Image Generation Diversity Issues and How to Tame Them

Supplementary Material

6. Toy Dataset

To illustrate the efficiacy of IRS we propose an experiment with a toy dataset. Specifically, we compare how IRS can be used to measure the inter and intra-class diversity. We model the two-dimensional distribution as a multi-modal Gaussian. The real dataset has six classes and five subclasses. We model synthetic diversity by progressively removing modes from the synthetic dataset. The results can be seen in Fig. 6. For both experiments the $\text{IRS}_{\infty,a}$ score is very close to the real value.

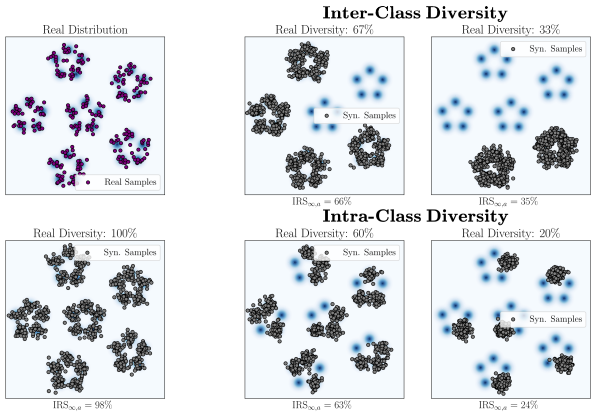

Figure 6. Inter and intra class diversity according to IRS on a toy dataset. The real pdf is shown in the background. Blue means high likelihood. We use 1000 each of real, test, and synthetic samples.

7. Model Rejection Based on IRS

Building on the methodology introduced in the main paper, we explore how IRS can provide additional insights into model diversity during training. IRS, which requires a minimal number of samples for computation, proves to be particularly useful for online evaluation methods such as rejecting early model checkpoints that lack sufficient diversity. This appendix analyzes IRS-based rejection through experiments, providing practical examples and additional interpretation of its efficacy. By defining a target diversity percentage IRS_d , we can leverage IRS to monitor progress towards achieving this goal. The core question IRS addresses is: 'How many images learnt from different training samples k_{min}will a model with this level of diversity generate at minimum?' To calculate this, we take Eqs. (7) and (8), and look for the minimum expected number of different learned images k_{min}:

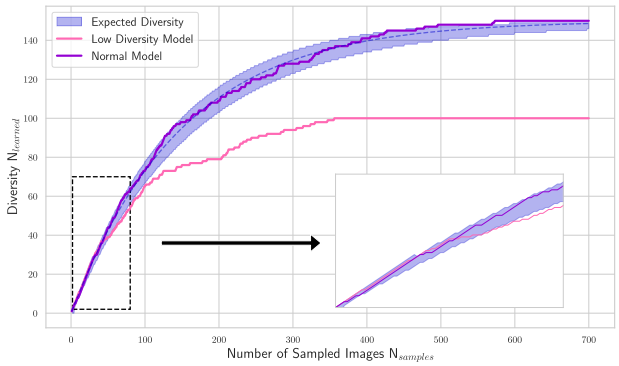

Figure 7. Detecting low diversity models. By leveraging the statistical properties of the training dataset, we can assess the diversity of the model during the sampling process. A low diversity model will saturate earlier and can be rejected quickly. In this example, we would reject the low diversity model after only generating 60 samples.

$$\text{k}_{\text{min}} = \arg\max_k \sum_k \text{P}(k, \text{N}_{sample}, \text{IRS}_d * \text{N}_{train}) < \alpha_e \tag{9}$$

Fig. 7 illustrates this using a toy example with 800 distinct training values. To get ground truth values for diversity, we simulate generation as random sampling from a set of integers. As shown in Fig. 7, a toy example with 800 training images demonstrates how low-diversity models quickly converge to a high number of duplicates. By sampling as few as 60 images, we can confidently reject models failing to meet the diversity threshold. This highlights IRS as a reliable tool for early-stage model evaluation.

We additionally investigate the efficacy of IRS by examining its accuracy according to Eqs. (6) to (8). Figure 8 illustrates how predicted IRS values vary with model diversity. The initial estimate with $\alpha = 0.01$, which evaluates the diversity of $1/100 * \text{N}_{train}$ samples, is reasonably close to the true value, but its confidence level suggests this may be due to chance. At $\alpha = 0.1$ the estimates become increasingly confident around the ground truth value.

Next we illustrate the probability distribution function and cumulative density function for five different training sizes in Fig. 9. The example uses the real formula for Stirling's number of the second kind instead of the estimate introduced in Eq. (5) in the main paper. For $\text{N}_{train} = 32$ it is expected that the amount of different images that were sampled after 15 samples lies between 10 and 14. If we observe more duplicates than this we would reject the hypothesis

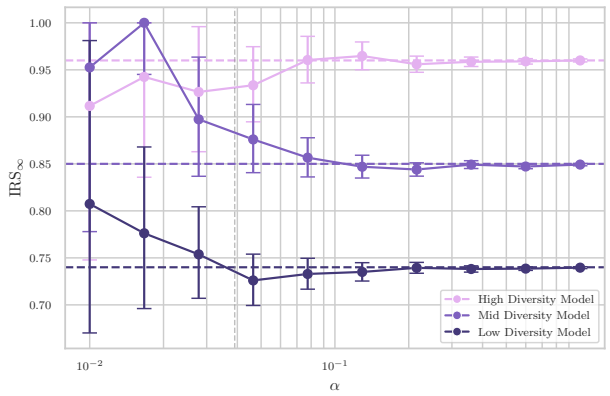

Figure 8. IRS_∞, indicated by dots, and confidence intervals $\text{IRS}_{\infty,L}$ and $\text{IRS}_{\infty,U}$ for three different simulated ImageNet models. By increasing the number of observations (α), we also increase the confidence of the IRS_∞ predictions. The black dashed line indicates the ground truth and the vertical line the value for 50k samples.

that this model has the potential to generate $\text{N}_{train} = 32$ different images. For larger training dataset sizes the number of expected different generated images increases. *E.g.*, for $\text{N}_{train} = 39$ we would even reject a model that only produces 10 different images.

8. Further Visual Results

To illustrate the image retrieval results, we present visualizations for ImageNet-512 in Fig. 10. For a randomly selected subset of images, we compute the image correspondences based on the predictions $\mathcal{P}$ from all benchmark models discussed in Sec. 4. The results highlight how certain models, such as DINOv2 and Inception, retrieve images based on semantic similarity, while others, like the Random model, focus on general image composition (*e.g.*, a dog resembling a triangular shape against a white background, similar to a child on the beach). The final example demonstrates that most models retrieve near duplicates, but not all do so consistently. DINOv2, for instance, incorporates a deduplication step in its pipeline to minimize redundancy [34], which may influence these results. The third row in Fig. 10 presents an ambiguous case. A model that emphasizes image composition, such as DINOv2, may retrieve an image of a woman standing in front of a board. In contrast, a model that prioritizes visual similarity in human appearance, such as CLIP, retrieves another image featuring a similarly appearing woman. This trade-off has to be considered when using feature extractors to assess the performance of generative models with metrics such as FID, Precision, Recall, or IRS.

9. Image Retrieval Agreement and Consensus

In Sec. 4.1 we argued that due to the adjustment step introduced in Sec. 3.4 we can in theory choose any kind of feature extractor $\mathcal{F}$. In order to maximize the interpretability of IRS we use the extractor that has the best agreement with the consensus of all models. Therefore, we compute the correspondence prediction for each of the feature extractors for ImageNet. We ensemble the decision of each model and call every prediction where five or more models decide for the same image correspondance as consensus. Then we check how often each model agrees with the ensemble decision. Agreement computes how often two models agree with each other. Our goal is to see if we can measure how good a model's prediction aligns with the prediction of the ensemble. The consensus reached by the ensemble is then considered the ground truth. The results are shown in Tab. 1. There is a large discrepancy between the agreement of all of these models. Our expectation was that consesus is reached mostly by the same models. While this is true, the models that are most frequently part of the consensus, are not the models that showed the most diversity which we discuss in Sec. 14. DINOv2, for example, performed best in terms of diversity on ImageNet but got beaten on agreement by ConvNeXt, one of the worst models in terms of diversity. SwAV on the other hand shows extraordinary agreement with the consensus and is almost always agrees with the consensus, but the output features lack diversity. Next, we examine ImageNet and the number of feature extractors that agree on image correspondence, as shown in Fig. 11. We find that, for the majority of images, almost 40%, all models disagree with each other. This outcome aligns with expectations, as many images have multiple valid correspondences, as discussed in Sec. 8. The more models required to reach a consensus, the fewer samples meet this criterion. Consequently, we consider the consensus decision correct when at least five models agree, as this represents a majority decision. Fig. 11 further illustrates the level of agreement between models. Notably, data2vec and Random exhibit low agreement with all other models, whereas ConvNeXt shows the highest agreement with other models, particularly with CLIP and DINOv2. This shows that there is no real correlation between diversity and agreement and both have to be measured seperately.

10. Further Distance Metric Sensitivty Analysis

We consider two different measurements $\mathcal{P}$ to compute the distance between $f_x = \mathcal{F}(\mathbf{x_t})$ of the query image and all reference images $f_{x'} = \mathcal{F}(\mathbf{x'_t})$. The first one is the cosine distance derived from cosine similarity. It is used by many feature extractors directly such as [34]. Additionally, we consider the Euclidean distance between features which is

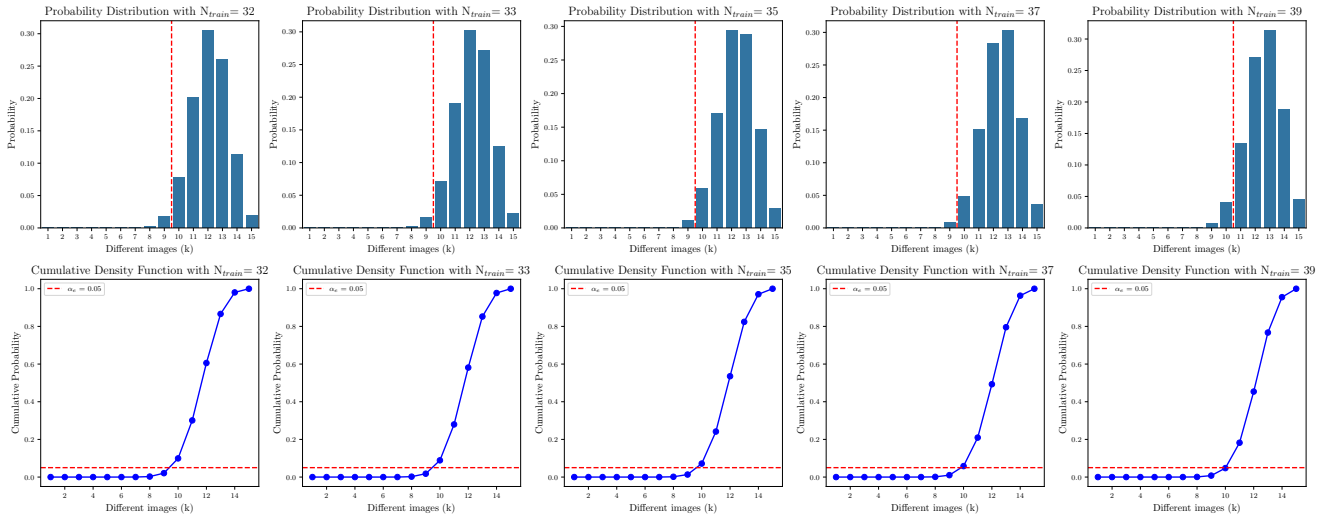

Figure 9. Illustration of the selection process of the threshold for a fixed number of samples with different trainingset sizes.

used by several metrics such as Precision, Recall [26]:

$$\begin{aligned}\mathcal{P}_{\text{Cosine}}(\mathbf{x_t}, \mathbf{x'_t}) &= d_{\text{Cosine}}(f_{\mathbf{x_t}}, f_{x'}) \\ &= 1 - \frac{f_{\mathbf{x_t}} \cdot f_{x'}}{\|f_{\mathbf{x_t}}\| \|f_{x'}\|}\end{aligned} \tag{10}$$

and:

$$\begin{aligned}\mathcal{P}_{\text{Euclidean}}(\mathbf{x_t}, \mathbf{x'_t}) &= d_{\text{Euclidean}}(f_x, f_{x'}) \\ &= \sqrt{\sum_{i=1}^{n} (f_x^i - f_{x'}^i)^2}\end{aligned} \tag{11}$$

We compute the distances on ImageNet over five folds with a fixed number ratio between the size of training and sampling set. The results are shown in Sec. 10 . Generally, we observe the same measurement gap. Irrespective of the measurement, all the feature extractors need much longer to converge to 100% expected diversity according to Eq. (1). Visual inspection of the retrieved images from randomly drawn samples yields that the results for each model are very similar. However, the performance of some of the models depends heavily on the used metric. data2vec, for example, did not show good performance for the Eucledian distance but the best performance for cosine distance. DINOv2 is the complete opposite. It outperformed all other models according to the Euclidean distance but is only mediocre according to cosine distance. We conclude that the choice of distance metric has a large influence on the relative performance between all feature extractors. It does not impact the underlying problem that all of them are far from reaching the expected performance. We decide to focus our experiments on the Eucledian distance due to its connection to established generative metrics [26].

	$\alpha = 2$		$\alpha = 6$	
Idealized	86.47		99.75	
	d_{cosine}	$d_{\text{euclidean}}$	d_{cosine}	$d_{\text{euclidean}}$
BYOL	67.46 ± 0.04	**67.93** ± 0.04	87.94 ± 0.04	**88.29** ± 0.09
CLIP	65.00 ± 0.03	**66.85** ± 0.03	85.07 ± 0.05	**87.45** ± 0.08
ConvNeXt	**64.56** ± 0.03	52.34 ± 0.07	**85.13** ± 0.09	68.30 ± 0.09
data2vec	**73.08** ± 0.09	40.75 ± 0.04	**93.26** ± 0.06	59.15 ± 0.07
DINOv2	66.57 ± 0.05	**66.64** ± 0.05	**87.31** ± 0.06	**87.32** ± 0.04
Inception	**65.75** ± 0.08	60.00 ± 0.05	**86.48** ± 0.04	79.71 ± 0.10
MAE	68.06 ± 0.04	**68.17** ± 0.06	88.13 ± 0.03	**88.22** ± 0.03
Random	69.65 ± 0.07	**71.29** ± 0.05	89.60 ± 0.06	**90.89** ± 0.06
SwAV	**70.31** ± 0.06	64.19 ± 0.08	**90.55** ± 0.03	84.38 ± 0.16

Table 4. Comparison of cosine and Euclidean distances averaged across five folds. For each model and α, the best $\mathcal{P}$ value is highlighted in bold.

11. Computational Requirement

To benchmark the proposed lacking diversity rejection method we use the method for the official ImageNet-512 train set with $\text{N}_{train} = 1281166$. We set the desired IRS to 80% and the probability of error to 5% with 50000 reference and synthetic samples each. Eq. (9) states that we reject the checkpoint for not being diverse enough if, after sampling, $\text{N}_{learned} < 48744$ images are learned (does not account for measurement gap). Computing this threshold takes roughly three seconds and does not depend on the images sampled. Computing $\mathcal{X}_{\text{learned}}$ takes longer and depends on the forward pass of $\mathcal{F}$. For Inception-v3 it roughly takes five minutes on a single Nvidia-A40 GPU. However, these features are also necessary to compute FID and other metrics so they are usually already available. If the pre-computed features are available, computing $\text{IRS}_{\infty,a}$ takes 40 seconds for ImageNet. Note that we do not need the entire synthetic

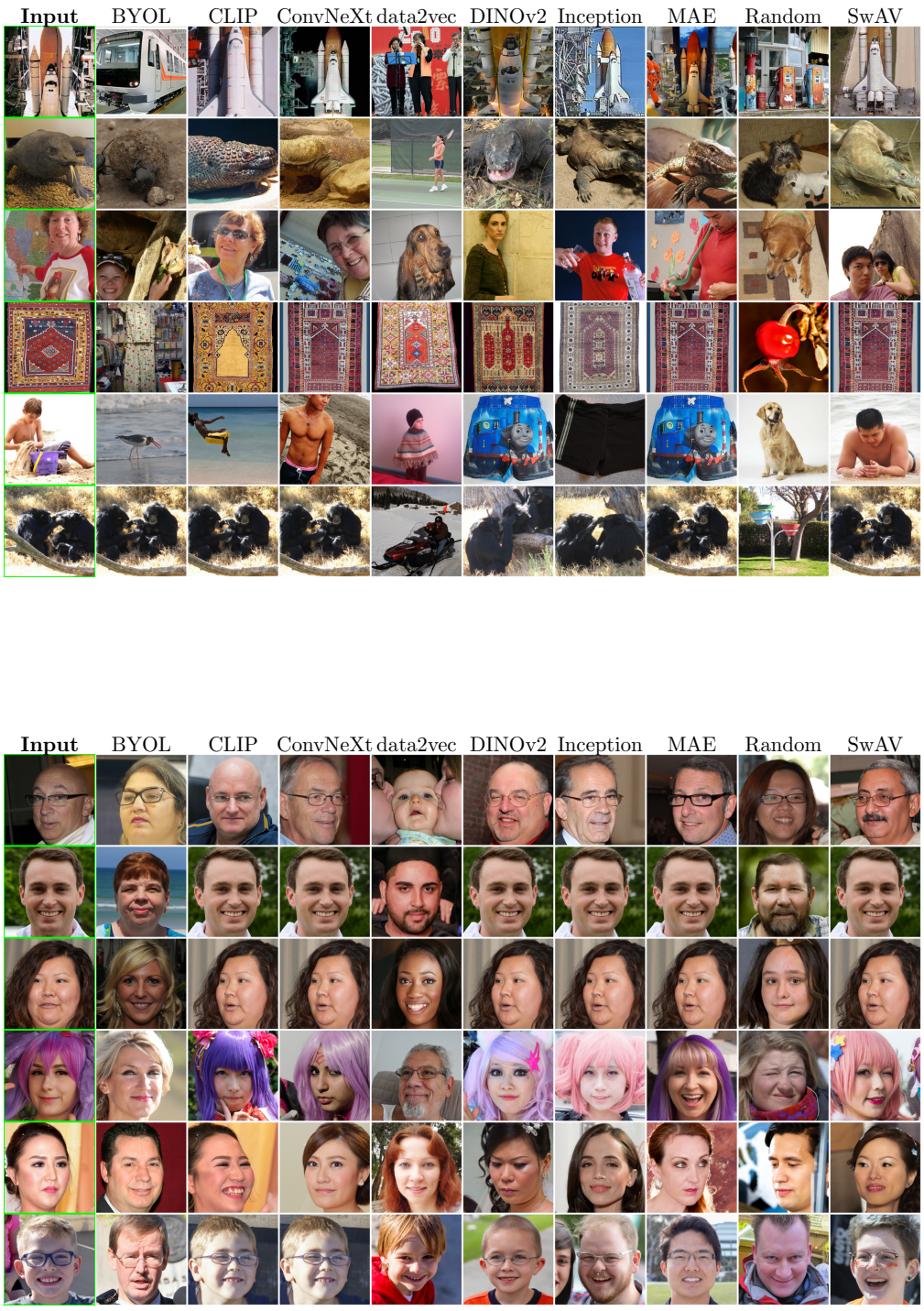

Figure 10. Qualitative results of image retrieval on ImageNet (Top) and FFHQ (bottom).

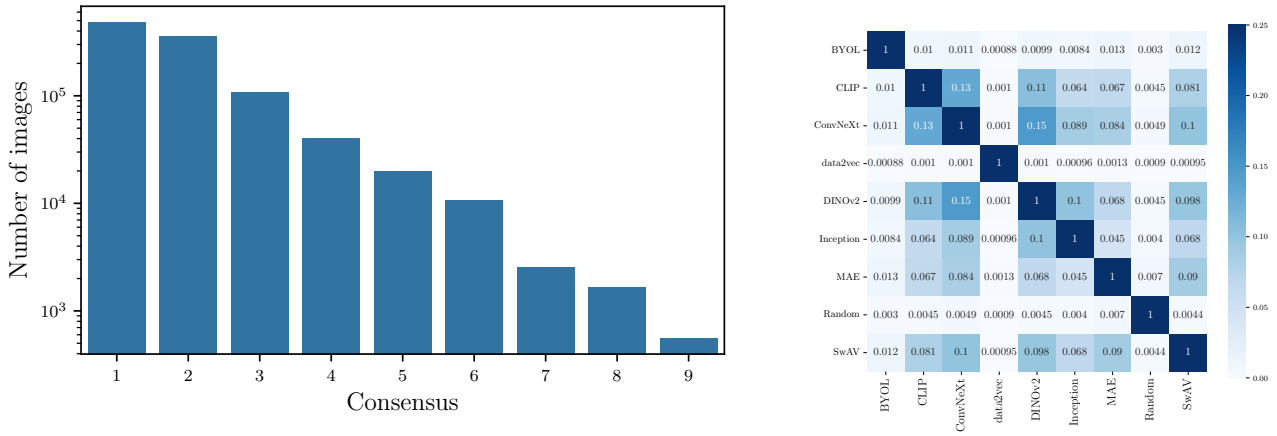

Figure 11. Comparison of Consensus and Agreement of different feature extractors for the ImageNet Dataset

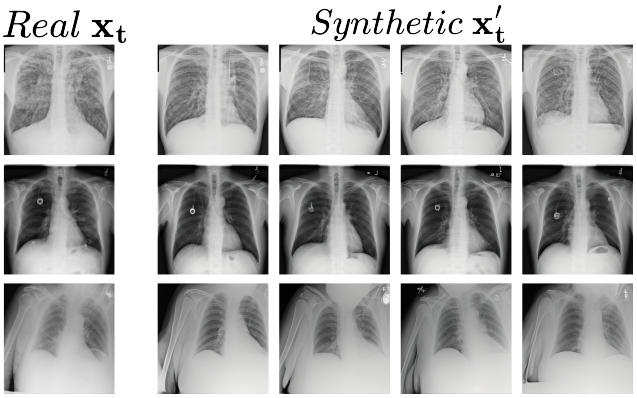

Figure 12. Random intra-class examples of real reference images and images generated by DiADM with different generation seeds for "No Finding" class.

dataset to compute an upper bound for $N_{learned}$ according to Eq. (2). If within the first 2000 samples we already observe over 1257 duplicates we can immmediately reject the model. For the smaller datasets like Dynamic, computing $IRS_{\infty,a}$ from precomputed features takes less than a second.

12. Visual Results DiADM

To illustrate why extracting features is effective for diverse data generation we show real and generated samples from DiADM (Fig. 12). From the feature vector of Inception, DiADM learned to pick up pseudo-conditional features skewness, brightness, presence of tubes and can be tasked to generate them with equivalent diversity. Small differences between the generated samples confirm that they are not simply memorized.

13. Results with Domain Specific Feature Extractors

In the next step, we analyze the impact of feature extractors on prediction performance. Specifically, we examine how the performance associated with the observed measurement gap changes when feature extractors are tailored to the dataset. For this analysis, we compare the privacy models proposed in [14, 40]. These models were trained for re-identification on the EchoNet dataset [35], utilizing a Siamese architecture where the input consists of two frames and the output predicts whether the frames originate from the same video [14]. This model can also be directly applied for image retrieval. The results, presented in Tab. 5, show that training models specifically for this dataset reduces the measurement gap. The models outperform all pre-trained feature extractors by more than eight percentage points. Furthermore, the findings suggest that the measurement gap can be minimized for specific datasets when necessary.

14. IRS_{real} Results

In Sec. 4.1 we explain the measurement gap stemming from feature extractors collapsing to smaller features spaces that lack expressiveness in terms of diversity. Quantitative prove for this is shown in Fig. 13 and Tab. 6. The best performing model across all datasets is the randomly initialized Inception-v3 suggested by [33]. If we compare this to the visual examples shown in Fig. 10, we see that this similarity seems to be more based on the general composure of the images than the semantics. BYOL pre-trained on ImageNet also performs well on all datasets. DINOv2 is the best performing model on ImageNet, which confirms the observations from [49].

α	2	16
Idealized	86.47	99.99
BYOL	69.91 ± 0.56	97.97 ± 0.48
CLIP	64.71 ± 0.50	95.61 ± 0.74
ConvNeXt	56.70 ± 1.05	89.58 ± 1.12
data2vec	56.54 ± 0.48	97.08 ± 0.62
DINOv2	63.13 ± 0.33	95.15 ± 0.63
Inception	60.62 ± 0.69	93.00 ± 1.04
MAE	64.57 ± 0.68	96.31 ± 0.81
Random	73.35 ± 0.71	98.75 ± 0.33
SwAV	59.20 ± 0.57	92.63 ± 0.88
Re-identification [40]	80.96 ± 0.34	99.61 ± 0.27
Re-identification Latent [14]	$\mathbf{81.91 \pm 0.75}$	$\mathbf{99.95 \pm 0.08}$

Table 5. IRS_α for EchoNet-Dynamic using domain specific re-identification models from [40] and [14].

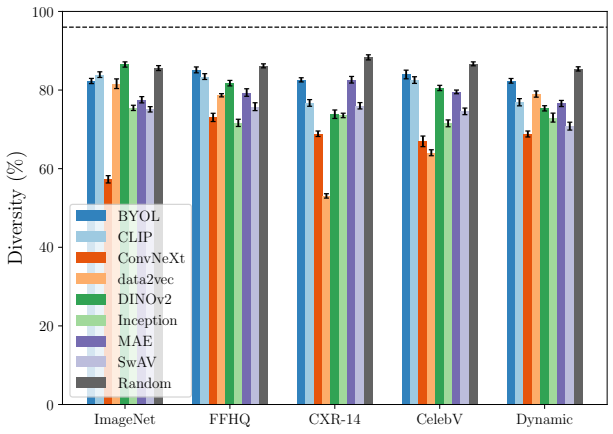

Figure 13. Measured diversity according to IRS of common feature extractors between real training and real test images. The dashed line indicates the idealized model $\text{IRS}_{\frac{7}{3}} = 96.98\%$.

15. Metrics Analysis and Comparison

15.1. IRS over FID in Measuring Diversity Insufficiency and Bias Amplification

Here, we analyze the properties of IRS and demonstrate its superiority over FID [20] in detecting diversity insufficiency and bias amplification in generative models.

Definition of FID. Follow [20], here let $\mathcal{X}_r$ and $\mathcal{X}_g$ denote the feature distributions of real and generated data, respectively. The FID is defined as:

$$\text{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + \text{Tr}\left(\mathbf{C}_r + \mathbf{C}_g - 2\left(\mathbf{C}_r\mathbf{C}_g\right)^{1/2}\right) \quad (12)$$

where $\boldsymbol{\mu}_r$, $\mathbf{C}_r$ and $\boldsymbol{\mu}_g$, $\mathbf{C}_g$ are the mean vectors and covariance matrices of the real and generated feature distributions, respectively. FID measures the distance between these two distributions under the assumption that they are Gaussian.

Definition of IRS. Let $\mathbf{x_t}$ be an image from a dataset consisting of N_{train} real images residing in image space $\mathcal{X} \in \mathbb{R}^{c\times h\times w}$. Unconditional generative models aim to learn the distribution $p_{data}(\mathbf{X})$ and sample N_{sample} synthetic images from it. Define $\text{N}_{learned}$ as the number of unique training samples retrieved by the generated samples. The IRS is:

$$\text{IRS}_\alpha = \frac{\text{N}_{learned}}{\text{N}_{train}} \quad (13)$$

where $\alpha = \frac{\text{N}_{sample}}{\text{N}_{train}}$ is the sampling rate. More details can be found in Eq. 2.

Theorem 1. *IRS exhibits higher statistical sensitivity than FID in detecting diversity insufficiency and bias amplification in generative models.*

Proof. *Diversity Insufficiency.* Assume the generative model can produce $K < \text{N}_{train}$ unique training samples. The expectation of IRS is:

$$\mathbb{E}[\text{IRS}_\alpha] = \frac{K}{\text{N}_{train}}\left(1 - \left(1 - \frac{1}{K}\right)^{\text{N}_{sample}}\right) \quad (14)$$

Since $K < \text{N}_{train}$ and $\left(1 - \frac{1}{K}\right)^{\text{N}_{sample}} > \left(1 - \frac{1}{\text{N}_{train}}\right)^{\text{N}_{sample}} \approx e^{-\alpha}$, it follows that:

$$\mathbb{E}[\text{IRS}_\alpha] < 1 - e^{-\alpha} \quad (15)$$

Thus, IRS effectively captures the reduction in diversity when $K < \text{N}_{train}$.

Conversely, FID measures the distance between feature distributions based on mean and covariance. Even if $K < \text{N}_{train}$, if the K samples are diverse in feature space, $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$ may remain close to $\boldsymbol{\mu}_r$ and $\mathbf{C}_r$, resulting in a low FID that fails to reflect the reduced diversity.

Bias Amplification. Suppose the generative model memorizes (*i.e.*, guided) $K \ll \text{N}_{train}$ training samples, effectively reproducing these samples. The expectation of IRS is:

$$\mathbb{E}[\text{IRS}_\alpha] = \frac{K}{\text{N}_{train}}\left(1 - \left(1 - \frac{1}{K}\right)^{\text{N}_{sample}}\right) \ll 1 \quad (16)$$

Given $K \ll \text{N}_{train}$, IRS approaches $\frac{K}{\text{N}_{train}}$, significantly lower than the ideal value of $1 - e^{-\alpha}$, thereby effectively indicating bias amplification.

In contrast, $\text{FID} \approx \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + \text{Tr}\left(\mathbf{C}_r + \mathbf{C}_g - 2\left(\mathbf{C}_r\mathbf{C}_g\right)^{1/2}\right)$. If the memorized K samples have feature statistics close to the biased data distribution, FID remains low, failing to detect bias amplification.

Model	ImageNet	FFHQ	ChestX-ray14	CelebV-HQ	Dynamic
BYOL	82.30 ± 0.66%	85.13 ± 0.75%	82.60 ± 0.53%	83.96 ± 1.10%	82.35 ± 0.59%
CLIP	83.92 ± 0.72%	83.43 ± 0.74%	76.72 ± 0.85%	82.54 ± 0.85%	76.90 ± 0.92%
ConvNeXt	57.29 ± 0.93%	73.04 ± 1.05%	68.88 ± 0.70%	66.95 ± 1.35%	68.82 ± 0.76%
data2vec	81.61 ± 1.23%	78.67 ± 0.41%	53.07 ± 0.56%	64.05 ± 0.76%	78.98 ± 0.79%
DINOv2	**86.50 ± 0.67%**	81.75 ± 0.70%	73.85 ± 1.08%	80.53 ± 0.67%	75.35 ± 0.67%
Inception	75.46 ± 0.67%	71.65 ± 0.92%	73.54 ± 0.57%	71.53 ± 0.87%	72.99 ± 1.14%
MAE	77.53 ± 0.80%	79.37 ± 0.96%	82.62 ± 0.84%	79.53 ± 0.47%	76.58 ± 0.78%
SwAV	75.10 ± 0.68%	75.77 ± 1.00%	76.00 ± 0.81%	74.58 ± 0.84%	70.82 ± 1.00%
Random	85.59 ± 0.63%	**86.16 ± 0.51%**	**88.32 ± 0.65%**	**86.67 ± 0.50%**	**85.38 ± 0.54%**

Table 6. IRS computation only using *real* data. Results give percentage of real samples retrieved using common feature extractors. The idealized scenario reaches $\text{IRS}_{\frac{7}{3}} = 96.98\%$. To make comparison across datasets easier results are presented for a fixed size of 3000 training images and 7000 test images on all dataset. Best results are bold and second best results underlined.

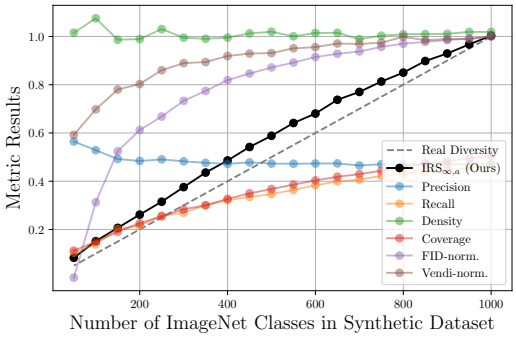

Figure 14. Measuring diversity of datasets by removing classes and computing IRS.

15.2. Comparison to Other Metrics

Alternative metrics, such as Precision, Recall, Density, and Coverage, are susceptible to issues arising from varying hyperparameter settings. We demonstrate this by setting the number of considered neighbors to 1 for these metrics and replicating the experiment shown in Fig. 5. The results, presented in Fig. 14, reveal that all metrics converge to very low diversity values, even when only real data is used. This indicates that incorrect hyperparameter configurations can prevent these metrics from converging to a diversity value of 1.0 for real data, significantly compromising the interpretability of the results. Note the key difference between recall and IRS in this example: For IRS, a single synthetic sample can correspond to only one real sample, whereas multiple synthetic samples can lie within the manifold spanned by the 1-NN manifold for recall.