photonics

MDPI

Article

A GenAI-Assisted Framework for Predictive Bandwidth Allocation in TWDM-PON

Andrew Fernando Pakpahan [1,*] and I-Shyan Hwang [2,*]

[1] Department of Informatics Engineering, Universitas Advent Indonesia, Bandung 40559, Indonesia
[2] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 32003, Taiwan
* Correspondence: andrew@unai.edu (A.F.P.); ishwang@saturn.yzu.edu.tw (I.-S.H.)

Abstract

Time and Wavelength Division Multiplexed Passive Optical Networks (TWDM-PONs) need adaptive resource allocation, but conventional mechanisms react only after traffic demand is reported. This paper proposes a Generative Artificial Intelligence (GenAI)-assisted framework that integrates Software-Defined Networking (SDN), virtual Optical Line Termination Hardware Abstraction (vOLTHA), GenAI-based traffic forecasting, and Dynamic Bandwidth Allocation (DBA), and separates the forecasting model from the allocation rules. A Python-based simulation evaluates one realization of the framework, a predictive DBA in which forecasted Assured Forwarding (AF) demand guides Optical Network Unit (ONU) service ordering and bounded bandwidth allocation, with static ONU-to-wavelength assignments. Without any training on PON data, the TimesFM 3 foundation model gave the lowest mean absolute error in predicting the current-cycle AF grant, 57.92% and 53.51% lower than Long Short-Term Memory (LSTM) for two traffic compositions, and a 20.87% lower error than LSTM in forecasting AF arrivals. At the network level, the proposed DBA reduced the mean AF packet delay by about 17.3% relative to Interleaved Polling with Adaptive Cycle Time (IPACT), while preserving aggregate throughput and introducing no Best Effort (BE) packet loss up to 80% offered load.

Keywords: TWDM-PON; Generative AI prediction; dynamic resource allocation; dynamic bandwidth allocation; software-defined networking

1. Introduction

The increasing complexity of modern network infrastructures and the surging demand for bandwidth-intensive applications have propelled a shift from static resource allocation paradigms toward more dynamic, intelligent mechanisms [1,2]. Generative Artificial Intelligence (GenAI), a branch of Machine Learning (ML) that includes generatively pre-trained foundation models, has emerged as a key enabler of this transformation [3]. By forecasting network traffic, GenAI supports proactive, data-driven decision-making, augmenting traditional network management techniques and paving the way for autonomous, self-optimizing systems [4,5].

In parallel, optical access networks have undergone significant evolution to meet escalating performance demands. Time and Wavelength Division Multiplexed Passive Optical Networks (TWDM-PONs), which combine time and wavelength multiplexing to increase throughput, are becoming central to next-generation broadband services [6,7]. However, efficiently managing time slots and multiple wavelengths in dynamically changing environments poses substantial challenges, particularly as user behavior and application

Check for updates

Received: 19 October 2026
Revised: 12 November 2026
Accepted: 17 November 2026
Published: 19 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

profiles become more unpredictable [8,9]. Traditional resource management strategies often rely on static or reactive approaches, which can lead to resource underutilization or congestion during peak periods.

Integrating GenAI-driven predictive analytics into TWDM-PON offers a way to move from reactive to proactive resource allocation. By using historical and real-time network telemetry, GenAI can forecast traffic demand and guide allocation decisions before congestion appears [1,3]. Executing such forecasts requires a programmable infrastructure. The Software-Enabled Broadband Access (SEBA) framework [10] decouples hardware from control and enables centralized programmability, providing a foundation for predictive dynamic resource allocation, which can later be extended to wavelength-level optimization and congestion mitigation.

This paper presents a framework that brings GenAI-based forecasting into TWDM-PON. We describe the required architecture and the interplay between SDN controllers, virtual Optical Line Termination Hardware Abstraction (vOLTHA), and Optical Network Units (ONUs). The framework is evaluated through GenAI-assisted predictive bandwidth allocation and ONU scheduling, while wavelength management is addressed at the architectural level.

2. Literature Review and Related Work

2.1. Evolution of Passive Optical Network

A Passive Optical Network (PON) is a point-to-multipoint fiber access technology in which an Optical Line Terminal (OLT) at the central office serves multiple ONUs at customer premises through a passive optical splitter, without active components along the transmission path.

PON standards have evolved from GPON (ITU-T G.984) [11] and EPON (IEEE 802.3ah) [12] to NG-PON2 (ITU-T G.989) [13] and 10G-EPON (IEEE 802.3av) [14]. NG-PON2 combines time and wavelength multiplexing over four wavelength channels to reach 40 Gbps, and IEEE 802.3ca specifies 25 and 50 Gbps EPON [15]. Recent studies demonstrate that WDM-PON systems can reach downstream and upstream rates of 640 Gbps and 160 Gbps, respectively, using colorless ONUs [16], and a bidirectional intensity-modulated/direct-detection (IM/DD) optical orthogonal frequency-division multiplexing (OFDM) WDM-PON using a Reflective Semiconductor Optical Amplifier (RSOA) as a colorless upstream transmitter has supported 16 × 100 Gbit/s downstream channels over 30 km of single-mode fiber [17].

2.2. Passive Optical Network with Software-Defined Enhancement

Bandwidth enhancements driven by diverse data-intensive applications necessitate the evolution of PON system architectures. A software-defined architecture offers the required agility and flexibility to manage expanding network services, enabling dynamic resource allocation [18,19], streamlined management, and responsive adaptation [20]. This results in PON systems that are both high-capacity and adaptable. Software-Defined Access Networks (SDAN) centralize network control for dynamic, programmable management. The SEBA platform [10], a scalable SDAN implementation, aligns with modern PON bandwidths, allowing operators to efficiently allocate resources and deliver responsive services, thereby fostering an adaptable broadband ecosystem.

Figure 1 illustrates the SEBA Architecture, which integrates SDN into a PON framework. SEBA decouples control functions from hardware, supports diverse access technologies, and routes traffic to the network backbone. Managed by a Carrier Automation Platform and interfaced via a SEBA Northbound Interface (NBI) Client, the Edge Cloud Orchestration layer unifies functions such as vOLTHA, Fabric, and BNG under an SDN Controller, en-

hancing network programmability. The infrastructure includes an Aggregation and Service Gateway directing traffic from a white-box OLT to multiple ONUs, representing customer connections. This demonstrates SDN's role in improving broadband access network control and efficiency through SEBA. Integrating SDN within a PON framework, exemplified by SEBA, provides the programmable control needed to manage time and wavelength resources in TWDM-PON.

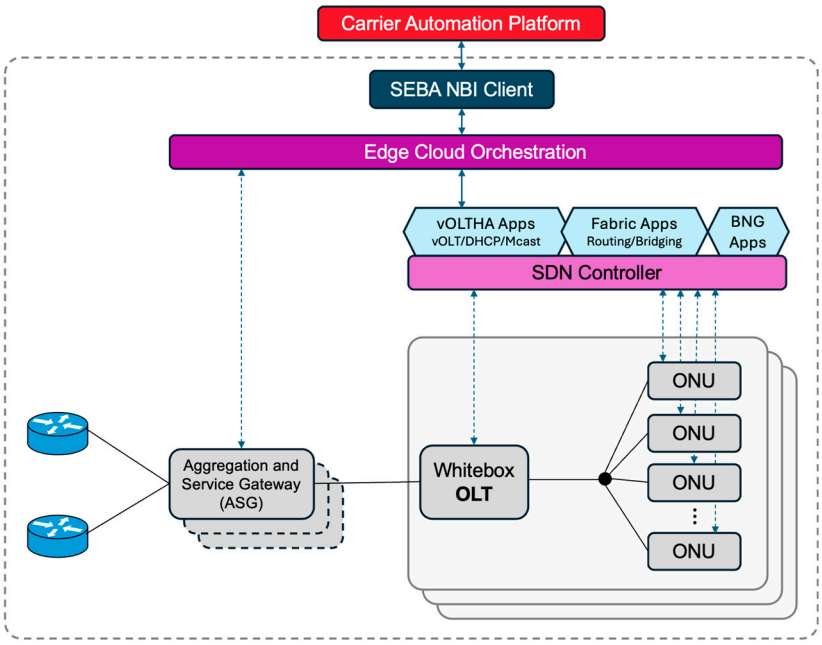

Figure 1. SEBA Architecture [10].

2.3. Generative AI in Networking

Generative AI (GenAI) refers to models that learn the distribution of their training data and can generate new samples from it. Building on earlier advances in neural network training and deep learning [21,22], generative models such as GANs [23] and VAEs [24] were followed by the Transformer architecture [25], which replaced recurrence with self-attention and underpins large pretrained models such as the GPT series [26–28] and native-AI network architectures built on them [29]. Diffusion models [30] and recent large language models [31,32] have extended this progress. In networking, GenAI has been applied at the physical layer for channel estimation and signal classification [33], at the network layer for generative content transmission [34], and to resource allocation, traffic prediction and autonomous operation in wireless, vehicular and edge networks [35–43]. Time-series foundation models extend this generative pretraining paradigm to forecasting, as discussed in Section 2.4.4. However, the application of GenAI to dynamic resource allocation in PONs remains largely unexplored, which motivates the framework proposed in this paper.

2.4. Comparing Machine Learning Paradigms for Network Optimization

This section compares four machine learning paradigms for network optimization, namely traditional ML, deep learning, Reinforcement Learning (RL), and GenAI, and summarizes them in Table 1.

Table 1. Comparison of Different Machine Learning Methodologies With Generative AI.

Aspect	**Traditional ML** [44–46]	**Deep Learning (LSTM)** [47,48]	**Reinforcement Learning (RL)** [39,49]	**Generative AI** [29,50–52]
Data Dependency	High reliance on historical and labeled datasets.	Requires labeled sequential data for training.	Requires interaction data for training through environment feedback.	Reduced dependency on task-specific data through pre-training.
Adaptability	Reactive to environmental changes; requires retraining.	Reactive; adapts to sequential changes but needs retraining for new conditions.	Proactively adapts through continuous learning from real-time interactions.	Can be applied without task-specific training or with light adaptation; performance on PON traffic must be validated.
Scalability	Limited in high-dimensional scenarios.	Scales with sequential data but requires significant computational resources.	Struggles with large state-action spaces; computationally intensive.	Scales with model size, but with higher inference cost.
Training Cost	Low; quick to train on structured datasets.	High; requires GPUs and extensive time for sequential data.	High; trial-and-error learning can be resource-intensive.	High pre-training cost, but usable without task-specific training.
Proactive Optimization	Limited to reactive adjustments based on historical data.	Predicts sequential outcomes but remains reactive.	Capable of proactive adaptation through policy optimization.	Provides forecasts that can support proactive allocation.
Main Limitation	Limited adaptability to rapidly changing traffic.	Requires network-specific training data and retraining.	Costly training and difficult scaling.	Inference latency and computational cost in millisecond-scale control loops.

2.4.1. Traditional ML Approaches

Traditional ML techniques, including XGBoost [44], Support Vector Machines (SVM) [45], and Decision Trees [46], have been used for traffic classification, anomaly detection, and performance prediction in optical access networks. They are computationally efficient and interpretable, but they rely on historical data and adapt poorly to rapidly changing traffic.

2.4.2. Deep Learning Approaches

Deep learning models, particularly Long Short-Term Memory (LSTM) networks [47,48], capture temporal dependencies in sequential data and have been applied to bandwidth prediction in optical access networks. They learn features automatically, but they must be trained on network-specific data and retrained when conditions change.

2.4.3. Reinforcement Learning-Based Optimization

RL methods learn allocation policies by interacting with the network environment [39,49]. They can adapt to changing conditions, but training is costly and scaling to large state and action spaces remains difficult.

2.4.4. Generative AI Frameworks

In this paper, GenAI refers to generatively pretrained foundation models [29,50], which learn to generate the continuation of a sequence and are applied to downstream tasks with little or no task-specific training. Time-series foundation models such as TimeGPT [51] and TimesFM [52] follow this paradigm and produce probabilistic forecasts. In the proposed framework, GenAI is used for forecasting, not for generating synthetic traffic. Pretrained generative forecasting models can reduce dependence on task-specific training data, but they have higher inference cost and their advantage depends on the dataset and metric; their suitability for TWDM-PON must therefore be established empirically.

Table 1 summarizes the strengths and limitations of the four paradigms in terms of data dependency, adaptability, scalability, training cost, proactive optimization, and main limitation.

3. Proposed Architectures

3.1. System Overview

Figure 2 illustrates the proposed transformation of TWDM-PON architecture from a traditional static fiber-optic network to a highly intelligent, AI-enhanced system. As a next-generation broadband technology, TWDM-PON utilizes multiple wavelengths $(\lambda_1, \lambda_2, \ldots, \lambda_8)$ to efficiently support upstream and downstream communication. The first four wavelengths $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are designated for downstream communication, while the remaining four $(\lambda_5, \lambda_6, \lambda_7, \lambda_8)$ are used for upstream communication. This architecture incorporates a Generative AI (GenAI) model to enhance dynamic resource allocation and performance optimization. At the Operator Central Office, an Optical Line Terminal (OLT) is connected to a Passive Splitter, which distributes the optical signals to multiple Colorless ONUs at the customer premises. These ONUs interface with a variety of end-user devices, both wired and wireless. The system's intelligence is driven by ONOS, an SDN controller, which manages and orchestrates the network through the SEBA Northbound Interface. The vOLTHA application provides hardware abstraction, enabling software-defined control of the OLT. Details of each component and their functionalities will be explained below.

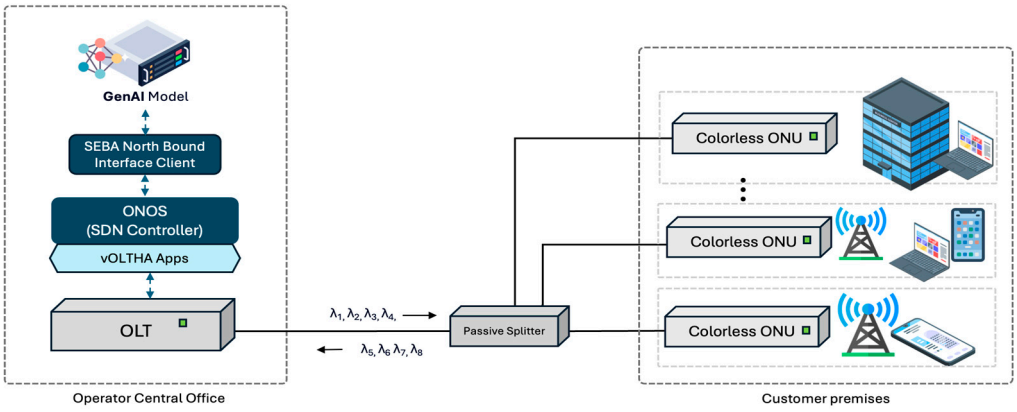

Figure 2. General architecture of the proposed GenAI enhanced TWDM-PON.

3.2. Optical Line Terminal (OLT)

At the operator's central office, the OLT orchestrates key optical functions. It employs WDM components, transceivers, and a Dynamic Bandwidth Allocation (DBA) module for efficient bandwidth distribution, alongside traffic shaping, QoS management, synchronization, and power backup systems. With vOLTHA's hardware abstraction, the OLT is treated as a logical switch, enabling ONUs to appear as virtual ports. This abstraction streamlines

interactions with the SDN controller, allowing traffic flows, bandwidth allocation, and QoS policies to be dynamically managed via protocols like OpenFlow.

3.3. Passive Splitter and Optical Network Units (ONUs)

Signals from the OLT are distributed to multiple ONUs through a passive optical splitter. In the proposed TWDM-PON architecture, the ONUs are colorless and tunable, meaning that they are not permanently associated with a fixed wavelength. Instead, each ONU can be configured to transmit on different upstream wavelengths according to the assignment determined by the network control mechanism. This colorless operation can be realized using Reflective Semiconductor Optical Amplifiers (RSOAs) for wavelength-flexible upstream transmission [17], together with appropriate wavelength-control functions. The ability to reassign ONUs among wavelengths can support dynamic load balancing among wavelengths; this capability is not evaluated in this study. A Media Access Control (MAC) layer then coordinates framing, synchronization, and upstream transmission according to the wavelength and transmission opportunity assigned by the OLT.

3.4. Software-Defined Enhancements

Integrating SDN and vOLTHA transforms PON management by abstracting physical components and centralizing control. The SDN controller, such as ONOS, orchestrates DBA and flow management through OpenFlow, while vOLTHA standardizes interactions with diverse OLTs and ONUs. By treating the PON as a virtual switch, bandwidth allocation, traffic flow optimization, and rapid adaptation become policy-driven, modular, and scalable. This software-defined framework is well-suited to meet the evolving requirements of TWDM-PON systems.

3.5. GenAI Integration

The proposed GenAI-driven dynamic resource allocation framework integrates predictive intelligence into the TWDM-PON control process. GenAI analyzes historical and current network information to estimate future traffic conditions, while the SDN controller, vOLTHA, and OLT mechanisms translate these predictions into feasible resource-allocation actions.

In TWDM-PON, dynamic resource allocation involves several coordinated decisions, including bandwidth allocation, ONU service scheduling, transmission-timeslot allocation, wavelength assignment, and QoS-aware resource distribution. GenAI predictions can support these functions by providing information about future demand before it becomes fully visible through conventional network reporting. The framework therefore enables resource allocation to move from purely reactive operation toward proactive, prediction-driven optimization. The architecture separates the GenAI prediction layer from the resource-allocation mechanisms. GenAI provides future-demand estimates, while the underlying control functions enforce protocol, capacity, and QoS constraints. This modular design also allows the prediction model to be replaced by more capable models as they become available.

In the implementation evaluated in this study, the broader framework is realized through GenAI-assisted predictive DBA, where the predicted demand influences ONU service ordering and upstream bandwidth allocation, while ONU-to-wavelength assignments remain static. The detailed DBA mechanism is presented in Section 4.4.

4. Proposed Operations

Figure 3 illustrates how GenAI integrates into the TWDM-PON ecosystem, interacting closely with ONUs, the OLT, the SDN Controller, and a virtualized abstrac-

tion layer vOLTHA. This integrated framework is designed to enable predictive and adaptive resource allocation.

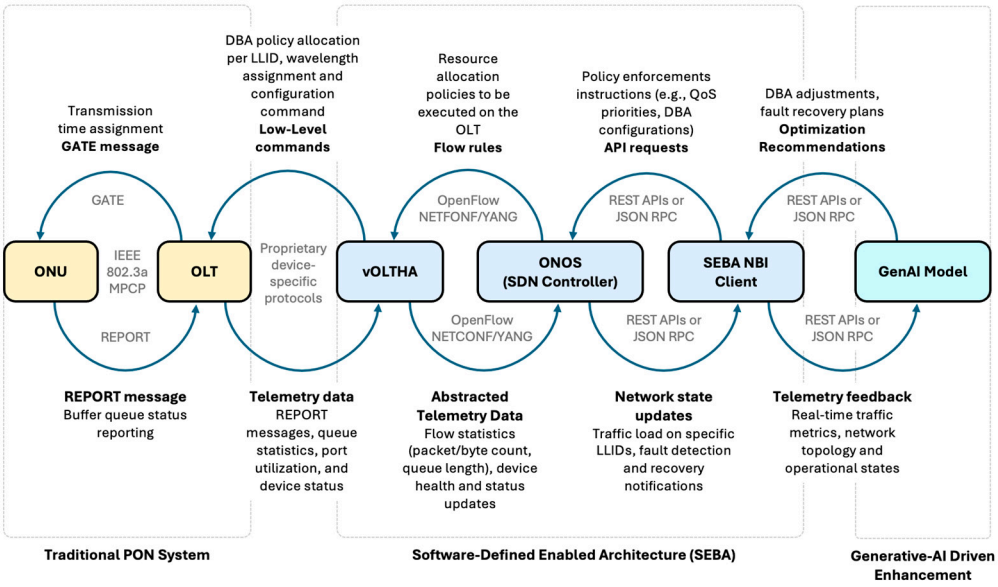

Figure 3. Generative-AI-driven TWDM-PON operations.

4.1. ONU-OLT Communication and Resource Allocation

In each polling cycle, every ONU reports the occupancy of its Expedited Forwarding (EF), Assured Forwarding (AF) and Best Effort (BE) queues, tied to its Logical Link Identifiers (LLIDs), in a REPORT message of the IEEE 802.3ah Multi-Point Control Protocol (MPCP), and the OLT replies with a GATE message that specifies the transmission start time, grant length and, in TWDM-PON, the upstream wavelength. In conventional DBA, the grant is computed only from the latest REPORT, so packets arriving between the REPORT and the granted transmission must wait for the next cycle. In the proposed framework, the GenAI forecast is inserted between REPORT reception and GATE generation: the DBA combines the reported queue state with the forecasted AF arrivals to determine the ONU service order and the AF grant, while the EF grant and the reported AF demand remain protected.

With GenAI in the loop, traditional reactive allocation is replaced by predictive adjustments. The Edge GenAI model, informed by real-time telemetry data, anticipates surges in bandwidth demand. For example, if the GenAI model predicts heavy streaming traffic in certain ONUs, the OLT proactively modifies its GATE assignments, so that the anticipated traffic can be served with reduced waiting time.

4.2. vOLTHA Abstraction and SDN Controller Integration

vOLTHA [53] presents the entire PON, consisting of the OLT and its ONUs, to the SDN controller as a single programmable OpenFlow switch. The OLT uplink, a network-to-network interface (NNI), appears as the switch uplink port, and each ONU user port, a user network interface (UNI), appears as a logical switch port. Vendor-specific adapters translate the abstract flows and meters configured by the SDN controller into device commands: an OLT adapter configures the OLT, and ONU adapters configure the ONUs through the ONU Management and Control Interface (OMCI). vOLTHA also handles ONU discovery and activation, provisions the logical links and queues associated with each service, and applies bandwidth profiles (committed, assured and excess information rates) that bound the grants issued by the DBA. In the opposite direction, it collects device events and performance

counters, such as port utilization and queue statistics, and publishes them to the SDN controller and the GenAI telemetry pipeline. As a result, the SDN controller and the GenAI component interact with a uniform logical switch and not with vendor-specific OLT and ONU interfaces.

Telemetry data collected at the OLT level, such as port utilization, queue lengths, and device health, are abstracted by vOLTHA and forwarded to the SDN Controller. The Edge GenAI model processes this data locally to forecast traffic for the DBA. Concurrently, high-level optimization strategies are coordinated with the Global GenAI model to ensure alignment with overarching network performance objectives.

Two GenAI roles are distinguished. The Edge GenAI model runs at the central office close to the OLT and provides per-cycle traffic forecasts to the DBA. The Global GenAI model runs in the operator cloud, aggregates telemetry across OLTs and provides longer-term policy recommendations to the SDN controller. In this study, only the Edge GenAI role is implemented and evaluated.

4.3. GenAI-Enhanced Telemetry Feedback Loop

A continuous feedback loop underpins the GenAI-driven framework, where GenAI models hosted at the operator's central office or cloud data centers process real-time telemetry to extract actionable insights. These models access context-specific knowledge bases, such as historical traffic records and device capability profiles, enabling swift and informed decision-making. By aggregating telemetry from multiple network segments, the GenAI models identify global trends, such as wavelength saturation during peak hours across various OLTs; this capability belongs to the general framework and is not evaluated in this study. Based on these analyses, the GenAI models issue policy recommendations, which are then translated into configuration changes by vOLTHA and the SDN Controller. This continuous loop ensures that local optimizations, such as adjusting bandwidth for individual ONUs, align with broader system objectives, thereby maintaining a harmonious balance across the entire TWDM-PON.

The framework operates on two timescales. The fast loop runs at the OLT in every DBA cycle (on the order of milliseconds): REPORT and grant history are supplied to the forecasting model, and the forecast is used by the DBA in the same decision; the inference time this requires is reported in Section 5. The slow loop runs at the SDN controller over seconds to minutes: aggregated telemetry is used to update QoS policies and bandwidth profiles, to retrain or replace the forecasting model and, in future extensions, to rebalance ONUs among wavelengths. The evaluation in Section 5 covers the fast loop only.

4.4. GenAI-Assisted Predictive Dynamic Bandwidth Allocation

DBA is one of the most important resource-allocation mechanisms in PONs because it determines how shared upstream capacity is distributed among multiple ONUs and traffic classes. Its effectiveness directly affects bandwidth utilization, packet delay, fairness, and QoS performance. In the proposed GenAI-driven resource-allocation framework, GenAI contributes predictive intelligence to this process by estimating traffic demand that is not yet visible in the latest REPORT information. These predictions are then translated by the DBA into proactive scheduling and bandwidth-allocation decisions. In this sense, the DBA provides the operational mechanism through which GenAI-driven prediction influences actual network resource allocation.

The DBA differentiates upstream traffic into Expedited Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE) QoS classes. Previous predictive DBA studies have similarly incorporated traffic-class information together with historical REPORT and GATE behavior to improve bandwidth estimation and scheduling decisions [47,48]. Among these

classes, AF is selected as the prediction target in this study. EF traffic is comparatively regular and is directly protected by the priority policy, whereas BE traffic has lower service priority and is served using residual capacity. AF lies between these two classes: it is sufficiently bursty to benefit from prediction while still having QoS requirements that make delay reduction important.

Conventional REPORT-based DBA determines upstream grants from the queue state observed when an ONU generates its REPORT. Packets arriving after the REPORT snapshot but before the ONU's scheduled transmission opportunity are therefore unavailable to a purely reactive scheduler. The proposed framework addresses this limitation using a modular GenAI forecasting component that estimates near-future AF demand before the corresponding packets are reported. TimesFM 3 [52] is used as the forecasting model in the current implementation, while the framework itself remains model-agnostic so that the forecasting component can be replaced by more capable models in future implementations. The resulting GenAI forecast is then used by the DBA to refine ONU service ordering and bandwidth allocation while preserving QoS priorities and resource constraints.

Let i denote the ONU index, N the number of ONUs, and t the DBA decision-cycle index. The traffic-class index q takes the values EF, AF, and BE. Class and stage superscripts are labels rather than exponents; a hat denotes a prediction. Each ONU contributes a 50-cycle causal history whose seven-variable observation is

$$\begin{aligned} \mathrm{x}_{i,t} = \big[& R^{\mathrm{EF}}_{i,t},\ R^{\mathrm{AF}}_{i,t},\ R^{\mathrm{BE}}_{i,t}, \\ & G^{\mathrm{EF}}_{i,t-1},\ G^{\mathrm{AF}}_{i,t-1},\ G^{\mathrm{BE}}_{i,t-1},\ \Delta t_{i,t} \big] \end{aligned} \tag{1}$$

In (1), $R^{q}_{i,t}$ is the queue demand reported for class q by ONU i at decision cycle t, and $G^{q}_{i,t-1}$ is the grant actually applied in the preceding cycle; both are measured in bits. The interval $\Delta t_{i,t}$ is the elapsed time between successive REPORT receipts from that ONU, in seconds. $\mathbf{x}_{i,t}$ denotes the feature vector. Applied grants represent allocated capacity and need not equal transmitted packet bits.

The GenAI predictor receives the causal AF REPORT-balance proxy as its target history and the seven variables in (1) as past-only covariates. The proxy is an observable estimate derived from REPORT changes and previous grants, rather than a directly observed arrival history. The predictor produces a nonnegative median (P50) proxy-rate forecast, denoted by $\hat{r}^{\mathrm{AF}}_{i,t}$, in bit/s, which is interpreted as an estimate of near-future AF arrival rate. The forecasting model remains fixed during closed-loop DBA operation, and no future packet arrivals or unreported live queue states are supplied to the predictor.

At each decision, a feasible REPORT-based allocation is computed using IPACT. Let $B^{q}_{i,t}$ denote its grant to traffic class q, in bits. This baseline defines the resource envelope and provides the conventional reactive comparison. The GenAI forecast influences service ordering and the allocation of AF reserve within that envelope.

Let $h^{\mathrm{ref}}_{i,t}$ be the interval, in seconds, from the exact REPORT-generation snapshot to the preliminary AF service start in the reference schedule. The predicted AF volume over this horizon is

$$\hat{A}^{\mathrm{AF,ref}}_{i,t} = \hat{r}^{\mathrm{AF}}_{i,t}\, h^{\mathrm{ref}}_{i,t} \tag{2}$$

In (2), $\hat{A}^{\mathrm{AF,ref}}_{i,t}$ is the estimated volume of AF arrivals, in bits, over the reference horizon; the label ref identifies the preliminary schedule. This interval includes traffic that is unreported at the decision time as well as traffic expected before service begins. The scheduler uses this estimate to calculate

$$S_{i,t} = w_{\mathrm{EF}}\, R^{\mathrm{EF}}_{i,t} + R^{\mathrm{AF}}_{i,t} + \hat{A}^{\mathrm{AF,ref}}_{i,t} \tag{3}$$

Here, $S_{i,t}$ is the scheduling-priority score for ONU i at cycle t, measured in bits. It ranks ONUs and is not itself a grant. The dimensionless EF weight w_{EF} is fixed at 2. The

other two terms have unit weight. Within each static wavelength group, ONUs are sorted by decreasing score after applying the BE urgency rule; rotating priorities resolve ties. Through (3), the forecast therefore adds anticipated AF demand to the currently reported queue state.

Reordering changes the planned service time. Let $h_{i,t}^{\text{final}}$ denote the updated REPORT-snapshot-to-AF-service interval, in seconds. The forecasted AF volume for the final schedule is

$$\hat{A}_{i,t}^{\text{AF,final}} = \hat{r}_{i,t}^{\text{AF}}\, h_{i,t}^{\text{final}} \tag{4}$$

The superscript final identifies the reordered schedule. The raw forecast rate is reused in (4) because the causal inputs have not changed; the new horizon changes the predicted volume. Thus, the two scheduling stages do not require two independent TimesFM 3 forecasts for identical histories.

In summary, the forecasting model outputs the AF arrival rate, in bit/s. This rate is multiplied by the interval from the REPORT to the planned AF service start, in seconds, to obtain the expected AF volume, in bits. This volume is used for ONU ordering in (3) and sets the bounded AF credit defined below. This forecast is distinct from the current-cycle grant prediction evaluated in the following chapter.

The additional AF credit is

$$C_{i,t}^{\text{AF}} = \min\left(\hat{A}_{i,t}^{\text{AF,final}},\ C_{i,t}^{\text{AF,max}}\right) \gamma_{i,t} \tag{5}$$

In (5), $C_{i,t}^{\text{AF}}$ is the additional AF allocation, in bits, and $C_{i,t}^{\text{AF,max}}$ is its permitted maximum. The cap is the smaller of 25% of the baseline BE grant and the capacity remaining above the protected BE floor. That floor preserves at least 75% of baseline BE and, when the baseline can support it, at least the smaller of the reported BE demand and the configured maximum packet size. If baseline BE cannot support that packet allowance, its entire grant is protected.

The dimensionless factor $\gamma_{i,t}$ is a deterministic BE-protection factor, not a statistical measure of forecast confidence. Define the reported BE occupancy ratio as

$$\rho_{i,t}^{\text{BE}} = \text{clip}\left(\frac{R_{i,t}^{\text{BE}}}{Q_{max}^{\text{BE}}},\ 0,\ 1\right) \tag{6}$$

Here, Q_{max}^{BE} is the configured BE queue capacity, in bits, and clip in (6) limits its argument to the interval $[0, 1]$. Therefore, $\rho_{i,t}^{\text{BE}}$ is a dimensionless occupancy ratio. The protection factor is

$$\gamma_{i,t} = \begin{cases} 1 - \rho_{i,t}^{\text{BE}}, & a_{i,t}^{\text{BE}} < a_{max} \\ 0, & a_{i,t}^{\text{BE}} \geq a_{max} \end{cases} \tag{7}$$

The variable $a_{i,t}^{\text{BE}}$ counts consecutive rounds without a packet-capable BE grant opportunity while BE demand is reported; a_{max} is the urgency threshold of two rounds. It tracks grant opportunities, not observed packet waiting time. Equation (7) suppresses predictive AF credit as BE occupancy increases and sets it to zero at the urgency threshold. Such ONUs receive ordering priority, although insufficient baseline capacity can still prevent a packet-capable BE grant.

The final applied class grants are

$$G_{i,t}^{\text{EF}} = B_{i,t}^{\text{EF}} \tag{8}$$

$$G_{i,t}^{\text{AF}} = B_{i,t}^{\text{AF}} + C_{i,t}^{\text{AF}} \tag{9}$$

$$G_{i,t}^{\text{BE}} = B_{i,t}^{\text{BE}} - C_{i,t}^{\text{AF}} \tag{10}$$

In (8)–(10), $G_{i,t}^{q}$ denotes the final grant to class q, in bits. The EF grant and the baseline reported AF allocation are protected. The AF credit is transferred from BE, so per-ONU grant conservation gives

$$\sum_{q} G_{i,t}^{q} = \sum_{q} B_{i,t}^{q} \tag{11}$$

The summation in (11) covers EF, AF, and BE. The total ONU grant remains within its configured cap, and the planned schedule retains the same-state baseline scheduling envelope. During flexible burst service, BE may reclaim unused contiguous AF capacity if the predicted demand does not materialize. The final GATE messages specify the resulting grants and transmission opportunities, and subsequent decisions use the resulting REPORT and applied-grant history.

The GenAI contribution is the forecast of unreported AF demand used for proactive ONU ordering and bounded AF reservation. ONU-to-wavelength assignments remain static. The forecast supplies the future-demand estimate, while the allocation rules enforce class protection and resource constraints; network-level benefits are assessed from the measured simulation results.

4.5. Policy-Driven Optimization and Scalability

The proposed GenAI-driven dynamic resource allocation framework combines network telemetry, configuration information, and predictive modeling to support proactive resource management in TWDM-PON. Network policies define QoS requirements, resource constraints, and operational objectives, while the SDN controller and vOLTHA provide the programmable control functions required to enforce the resulting decisions. GenAI complements these mechanisms by predicting future network demand, allowing resource-allocation decisions to consider not only the current network state but also anticipated traffic conditions.

The framework uses two complementary sources of network information, as illustrated in Figure 4. Time-series data represent the dynamic network state, including traffic demand, queue occupancy, bandwidth utilization, REPORT/GRANT history, and other operational measurements. Configuration and topology data, represented through NETCONF/YANG or equivalent management models, describe relatively static information such as ONU connectivity, wavelength assignments, device capabilities, and network-resource constraints. Together, these sources provide both temporal and structural context for resource optimization.

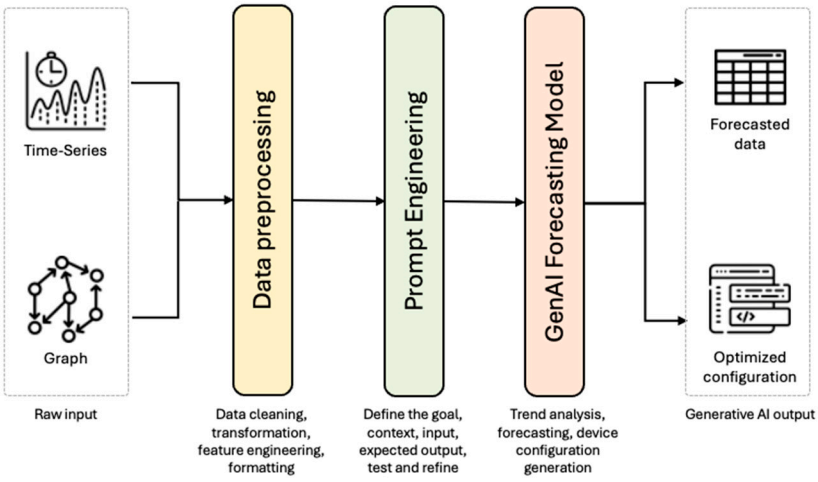

Figure 4. Generalized view of GenAI pipeline processing.

4.5.1. Data Preprocessing

Raw network information is first transformed into a structured representation suitable for the GenAI forecasting component. Time-series measurements are synchronized, normalized, and organized into causal historical sequences, while configuration information is processed to identify relevant resource and topology constraints. For predictive DBA, for example, the historical observations include traffic-class REPORT information, previous grants, and REPORT intervals. This stage ensures that only information available at the current decision time is supplied to the forecasting model.

4.5.2. Feature and Covariate Engineering

The processed measurements are then converted into target histories and relevant covariates. Traffic demand serves as the principal forecasting target, while related observations such as queue states, previous resource allocations, traffic-class information, and temporal intervals provide contextual information. This representation allows the forecasting component to capture relationships between recent network behavior and near-future resource demand.

4.5.3. GenAI Foundation Model

The processed time-series information is supplied to a GenAI time-series foundation model to estimate future network demand. In the implementation evaluated in this study, TimesFM 3 [52] is used to forecast near-future AF demand; however, the framework is model-agnostic and can incorporate more capable forecasting models as they become available. The GenAI model therefore functions as a predictive intelligence layer rather than directly controlling the PON.

The forecast can support multiple resource-allocation functions, including bandwidth allocation, ONU service ordering, transmission-timeslot allocation, wavelength assignment, and QoS-aware resource distribution. In the evaluated implementation, the prediction is specifically applied to ONU scheduling and DBA bandwidth allocation, while ONU-to-wavelength assignments remain static.

4.5.4. Resource Optimization and SDN Integration

The predicted network demand is provided to the resource-allocation logic, which converts the forecast into feasible control decisions while enforcing capacity, QoS, and protocol constraints. Depending on the optimization objective, these decisions may include adjusted bandwidth grants, ONU scheduling priorities, transmission windows, or wavelength assignments. For example, predicted congestion on one wavelength could support future wavelength rebalancing among tunable ONUs, whereas predicted short-term traffic demand can be incorporated directly into DBA decisions.

The resulting resource decisions are then enforced through the SDN/vOLTHA control framework and the corresponding OLT functions. This separation between prediction, optimization, and execution preserves the reliability of conventional network-control mechanisms while enabling GenAI to drive more proactive resource allocation. It also improves scalability because the forecasting model, optimization policy, and underlying network-control functions can evolve independently.

5. System Performance

The proposed GenAI-driven resource-allocation framework was evaluated using a Python-based TWDM-PON simulation. The system consists of 64 ONUs sharing two upstream wavelengths at 1 Gbps, with OLT–ONU distances of 10–20 km, a 10-Mb ONU buffer, and a maximum transmission cycle time of 1.5 ms. For the evaluation, two of the four

upstream wavelengths in Figure 2 are simulated. The guard time, ONU tuning time, DBA computation time, and control-message length are also specified in Table 2. Three DiffServ traffic classes are considered: EF, AF, and BE. Following the traffic-generation model in [47], EF traffic is generated as constant-frame-rate traffic using a T1 circuit-emulated source with IP/UDP encapsulation, one frame every 125 µs, and a fixed packet size of 70 bytes. AF and BE traffic are generated using self-similar, long-range-dependent processes with a Hurst parameter of 0.7 and packet sizes uniformly distributed between 64 and 1518 bytes. Different EF:AF:BE traffic compositions and offered-load conditions are used to evaluate the behavior of the DBA mechanism under heterogeneous traffic conditions.

Table 2. Simulation parameters.

Parameters	Value
Number of ONUs	64
Number of upstream wavelengths	2
Up/down link rate	1 Gbps
OLT–ONU distance	10–20 km
ONU buffer size	10 Mb
Per-ONU grant limit	46,875 bits
Maximum transmission cycle time	1.5 ms
Guard time	5 µs
ONU tuning time	100 ns
DBA computation time (excluding forecast inference)	10 µs
Control message length	64 bytes

The evaluation is conducted at two complementary levels. First, the prediction models are compared on prediction of the current-cycle AF grant, GRANT(t), using MAE, RMSE, and R^2, and on forecasting of the AF arrivals used by the DBA. AF is selected as the prediction target because it combines bursty traffic characteristics with QoS requirements that make proactive estimation beneficial, while EF follows a comparatively regular traffic pattern and BE is served after the higher-priority classes. Second, the impact of the resulting prediction-assisted DBA is evaluated at the network level using packet delay, throughput, and packet-loss performance. This separation allows prediction accuracy and network performance to be reported separately.

To assess the contribution of GenAI-based prediction to the proposed resource-allocation framework, TimesFM 3 [52] was compared with LSTM [47], Informer [54], PatchTST [55], and Sundial with a supervised adapter [56]. LSTM was included as an established recurrent baseline for network-traffic prediction, Informer and PatchTST as representative Transformer-based forecasting models, and Sundial as a recent generative time-series foundation model. This comparison therefore evaluates TimesFM 3 against both conventional task-specific predictors and contemporary foundation-model approaches.

Table 3 reports a model-comparison test that follows the grant-prediction approach in [47]. Each model predicts the current-cycle AF grant, GRANT(t), in bits, from a 50-cycle history of REPORT and grant values. The target is the grant of the current cycle, and the latest REPORT is among the inputs, so this test does not involve a forward forecast horizon. In the dataset used, the AF grant is given by

$$\mathrm{GRANT}_i(t) = \min\left(R_{i,t}^{\mathrm{AF}},\ \max\left(0,\ L_{\max} - R_{i,t}^{\mathrm{EF}}\right)\right) \tag{12}$$

where $R_{i,t}^{\mathrm{AF}}$ and $R_{i,t}^{\mathrm{EF}}$ are the reported AF and EF demands defined in (1), and $L_{\max} = 46{,}875$ bits is the per-ONU grant limit. The rule in (12) is the same priority allocation that gives the baseline AF grant $B_{i,t}^{\mathrm{AF}}$ in Section 4.4. Although this rule is known, the

test follows [47] and shows how closely each model family reproduces this allocation behavior from the REPORT and grant history, which is the setting used in earlier prediction-based DBA studies.

Table 3. Prediction Results of the Current-Cycle AF Grant, GRANT(t), for Different Models.

Mix	Method	Samples	MAE (Bits)	RMSE (Bits)	R^2
1:3:6	LSTM	10240	905.772	2257.172	0.928031
1:3:6	Informer	10240	929.472	1776.108	0.955439
1:3:6	PatchTST	10240	4455.052	5148.552	0.625556
1:3:6	Sundial + adapter	10240	2352.673	4120.112	0.760208
1:3:6	TimesFM 3	10240	381.162	869.958	0.989309
2:4:4	LSTM	10240	1404.396	2810.413	0.926127
2:4:4	Informer	10240	758.617	1210.750	0.986289
2:4:4	PatchTST	10240	6539.733	7410.175	0.486425
2:4:4	Sundial + adapter	10240	2243.120	3812.890	0.864026
2:4:4	TimesFM 3	10240	652.925	1567.080	0.977032

The prediction data were generated in separate 10 s simulation runs for each traffic mix and load, using independent random seeds for training (128), validation (129) and testing (130). Each mix provided 20,480 training, 10,240 validation and 10,240 test samples, drawn evenly from all loads and ONUs, with targets taken after a 3 s warm-up. Every model received the same 50-cycle window of the seven variables in (1). LSTM, Informer and PatchTST were trained from scratch, with the best model selected on the validation set by early stopping. Sundial used its frozen pretrained backbone with a trained adapter. TimesFM 3 used its frozen pretrained weights without fine-tuning, with the past AF grants as the target history and the seven variables as past covariates. Each model used its own input scaling, and model initialization used seed 42. All models were run on a MacBook Air with an Apple M4 chip (10-core CPU, 10-core integrated GPU, 24 GB memory), using the GPU. The settings of each model are listed in Table 4.

Table 4. Settings of the models in the current-cycle grant prediction test.

Model	Training Mode	Parameters	Main Settings
LSTM [47]	Trained from scratch	13,473	Two 32-unit layers, dropout 0.1; RMSprop, learning rate 0.001, batch size 256, up to 50 epochs, early-stopping patience 7
Informer [54]	Trained from scratch	133,441	Model dimension 64, 4 heads, feed-forward dimension 128, 2 encoder layers, 1 decoder layer, dropout 0.1; Adam, learning rate 0.001, batch size 256, up to 50 epochs, patience 7
PatchTST [55]	Trained from scratch	68,939	Patch length 10, stride 5, model dimension 64, 4 heads, 2 layers, feed-forward dimension 128
Sundial + adapter [56]	Frozen pretrained backbone; trained adapter	128.3 million (frozen); 345,537 (adapter)	sundial-base-128m backbone; adapter 5397→64→1, dropout 0.1; Adam, learning rate 0.001, batch size 256, up to 50 epochs, patience 7
TimesFM 3 [52]	Pretrained, no fine-tuning	330.7 million	Official pretrained checkpoint (20 layers, model dimension 1280, 16 heads); batch size 16; median (P50) output

Let y_k be the actual AF grant of test sample k, $\hat{y}_k$ its prediction, $\overline{y}$ the mean of the actual values, and n the number of test samples. The metrics are

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k| \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2} \tag{14}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{n}(y_k - \overline{y})^2} \tag{15}$$

where all sums in (13)–(15) run over $k = 1, \ldots, n$. MAE and RMSE are expressed in bits; R^2 is dimensionless, equals 1 for a perfect prediction, and is 0 for a predictor that always outputs the mean.

Grant prediction was tested on two traffic compositions, with EF:AF:BE ratios of 1:3:6 and 2:4:4; the network evaluation also includes the 1:6:3 mix. All methods were evaluated using the same 10,240 held-out samples per mix and predicted the current-cycle AF allocation, GRANT(t).

As shown in Table 3, TimesFM 3 achieved the lowest MAE in both traffic mixes. For the 1:3:6 mix, its MAE decreased from 905.772 bits for LSTM [47] to 381.162 bits, corresponding to a 57.92% reduction. TimesFM 3 also obtained the lowest RMSE of 869.958 bits and the highest R^2 of 0.9893, indicating that the GenAI-based foundation model can estimate the required AF allocation with substantially lower prediction error under this traffic composition.

For the 2:4:4 mix, TimesFM 3 achieved an MAE of 652.925 bits, representing a 53.51% reduction relative to LSTM and a 13.93% reduction relative to Informer [54], which obtained the next-lowest MAE of 758.617 bits. Informer achieved the lowest RMSE of 1210.750 bits and the highest R^2 of 0.9863, while TimesFM 3 achieved an RMSE of 1567.080 bits and an R^2 of 0.9770. This shows that the relative advantage depends on the error metric, although TimesFM 3 maintained the lowest average absolute allocation error across both traffic compositions.

These results highlight the benefit of incorporating a pretrained time-series foundation model into the proposed GenAI-driven TWDM-PON framework. Compared with the conventional LSTM-based approach in [47], which learns temporal characteristics primarily from network-specific training data, TimesFM 3 leverages pretrained temporal representations together with PON-specific causal observations. This allows the model to predict the AF grant with the lowest MAE in both traffic compositions, although the ranking under RMSE depends on the traffic mix.

The measured contribution of this test is therefore the accuracy of current-cycle AF grant prediction.

For the system-level evaluation, the LSTM and TimesFM 3 forecasters predict the AF arrivals described in Section 4.4. For each traffic mix and load, the LSTM was trained on 1024 samples and validated on 512 samples from 10 s runs with seeds 128 and 129, and both forecasters were tested on 512 held-out samples from seed 130. The LSTM uses the same 50-cycle window of the seven variables in (1) together with the planned AF service interval; it has two 32-unit layers with dropout 0.1 and was trained with RMSprop (learning rate 0.001, batch size 256, seed 42). TimesFM 3 was again used with its pretrained weights and no fine-tuning, with the REPORT-balance proxy of Section 4.4 as its target history and the seven variables as past covariates.

The forecast used by the DBA was evaluated separately on held-out data, by comparing the predicted AF volume with the AF traffic that actually arrived between the REPORT

and the planned AF service start. TimesFM 3 gave a lower mean absolute error than LSTM in all 30 traffic-mix and load settings, by 20.87% on average. The errors of both models remain large relative to the traffic variation, because arrivals over such a short interval are highly variable.

One forecasting call per DBA decision required a median of 729 ms for TimesFM 3 and 2.6 ms for LSTM, compared with the 1.5 ms cycle budget, on the hardware described above. Inference time was not charged to simulated time; the network-level results therefore assume that the forecast is available at decision time.

Based on these comparisons, TimesFM 3 is used as the GenAI forecasting model in the system-level evaluation. Its forecast, described in Section 4.4, is incorporated into the proposed DBA mechanism, and the resulting network performance is assessed in terms of packet delay, throughput, and packet loss in the following subsections.

5.1. Mean Packet Delay

The results in Figures 5–8 evaluate the mean packet-delay behavior of IPACT, LSTM-assisted scheduling, and GenAI-assisted scheduling across the EF:AF:BE traffic mixes of 1:3:6, 2:4:4, and 1:6:3. All configurations use two upstream wavelengths. The curves remain closely grouped at low and moderate loads, while differences between service classes and scheduling methods become more visible as the offered load approaches system capacity.

EF traffic in Figure 5 maintains low delay over most of the evaluated load range. At 10% offered load, mean EF delay is approximately 0.29–0.32 ms, and all methods remain at or below approximately 1.10 ms through 70% load. Delay then increases to approximately 1.82–2.09 ms at 80% load and approaches 2.31–2.50 ms at full load. GenAI records lower EF delay than IPACT in 49 of the 60 paired cases, with an average reduction of 3.15%. However, this advantage is concentrated at low and moderate loads. At 80% load, GenAI produces slightly higher EF delay than IPACT for all three traffic mixes, while LSTM provides the lowest EF delay in several high-load cases. Improved AF prediction therefore does not automatically minimize EF delay because the predictor changes ONU transmission order while EF grant sizes remain controlled by the priority allocator. The EF results demonstrate a scheduling trade-off rather than uniform GenAI superiority.

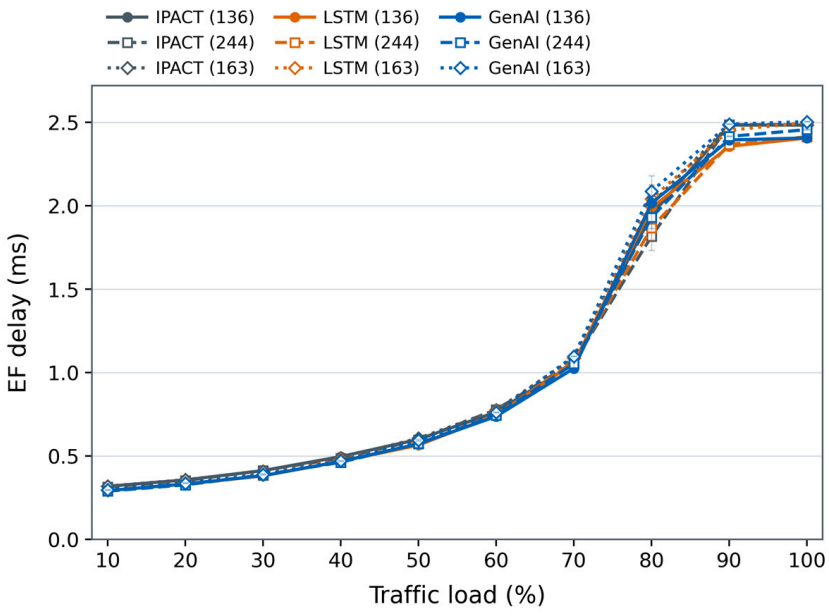

Figure 5. Comparison of EF Mean Packet Delay in Different Traffic Scenarios.

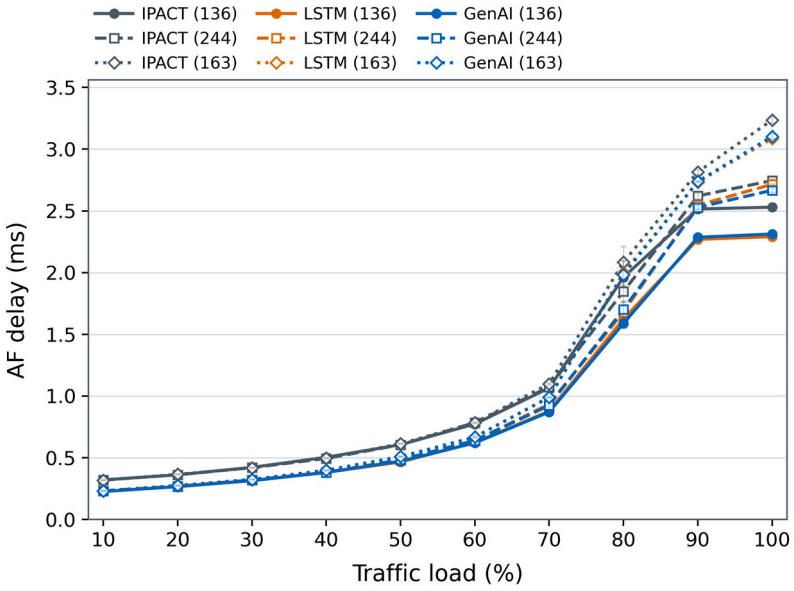

Figure 6. Comparison of AF Mean Packet Delay in Different Traffic Scenarios.

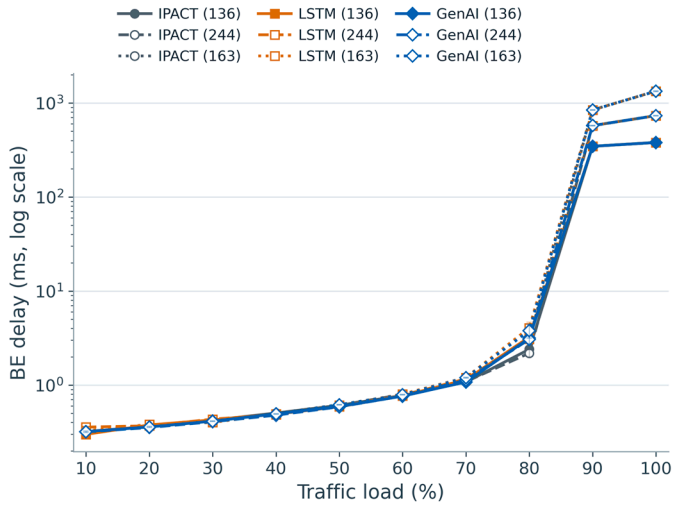

Figure 7. Comparison of BE Mean Packet Delay in Different Traffic Scenarios.

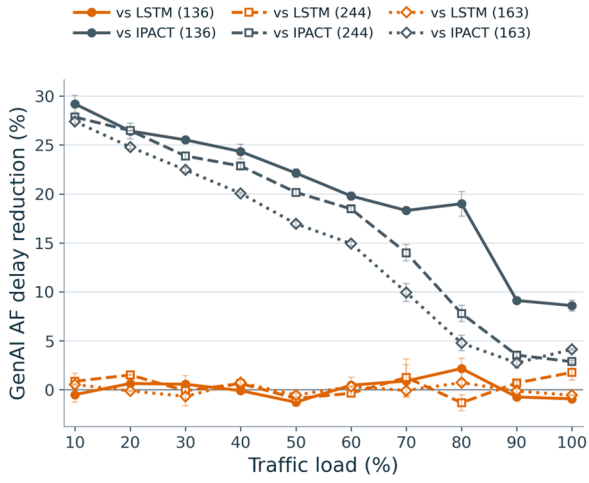

Figure 8. AF Delay Reduction of GenAI Relative to IPACT and LSTM in Different Traffic Scenarios.

AF traffic in Figure 6 shows the clearest benefit from predictive scheduling because AF is the class targeted by the proposed scheduler. At 10% load, LSTM and GenAI reduce AF delay from approximately 0.32 ms under IPACT to approximately 0.23 ms. Through 70% load, their AF delay remains below approximately 1.00 ms, compared with up to approximately 1.10 ms for IPACT. At full load, GenAI records approximately 2.31 ms for mix 1:3:6, 2.67 ms for mix 2:4:4, and 3.10 ms for mix 1:6:3. The corresponding IPACT delays are approximately 2.53, 2.75, and 3.24 ms.

Across all 60 paired cases, GenAI produces lower AF delay than IPACT in every case, with an average reduction of 17.28%. This advantage persists in the saturated region, where GenAI keeps the AF delay below IPACT by 4.8–19.1% at 80% load and by 2.7–9.1% at 90–100% load. The AF graph therefore shows a strong and consistent advantage of the proposed scheduling over IPACT.

BE traffic in Figure 7, which receives service after EF and AF, is the most sensitive to heavy load. Mean BE delay remains below approximately 1.20 ms through 70% offered load. At 80% load, it increases to approximately 3.07–3.79 ms under GenAI, compared with approximately 2.19–3.07 ms under IPACT. Beyond this point, BE delay rises sharply into the hundreds of milliseconds. Under GenAI, it reaches approximately 346–843 ms at 90% load and approximately 380–1335 ms at full load, depending on the traffic mix.

This sharp transition indicates substantial queue accumulation for the lowest-priority class as the system approaches saturation. GenAI has lower BE delay than LSTM in 43 of 60 cases, with an average reduction of 1.63%, but it does not provide a consistent BE-delay advantage over IPACT. This behavior is consistent with a priority scheduler that protects EF and AF traffic first and assigns the remaining capacity to BE traffic.

Figure 8 presents the AF-delay reduction achieved by GenAI relative to IPACT and LSTM across the three traffic mixes. The consistently positive curves against IPACT show that GenAI reduces AF delay throughout the complete load range, with the largest benefit at low and moderate loads and a smaller but persistent advantage as the system approaches saturation. This pattern shows the benefit of the proposed demand-based ONU ordering before congestion becomes severe. In contrast, the curves relative to LSTM remain close to zero and alternate between small positive and negative values, indicating that both prediction-based schedulers provide broadly comparable AF-delay performance. The main GenAI benefit is therefore its clear and robust improvement over conventional IPACT while maintaining performance competitive with LSTM across different traffic compositions. Together with the throughput results, the graph shows that GenAI improves the delay of its targeted traffic class without reducing aggregate carried traffic, supporting its use as a predictive scheduling enhancement rather than establishing uniform superiority over every forecasting method.

Collectively, these results show that the principal measured advantage of GenAI-assisted scheduling is its reduction in AF delay relative to IPACT. EF delay remains controlled through moderate load but contains high-load exceptions, while BE and aggregate delay deteriorate sharply after the system enters the saturated region. The results support selective, traffic-class-specific benefits rather than a claim that GenAI always provides the lowest delay.

5.2. Throughput

Figure 9 presents the aggregate system throughput of IPACT, LSTM-assisted scheduling, and GenAI-assisted scheduling across the three EF:AF:BE traffic mixes. Throughput increases almost linearly with the offered load up to approximately 80% and then levels off as the two-wavelength upstream system approaches saturation. The close overlap among the curves shows that traffic composition and prediction method have little effect on the

total amount of successfully transmitted data. Although GenAI changes the ONU service order using predicted AF demand, it maintains throughput comparable to IPACT and LSTM throughout the complete load range. At high loads, the GenAI curves remain comparable to or slightly above IPACT, indicating that predictive scheduling does not introduce an aggregate-capacity penalty when the network is heavily loaded.

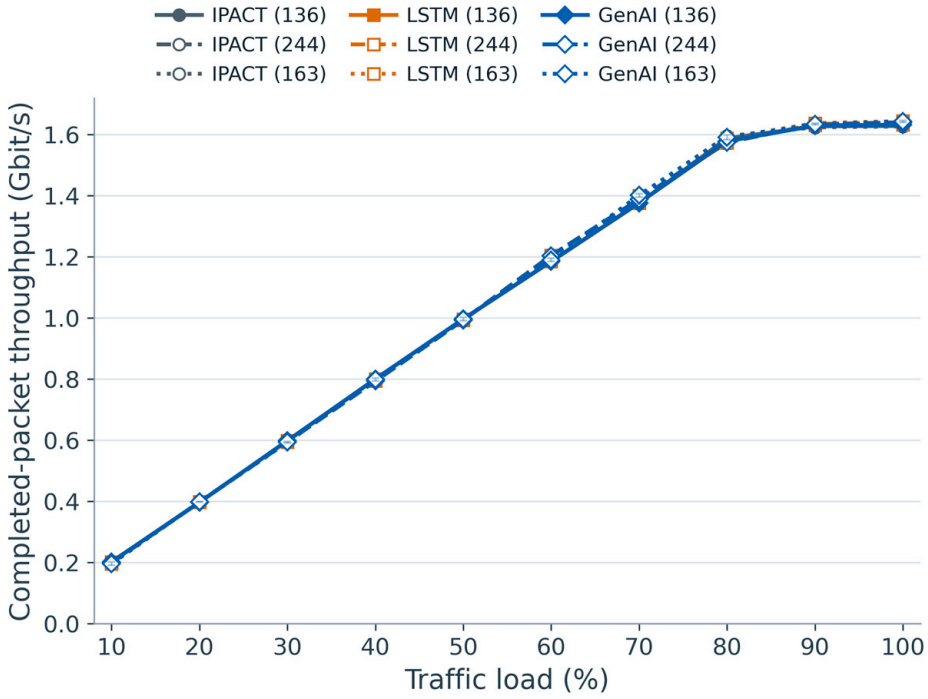

Figure 9. Comparison of Throughput in Different Traffic Scenarios.

This result becomes more meaningful when considered together with the delay graphs. GenAI preserves the available system throughput while providing its clearest improvement in AF delay. Its advantage is therefore a better distribution of transmission opportunities among traffic classes without reducing overall carried traffic. Since the throughput curves are very close, the graph should be interpreted as evidence of throughput preservation rather than substantial throughput superiority.

5.3. Packet Loss Ratio

Figure 10 presents the BE packet-drop ratio for IPACT, LSTM-assisted scheduling, and GenAI-assisted scheduling across the three EF:AF:BE traffic mixes. No BE packet loss is observed from 10% through 80% offered load for any evaluated method, while drops begin at 90% load and increase sharply at full load because the remaining wavelength capacity is insufficient to serve all accumulated BE traffic after the higher-priority EF and AF queues have been scheduled. GenAI preserves the same loss-free operating range as IPACT and LSTM, demonstrating that prediction does not cause premature BE loss at low or moderate loads. Under saturation, GenAI maintains a drop ratio comparable to LSTM and reduces BE loss relative to IPACT in the more congested 2:4:4 and 1:6:3 mixes, while all three methods remain closely grouped for mix 1:3:6. Together with the throughput result, this shows that GenAI can preserve aggregate carried traffic while distributing constrained transmission opportunities more effectively under demanding loads.

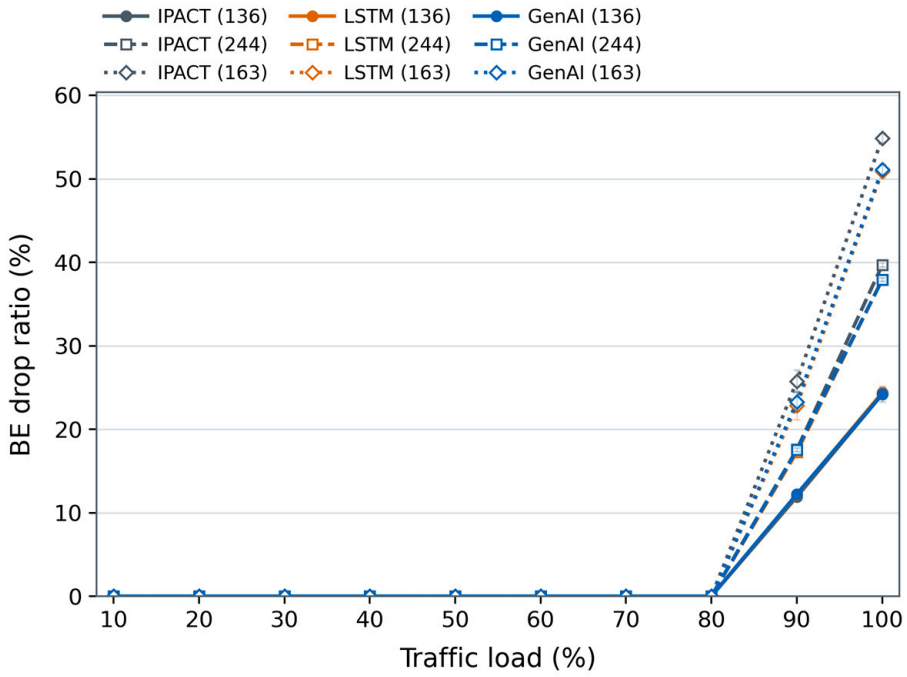

Figure 10. Comparison of Drop Probability in Different Traffic Scenarios.

6. Conclusions

This paper proposed a GenAI-assisted framework for bandwidth allocation in TWDM-PON, in which the DBA serves ONUs in order of their demand and a forecasting model estimates the AF traffic that has not yet been reported. The simulation results show three points. First, the proposed DBA reduced the mean AF packet delay by about 17.3% compared with IPACT, in all 60 evaluated cases, with the largest gains at low and moderate loads. Second, this was achieved without losing capacity: total throughput was unchanged and no BE packets were lost up to 80% offered load. Third, the GenAI forecaster, used without any training on PON data, gave the lowest mean absolute error: 57.92% and 53.51% lower than LSTM in predicting the current-cycle AF grant for the two traffic compositions, and 20.87% lower than LSTM in forecasting AF arrivals.

A key challenge for practical deployment is the inference time of the forecasting model. DBA operates at millisecond-scale cycles, so forecasts must be produced within the available scheduling interval. The modular structure of the framework allows the current forecasting model to be replaced by lighter or more capable models as time-series foundation models evolve. Our near-future work will therefore study lightweight, domain-specific forecasting models for PON traffic, using model compression, quantization, pruning and hardware acceleration, and will extend the framework to joint predictive wavelength and bandwidth allocation, in which predicted traffic and wavelength loads guide the reassignment of tunable ONUs.

Author Contributions: A.F.P. prepared the manuscript and conducted the simulation; I.-S.H. initiated the research and edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 114-2221-E-155-053.

Data Availability Statement: The simulation code, configuration files, random seeds, generated datasets and raw results supporting this study are openly available at https://github.com/

andrewpakpahan/genai-dba-twdm-pon (accessed on 12 November 2026). The pretrained TimesFM 3 and Sundial models are available from their original authors [52,56].

Acknowledgments: We wish to acknowledge the anonymous referees who gave precious suggestions to improve the work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AF	Assured Forwarding
AI	Artificial Intelligence
BE	Best Effort
BNG	Broadband Network Gateway
CPU	Central Processing Unit
DBA	Dynamic Bandwidth Allocation
DiffServ	Differentiated Services
EF	Expedited Forwarding
EPON	Ethernet Passive Optical Network
GAN	Generative Adversarial Network
GenAI	Generative Artificial Intelligence
GPON	Gigabit Passive Optical Network
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
IM/DD	Intensity-Modulated/Direct-Detection
IP	Internet Protocol
IPACT	Interleaved Polling with Adaptive Cycle Time
LLID	Logical Link Identifier
LSTM	Long Short-Term Memory
MAC	Media Access Control
MAE	Mean Absolute Error
ML	Machine Learning
MPCP	Multi-Point Control Protocol
NBI	Northbound Interface
NETCONF	Network Configuration Protocol
NG-PON	Next-Generation Passive Optical Network
NNI	Network-to-Network Interface
OFDM	Orthogonal Frequency-Division Multiplexing
OLT	Optical Line Terminal
OMCI	ONU Management and Control Interface
ONOS	Open Network Operating System
ONU	Optical Network Unit
PON	Passive Optical Network
QoS	Quality of Service
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RSOA	Reflective Semiconductor Optical Amplifier
SDAN	Software-Defined Access Network
SDN	Software-Defined Networking
SEBA	Software-Enabled Broadband Access
SVM	Support Vector Machine
TWDM-PON	Time and Wavelength Division Multiplexed Passive Optical Network
UDP	User Datagram Protocol
UNI	User Network Interface
VAE	Variational Autoencoder
vOLTHA	virtual Optical Line Termination Hardware Abstraction
WDM	Wavelength Division Multiplexing
WDM-PON	Wavelength Division Multiplexed Passive Optical Network
YANG	Yet Another Next Generation

References

1. Sengar, S.S.; Hasan, A.B.; Kumar, S.; Carroll, F. Generative artificial intelligence: A systematic review and applications. *Multimed. Tools Appl.* **2024**, *84*, 23661–23700. [CrossRef]
2. Du, H.; Niyato, D.; Kang, J.; Xiong, Z.; Zhang, P.; Cui, S.; Shen, X.; Mao, S.; Han, Z.; Jamalipour, A.; et al. The age of generative AI and AI-Generated everything. *IEEE Netw.* **2024**, *38*, 501–512. [CrossRef]
3. Hagos, D.H.; Battle, R.; Rawat, D.B. Recent advances in generative AI and large language models: Current status, challenges, and Perspectives. *IEEE Trans. Artif. Intell.* **2024**, *5*, 5873–5893. [CrossRef]
4. Bariah, L.; Zhao, Q.; Zou, H.; Tian, Y.; Bader, F.; Debbah, M. Large generative AI models for telecom: The next big thing? *IEEE Commun. Mag.* **2024**, *62*, 84–90. [CrossRef]
5. Vu, T.-H.; Jagatheesaperumal, S.K.; Nguyen, M.-D.; Huynh, N.V.; Kim, S.; Pham, Q.-V. Applications of generative AI (GAI) for mobile and wireless networking: A survey. *IEEE Internet Things J.* **2024**, *12*, 1266–1290. [CrossRef]
6. Liu, G.; Zhang, K.; Chen, X.; Lu, H.; Guo, J.; Yin, J.; Proietti, R.; Zhu, Z.; Ben Yoo, S.J. Hierarchical learning for cognitive end-to-end service provisioning in multi-domain autonomous optical networks. *J. Light. Technol.* **2019**, *37*, 218–225. [CrossRef]
7. Pakpahan, A.F.; Hwang, I.-S. Flexible access network multi-tenancy using NFV/SDN in TWDM-PON. *IEEE Access* **2023**, *11*, 42937–42948. [CrossRef]
8. Assi, C.; Maier, M.; Shami, A. Toward quality of service protection in ethernet passive optical networks: Challenges and solutions. *IEEE Netw.* **2007**, *21*, 12–19. [CrossRef]
9. Sun, T.; Sun, B.; Jiang, Z.; Hao, R.; Xie, J. Traffic flow online prediction based on a generative adversarial network with multi-source data. *Sustainability* **2021**, *13*, 12188. [CrossRef]
10. Das, S. From CORD to SDN Enabled Broadband Access (SEBA) [Invited Tutorial]. *J. Opt. Commun. Netw.* **2021**, *13*, A88–A99. [CrossRef]
11. ITU-T Recommendations G-984 Series. Available online: https://www.itu.int/rec/T-REC-G/en (accessed on 1 August 2026).
12. *IEEE Std 8023ah-2004*; IEEE Standard for Information Technology—Local and Metropolitan Area Networks—Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks. IEEE: Piscataway, NJ, USA, 2004; pp. 1–640. [CrossRef]
13. G.989.1: 40-Gigabit-Capable Passive Optical Networks (NG-PON2): General Requirements. Available online: https://www.itu.int/rec/T-REC-G.989.1/en (accessed on 17 August 2026).
14. *IEEE Std 802.3av-2009*; IEEE Standard for Information Technology—Local and Metropolitan Area Networks—Specific Requirements—Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 1: Physical Layer Specifications and Management Parameters for 10 Gb/s Passive Optical Networks. IEEE: Piscataway, NJ, USA, 2009. [CrossRef]
15. *IEEE Std 8023ca-2020*; IEEE Standard for Ethernet Amendment 9: Physical Layer Specifications and Management Parameters for 25 Gb/s and 50 Gb/s Passive Optical Networks. IEEE: Piscataway, NJ, USA, 2020; pp. 1–267. [CrossRef]
16. Chenika, A.; Derkaoui, M.; Mazouzi, A. Demonstration of 640 Gbps/160 Gbps WDM-PON system based on colorless ONUs. *J. Opt. Commun.* **2024**, *47*, 203–208. [CrossRef]
17. Alhalabi, M.; Taşpınar, N.; El-Nahal, F.I. Bidirectional intensity modulated/direct detection optical OFDM WDM-PON system. *Optoelectron. Lett.* **2023**, *19*, 83–87. [CrossRef]
18. Pakpahan, A.F.; Hwang, I.-S. Enabling flexible software-defined energy-efficient orchestration in TWDM-PON. *J. Internet Technol.* **2020**, *21*, 2.
19. Pakpahan, A.F.; Hwang, I.-S.; Nikoukar, A. OLT energy savings via software-defined dynamic resource provisioning in TWDM-PONs. *J. Opt. Commun. Netw.* **2017**, *9*, 1019–1029. [CrossRef]
20. Pakpahan, A.F.; Hwang, I.-S.; Yu, Y.-M.; Hsu, W.-H.; Liem, A.T.; Nikoukar, A. Novel elastic protection against DDF failures in an enhanced software-defined SIEPON. *Opt. Fiber Technol.* **2017**, *36*, 51–62. [CrossRef]
21. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
22. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
23. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
24. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**, arXiv:1312.6114. [CrossRef]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023. [CrossRef]
26. Bengesi, S.; El-Sayed, H.; Sarker, M.K.; Houkpati, Y.; Irungu, J.; Oladunni, T. Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access* **2024**, *12*, 69812–69837. [CrossRef]
27. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. *Language Models Are Unsupervised Multitask Learners*; OpenAI: San Francisco, CA, USA, 2019.

28. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
29. Chen, Y.; Li, R.; Zhao, Z.; Peng, C.; Wu, J.; Hossain, E.; Zhang, H. NetGPT: A native-AI network architecture beyond provisioning personalized generative services. *arXiv* **2024**, arXiv:2307.06148. [CrossRef]
30. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 8780–8794. Available online: https://papers.nips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html (accessed on 26 November 2024).
31. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A comprehensive survey of AI-Generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv* **2023**, arXiv:2303.04226. [CrossRef]
32. Hu, H.; Shang, Y.; Xu, G.; He, C.; Zhang, Q. Can GPT-O1 kill all bugs? An evaluation of GPT-family LLMs on QuixBugs. *arXiv* **2024**, arXiv:2409.10033. [CrossRef]
33. Van Huynh, N.; Wang, J.; Du, H.; Hoang, D.T.; Niyato, D.; Nguyen, D.N.; Kim, D.I.; Letaief, K.B. Generative AI for physical layer communications: A survey. *IEEE Trans. Cogn. Commun. Netw.* **2024**, *10*, 706–728. [CrossRef]
34. Thorsager, M.; Leyva-Mayorga, I.; Soret, B.; Popovski, P. Generative network layer for communication systems with artificial intelligence. *IEEE Netw. Lett.* **2024**, *6*, 82–86. [CrossRef]
35. Khoramnejad, F.; Hossain, E. Generative AI for the optimization of next-generation wireless networks: Basics, state-of-the-Art, and open challenges. *arXiv* **2024**, arXiv:2405.17454. [CrossRef]
36. Xu, M.; Niyato, D.; Kang, J.; Xiong, Z.; Guo, S.; Fang, Y.; Kim, D.I. Generative AI-enabled mobile tactical multimedia networks: Distribution, generation, and perception. *arXiv* **2024**, arXiv:2401.06386. [CrossRef]
37. Zhang, R.; Xiong, K.; Du, H.; Niyato, D.; Kang, J.; Shen, X.; Poor, H.V. Generative AI-Enabled vehicular networks: Fundamentals, framework, and case Study. *IEEE Netw.* **2024**, *38*, 259–267. [CrossRef]
38. Liu, C.; Yang, S.; Xu, Q.; Li, Z.; Long, C.; Li, Z.; Zhao, R. Spatial-temporal large language model for traffic prediction. *arXiv* **2024**, arXiv:2401.10134. [CrossRef]
39. Nie, L.; Wang, X.; Zhao, Q.; Shang, Z.; Feng, L.; Li, G. Digital twin for transportation big data: A reinforcement learning-based network traffic prediction approach. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 896–906. [CrossRef]
40. Gu, H.; Zhao, L.; Han, Z.; Zheng, G.; Song, S. AI-enhanced cloud-edge-terminal collaborative network: Survey, applications, and future directions. *IEEE Commun. Surv. Tutor.* **2024**, *26*, 1322–1385. [CrossRef]
41. Karapantelakis, A.; Alizadeh, P.; Alabassi, A.; Dey, K.; Nikou, A. Generative AI in mobile networks: A survey. *Ann. Telecommun.* **2024**, *79*, 15–33. [CrossRef]
42. Du, H.; Zhang, R.; Liu, Y.; Wang, J.; Lin, Y.; Li, Z.; Niyato, D.; Kang, J.; Xiong, Z.; Cui, S.; et al. Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization. *arXiv* **2024**, arXiv:2308.05384. [CrossRef]
43. Huang, Y.; Xu, M.; Zhang, X.; Niyato, D.; Xiong, Z.; Wang, S.; Huang, T. AI-generated network design: A diffusion model-based learning approach. *IEEE Netw.* **2024**, *38*, 202–209. [CrossRef]
44. Cao, B.; Zheng, X.; Yuan, K.; Qin, D.; Hong, Y. Dynamic bandwidth allocation based on adaptive predictive for low latency communications in changing passive optical networks environment. *Opt. Fiber Technol.* **2021**, *64*, 102556. [CrossRef]
45. Luo, Y.; Zhang, C.; Wang, X.; Liang, X.; Qiu, K. Robust key update with controllable accuracy using support vector machine for secure OFDMA-PON. *J. Light. Technol.* **2023**, *41*, 4663–4671. [CrossRef]
46. Singh, K.; Varma, P.R.K.; Singh, R.; Kaur, R. Predicting the performance of broadband passive optical networks using machine learning. *J. Opt. Commun.* **2023**, *45*, s1287–s1294. [CrossRef]
47. Liem, A.T.; Hwang, I.-S.; Ganesan, E.; Taju, S.W.; Sandag, G.A. A novel temporal dynamic wavelength bandwidth allocation based on long-short-term-memory in NG-EPON. *IEEE Access* **2023**, *11*, 82095–82107. [CrossRef]
48. Ganesan, E.; Liem, A.T.; Hwang, I.-S.; Ab-Rahman, M.S.; Taju, S.W.; Sheikh, M.N.A. LSTM-based DWBA prediction for tactile applications in optical access network. *Photonics* **2023**, *10*, 37. [CrossRef]
49. Ruan, L.; Wong, E.; Zhu, H. Reinforcement learning-based bandwidth decision in optical access networks: A study of exploration strategy and time with confidence guarantee. *IEEE Trans. Commun.* **2024**, *72*, 6943–6955. [CrossRef]
50. Wu, D.; Wang, X.; Qiao, Y.; Wang, Z.; Jiang, J.; Cui, S.; Wang, F. NetLLM: Adapting large language models for networking. In *Proceedings of the ACM SIGCOMM 2024 Conference*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 661–678. [CrossRef]
51. Garza, A.; Challu, C.; Mergenthaler-Canseco, M. TimeGPT-1. *arXiv* **2024**, arXiv:2310.03589. [CrossRef]
52. Google. TimesFM-3: A Zero-Shot Foundation Model for Multivariate Forecasting. Available online: https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/ (accessed on 21 September 2026).
53. Open Networking Foundation. VOLTHA: Virtual OLT Hardware Abstraction. Available online: https://docs.voltha.org (accessed on 11 November 2026).

54. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv* **2020**, arXiv:2012.07436. Available online: https://arxiv.org/abs/2012.07436v3 (accessed on 21 September 2026).
55. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv* **2022**, arXiv:2211.14730. Available online: https://arxiv.org/abs/2211.14730v2 (accessed on 21 September 2026).
56. Liu, Y.; Qin, G.; Shi, Z.; Chen, Z.; Yang, C.; Huang, X.; Wang, J.; Long, M. Sundial: A family of highly capable time series foundation models. *arXiv* **2025**, arXiv:2502.00816. Available online: https://arxiv.org/abs/2502.00816v4 (accessed on 21 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.