FGM-HD: Boosting Generation Diversity of Fractal Generative Models through Hausdorff Dimension Induction

Haowei Zhang[1,2], Yuanpei Zhao[1,2], Ji-Zhe Zhou[1,2]* , Mao Li[1,2]

[1]College of Computer Science, Sichuan University, China
[2]Engineering Research Center of Machine Learning and Industry Inteligence, Ministry of Education of China
{zhanghaowei1, zhaoyuanpei}@stu.scu.edu.cn, {jzzhou, limao}@scu.edu.cn

Abstract

Improving the diversity of generated results while maintaining high visual quality remains a significant challenge in image generation tasks. Fractal Generative Models (FGMs) are efficient in generating high-quality images, but their inherent self-similarity limits the diversity of output images. To address this issue, we propose a novel approach based on the Hausdorff Dimension (HD), a widely recognized concept in fractal geometry used to quantify structural complexity, which aids in enhancing the diversity of generated outputs. To incorporate HD into FGM, we propose a learnable HD estimation method that predicts HD directly from image embeddings, addressing computational cost concerns. However, simply introducing HD into a hybrid loss is insufficient to enhance diversity in FGMs due to: 1) degradation of image quality, and 2) limited improvement in generation diversity. To this end, during training, we adopt an HD-based loss with a monotonic momentum-driven scheduling strategy to progressively optimize the hyperparameters, obtaining optimal diversity without sacrificing visual quality. Moreover, during inference, we employ HD-guided rejection sampling to select geometrically richer outputs. Extensive experiments on the ImageNet dataset demonstrate that our FGM-HD framework yields a 39% improvement in output diversity compared to vanilla FGMs, while preserving comparable image quality. To our knowledge, this is the very first work introducing HD into FGM. Our method effectively enhances the diversity of generated outputs while offering a principled theoretical contribution to FGM development.

Introduction

Generative models have achieved remarkable progress in recent years, particularly in the image synthesis domain. Approaches such as generative adversarial networks (GANs) (Goodfellow et al. 2014; Wiatrak, Albrecht, and Nystrom 2019), variational autoencoders (VAEs) (Kingma, Welling et al. 2013; Van Den Oord, Vinyals et al. 2017), diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Shen et al. 2025), and autoregressive and flow-based models (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016; Kingma and Dhariwal 2018; You et al. 2022), have demonstrated the ability to produce high-fidelity and semantically coherent images. However, maintaining an adequate balance between image **quality** and **diversity** remains a fundamental challenge. Although many models excel at generating visually appealing outputs, they often fail to capture the full variability of the underlying data distribution.

Fractal Generative Models (FGMs) (Li et al. 2025) offer a unique architecture for high-quality image generation by leveraging recursive self-similarity, an intrinsic property of fractals. By repeatedly applying a compact generative module across multiple scales, FGMs can efficiently generate complex, globally consistent, and high-resolution images. This modular recursive design makes FGMs particularly suitable for tasks requiring structural coherence and visual richness. Nevertheless, the same self-similar structure that ensures global consistency can also result in repetitive patterns and insufficient diversity in generated results.

To address the challenge of limited diversity in FGMs, we introduce the **Hausdorff Dimension (HD)** (Hausdorff 1918) as a geometric indicator for structural complexity. As a numerical concept in fractal geometry, HD quantifies the variation of spatial detail across scales, with higher values typically reflecting greater structural richness. This makes HD particularly suitable for guiding the generation of diverse and intricate outputs.

Nevertheless, conventional HD estimation techniques, such as the box counting method (Mandelbrot 1983), are computationally expensive and not easily compatible with large-scale training pipelines. In addition, HD measures structural complexity, and deep layer embeddings learned by neural networks have been proven to capture this structure effectively (Valle et al. 2022; Werbos 2002). Therefore, at the commencement, we propose a learning-based HD estimation method for efficient HD prediction directly from image embeddings to reduce computational costs while maintaining high accuracy, making it feasible for integration into training pipelines.

However, simply introducing HD into a hybrid loss function is insufficient for improving generation diversity, as it leads to degraded image quality and limited improvement in generation diversity. We observed that in the hybrid loss function, the relative weighting of each component is critical: during the early training phase, image quality is typically suboptimal and HD estimates are unreliable,

*Corresponding authors: Ji-Zhe Zhou, Mao Li

necessitating a focus on quality while gradually increasing the weight of diversity as training progresses, without sacrificing visual quality. Therefore, an optimal balance between HD loss and generative loss is a key factor for effective training. To address this, we propose the **Monotonic Momentum-Driven Scheduling (MMDS)** strategy, which dynamically adjusts the influence of HD loss over time. The MMDS strategy ensures that the model first focuses on high-quality structure and progressively incorporates diversity, providing a smoother transition and more stable optimization without compromising generation quality.

To further enhance output diversity during inference, we introduce an HD-guided sampling strategy. Leveraging the recursive structure of FGMs, which naturally supports the generation of multiple candidate patches in parallel, we retain only those outputs whose estimated HD exceeds a predefined threshold. This post hoc selection process filters out structurally simple images, resulting in a final output set with greater geometric richness and perceptual variety, achieved without modifying the underlying generative architecture.

Experiments on the ImageNet dataset indicate that our approach significantly improves diversity while maintaining visual quality. In particular, our method achieves a 39% improvement in Recall compared to the vanilla FGM, highlighting HD's effectiveness as both a training signal and a sampling criterion. To our knowledge, this is the first systematic effort to introduce HD into fractal generative modeling. Additionally, the MMDS strategy for dynamic weight optimization offers a generalizable framework that can be applied to other models utilizing hybrid loss functions, thereby extending its applicability across a broader range of generative tasks.

In summary, our main contributions are as follows.

- **Introducing Hausdorff Dimension for Diversity Enhancement:** We are the first to introduce HD into the FGM framework, proposing a set of methods during both training and inference, including the MMDS and Sampling Strategy, to enhance generation diversity while maintaining high image quality. Our work alleviates the self-similarity limitation inherent in fractal-based generation.
- **Developing a learnable and efficient HD estimation module:** To overcome the inefficiency of numerical HD estimation methods, we design a learnable HD predictor that operates directly on image embeddings, enabling fast and accurate HD prediction and supporting scalable integration into generative frameworks.
- **Monotonic Momentum-Driven Scheduling (MMDS) strategy:** We introduce MMDS, a strategy that dynamically adjusts the influence of HD loss during training to balance visual quality and structural diversity, ensuring stable optimization. MMDS can also be extended to models with hybrid loss functions, providing a systematic approach to optimize the balance between loss components in diverse generative tasks.

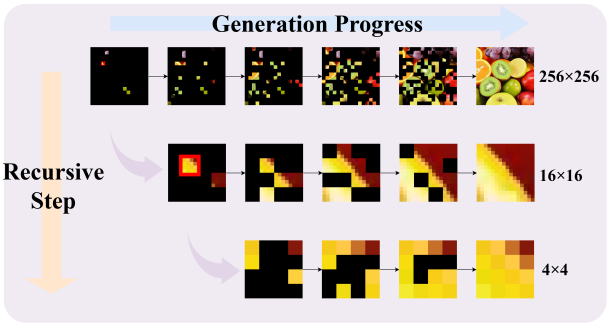


Figure 1: Overview of the FGM process, where the image is generated from sparse patches to 256×256 resolution through recursive refinement at smaller scales (e.g., 16×16 and 4×4 blocks). A shared generative module is reused across scales, capturing global and fine-grained structural details.

Related Work

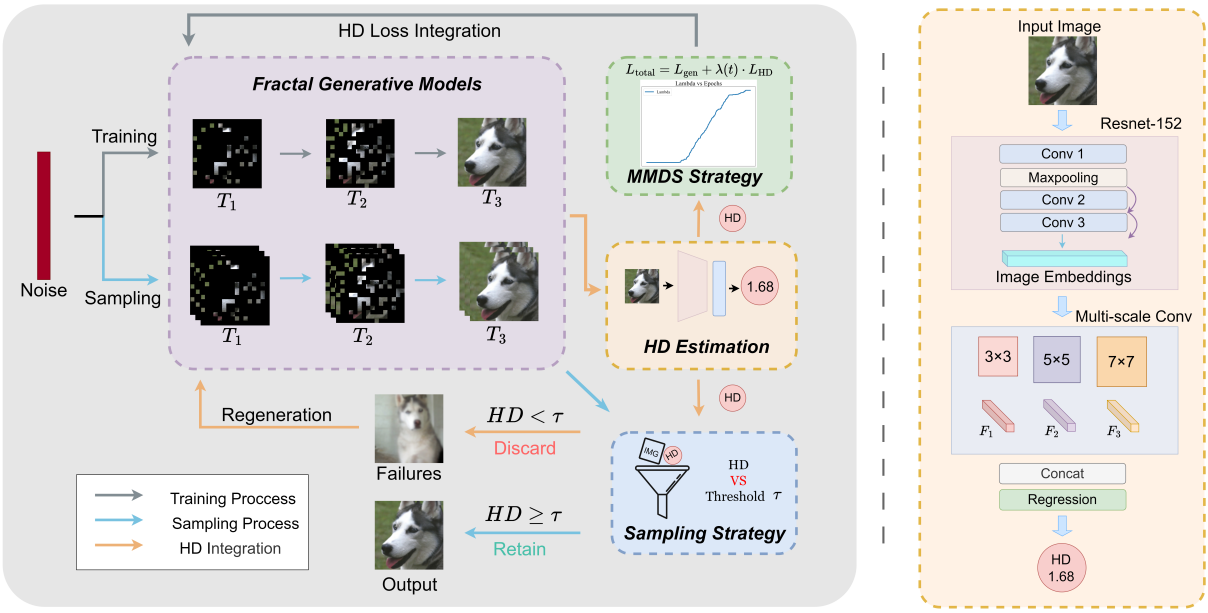
Generative Models

Generative models have made significant strides in image generation tasks. For instance, *GANs* (Goodfellow et al. 2014) have enabled high-quality image generation through adversarial learning and have been widely applied in various tasks. Similarly, *VAEs* (Kingma, Welling et al. 2013) leverage probabilistic modeling for generating images through latent variable distributions. More recently, *diffusion models* (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Shen et al. 2025) have gained popularity for generating high-quality and diverse images by iteratively denoising samples. Moreover, *autoregressive and flow-based models* (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016; Kingma and Dhariwal 2018; You et al. 2022) excel in generating high-quality images with strong mode coverage. Despite these advancements, a key challenge remains achieving a balance between image quality and diversity (Lan et al. 2024; Li et al. 2021a), as existing models struggle to generate sufficiently varied samples that comprehensively reflect the underlying data distribution.

Fractal Generative Models

Fractal-based generative models leverage the recursive and self-similar properties of fractals to synthesize complex visual patterns. Classical approaches such as iterated function systems (IFS) (Barnsley 2014) and Lindenmayer systems (L-systems) (Prusinkiewicz and Lindenmayer 2012) generate intricate structures through predefined rules, offering mathematical simplicity but lacking adaptability. Recent advances have incorporated neural networks into fractal frameworks, notably *FGMs* (Li et al. 2025), employing recursive atomic modules to progressively refine images across multiple scales. This design enables FGMs to efficiently generate high-quality and visually coherent images with minimal architectural complexity.

Although FGMs capture structural detail and visual quality, their recursive nature limits output diversity, leading



(a) Training and Sampling Strategies

(b) Hausdorff Dimension Estimation

Figure 2: Overview of the proposed FGM-HD framework. (a) **Training and Sampling Strategies:** During training (gray line), input noise is recursively processed by the FGM (purple section), and the generated images are evaluated by the HD estimation module to compute HD loss. Then, the HD loss is dynamically weighted by the MMDS strategy (green section) through $\lambda(t)$ to balance image quality and structural diversity during optimization. During inference (blue line), a batch of samples is generated from noise via FGM and passed through the Sampling Strategy (blue section). Only geometrically richer outputs with HD values above the threshold (τ) are retained while others are discarded and regenerated, ensuring structurally diverse outputs without modifying the generator architecture. (b) **Hausdorff Dimension Estimation:** The HD estimation (yellow section) is performed using a multi-scale convolutional network built upon the ResNet152 architecture, enabling accurate and efficient HD prediction directly from image embeddings.

to redundancy and hindering the representation of complex data distributions. Thus, enhancing diversity remains a critical challenge for improvement.

Hausdorff Dimension

The HD concept (Hausdorff 1918) was originally introduced to measure the complexity of geometric sets, enabling a precise characterization of self-similar or irregular structures. For instance, Khrulkov and Oseledets (2019) showed that generative models can approximate data supported on low-dimensional manifolds, with HD providing a useful measure of their expressive capacity. Simsekli et al. (2020) demonstrated that HD serves as a proxy for model generalization, revealing a strong correlation between geometric complexity and generalization ability. More recently, Li et al. (2021b) proposed Hausdorff GAN, which incorporates HD into the training objective to align the intrinsic dimensionality of real and generated data, resulting in improved output quality and diversity. Leveraging HD, FGMs can generate more diverse outputs while maintaining structural integrity, addressing the limitations of traditional fractal self-similarity.

Methods

Preliminary of Fractal Generative Models (FGMs)

Inspired by the recursive and self-similar nature of fractals, FGMs generate complex patterns by repeatedly applying the same generative module at different scales, enabling them to capture the data distribution's global and local features. This recursive design allows FGMs to generate high-quality images efficiently while maintaining multiscale structural details, making them suitable for real-time inference and deployment on resource-constrained devices. An overview of the generation process is illustrated in Figure 1 consisting of the following steps:

- **Initial Generation.** First, a low-resolution image is generated using a base generative model. We use the masked autoregressive model (MAR) (Li et al. 2024) for continuous value autoregression, employing a diffusion-based loss function to avoid discrete tokenization. This serves as the foundation for further refinement.
- **Recursive Refinement.** The image is progressively refined through the recursive application of the generative module. Initially, the image is divided into coarse regions, which are subdivided further at each level. Then,

the module synthesizes higher-resolution content, capturing global structure and local details. At level l, the image I_l is generated from I_{l-1} as follows:

$$I_l = Module(I_{l-1}, \theta_l), \tag{1}$$

where θ_l represents the parameters of the generative module at level l.

- **Final Output.** After several recursive steps, a high-resolution image is produced, characterized by intricate details that emerge through the recursive process.

Hausdorff Dimension Estimation

The widely used *box counting method* (Mandelbrot 1983) for HD estimation is accurate but computationally expensive for high-resolution images and sensitive to noise and edge complexities, limiting its applicability for modern image analysis (Gneiting, Ševčíková, and Percival 2012). HD reflects an image's structural features, with intricate textures and self-similar patterns generally exhibiting higher values (Napolitano, Ungania, and Cannata 2012). *Convolutional neural networks (CNNs)*(LeCun et al. 1989), due to learning hierarchical representations, are well-suited to capture these structural features for efficient HD estimation (Valle et al. 2022).

To address the limitations of the box counting method and leverage CNNs, we propose a novel method that directly extracts dimensionality features from image embeddings. Our approach utilizes the *ResNet152* (He et al. 2016) architecture, enhanced with a **multi-scale convolutional module**, as shown in Figure 2 (b). This model leverages the deep feature extraction capabilities of ResNet152, enhanced by multi-scale convolutions, to capture image features across different scales, which is then used in a regression layer for efficient and accurate HD prediction. Specifically, our method achieves superior HD estimation precision while significantly reducing computational overhead, making it more suitable for real-world applications.

Framework

Training Strategy

To balance image quality and structural diversity during training, we formulated the learning objective as a hybrid loss function that combines the base generative loss and the *Hausdorff Dimension loss (HD loss)*, which is defined as:

$$L_{\text{HD}} = \left|\text{HD}_{\text{gen}} - \text{HD}_{\text{target}}\right|, \tag{2}$$

where HD_{gen} denotes the estimated HD of a generated image, and $\text{HD}_{\text{target}}$ represents the class-specific median HD from the training dataset.

To incorporate HD loss into training in a stable and effective manner, we introduced a novel scheduling strategy termed **Monotonic Momentum-Driven Scheduling (MMDS)**, which defines a time-dependent weighting function $\lambda(t)$ that gradually increases during training to control the influence of HD loss. The overall training objective becomes:

$$L_{\text{total}} = L_{\text{gen}} + \lambda(t) \cdot L_{\text{HD}}, \tag{3}$$

Algorithm 1: Monotonic Momentum-Driven Scheduling (MMDS) Strategy

Require: Epochs E, momentum $\mu \in [0, 1]$, scale factor γ
Initialize $\lambda \leftarrow 0$, $m \leftarrow 0$, $L_{\text{prev}} \leftarrow 0$
for epoch e to E **do**
  Compute validation loss L_{val}
  $\Delta L \leftarrow \max(0, L_{\text{prev}} - L_{\text{val}})$
  $m \leftarrow \mu \cdot m + (1 - \mu) \cdot \gamma \cdot \Delta L$
  $\lambda \leftarrow \lambda + m$
  $L_{\text{prev}} \leftarrow L_{\text{val}}$
  Compute $L_{\text{total}} = L_{\text{gen}} + \lambda \cdot L_{\text{HD}}$
  Backpropagate and update model
end for

where L_{gen} is the primary generative loss, and $\lambda(t)$ is a non-negative, monotonically increasing coefficient.

As shown in Figure 3, in the early stages of training, the model generates noisy images, leading to HD estimates with large deviations and high variance, making HD estimation highly inaccurate and unstable. As training progresses, with the generated images improved, HD estimation becomes more reliable and it is feasible to gradually introduce HD loss to enhance structural diversity without compromising image quality. To implement this progression, $\lambda(t)$ is constructed via a momentum-driven accumulation scheme, inspired by the SGD (Robbins and Monro 1951). This scheme ensures that $\lambda(t)$ increases monotonically, with its rate of growth controlled by the accumulated momentum. This momentum-driven approach reflects the training state, allowing for smooth and controlled growth over time, which avoids abrupt changes in HD loss weighting, outlined in Algorithm 1. Compared to static schedules, such as exponential or linear increases, MMDS is better suited to dynamically adjust to the model's evolving requirements during training. As demonstrated in our experiments in Figure 5, MMDS results in a smoother loss trajectory and improved convergence stability while effectively balancing image quality and structural diversity throughout training.

Sampling Strategy

To further promote structural diversity in FGM's outputs, we introduced a post-processing strategy termed *Hausdorff Dimension sampling (HD sampling)*. This method operates after initial sample generation and selectively retains images that exhibit sufficient structural complexity, as measured by their estimated HD. The complete filtering procedure is outlined in Algorithm 2.

Specifically, once a batch of candidate images is generated, we computed each sample's HD and compared it with a predefined threshold. Only samples with HD values exceeding this threshold are retained, while others are discarded and regenerated from the beginning, shown in Figure 2 (a).

By filtering out low-HD images, this method effectively suppresses overly smooth, repetitive, or degenerate outputs, resulting in a final sample set that is visually diverse and structurally rich. As this process is applied post-generation, it incurs no additional computational cost during training

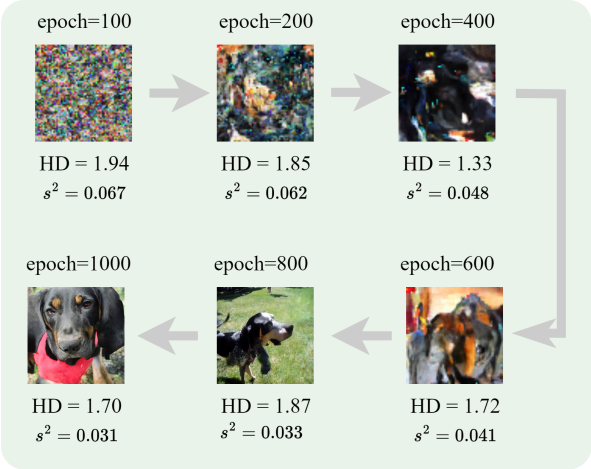


Figure 3: Evolution of image quality and HD variance across training epochs. Early-stage generations are noisy with unstable HD values, while later epochs yield high-quality images with more reliable HD estimation.

Algorithm 2: HD Sampling Strategy

Require: Threshold T_{HD}, batch of generated samples $\{I_1, I_2, ..., I_N\}$

Initialize empty list of selected samples S
for each generated sample I_i in $\{I_1, I_2, ..., I_N\}$ **do**
  Compute $\text{HD}_{\text{generated}}(I_i)$
  if $\text{HD}_{\text{generated}}(I_i) \geq T_{\text{HD}}$ **then**
    Add I_i to selected list S
  else
    Regenerate I_i from scratch
  end if
end for
Return selected list S

and remains fully compatible with any pre-trained generative model.

Experiments

To validate our proposed FGM-HD framework's effectiveness, we conducted the following experiments: a) evaluating the performance of HD estimation network, b) assessing the overall effectiveness of the FGM-HD framework. All experiments were conducted on NVIDIA RTX 4090 GPUs, using the PyTorch framework for deep learning.

Dataset

Fractal Dataset. First, we constructed a dedicated dataset to systematically study the HD behavior in controlled settings. This dataset consists of 1,200 images evenly divided into three categories: (i) a collection of canonical fractal images (e.g., Sierpinski triangle and Koch snowflake) with known theoretical HD; (ii) a diverse set of fractal patterns generated using IFS (Barnsley 2014) with varying parameters to create a wide range of structural complexities; and (iii) FractalDB images (Kataoka et al. 2020) with HD values that are approximated using the box counting method. This balanced composition enables the validation of HD estimation accuracy and benchmarking of generative models on data with clearly defined fractal complexity.

Generation Dataset. Additionally, we evaluated our HD-based methods on the ImageNet-1000 dataset (Deng et al. 2009), which contains high-resolution images from 1,000 diverse categories, and all the images are resized to 256×256 for consistency. Using our HD estimation network, we analyzed structural complexity across categories and applied class-wise HD targets during training. The images generated by our proposed **FGM-HD** are compared with outputs from publicly available pre-trained generative models.

Method	Type	Error ↓	Time(s) ↓
Box Counting	Non-learning	**0.002**	4.70
Power Spectrum		0.079	3.41
Perimeter-Area		0.137	6.85
Sandbox Method		0.094	5.73
ResNet152	Learning	0.012	0.40
Ours		0.005	**0.32**

Table 1: Comparison of HD estimation methods, along with their accuracy (Error) and runtime (Time).

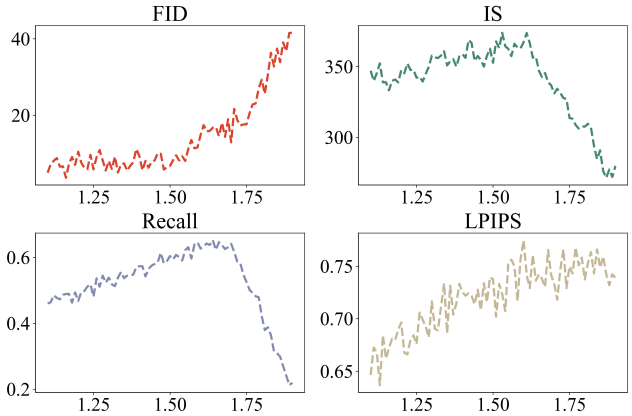


Figure 4: Performance trends of evaluation metrics under varying HD thresholds.

Evaluation Metrics

To comprehensively evaluate our HD-based strategy effectiveness, we adopted four widely used metrics to jointly assess the diversity and quality of the generated results. We used the *Fréchet inception distance (FID)* (Heusel et al. 2017) to measure distributional similarity to real images, and *Inception Score (IS)* (Salimans et al. 2016) to capture both semantic clarity and variety. To assess structural diversity, we used *Recall* (Kynkäänniemi et al. 2019), which evaluates the coverage of the target distribution, and *LPIPS* (Zhang

Model	Type	FID ↓	IS ↑	Recall ↑
BigGAN-deep	GAN	6.95	198.2	0.28
GigaGAN		3.45	225.5	0.61
StyleGAN-XL		2.30	265.1	0.53
ADM	Diffusion	4.59	186.7	0.52
Simple Diffusion		3.54	205.3	0.56
VDM++		2.12	267.7	–
SiD2		**1.38**	–	–
DiffiT		1.73	276.5	0.62
JetFormer	AR+Flow	6.64	–	0.56
RCG	MAGE	2.15	253.4	0.53
FGM	Fractal	6.15	348.9	0.46
FGM-HD(Ours)		6.21	**367.1**	**0.64**

Table 2: Quantitative evaluation of pixel-level generative models on ImageNet 256×256.

Variant	FID ↓	IS ↑	Recall ↑	LPIPS ↑
FGM (baseline)	6.15	348.9	0.46	0.64
+ Fixed HD Loss only	6.22	333.7	0.47	0.65
+ MMDS only	**6.04**	361.7	0.51	0.69
+ HD Sampling only	6.78	357.9	0.58	0.73
+ MMDS & Sampling	6.21	**367.4**	**0.64**	**0.76**

Table 3: Ablation study of HD-based components. Each component contributes independently to the overall performance, with the combination of MMDS and Sampling Strategy achieving the best balance between diversity and visual quality.

et al. 2018), which measures the perceptual difference between generated image pairs, providing a direct evaluation of perceptual diversity. These metrics offer a holistic evaluation of the HD loss and HD sampling method's effect on generative quality and diversity.

Comparison of HD Estimation Methods

To assess the accuracy and computational efficiency of different HD estimation techniques, we compared several widely used classical methods and deep learning approaches. Specifically, we evaluated *box counting*, *power spectrum* (Pentland 1984), *perimeter-area scaling* (Russ 1994), and *sandbox method* (Bandt 1991), as well as deep convolutional regression models based on *ResNet152* and our proposed *multi-scale convolutional network*. All methods were tested on our proposed fractal dataset. Table 1 presents each method's average estimation error and runtime per image.

The results indicate that box counting remains the most precise technique partially because it was used to label a part of the dataset. In contrast, the other methods demonstrated higher error and longer runtime. Our multi-scale convolutional network achieves comparable accuracy with faster inference. In practice, especially when using HD to improve output diversity, computational efficiency is often more important than maximum precision. Therefore, a moderate reduction in accuracy is acceptable if the estimates remain consistent.

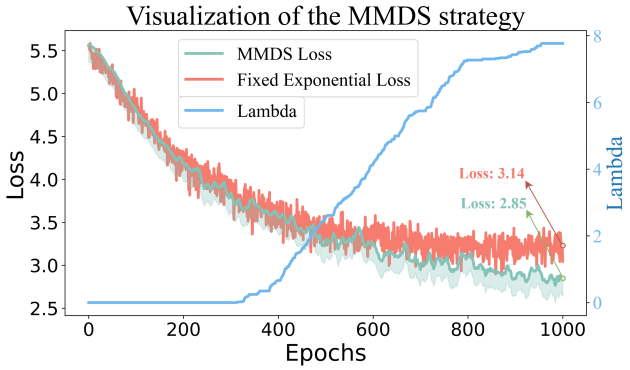


Figure 5: Visualization of the MMDS strategy. The blue curve shows the variation of λ optimized by MMDS, the red curve represents the fixed exponential loss (final value: 3.14), and the green curve shows the MMDS loss (final value: 2.85). MMDS Strategy leads to a smoother and lower loss trajectory during training.

Effectiveness of Monotonic Momentum-Driven Scheduling (MMDS)

To validate our MMDS strategy's effectiveness, we visualized the evolution of $\lambda(t)$ and the corresponding training loss in Figure 5. The blue curve depicts the momentum-driven schedule, which begins increasing gradually at epoch 300 and plateaus beyond 800. The red and green curves represent the training losses under two different $\lambda(t)$ schedules: the red curve corresponds to a fixed exponential schedule end up to 3.14, while the green curve illustrates our MMDS strategy end up to 2.85. Training was stopped after 1,000 epochs because further training did not significantly improve either quality or diversity. Since MMDS involves a monotonic increase in HD loss, this stopping criterion prevents overfitting by ensuring a balanced focus on visual quality and diversity. The optimal balance between these factors was reached after 1,000 epochs, allowing efficient convergence.

Compared to the fixed exponential schedule, the momentum-driven $\lambda(t)$ yields a significantly smoother and lower loss trajectory. Despite the increasing contribution of HD loss over time, the smoothed loss under dynamic weighting exhibits consistent and stable descent. This indicates that MMDS effectively introduces structural diversity without disrupting training convergence. These results demonstrate that the proposed scheduling mechanism successfully balances structural complexity and optimization stability in generative modeling.

HD Sampling Threshold

To evaluate the of HD sampling effect, we applied varying HD thresholds to the generated samples and monitored performance using FID, IS, Recall, and LPIPS, as shown in Figure 4. As the threshold increases from 1.1 to around

Figure 6: Comparison of generated 256×256 images between vanilla FGM and FGM-HD across representative ImageNet classes.

1.7, FID remains low around a threshold of 1.55 before rising sharply, while IS improves moderately but declines beyond 1.6. Recall steadily increases until 1.7 before dropping. LPIPS, which measures perceptual similarity, increases with the threshold, reflecting improved image quality. However, excessively high values may indicate perceptual distortions, suggesting a trade-off between quality and diversity.

These results suggest that HD thresholds between 1.55 and 1.60 provide an effective balance, preserving sample quality while enhancing structural diversity. This demonstrates the applicability of HD sampling as a targeted post-processing strategy.

Comparison with Baseline Generative Models

To evaluate our proposed HD-based enhancement's effectiveness during the challenging pixel-by-pixel image generation task, we compared several variants of our FGM-HD framework with representative baseline generative models, including *GANs* (Brock, Donahue, and Simonyan 2018; Sauer, Schwarz, and Geiger 2022; Kang et al. 2023), *diffusion models* (Dhariwal and Nichol 2021; Kingma et al. 2021; Hoogeboom et al. 2024; Hatamizadeh et al. 2024), and other novel methods (Li, Katabi, and He 2024; Tschannen, Pinto, and Kolesnikov 2024). The results are summarized in Table 2.

The experimental results demonstrate that incorporating HD loss during training and HD sampling during inference significantly improves the diversity of the generated results. Specifically, HD loss encourages higher structural complexity, which leads to improved recall and perceptual variation. HD sampling, as an effective post-processing mechanism, further enhances diversity without altering the model architecture. Overall, this combination achieves a 39% improvement in diversity metrics, while maintaining image quality, as summarized in Table 3.

Qualitative Analysis

To qualitatively demonstrate our method's capacity to generate diverse and high-quality images, Figure 6 compares images generated by vanilla FGM (left) and FGM-HD (right). The images span a wide range of semantic categories and exhibit rich structural variation, with noticeable improvements in object shape, color, pose, and background after incorporating HD. These results visually support the improvements in diversity metrics, highlighting the model's ability to synthesize realistic, class-consistent images with enhanced visual diversity. All the images are generated at a resolution of 256×256 in a pixel-by-pixel manner.

Conclusion

We proposed a novel HD-guided framework to enhance the structural diversity of FGMs without compromising visual quality. During training, we incorporated HD-based loss with MMDS strategy, which dynamically adjusts the influence of HD loss, ensuring an optimal balance between image quality and structural diversity. During inference, we applied HD-guided rejection sampling to retain only geometrically richer outputs, further promoting diversity. Additionally, we introduced an efficient and learnable HD estimation method that directly predicts HD from image embeddings, significantly improving computational efficiency and accuracy. Our approach demonstrated a 39% improvement in generation diversity compared to vanilla FGM while maintaining competitive image quality. By introducing HD into the FGM framework, we provide a principled method to enhance diversity without sacrificing quality. Furthermore, the MMDS strategy offers a generalizable optimization technique for hybrid-loss models. Future work includes applying this framework to conditional and multi-modal generative models, and developing perceptually aligned HD estimation methods for large-scale image synthesis.

Acknowledgments

This work was supported in part by the National Major Scientific Instruments and Equipments Development Project of National Natural Science Foundation of China under Grant 62427820, in part by the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation under Grant 2024NSFTD003, in part by the Fundamental Research Funds for the Central Universities under Grant 1082204112364, in part by the Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music(Grant No. 22DMAKL04). Numerical computations were supported by Chengdu Haiguang Integrated Circuit Design Co., ltd. with HYGON K100AI DCU units.

References

Bandt, C. 1991. Deterministic fractals and fractal measures.

Barnsley, M. F. 2014. *Fractals everywhere*. Academic press.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Gneiting, T.; Ševčíková, H.; and Percival, D. B. 2012. Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statistical Science*, 247–277.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hatamizadeh, A.; Song, J.; Liu, G.; Kautz, J.; and Vahdat, A. 2024. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, 37–55. Springer.

Hausdorff, F. 1918. Dimension und äußeres Maß. *Mathematische Annalen*, 79(1): 157–179.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hoogeboom, E.; Mensink, T.; Heek, J.; Lamerigts, K.; Gao, R.; and Salimans, T. 2024. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10124–10134.

Kataoka, H.; Okayasu, K.; Matsumoto, A.; Yamagata, E.; Yamada, R.; Inoue, N.; Nakamura, A.; and Satoh, Y. 2020. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*.

Khrulkov, V.; and Oseledets, I. 2019. Universality theorems for generative models. *arXiv preprint arXiv:1905.11520*.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in neural information processing systems*, 34: 21696–21707.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32.

Lan, G.; Xiao, S.; Yang, J.; and Wen, J. 2024. Generative model perception rectification algorithm for trade-off between diversity and quality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13328–13336.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

Li, M.; Lv, J.; Tang, C.; Wang, J.; Lai, Z.; and Huang, Y. 2021a. Combination of certainty and uncertainty: Using FusionGAN to create abstract paintings. *Neural Networks*, 144: 443–454.

Li, T.; Katabi, D.; and He, K. 2024. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37: 125441–125468.

Li, T.; Sun, Q.; Fan, L.; and He, K. 2025. Fractal generative models. *arXiv preprint arXiv:2502.17437*.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

Li, W.; Liang, Z.; Ma, P.; Wang, R.; Cui, X.; and Chen, P. 2021b. Hausdorff GAN: Improving GAN generation quality with Hausdorff metric. *IEEE Transactions on Cybernetics*, 52(10): 10407–10419.

Mandelbrot, B. B. 1983. The fractal geometry of nature/Revised and enlarged edition. *New York*.

Napolitano, A.; Ungania, S.; and Cannata, V. 2012. Fractal dimension estimation methods for biomedical images. In *MATLAB-A Fundamental Tool for Scientific Computing and Engineering Applications-Volume 3*. IntechOpen.

Pentland, A. P. 1984. Fractal-based description of natural scenes. *IEEE transactions on pattern analysis and machine intelligence*, (6): 661–674.

Prusinkiewicz, P.; and Lindenmayer, A. 2012. *The algorithmic beauty of plants*. Springer Science & Business Media.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Russ, J. C. 1994. *Fractal surfaces*. Springer Science & Business Media.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Shen, H.; Zhang, J.; Xiong, B.; Hu, R.; Chen, S.; Wan, Z.; Wang, X.; Zhang, Y.; Gong, Z.; Bao, G.; et al. 2025. Efficient Diffusion Models: A Survey. *arXiv preprint arXiv:2502.06805*.

Simsekli, U.; Sener, O.; Deligiannidis, G.; and Erdogdu, M. A. 2020. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33: 5138–5151.

Tschannen, M.; Pinto, A. S.; and Kolesnikov, A. 2024. JetFormer: An autoregressive generative model of raw images and text. *arXiv preprint arXiv:2411.19722*.

Valle, D.; Wagemakers, A.; Daza, A.; and Sanjuán, M. A. 2022. Characterization of fractal basins using deep convolutional neural networks. *International Journal of Bifurcation and Chaos*, 32(13): 2250200.

Van Den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, 1747–1756. PMLR.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Werbos, P. J. 2002. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10): 1550–1560.

Wiatrak, M.; Albrecht, S. V.; and Nystrom, A. 2019. Stabilizing generative adversarial networks: A survey. *arXiv preprint arXiv:1910.00927*.

You, T.; Kim, S.; Kim, C.; Lee, D.; and Han, B. 2022. Locally hierarchical auto-regressive modeling for image generation. *Advances in Neural Information Processing Systems*, 35: 16360–16372.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.