2024 IEEE International Conference on High Performance Computing and Communications (HPCC) | 979-8-3315-4046-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/HPCC64274.2024.00206

6Diffusion-LM:IPv6 address generation method based on diffusion-LM

1st Xinyi Zhao
School of Computer Engineering and Science
Shanghai University
Shanghai, China
zhaoxinyi@shu.edu.cn

2nd Huahu Xu*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
huahuxu@shu.edu.cn

3rd Ruiping Xing
School of Computer Engineering and Science
Shanghai University
Shanghai, China
xrp@shu.edu.cn

4th Yiqin Gao
High Performance Computing Center
Shanghai Jiao Tong University
Shanghai, China
gaoyiqin95@sjtu.edu.cn

5th Jingkun Xu
College of Professional Studies
Northeastern University
Massachusetts, United States
jingkun-xu-official@outlook.com

***Abstract*—IPv6 is instrumental in the ultra-large-scale intelligent computing interconnection system, yet its integration is not without challenges. The vast IPv6 address space renders traditional brute-force scanning methods infeasible, with the considerable time and resource consumption severely impacting the integration of supercomputing capabilities. This also affects the accuracy and efficiency required in high-performance computing environments. Consequently, it becomes necessary to develop new scanning technologies to address the unique challenges presented by the expansive IPv6 address space. Our novel approach, 6Diffusion-LM, transforms IPv6 scanning by fusing diffusion and linguistic models. Utilizing the Transformer architecture, it excels at extracting key features from IPv6 addresses and employs clustering algorithms to organize them effectively. Building upon the BERT pre-trained language model, 6Diffusio-LM integrate a noise mechanism that encapsulates the inherent randomness and unpredictability inherent in IPv6 address generation. The model then refines this process by progressively eliminating noise to yield precise and clear IPv6 addresses. Additionally, our proprietary embedding method enhances the generation process, ensuring higher quality addresses. Our experiments demonstrate that 6Diffusion-LM surpasses conventional methods, boasting a remarkable hit rate improvement to 43.53%**

***Index Terms*—IPv6,Diffusion-LM,Transformer,Scanning**

I. Introduction

In the ultra-large-scale intelligent computing interconnection system composed of heterogeneous hybrid intelligent centers and supercomputing Internet, IPv6 plays a key role in expanding the network scale and providing sufficient address resources [1], and at the same time provides a solid foundation for the new generation of intelligent computing and large-scale data transmission. With the rapid increase in the number of devices and service accesses globally, the vast address space provided by IPv6 can effectively support the interconnection of large-scale devices and significantly improve the scalability of the network to meet the needs of complex computing tasks.

However, the current widespread application of IPv6 still faces challenges. Due to the extremely large IPv6 address space, the traditional brute-force scanning method is no longer feasible in practice, and the process would take millions of years with current network speeds and computing power. This greatly limits the ability of network researchers to perform comprehensive measurements and monitoring of IPv6 networks.

In early explorations, researchers relied heavily on passive measurements [2]- [3]and active scanning [4]- [5]. However, the limitations of passive methods and their dependence on the strategic deployment of monitors make fast IPv6 scanning particularly urgent in the research field.Although previous research has proposed a variety of technical solutions, the inherent characteristics of IPv6 networks still pose challenges to algorithm design. The recently proposed target generation algorithms [6]- [7] have used active IPv6 addresses as seed addresses by learning their salient features in order to generate candidate addresses with high probability of activity. Despite the achievements in active IPv6 address detection, comprehensive and effective discovery of active IPv6 addresses still faces challenge, especially in the low hit rate. Although existing research has explored statistical and semantic information, it still fails to fully utilize inter-address correlation and internal structure, resulting in a low match between generated addresses and those actually use.

To address these challenges, we propose a semantic information-based IPv6 address generation method, 6Diffusion-LM, inspired by natural language processing techniques [8].The 6Diffusion-LM system consists of two core modules: a seed classifier and an address generator. The classifier categorizes IPv6 addresses into different semantic classes or patterns and the generator learns IPv6 address generation patterns based on the Diffusion-LM architecture.

The main contributions of this paper are as follows:

(1)An innovative 6Diffusion-LM architecture is proposed to apply the diffusion model to the IPv6 address generation task, which gradually generates reasonable IPv6 addresses from random noise through a denoising process.

979-8-3315-4046-3/24/$31.00 ©2024 IEEE
DOI 10.1109/HPCC64274.2024.00206

(2)Improved the seed classification algorithm by introducing the powerful feature extraction capability of the language model, treating IPv6 address as a special language, and learning the hidden representation of the address using the pre-trained BERT model.

(3)The quality of generated addresses is significantly improved. Experimental results show that 6Diffusion-LM outperforms other target generation algorithms in several metrics.

Section II, we survey the pertinent literature and outline the novel aspects of our research methodology. Section III details the technical implementation of our approach. Section IV evaluates the performance, demonstrating its effectiveness and efficiency. Finally, Section V synthesizes the paper's content, summarizing our contributions and discussing broader implications.

II. Preliminaries

A. IPv6 Address

IPv6 addresses are 128 bits long and can be represented as eight groups of hexadecimal numbers. Each group of numbers is a number obtained by converting binary to hexadecimal, often called a nybble, and each group is separated by a colon. To give a typical IPv6 address: 2001:0DB8:0000:0023:0008:0800:200C:417A.

Each IPv6 address consists of three parts: a global routing prefix, a local subnet identifier, and an interface identifier (IID) [9]. The global routing prefix is used to determine the routing of traffic over the Internet, directing packets to the destination Local Area Network (LAN). In contrast, IIDs are used to identify individual host interfaces within a LAN, and their assignment is more flexible than global prefixes According to the IETF's RFC 7707 standard, the IID portion can be divided into the following typical patterns [10]. These patterns specify how the IIDs are generated and the semantic information they contain.

- Low Byte. All IID bytes are set to zero.
- Embedded IPv4. IPv4 addresses can be mapped to the lower 32 bits of the IID, i.e. the last 4 bytes of the IID.
- Embedded Port. The service port number is encoded in the lowest byte of the IID, while all other bytes of the IID are set to zero.
- Randomization [11]. Addresses may be configured using stateless address autoconfiguration SLAAC, which generates a pseudo-randomized 64-bit IID, replacing the traditional MAC address-based IID.

The structure of IPv6 addresses is complex and contains multiple semantic fields. For the target IPv6 address generation task, the generation algorithm needs to be able to focus on a specific address pattern and generate IPv6 addresses that satisfy the characteristics of the pattern, rather than trying to cover the entire huge IPv6 address space. Classifying addresses into different pattern categories can reduce the difficulty of the generation task and improve the relevance and effectiveness of the generated addresses. By focusing on a specific address pattern of interest, the generation algorithm can learn the address distribution and structural characteristics of the pattern, and thus generate more reasonable IPv6 addresses that match the target pattern, instead of facing the huge pressure of the entire address space.

B. Diffusion-LM

The Diffusion Language Model (Diffusion-LM) is a novel approach to language generation, inspired by the success of diffusion models in the field of image generation [12]. By adapting the principles of these models to text generation [8], Diffusion-LM achieves the production of high-quality text sequences through a step-by-step denoising process that starts from random noise.

The integration of autoregressive properties within the language modeling component, combined with the iterative diffusion process, enables Diffusion-LM to produce high-quality, controllable text outputs. This level of precision and flexibility is particularly advantageous for tasks like IPv6 address generation, where the model must generate a diverse range of valid and realistic-looking addresses.

Much like text generation, the task of generating IPv6 addresses involves creating structured and sequential outputs that adhere to specific formats and constraints. Both tasks require the model to learn and replicate the fundamental patterns and structures of the target data, whether these are natural language sentences or valid IPv6 addresses.

III. Realated Work

Previous work on active IPv6 address detection can be divided into two main categories: (1) analyzing seed addresses for address pattern mining, and (2) designing algorithms for generating candidate targets and scanning them.

A. Mining address patterns from seed addresses

RFC 7077 [10] identifies common address assignment schemes and potential administrator configuration conventions, suggesting that IPv6 addresses follow specific patterns. Gasser et al [13].used an entropy clustering method to classify known addresses into six pattern classes, showing a strong correlation with configuration schemes. Cui et al [7].employed IPv62Vec to explore the semantic information of IPv6 addresses. This method utilizes deep learning to map addresses into a contiguous vector space, forming clusters of semantically similar addresses. This representation learning reveals the intrinsic structure and regularity of the IPv6 address space.

These studies provide a foundation for understanding IPv6 address allocation patterns and inform the target generation strategies discussed in this paper.

B. Design for target generation and scanning

(1)Traditional Algorithms:

Ullrich et al [14]. proposed the target generation method, a recursive address generation algorithm that synthesizes new addresses from existing data. This method enhances the pool of active addresses by leveraging known information. Foremski et al [6].introduced Entropy/IP, which uses Bayesian networks to model statistical relationships

between different entropy values, highlighting correlations between IPv6 addresses. Murdock et al [15].developed 6Gen, an algorithm that identifies high-density network areas by analyzing the spatial distribution of known active addresses and then generates new addresses to improve scanning efficiency and hit rates. Liu et al [16]. created 6Tree, which constructs an efficient spatial tree structure that dynamically adjusts detection direction, significantly enhancing detection efficiency compared to 6Gen. Song et al [17]. presented the DET algorithm, which uses the Distributed Hash Cluster (DHC) framework to build a spatial tree and identify division points with minimal entropy, thereby optimizing the scanning process.

(2)Deep learning methods.:

Cui et al [18]. introduced the 6GCVAE model, which applies deep learning to address generation using a variational autoencoder. Although 6GCVAE improves generation efficiency compared to Entropy/IP, its hit rate remains lower. Hou et al [19].proposed 6Hit, which incorporates spatial node slicing and spatial repartitioning to optimize detection direction, achieving a hit rate of 11.5 percent. Cui et al [7]. developed 6VecLM, using the Word2Vec algorithm to learn intrinsic sequential relationships of IPv6 addresses. This method generates new addresses by identifying patterns and structures within the addresses.

Existing research provides valuable insights into the patterns and characteristics of IPv6 address allocation. However, challenges such as low hit rates indicate that there is still room for further improvement.

IV. 6DIFFUSION-LM

A. Overview of 6Diffusion-LM

6Diffusion-LM is an IPv6 target address generation architecture. Its structure is shown in FigI. It consists of a seed classifier and an address generator. The seed classifier categorizes IPv6 addresses into different patterns. The obtained information is sent to the subsequent modules. The address generator integrates a Transformer-based diffusion language model Diffusion-LM and a post-processor that evaluates the output of the model. The Address Generator is the core of the IPv6 Diffusion-LM architecture. It integrates a Transformer-based diffusion language model, called Diffusion-LM, and a post-processor.Diffusion-LM is a powerful generative model that utilizes the diffusion process to incrementally generate real and valid IPv6 addresses. The post-processor module then evaluates the output of Diffusion-LM to ensure that the generated addresses conform to the desired format and constraints.

B. Seeds Classificatiion

The seed classifier is tasked with clustering the raw IPv6 addresses and subsequently assigning address classes to their respective patterns. Unlike methods that first identify patterns before segmenting addresses, the seed classifier may introduce a multitude of patterns.

We utilize the Transformer architecture to extract features from IPv6 addresses and apply dimensionality reduction to improve clustering efficiency. We then employ the MiniBatchKMeans algorithm to cluster the reduced-dimensional feature vectors, grouping similar IPv6 addresses into the same cluster. This method facilitates the discovery of patterns and structures within the IPv6 address space.

The seed classifier maps IPv6 addresses into a low-dimensional vector space, enabling effective clustering and analysis. Each IPv6 address is conceptualized as a ”sentence” composed of nybble indices. Initially, the seed classifier preprocesses the raw IPv6 address data by converting each address into a sequence of 32 nested indices, each consisting of a hexadecimal digit (0-9 or a-f) paired with a corresponding positional index (0-9 or a-v). This representation maintains the semantic structure of IPv6 addresses and is suitable for Transformer model training.

After preprocessing, the seed classifier uses these ”sentences” to train the Transformer model, which learns distributed representations of words (or word embeddings), placing semantically similar words closer together in the vector space. In this context, the model learns distributed representations of nested indices, capturing the intrinsic structure and similarities of IPv6 addresses.

For the clustering of IPv6 addresses, the seed classifier applies the MiniBatchKMeans algorithm. MiniBatchKMeans updates cluster centers using a small random batch of data in each iteration, rather than the entire dataset, which makes it more efficient for large-scale data. This approach speeds up the convergence of the algorithm while maintaining clustering quality.

The novelty of the seed classifier lies in its application of the Transformer and clustering algorithms to the analysis of IPv6 addresses, a field previously unexplored. By treating each address as a semantically rich ”sentence,” the method can capture structural similarities between addresses, something not possible with traditional prefix-based or bit pattern-based methods. Furthermore, the seed classifier provides a flexible and interpretable approach to discovering potential patterns and anomalies in the IPv6 address space through the use of Mini.

C. Address Generator

Our model uses the Transformer architecture to model the denoising process, effectively capturing both long- and short-term dependencies in address sequences through a self-attention mechanism. This approach enables our model to generate high-quality IPv6 addresses while overcoming the limitations of traditional language models. In the next subsections, we provide a detailed description.

a) IPv6Tokenizer:

Our model employs a custom-built IPv6Tokenizer to convert IPv6 addresses into numerical sequences, which are then mapped to a continuous vector space through an embedding layer. This embedding efficiently captures the components of the IPv6 addresses, providing the Transformer layer with meaningful inputs. This approach enables the model to better comprehend and manipulate the structure and characteristics of IPv6 addresses.

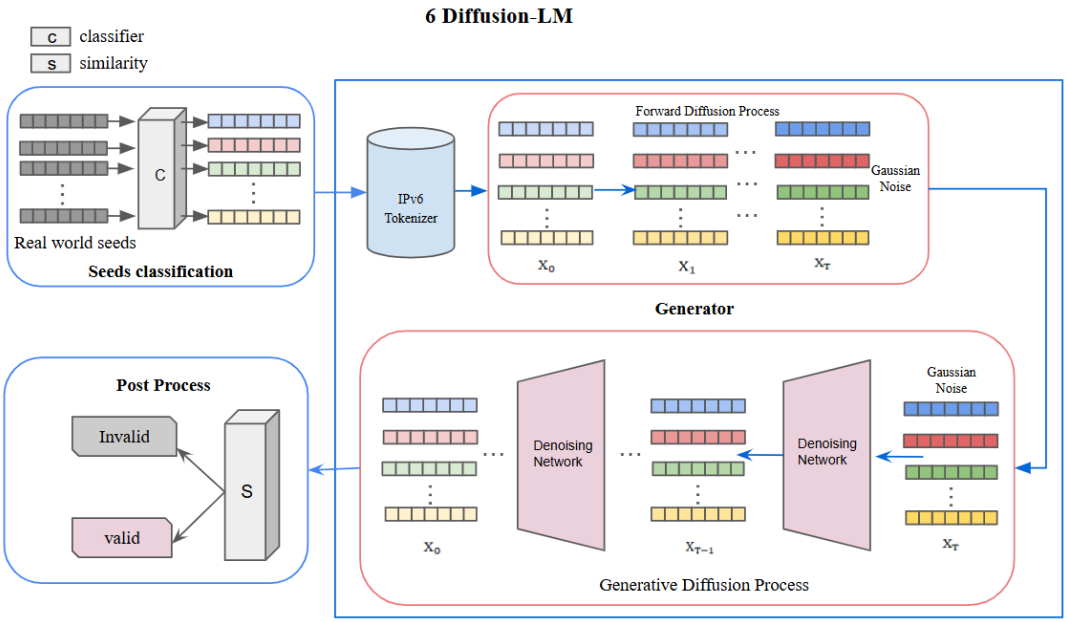

Fig. 1. Overview of 6Diffusion-LM.

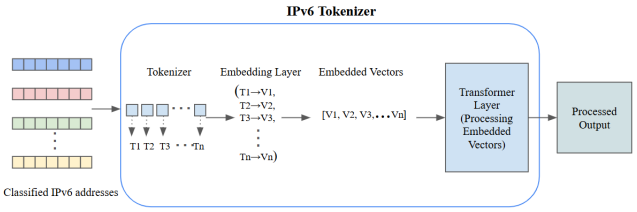

Fig. 2. Overview of IPv6Tokenzier.

For each IPv6 address, we begin with a thorough format and integrity check. This involves expanding shorthand representations to their full form and verifying the address's validity. Any addresses that fail this validation are discarded. The IPv6 tokenizer processes the addresses by tokenizing and encoding them into a sequence of tag IDs. The tokenizer uses the colon (:) as a delimiter to segment the input address into parts, assigning a unique tag to each character within these segments. If a segment is empty, a special token is used. During training and generation, these tag IDs serve as the model's input. This tokenization strategy allows the model to learn the structure and patterns of valid IPv6 addresses and to generate authentic addresses during the sampling process.

b) Diffusion Process:

In this process, the BERT-based diffusion model employs an innovative approach to generate valid IPv6 addresses. The diffusion model first introduces noise based on the BERT pre-trained language model to simulate the randomness and uncertainty in IPv6 address generation. Then, the model gradually recovers clear and accurate IPv6 addresses by gradually removing the noise. By controlling the increase and decrease of noise, the model is able to learn how to generate valid addresses that comply with the IPv6 specification.

1)Forward Noise Addition Process:

We initiate the process by generating a Gaussian noise vector that matches the dimensions of the IPv6 address vector, signifying the starting point of the diffusion process. We incorporate two encoding strategies: position encoding and time encoding. Position encoding is a fixed vector that captures the positional data of each character within the IPv6 address. In contrast, time encoding is a vector that is learnable and signifies the current time step. These encodings are fused with the original IPv6 address vector to create the ultimate input representation. At each time step, we compute a weighted sum of the current address vector and the noise vector based on a predetermined ratio. As the time step progresses, the influence of the noise vector intensifies, while the impact of the address vector wanes, mimicking the transition from the original address vector to a vector dominated by noise. This transition can be mathematically represented as:

$$\mathbf{z}_t = \alpha_t \mathbf{x}_t + (1 - \alpha_t)\mathbf{n} \tag{1}$$

Here, $\mathbf{z}_t$ denotes the noise-address vector at time step t, $\mathbf{x}_t$ is the current address vector, $\mathbf{n}$is the Gaussian noise vector, and α_t is the weight assigned to the address vector at time step t, which decreases from 1 to 0 as the time step advances. At each time step, the noise-address vector $\mathbf{z}_t$ is fed into the BERT model for encoding. The BERT model leverages this input to generate a hidden space representation, capturing the semantic and structural attributes of the noise-address vector. This iterative process of noise addition and BERT encoding continues until the maximum predefined

time step is reached. At the final time step, the address vector is entirely substituted by noise, yielding a pure noise vector. Utilizing a learnable time-step embedding vector, the model can tailor its behavior and outputs to the ongoing denoising phase. This time-step embedding facilitates the model's understanding of the transformation between noise and the original address, enabling it to progressively refine the address to a clearer and more logical form during generation. Our model integrates time-step embedding with input and position embeddings, allowing the Transformer layer to consider the evolving nature of the diffusion process. During training, the model strives to minimize the cross-entropy loss between the generated and actual addresses, while also incorporating a regularization loss to guide certain components of the generated address towards specific target values. This is encapsulated in the following equation:

$$L = L_{CE} + \lambda L_{reg} \tag{2}$$

where L_{CE} is the cross-entropy loss, L_{reg} is the regularization loss, and λ is the regularization coefficient.

2)Generative Denoising Process:

Commencing with the pure noise vector acquired from the terminal stage of the addition process, we employ it as the starting point for our denoising routine. During each epoch t of this process, the current noise vector is fed into the BERT model, which forecasts the subsequent denoised address vector $\mathbf{x}_{t-1}$ by interpreting the noise vector's hidden space representation. The BERT model is tasked with approximating the transformation $f(\mathbf{z}_t;\theta) \rightarrow \mathbf{x}_{t-1}$ for the denoising phase, where θ denotes the BERT model's parameters. To quantify the noise, we align the BERT model's predictions $\hat{\mathbf{x}}_{t-1}$ with the ongoing noise vector $\mathbf{z}_t$. This is achieved through the computation of the mean square error loss (MSE), defined as:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{x}}_{t-1}^{(i)} - \mathbf{z}_t^{(i)})^2 \tag{3}$$

Here, n signifies the vector's dimensionality, while $\hat{\mathbf{x}}_{t-1}^{(i)}$ and $\mathbf{z}_t^{(i)}$ represent the ith components of the predicted denoised address vector and the current noise vector, respectively. Using this estimated noise, we refine the address vector by subtracting the approximated noise from the present noise vector:

$$\mathbf{x}_{t-1} = \mathbf{z}_t - \Delta\mathbf{z}_t \tag{4}$$

In this equation, $\Delta\mathbf{z}_t$ denotes the noise estimated at time step t. This method mimics the stepwise restoration of the original IPv6 address vector from the noisy data. The cycle of denoising, noise estimation, and noise subtraction continues until we reach the predetermined maximum number of time steps, T. At the conclusion of this process, we are left with a pristine IPv6 address vector $\mathbf{x}_0$. During the inference phase, our model initiates with a randomly generated noise vector and progressively refines the IPv6 addresses through iterative denoising. At each stage, the model introduces a calculated degree of noise to ensure the variety of the generated addresses while confirming their validity and adherence to the IPv6 format specifications. This can be expressed as:

$$\mathbf{z}_t = \mathbf{z}_{t-1} - g(\mathbf{z}_{t-1};\theta) \tag{5}$$

Here, $g(\cdot;\theta)$ is the denoising function, parameterized by the BERT model. By engaging the BERT model in the encoding and prediction of address vectors at each time step, our approach empowers the model to discern the underlying structure and generative patterns of IPv6 addresses. The resultant BERT-based diffusion model is adept at producing IPv6 addresses that are both authentic and varied.

3)Output Layer:

The output of the Transformer decoder is converted into an IPv6 address probability distribution through a linear layer followed by a softmax function. The linear layer adjusts the decoder's output to align with the vocabulary size of the IPv6 address generator. Subsequently, the softmax function transforms these scaled outputs into a valid probability distribution, ensuring that all elements are between 0 and 1 and that they sum to 1. This conversion is crucial for generating realistic and valid IPv6 addresses.

Additionally, we incorporate a regularization term designed to guide the generated address segments towards specific target values, thereby enhancing the quality and legitimacy of the addresses. This regularization term helps maintain the structural integrity and practical applicability of the generated IPv6 addresses.

During training, the output layer is optimized in conjunction with the rest of the model to minimize both cross-entropy loss and regularization loss. The cross-entropy loss focuses on improving the model's accuracy in predicting the probabilities of each token, ensuring that the generated sequences are as close as possible to real IPv6 addresses. Meanwhile, the regularization loss ensures that the generated addresses adhere to predefined quality standards and patterns.

These cross-entropy and regularization are combined to form the model's total loss function. By jointly minimizing these losses, we ensure that the generated IPv6 addresses are not only syntactically and semantically accurate but also adhere to the practical constraints and standards necessary for real-world application. This dual optimization approach enhances both the theoretical robustness and practical utility of our IPv6 address generation model.

D. Post-processing Program

The output of the generator is decoded and then organized into standard IPv6 addresses for similarity calculation as shown in Equation 6 to ensure that the generated addresses have a certain similarity with the seed addresses. If the similarity does not meet a certain threshold address will be discarded. Experiments have verified that the efficiency in terms of quality of generated addresses can be improved.

In Equation 1, the output of our generator undergoes decoding and subsequent organization into standard IPv6 addresses, a crucial step aimed at ensuring that the generated addresses maintain a certain level of similarity

with the seed addresses. This similarity calculation process is instrumental in assessing the fidelity of the generated addresses to the original dataset. If the similarity falls below a predetermined threshold, indicating a lack of resemblance to the seed addresses, the respective addresses are discarded from the final output.

Through extensive experimentation, we validate the effectiveness of this approach in enhancing the efficiency and quality of the generated addresses. This methodology not only improves the overall quality of the generated addresses but also enhances the reliability and utility of the generated dataset for various applications, including network analysis, security assessment, and protocol development.

$$Similarity(a, s) = 1 - \frac{Hamming\_\text{dist}(a, s)}{128} \quad (6)$$

V. Evaluation

In this section, we will present our experimental setup and all experimental results to demonstrate the performance of 6Diffusion-LM.

A. Experimental Setup

a) Dataset:

We use the IPv6 Hitlist dataset, a library of addresses collected and made publicly available by Gasser et al [13]. We use a dataset of 9 million IPv6 addresses generated on November 25, 2023 for our experimental analysis. Table I shows the details of this dataset

TABLE I
DETAILS OF THE DATASET USED IN THE PAPER.

Dataset	***Time***	***Description***	***Seeds***
IPv6 hitlist	November 25,2023	responsive-addresses	9063318
		aliased-prefixes	1709573
		non-aliased-prefixes	90738811

b) Validation Methods:

We use xmap [20] to perform a scan of the generated addresses over multiple protocols including TCP/80, ICMPv6, TCP/443, etc. Addresses will be considered as active targets if any of the scanning protocols get a response. We perform scans at different time intervals to mitigate the errors present in a single measurement. We chose 2 days and 5 days and performed multiple scans and if the address responded within the specified time then we consider it as active. We categorize them as "2day-active" and "5day-active" according to the time interval.

c) Assessment of indicators:

To evaluate the quality of the generated candidate sets comprehensively, we adopt a suite of evaluation metrics, which are designed to quantify diverse characteristics and aspects.

1)Hit Rate: The Hit Rate can be defined as the ratio of active addresses generated by the model to the total number of generated addresses. It is used as a measure of the accuracy of the model in predicting active addresses, demonstrating the model's ability to generate such addresses. If Total_number represents the real active target set in the IPv6 address space and Hit_number represents the real set of alias addresses, then the hit rate can be defined as follows:

$$\text{Hit\_rate} = \frac{\text{Hit\_num}}{\text{Total\_num}} \quad (7)$$

2)Generation Rate: The Generation Rate is used to assess the diversity of the model's generation capabilities. It can be defined as the ratio of model-generated active addresses that are not in the seed set to the total number of generated addresses. The generation rate can be expressed as follows:

$$\text{Gen\_rate} = \frac{\text{Gen\_num}}{\text{Total\_num}} \quad (8)$$

3)Novelty.: The Novelty quality metric assesses the dissimilarity between the generated set Gen and the seed set Seed,serving as an indicator of the algorithm's capability to generate novel address sequences. Given a seed set seed,the novelty of the generated set Gen an be measured as follows:

$$\text{Novelty} = \frac{1}{|\text{Gen}|} \sum_{i=1}^{|\text{Gen}|} \left(1 - \max_{j=1}^{|\text{Seed}|} \varphi(\text{Gen}_i, \text{Seed}_j)\right) \quad (9)$$

d) Model setup.:

6The Diffusion-LM architecture consists of three layers, with the hidden layer dimension in the transformer model configured as 256.The multi-head self-attention mechanism in the Transformer is equipped with 8 heads, providing a strong attention to different parts of the input data. In addition, the hidden layer dimension of the feed-forward neural network in the Transformer is set to 1024 to ensure sufficient capacity to learn complex patterns.

To demonstrate the validity of seed categorization and post-processing, we refer to the variant without seed categorization but with post-processing as 6Diffusion-LM-PP, and the variant without post-processing but with seed categorization as 6Diffusion-LM-SC.This distinction allows us to clearly demonstrate the generation of high-quality IPv6 addresses with the improvements in the post-processing step. This distinction allows us to clearly demonstrate the improvement of the post-processing step in generating high-quality IPv6 addresses.

B. Quality of generated addresses

We continue to use the hit rate and generation rate used in previous methods as validation metrics. These metrics allow for a more accurate evaluation of the IPv6 addresses generated by 6Diffusion-LM, thus proving that it outperforms previous research results. By evaluating both the quantity and quality of candidate sets, our approach provides a clear and accurate measure of the efficiency of IPv6 address generation.

a) Baselines.:

The baselines used in our comparison experiments mainly consisted of traditional and deep learning algorithms.

1)Traditional Algorithms: We chose Entropy/IP [6], 6Gen [15], 6Tree [16], and DET [17] as the benchmarks for our study. Entropy/IP and 6Tree utilize entropy information in address segments or spatial trees to guide the exploration

process while 6Tree constructs an entropy-based spatial tree to guide exploration. In contrast, 6Gen uses an algorithmic analysis-based approach to explore the target space by discovering active address clusters.The DET algorithm, similar to Entropy/IP and 6Tree, is an analysis technique that relies on the dynamic behavior of the target space, using behavioral characteristics to guide the exploration process. In this study, we utilized Entropy/IP, 6Tree and DET as benchmarks for both approaches. For 6Gen, since there is no publicly available code, we reimplemented its approach based on the algorithmic description provided in the paper.

2)Deep Learning Algorithms Algorithms: 6GCVAE [18], 6VecLM [7] are three representative works. 6GCVAE utilizes a gated convolutional variational autoencoder (VAE) to construct a generative model for generating new targets by learning a latent representation of IPv6 addresses. In contrast, 6VecLM constructs an IPv6 language model for predicting and generating address sequences using converter architectures and softmax temperature techniques.In this study, we replicated the code of 6GCVAE and 6VecLM and benchmarked these approaches.

TABLE II
QUALITY OF GENERATING TARGETS WITH 500K PROBS.

Category	*Method*	*Hit_rate*	*Gen_rate*	*Novelty*
Traditional	Entropy/IP	11.15%	7.12%	11.93
	6Gen	16.79%	10.25%	10.84
	6Tree	26.88%	23.96%	10.93
	DET	31.12%	12.01%	11.46
Deep Learning	6GCVAE	15.63%	11.76%	12.07
	6VecLM	35.16%	9.20%	11.56
Our Approach	6Diffusion-LM-PP	37.91%	30.26%	13.13
	6Diffusion-LM-SC	34.15%	**37.73%**	**14.96**
	6Diffusion-LM	**43.53%**	35.97%	13.76

b) Target generation.:

To rigorously assess the effectiveness of our target generation algorithms, we have implemented a meticulous evaluation approach to measure the quality of the produced candidate sets. TableII displays the hit and generation rates of various algorithms, including those derived from the IPv6Hitlist dataset, and contrasts them with the outcomes of other deep learning algorithms trained on the identical dataset. Our methods demonstrate marked improvements over conventional algorithms such as Entropy/IP and 6Tree. Moreover, our technique surpasses the language modeling-based 6VecLM, the gated convolutional VAE-based 6GCVAE, with notable enhancements in both hit and generation rates.

In our extensive evaluation, we utilized an integrated strategy to accurately quantify the strengths of the generated candidate sets. Through stringent testing with the Probing IPv6Hitlist dataset, we conducted a detailed comparative analysis of the hit and generation rates across different algorithms. To establish a credible benchmark, we also incorporated the results from deep learning algorithms trained on this dataset. Our algorithm significantly outperforms traditional methods. Notably, our method not only exceeds these traditional technologies in terms of hit rate but also achieves higher rates of generation and novelty, underscoring the efficacy and practicality of our approach in generating premium IPv6 addresses.

In the comparison against 6VecLM and 6GCVAE our method has achieved substantial progress on both the pivotal metrics of hit rate and generation rate. This not only confirms the state-of-the-art nature of our diffusion-based IPv6 address generation technology but also highlights its vast potential for practical applications.

Comparing 6Diffusion-LM-PP, 6Diffusion-LM-SC, and 6Diffusion-LM, it is clear that the inclusion of seed classifiers and post-processors significantly improves the hit rate. However, as can be seen in the table, the post-processor reduces the generation rate and novelty to some extent. This may be due to the application of certain similarity-based filtering criteria.

Figure visually contrasts the performance of Entropy/IP, DET, 6VecLM, 6GCVAE, 6Gen, 6Tree and our proposed 6Diffusion-LM method in terms of hit rates. This graphical representation clearly delineates the performance disparities among the algorithms, further accentuating the significant progress in IPv6 address generation made possible by our innovative approach.

These results compellingly advocate for our algorithm, highlighting its potential to revolutionize the efficiency and effectiveness of network and cybersecurity applications in the realm of IPv6 address generation.

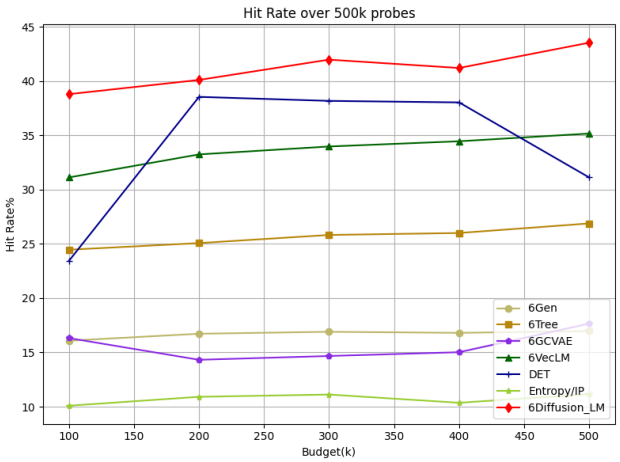

Fig. 3. Hit rate over 500k probes.

C. Analysis of active addresses.

We analyzed the generated IPv6 addresses in detail. By comparing these addresses with a real IPv6 address dataset, we found that our model generates addresses that are highly consistent with the real distribution. After classifying the IID (Interface Identifier) types of the generated addresses, we observed that the proportions of Low-Byte, Embedded-IPv4, Randomized, and EUI-64 categories closely match the distributions found in real data. This indicates that our model is not only capable of generating syntactically

and semantically correct IPv6 addresses but also effectively captures the underlying distribution patterns of different IID types.

Additionally, we analyzed the activity of IPv6 addresses generated by 6Diffusion-LM over time,Table III shows this in detail. By monitoring these addresses, we discovered that most of the generated addresses remained active for at least a week, mirroring the active periods observed in real data. This consistency in temporal stability further demonstrates that the addresses generated by our model resemble real addresses not just in structure but also in behavior over time.

In conducting an exhaustive longitudinal analysis, we noted that the proportions of different classes of IPv6 addresses remained stable throughout the observation period. There were no significant trends or decreases in any particular class, underscoring the robustness of our model in maintaining the diversity and consistency of address types over time. These observations provide valuable insights into the global IPv6 address allocation landscape, highlighting our model's ability to generate realistic and temporally stable IPv6 addresses.

TABLE III
IIDs Pattern Analysis

Active days	Different Patterns			
	IPs	***Low-Byte***	***Embedded-IPv4***	***Other***
1day-activate	100%	17.28%	14.32%	68.40%
2day-activate	89.02%	16.93%	13.09%	59.18%
5day-activate	78.36%	14.43%	11.97 %	51.96%

VI. Conclusion

In this paper, we propose an efficient system called 6Diffusion-LM for generating active IPv6 addresses. This system leverages the diffusion model for IPv6 address generation, combining classification and iterative optimization to produce realistic and active addresses.

The 6Diffusion-LM system begins by classifying existing IPv6 addresses into various categories. Following this classification, it employs a diffusion process to model the IPv6 address space. This process involves starting with random noise and iteratively refining it to generate target addresses that closely mimic real-world distributions. Experimental results demonstrate that 6Diffusion-LM significantly outperforms existing methods in terms of hit rate.

In conclusion, 6Diffusion-LM represents a substantial advancement in the field of IPv6 address generation. By integrating sophisticated classification and diffusion techniques, it achieves a higher hit rate than existing models, providing a robust tool for enhancing IPv6 address detection and allocation strategies.We will further apply the system to the ultra-large-scale intelligent computing interconnection system composed of heterogeneous hybrid intelligence centers and supercomputing internet, to help the efficient operation and management of the system.

References

[1] F. Hilal, P. Sattler, K. Vermeulen, and O. Gasser, "A first look at ipv6 hypergiant infrastructure," *Proc. ACM Netw.*, vol. 2, no. CoNEXT2, jun 2024. [Online]. Available: https://doi.org/10.1145/3656300

[2] D. Plonka and A. Berger, "Temporal and spatial classification of active ipv6 addresses," in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 509–522.

[3] J. Czyz, K. Lady, S. G. Miller, M. Bailey, M. Kallitsis, and M. Karir, "Understanding ipv6 internet background radiation," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 105–118.

[4] O. Gasser, Q. Scheitle, S. Gebhard, and G. Carle, "Scanning the ipv6 internet: towards a comprehensive hitlist," *arXiv preprint arXiv:1607.05179*, 2016.

[5] J. Czyz, M. Luckie, M. Allman, M. Bailey *et al.*, "Don't forget to lock the back door! a characterization of ipv6 network security policy," in *Network and Distributed Systems Security (NDSS)*, 2016.

[6] P. Foremski, D. Plonka, and A. Berger, "Entropy/ip: Uncovering structure in ipv6 addresses," in *Proceedings of the 2016 Internet Measurement Conference*, 2016, pp. 167–181.

[7] T. Cui, G. Xiong, G. Gou, J. Shi, and W. Xia, "6veclm: Language modeling in vector space for ipv6 target generation," in *Machine Learning and Knowledge Discovery in Databases: Applied Data Science Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part IV*. Springer, 2021, pp. 192–207.

[8] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4328–4343, 2022.

[9] R. Hinden and S. Deering, "Ip version 6 addressing architecture," Tech. Rep., 2006.

[10] T. Chown and F. Gont, "Network reconnaissance in ipv6 networks," *RFC 7707*, pp. 1–38, 2016.

[11] S. Thomson, T. Narten, and T. Jinmei, "Rfc 4862: Ipv6 stateless address autoconfiguration," 2007.

[12] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[13] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczyński, S. D. Strowes, L. Hendriks, and G. Carle, "Clusters in the expanse: Understanding and unbiasing ipv6 hitlists," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 364–378.

[14] J. Ullrich, P. Kieseberg, K. Krombholz, and E. Weippl, "On reconnaissance with ipv6: a pattern-based scanning approach," in *2015 10th International Conference on Availability, Reliability and Security*. IEEE, 2015, pp. 186–192.

[15] A. Murdock, F. Li, P. Bramsen, Z. Durumeric, and V. Paxson, "Target generation for internet-wide ipv6 scanning," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 242–253.

[16] Z. Liu, Y. Xiong, X. Liu, W. Xie, and P. Zhu, "6tree: Efficient dynamic discovery of active addresses in the ipv6 address space," *Computer Networks*, vol. 155, pp. 31–46, 2019.

[17] G. Song, J. Yang, Z. Wang, L. He, J. Lin, L. Pan, C. Duan, and X. Quan, "Det: Enabling efficient probing of ipv6 active addresses," *IEEE/ACM Transactions on Networking*, vol. 30, no. 4, pp. 1629–1643, 2022.

[18] T. Cui, G. Gou, and G. Xiong, "6gcvae: Gated convolutional variational autoencoder for ipv6 target generation," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*. Springer, 2020, pp. 609–622.

[19] B. Hou, Z. Cai, K. Wu, J. Su, and Y. Xiong, "6hit: A reinforcement learning-based approach to target generation for internet-wide ipv6 scanning," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[20] X. Li, B. Liu, X. Zheng, H. Duan, Q. Li, and Y. Huang, "Fast ipv6 network periphery discovery and security implications," in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2021, pp. 88–100.