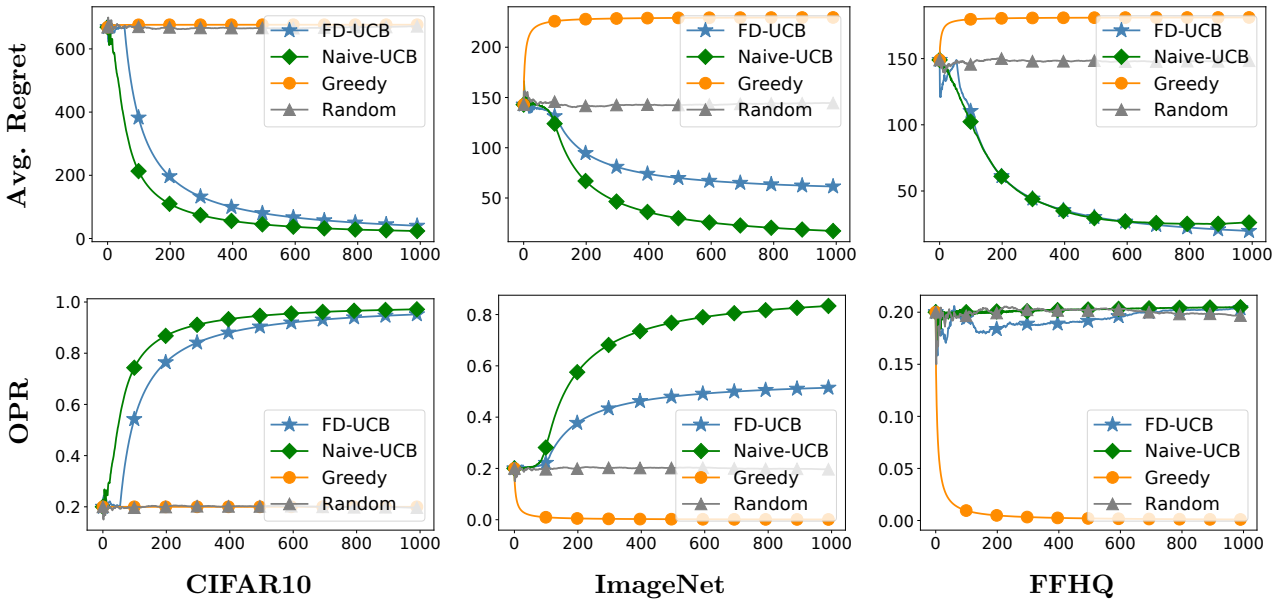

Figure 8: Online FD-based evaluation and selection among standard generative models: The image data embeddings are extracted by DINOv2-ViT-L/14. Results are averaged over 20 trials.

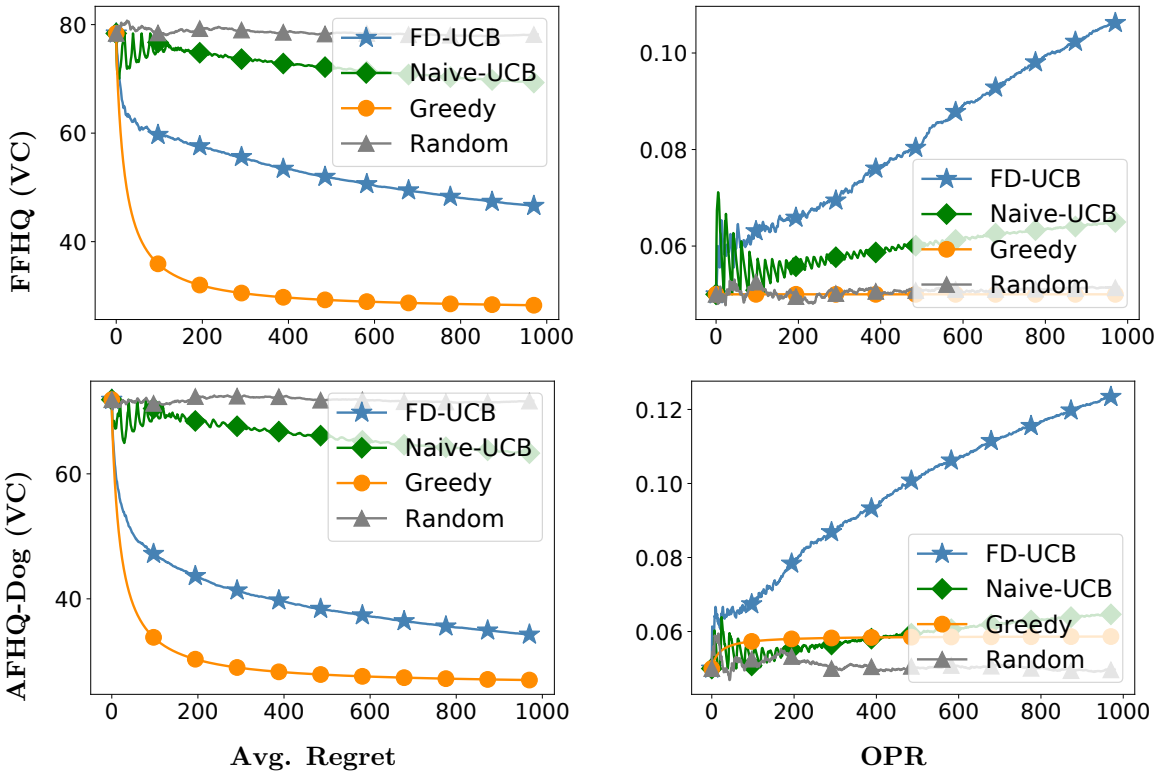

Figure 9: Online FD-based evaluation and selection among variance-controlled (VC) models: The image data embeddings are extracted by InceptionV3.Net. Results are averaged over 20 trials.

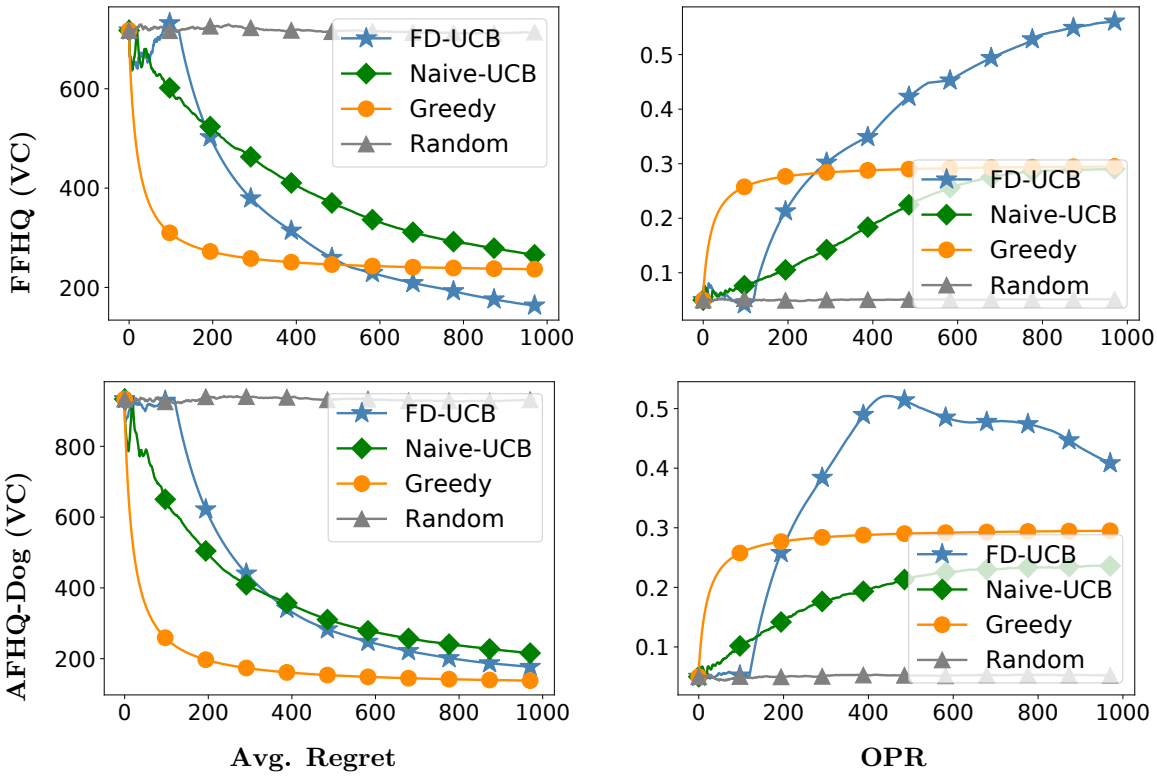

Figure 10: Online FD-based evaluation and selection among variance-controlled (VC) models: The image data embeddings are extracted by DINOv2. Results are averaged over 20 trials.

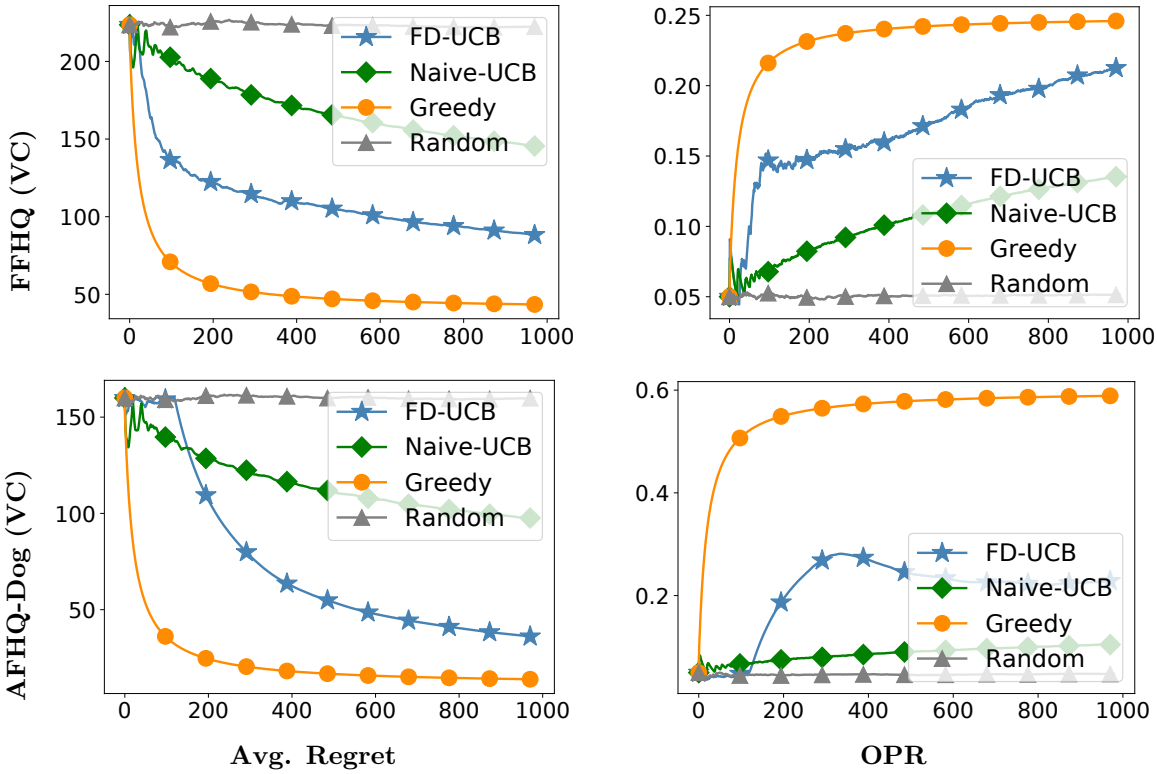

Figure 11: Online FD-based evaluation and selection among variance-controlled (VC) models: The image data embeddings are extracted by CLIP. Results are averaged over 20 trials.

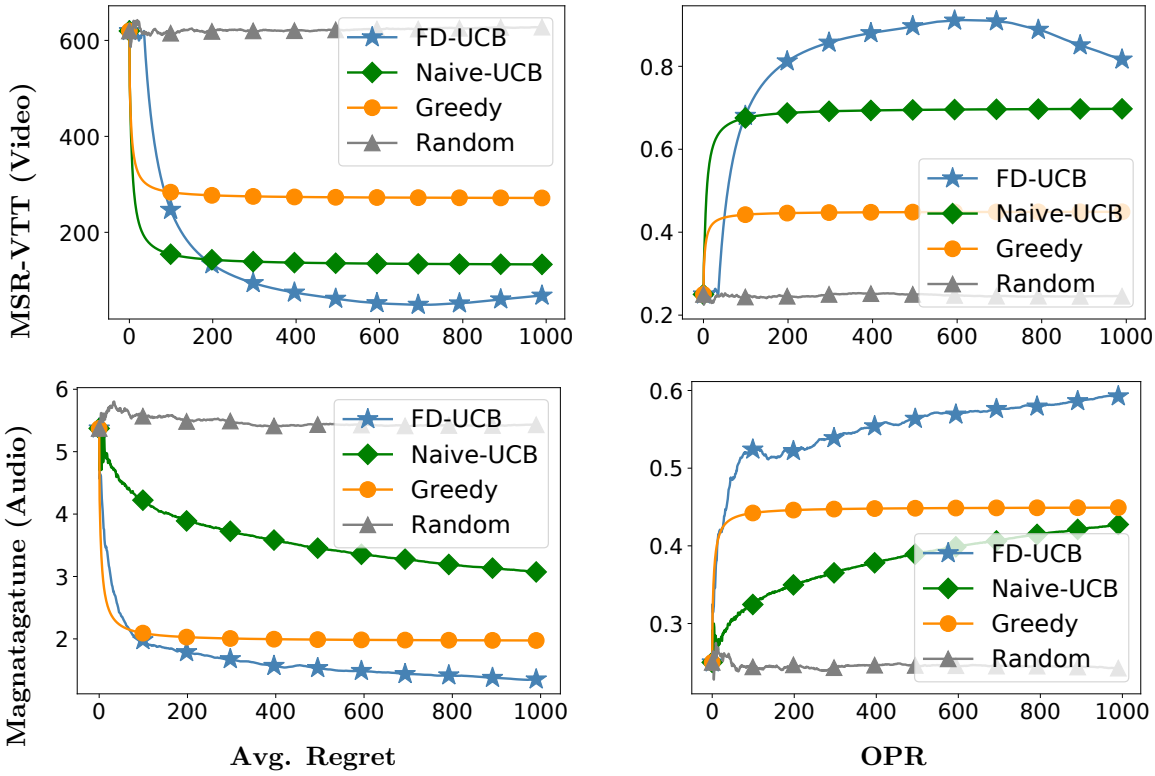

Figure 12: Online evaluation and selection on video and audio data: The video data embeddings are extracted by I3D, and the audio data embeddings are extracted by VGGish. Results are averaged over 20 trials.

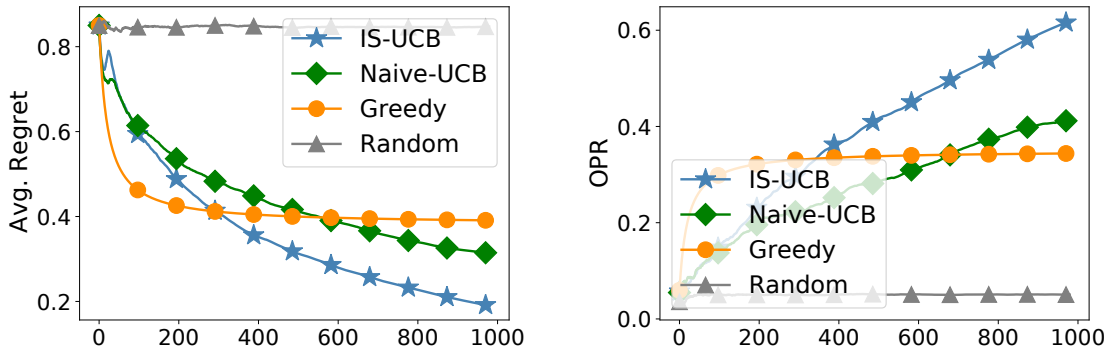

Figure 13: Online IS-based evaluation and selection among variance-controlled models on the AFHQ Dog dataset: IS-UCB can identify models that generate images with more diversity. Results are averaged over 20 trials.

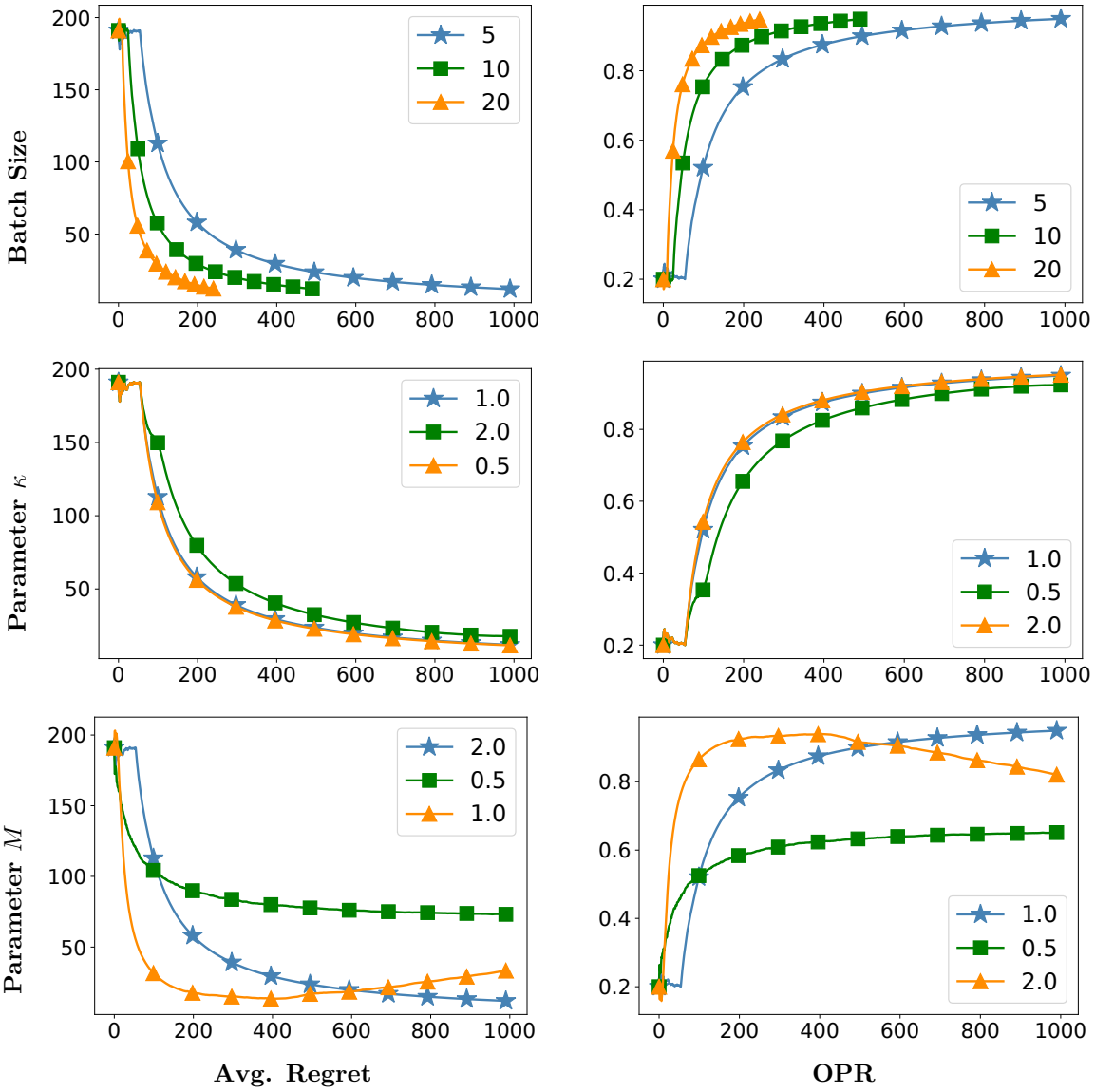

Figure 14: Ablation study on the hyperparameters of FD-UCB on the CIFAR10 dataset: The image data embeddings are extracted by CLIP. Results are averaged over 20 trials.

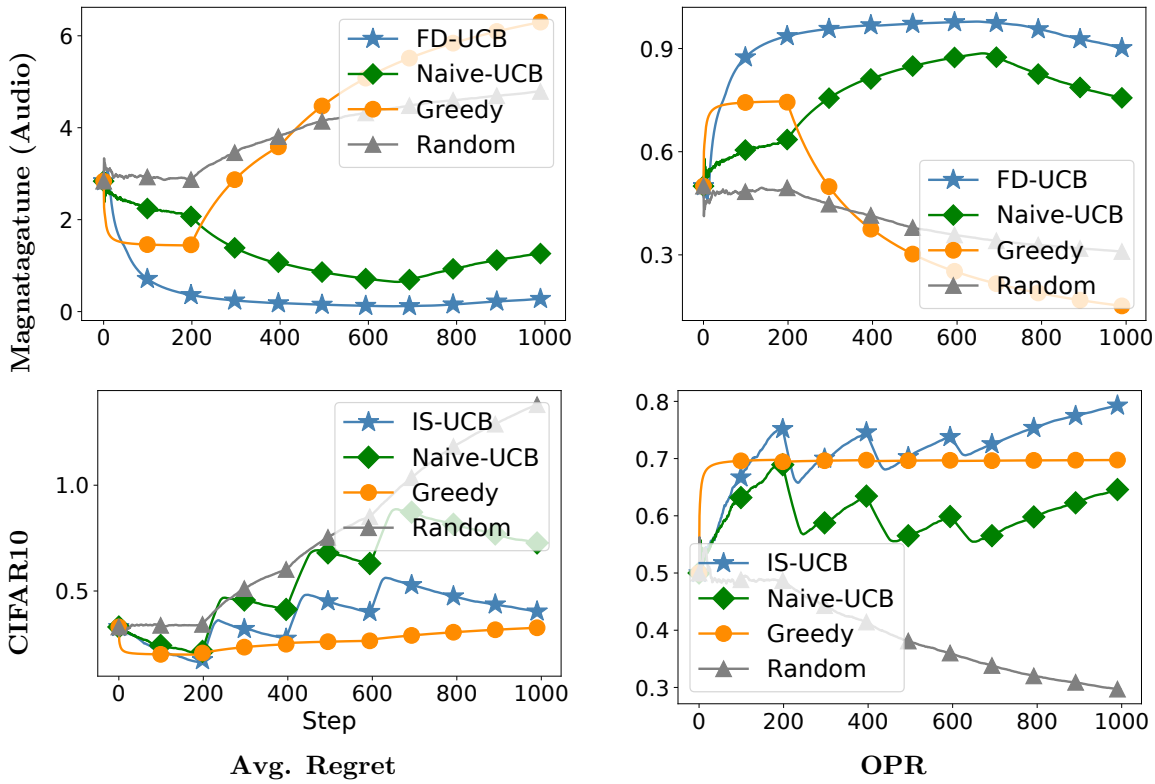

Figure 15: Adaptation to new models: A new model is introduced after each 200 steps. Results are averaged over 20 trials.