A Physics-Guided Learning Framework for Accurate and Generalizable EDFA Gain Spectrum Modeling with FourierKAN [Invited]

Qizhi Qiu†, Yukang Zhu†, Shaowen Xiang†, Yihao Zhang, Jingyu Wen, Ziheng Zhang, Weisheng Hu, and Qunbi Zhuge, *Fellow, Optica*

***Abstract*—Accurate erbium-doped fiber amplifier (EDFA) gain spectrum modeling is essential for constructing reliable physical-layer digital twins of optical networks. However, conventional machine-learning-based EDFA models usually rely on generic neural network architectures and may exhibit limited reliability under unseen operating configurations, channel-loading patterns, and device variations. In this paper, we propose a physics-guided Fourier Kolmogorov-Arnold Network (FourierKAN) framework for accurate and generalizable EDFA gain spectrum modeling. The proposed method constructs a physics-derived baseline gain spectrum from the target gain and tilt settings, and then uses FourierKAN to learn the residual spectral ripple on activated channels. A two-stage transfer-learning strategy is further introduced to adapt the model from a large-scale source EDFA dataset to the target domain with limited measurements. Experimental results on a real-world EDFA dataset show that the proposed method achieves an overall mean absolute error (MAE) of 0.097 dB, corresponding to an 89% error reduction compared with a deep neural network transfer-learning baseline. For unseen operating configurations, the proposed method achieves an MAE of 0.105 dB and reduces the error by 91%, demonstrating strong generalization capability. Ablation studies further verify the effectiveness of the physics-guided residual formulation, the FourierKAN backbone, and the two-stage transfer-learning workflow. These results indicate that incorporating physical knowledge into machine-learning-based EDFA modeling can significantly improve both modeling accuracy and generalization performance.**

***Index Terms*—Digital twin, erbium-doped fiber amplifier, gain spectrum modeling, FourierKAN, machine learning, optical networks, transfer learning.**

Manuscript received xxx. This work was supported by the Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University under Grant 21TQ1400213. *(†Qizhi Qiu, Yukang Zhu, and Shaowen Xiang contributed equally to this work.) (Corresponding author: Qunbi Zhuge.)*

Qizhi Qiu is with the State Key Laboratory of Photonics and Communications, School of Information Science and Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Photonic Communications Lab, Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia (e-mail: qizhi.qiu@sjtu.edu.cn).

Yukang Zhu was with the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA. He is now an Independent Researcher, Hangzhou 310030, Zhejiang, China (e-mail: yukangzhu@g.ucla.edu).

Shaowen Xiang, Yihao Zhang, Jingyu Wen, Ziheng Zhang, Weisheng Hu, and Qunbi Zhuge are with the State Key Laboratory of Photonics and Communications, School of Information Science and Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: starryspace@sjtu.edu.cn; yihao.zhang@sjtu.edu.cn; jingyu.wen@sjtu.edu.cn; zihengzhang@sjtu.edu.cn; wshu@sjtu.edu.cn; qunbi.zhuge@sjtu.edu.cn).

I. Introduction

The exponential growth of data-intensive technologies such as large language models (LLMs) and cloud computing imposes increasing demands on the capacity of optical networks [1]. To enhance network capacity and fully exploit reserved network margins, numerous intelligent management and control strategies have been proposed and extensively investigated [2]–[5]. These strategies heavily rely on digital twins (DTs) of optical networks, which can dynamically reflect the real-time state and physical information to support intelligent network management and control [6], [7]. Consequently, establishing an accurate DT based on physical-layer data is critical to the efficient and intelligent management of autonomous optical networks [8]. Within this context, constructing a DT for the erbium-doped fiber amplifier (EDFA) is particularly crucial. As a widely deployed amplification component, the EDFA directly determines the amplified spontaneous emission (ASE) noise introduced during signal transmission. It also significantly affects the optical power evolution along the transmission link, which dictates the magnitude of fiber nonlinear interference [9]. These factors constitute the primary impairments in long-haul optical transmission systems [10], which means that the precision of EDFA modeling directly determines the accuracy of quality-of-transmission (QoT) estimation [11], [12].

However, modeling the EDFA gain profile remains a highly challenging issue in the development of reliable DT models [13]. A straightforward solution is to describe the EDFA through physical models, such as the Giles model [14]. However, building an accurate EDFA model based on purely physical mechanisms is almost infeasible for commercial devices. First, the key erbium-doped fiber (EDF) parameters, such as the absorption and emission cross sections and doping concentration, are generally unavailable from commercial amplifier products. Second, the amplifier behavior is governed by multiple coupled physical effects, such as spectral hole burning (SHB), excited-state absorption, homogeneous upconversion, and component aging, which makes the theoretical characterization of the wavelength-dependent gain response highly complicated [15]. In addition, practical EDFAs usually incorporate internal control algorithms and electronic feedback mechanisms. These mechanisms reshape the pump power and gain dynamics under different operating conditions, while their effects are difficult to decouple from the intrinsic optical amplification process.

Therefore, machine learning (ML) techniques have been widely adopted to improve the accuracy of EDFA modeling. In [16], a neural-network-based EDFA gain model is proposed to learn the nonlinear mapping between the input power profile, amplifier operating conditions, and wavelength-dependent gain spectrum, demonstrating improved prediction accuracy compared with conventional analytical models. In [17], an ML-based EDFA gain model generalizable to multiple physical devices is investigated by training the model with measurement data collected from different EDFAs, showing enhanced robustness against device-to-device variations. In [18], an open EDFA gain spectrum dataset is released and benchmarked with several data-driven modeling methods, providing a common basis for evaluating EDFA gain prediction accuracy and cross-device generalization. In [19], a generalized transfer learning architecture is proposed for EDFA gain spectrum modeling, where a pretrained model is adapted to unseen EDFAs with only a small number of target-device samples. In [20], ML-based EDFA models are further incorporated into a cascaded learning framework for multi-span optical power spectrum prediction, enabling system-level modeling across multiple fiber spans and EDFAs.

Nevertheless, most existing ML-based approaches require large amounts of training data, which are difficult to acquire for each EDFA in practical deployment scenarios due to the constraints of measurement time and test equipment. In addition, their generalization to unseen operating configurations is not sufficiently guaranteed. When the input power profile, gain setting, or channel loading condition deviates from the training distribution, the prediction error may increase and degrade model reliability. To overcome these limitations, physical knowledge has been introduced into ML-based EDFA modeling. Physics-guided schemes [21]–[23] have been shown to improve data efficiency, modeling accuracy, and generalization capability. Therefore, incorporating physical knowledge into advanced ML modeling frameworks is regarded as a promising solution to the remaining challenges in EDFA gain modeling.

In this paper, we propose a physics-guided framework for accurate and generalizable EDFA gain spectrum modeling. A physics-derived baseline gain spectrum is constructed from the EDFA gain and tilt settings to provide a physically meaningful estimate of the global gain profile. Moreover, a Fourier Kolmogorov-Arnold Network (FourierKAN) [24], [25] is employed to predict the residual ripple with respect to this baseline. FourierKAN uses learnable sinusoidal basis functions as nonlinear activations, which is well suited for modeling continuous gain spectra and localized nonlinear effects [25], such as SHB. Transfer learning is further introduced to adapt the model to device-specific EDFA responses using limited data, thereby improving modeling accuracy and practical applicability.

On a real-world EDFA dataset covering aging, SHB effects, and unseen operating configurations [26], experimental results demonstrate that the proposed method achieves a mean absolute error (MAE) of 0.097 dB, corresponding to an 89% error reduction compared with an established transfer learning baseline [18]. For the unseen subset, the proposed method achieves an MAE of 0.105 dB, reducing the error by 91% over the same baseline. These results confirm the high modeling accuracy and strong generalization capability of the proposed framework. Further ablation studies verify the effectiveness of the physics-guided residual modeling, the FourierKAN-based network structure, and the two-stage transfer learning strategy. Notably, the proposed method ranked first in the OFC 2026 Machine Learning Challenge [26], and the source code has been publicly released in a GitHub repository.

The rest of the paper is organized as follows. Section II presents the proposed modeling framework, including the physics-guided formulation, the FourierKAN-based architecture, and the two-stage transfer learning strategy. Section III describes the experimental setup. Section IV presents the accuracy comparisons, ablation studies, and further discussions. Finally, Section V concludes the paper.

II. Principle

A. Modeling Framework and Physics-Derived Baseline

As shown in Fig. 1(a), the proposed framework is designed to model the EDFA gain spectrum over a 95-channel WDM system. For the i-th operating state, the input features include the target gain setting g_i, the target gain tilt τ_i, the total input and output powers, the input spectrum, the activated-channel mask, and device descriptors such as EDFA type and device index, consistent with prior EDFA gain spectrum learning tasks [16], [18], [23]. The desired output is the channel-wise gain spectrum.

$$\mathbf{G}_i = [G_{i,1}, G_{i,2}, \ldots, G_{i,95}], \tag{1}$$

where only activated channels are used for training and evaluation [18], [23]. We denote the binary activated-channel mask by $\mathbf{m}_i$, with $m_{i,j} = 1$ if the j-th wavelength channel is active and $m_{i,j} = 0$ otherwise. In contrast to component-level models that train one network per EDFA, our objective is a single global spectrum mapping model shared across EDFA types, device indices, channel-loading patterns, and operating states, which is aligned with cross-device EDFA modeling and few-shot adaptation studies [17]–[19].

The physics-derived baseline is constructed from the target gain and tilt settings, which specify the dominant control objective of the EDFA gain spectrum. For a commercial EDFA operating in automatic gain control (AGC) mode, the pump power is dynamically adjusted so that the monitored average gain tracks the prescribed target gain [27]. In addition, the gain flattening filters (GFFs) and other passive components shape the static wavelength-dependent response of the amplifier [28]. These mechanisms provide a physical basis for decomposing the measured gain spectrum into a dominant baseline term and a residual spectral correction.

Specifically, following the widely used EDFA model in [29], the signal propagation in the EDF can be described as

$$\frac{1}{P_\lambda(z)}\frac{dP_\lambda(z)}{dz} = \rho\Gamma_\lambda \left[\sigma_{e,\lambda}N_2(z) - \sigma_{a,\lambda}N_1(z)\right] - \alpha_s, \tag{2}$$

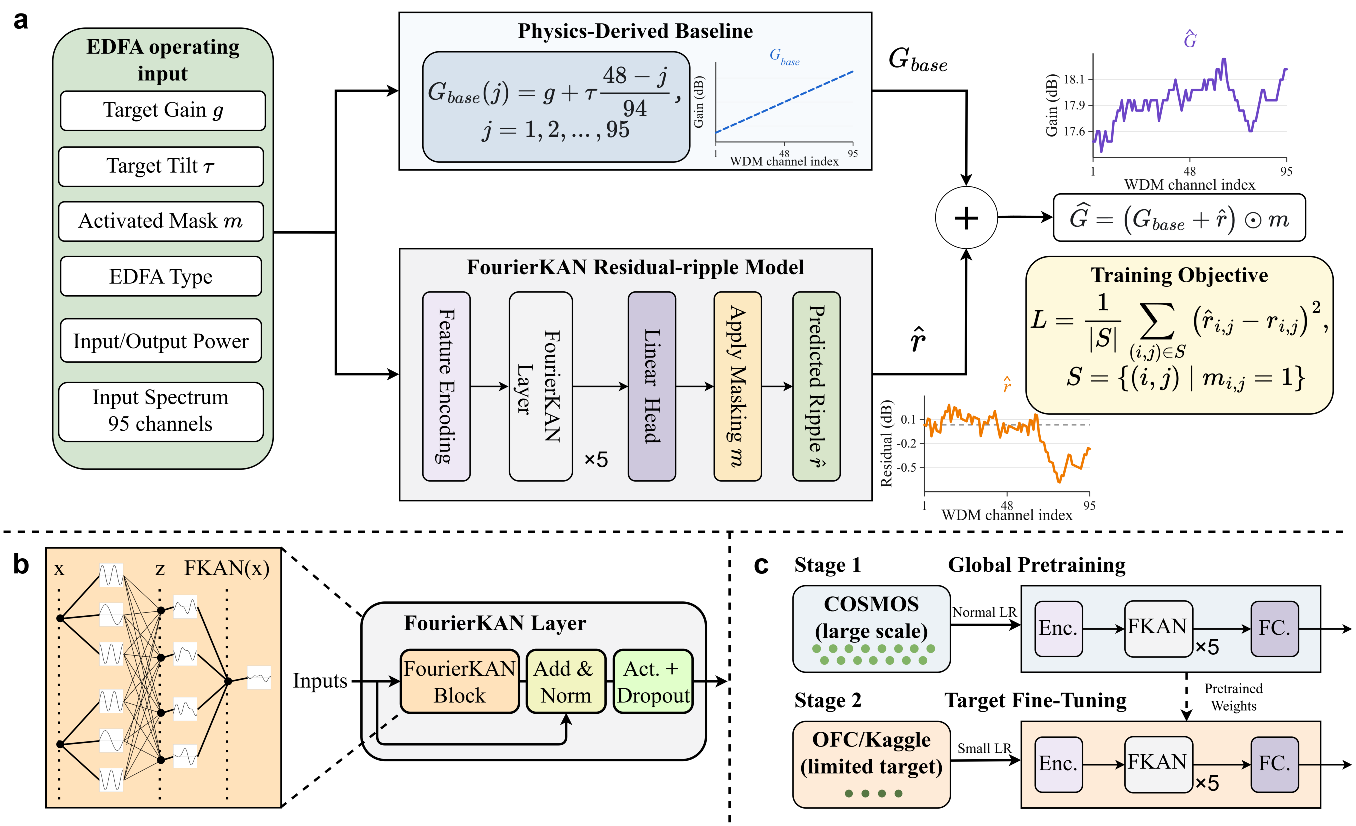

Fig. 1: Physics-guided gain spectrum modeling for an EDFA digital twin. (**a**) Physics-derived baseline and FourierKAN residual branches are combined with activated-channel masking to predict the gain spectrum. (**b**) FourierKAN layer for residual-ripple modeling. (**c**) Two-stage training with COSMOS-EDFA pretraining and OFC/Kaggle fine-tuning.

where ρ is the Er^{3+} ion concentration, α_s denotes the background signal loss, $\sigma_{a,\lambda}$ and $\sigma_{e,\lambda}$ are the absorption and emission cross sections at wavelength λ, and Γ_λ is the signal overlap factor. The terms $N_1(z)$ and $N_2(z)$ represent the fractional populations of the lower and upper energy levels at the fiber position z, respectively, which are determined by the local signal and pump powers. For a two-level EDFA model [14], these populations satisfy $N_1(z) + N_2(z) = 1$.

Substituting $N_1(z) = 1 - N_2(z)$ into Eq. (2) and integrating along the EDF length of the s-th stage, denoted as L_s, yields the gain in the dB domain as

$$G^{\mathrm{EDF}}_{s,\mathrm{dB}}(\lambda) = A(\lambda)\chi_s + B(\lambda)L_s, \tag{3}$$

where

$$\begin{aligned} \chi_s &= \int_0^{L_s} N_2(z)\,dz, \\ A(\lambda) &= \frac{10}{\ln 10}\rho\Gamma_\lambda\left(\sigma_{e,\lambda} + \sigma_{a,\lambda}\right), \\ B(\lambda) &= -\frac{10}{\ln 10}\left(\rho\Gamma_\lambda\sigma_{a,\lambda} + \alpha_s\right). \end{aligned} \tag{4}$$

Eq. (3) indicates that the EDF gain spectrum is controlled by a single scalar state variable χ_s, while $A(\lambda)$ and $B(\lambda)$ are fixed spectra determined by the erbium fiber parameters, which is consistent with the gain spectrum representation derived in [21].

For a multi-stage EDFA, the cascaded gains and the insertion responses of static components are additive in the dB domain. Denoting the static wavelength-dependent response after the s-th stage as $\mathrm{Loss}_s(\lambda)$, the overall gain is

$$G_{\mathrm{overall}}(\lambda) = \sum_{s=1}^{n}\left[A(\lambda)\chi_s + B(\lambda)L_s + \mathrm{Loss}_s(\lambda)\right]. \tag{5}$$

Assuming that the EDFs in different stages of the amplifier have the same specifications, $A(\lambda)$ and $B(\lambda)$ are shared by all stages. Therefore, Eq. (5) can be simplified as

$$G_{\mathrm{overall}}(\lambda) = A(\lambda)k + C(\lambda), \tag{6}$$

where

$$k = \sum_{s=1}^{n}\chi_s, \quad C(\lambda) = B(\lambda)\sum_{s=1}^{n}L_s + \sum_{s=1}^{n}\mathrm{Loss}_s(\lambda). \tag{7}$$

Here, k represents a global scalar state variable of the cascaded EDFs, while $C(\lambda)$ collects all wavelength-dependent static contributions. Hence, the GFF, couplers, isolators, fixed attenuators, and other passive elements are folded into the intercept spectrum $C(\lambda)$.

In AGC mode, the pump current is adjusted until the channel-averaged gain equals the target gain g_i. This control condition can be written as

$$\overline{G}_{\mathrm{overall}} = \overline{A}k_i + \overline{C} = g_i, \tag{8}$$

which gives

$$k_i = \frac{g_i - \overline{C}}{\overline{A}}. \tag{9}$$

Substituting Eq. (9) into Eq. (6), the gain at the j-th wavelength channel becomes

$$G_{\text{overall}}(\lambda_j) = \frac{A(\lambda_j)}{\overline{A}} g_i + C(\lambda_j) - \frac{A(\lambda_j)}{\overline{A}}\overline{C}. \tag{10}$$

When tilt control is enabled, the target gain command becomes wavelength dependent. For the 95-channel band considered in this work, it is written as

$$g_i(\lambda_j) = g_i + \tau_i \frac{48 - j}{94}, \quad j = 1, \ldots, 95, \tag{11}$$

where τ_i denotes the target tilt. By replacing g_i in Eq. (10) with the channel-dependent command $g_i(\lambda_j)$, we obtain

$$G_{\text{overall}}(\lambda_j) = \frac{A(\lambda_j)}{\overline{A}} g_i(\lambda_j) + C(\lambda_j) - \frac{A(\lambda_j)}{\overline{A}}\overline{C}. \tag{12}$$

By subtracting and adding $g_i(\lambda_j)$, the measured gain of the i-th sample at the j-th channel can be decomposed as

$$\begin{aligned} G_{i,j} = {} & g_i + \tau_i \frac{48 - j}{94} \\ & + \left[\frac{A(\lambda_j) - \overline{A}}{\overline{A}}\right] g_i(\lambda_j) \\ & + \left[C(\lambda_j) - \frac{A(\lambda_j)}{\overline{A}}\overline{C}\right]. \end{aligned} \tag{13}$$

The two correction terms that remain after subtracting the commanded gain profile have clear physical interpretations. The first correction is governed by $\frac{A(\lambda_j) - \overline{A}}{\overline{A}}$, which quantifies the deviation of the local gain from its band-averaged value. For a well-balanced EDF spectral response, this deviation is expected to remain close to zero. The second correction, $C(\lambda_j) - \frac{A(\lambda_j)}{\overline{A}}\overline{C}$, represents the remaining static spectral response after the band-averaged AGC constraint is imposed. In a well-designed commercial EDFA, this wavelength-dependent offset is expected to be suppressed by the GFF and other static spectral-shaping elements. However, it cannot be completely eliminated in practice due to manufacturing tolerances, fixed component responses, variable optical attenuator (VOA) deviations, aging, and the nonideal assumptions adopted in the above derivation.

Based on these observations, the two correction terms are defined as the residual spectral component $r_{i,j}$, which captures the device-specific deviation beyond the commanded gain profile. Accordingly, the baseline gain spectrum is defined as

$$G_{i,j}^{\text{base}} = g_i + \tau_i \frac{48 - j}{94}. \tag{14}$$

Consequently, the baseline $G_{i,j}^{\text{base}}$ represents the first-order gain profile directly specified by the amplifier control system through the target gain and tilt settings. The remaining difference between the measured gain and this commanded profile is treated as a residual spectral correction. To account for invalid or unloaded channels, the residual is defined in a masked form as

$$r_{i,j} = m_{i,j}\left(G_{i,j} - G_{i,j}^{\text{base}}\right), \tag{15}$$

where $G_{i,j}$ is the measured gain and $m_{i,j}$ denotes the channel-validity mask. This residual contains the device-specific spectral deviations that are not represented by the two global settings g_i and τ_i. Additional contributions may arise from unequal inversion redistribution among cascaded EDF stages, inter-stage saturation differences, SHB effects, and channel-loading-dependent operating-point variations. Since these effects are generally smooth in wavelength or limited in magnitude, the residual spectrum is well suited to a compact smooth function approximator, which motivates the Fourier-basis residual model introduced in the next subsection.

With the physics-derived baseline, the gain prediction is parameterized as a residual correction to Eq. (14), rather than as a direct regression of the absolute gain spectrum. Given the encoded EDFA operating conditions, the FourierKAN residual-ripple model predicts the masked residual $\hat{r}_{i,j}$. The final gain estimate is then reconstructed as

$$\hat{G}_{i,j} = m_{i,j}\left(G_{i,j}^{\text{base}} + \hat{r}_{i,j}\right). \tag{16}$$

Under this formulation, the physics-derived baseline corresponds to the zero-residual solution, while the FourierKAN-based residual model learns the remaining wavelength-dependent deviations.

B. FourierKAN-Based Residual-Ripple Modeling

The residual correction in Eq. (15) is learned by the FourierKAN-based model shown in Fig. 1(b). KANs replace fixed nodal activations with learnable function representations, and Fourier-parameterized variants provide compact sinusoidal transformations for modeling structured spectral variations [25]. In the proposed model, scalar operating quantities are standardized, categorical device descriptors are one-hot encoded, and both the input spectrum and the activated-channel mask are included in the feature encoding. The encoded feature vector is then processed by stacked FourierKAN layers and a linear prediction head to generate the channel-wise residual estimate $\hat{\mathbf{r}}_i$. For an input hidden representation $\mathbf{h}$, each FourierKAN layer first applies an affine transformation and then augments the transformed features using learnable sinusoidal basis functions,

$$\begin{gathered} \mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}, \\ \Phi(\mathbf{z}) = \mathbf{z} + \sum_{k=1}^{K} \left(\mathbf{a}_k \odot \sin(k\mathbf{z}) + \mathbf{c}_k \odot \cos(k\mathbf{z})\right), \end{gathered} \tag{17}$$

where K is the number of retained Fourier frequencies and $\odot$ denotes element-wise multiplication. Here, $\mathbf{a}_k$ and $\mathbf{c}_k$ are learnable Fourier coefficients. Each block applies the FourierKAN transformation followed by normalization, nonlinear activation, and dropout. Residual projections between adjacent blocks preserve low-order information, while the Fourier basis terms provide a compact representation for smooth wavelength-dependent residual corrections. This structure is well suited to modeling gain ripples and localized spectral deviations associated with SHB, channel-loading interactions, aging-induced drift, and other device-specific nonlinear responses. Related Fourier features and sinusoidal activations have also been shown to improve neural representations of frequency-rich signals [30], [31].

The model parameters are optimized using the masked mean-squared error between the predicted residual and the supervised residual target,

$$\mathcal{L}(\theta) = \frac{\sum_i \sum_{j=1}^{95} m_{i,j} \left[\hat{r}_{i,j} - r_{i,j}\right]^2}{\sum_i \sum_{j=1}^{95} m_{i,j}}. \tag{18}$$

The mask removes inactive channels from the loss calculation, which is consistent with the physical meaning of channel loading and with the evaluation protocol adopted in prior EDFA gain spectrum benchmarks [18], [23].

C. Two-Stage Transfer Learning

The same residual-ripple formulation is adopted in the two-stage transfer-learning workflow shown in Fig. 1(c). In the first stage, the model is pretrained on the large-scale COSMOS-EDFA dataset [18] after converting the data into a unified feature and label schema. This stage is intended to capture general gain spectrum regularities across heterogeneous EDFA types, device instances, operating configurations, and input spectral profiles. In the second stage, the pretrained model is fine-tuned on the OFC/Kaggle [26] target domain using a smaller learning rate, which allows the learned mapping to be moderately adjusted to the EDFA devices and operating states represented by the limited target-domain measurements. This two-stage strategy separates broad EDFA spectral learning from target-domain calibration. Specifically, large-scale pretraining provides a transferable initialization for gain spectrum modeling, while fine-tuning performs small device-specific updates with regularization techniques such as dropout to help reduce overfitting on the small target dataset. Following the standard source-to-target transfer-learning paradigm and recent EDFA-specific transfer studies [18]–[20], the proposed workflow supports effective and robust adaptation under limited labeled measurements.

III. Experimental Setup

A. Datasets

Two EDFA gain spectrum datasets were used to evaluate the proposed modeling strategy. The COSMOS-EDFA dataset was used as the source domain for pretraining, whereas the OFC/Kaggle dataset was used as the target domain for fine-tuning and testing.

COSMOS-EDFA Dataset. The COSMOS-EDFA dataset [18] was adopted as the large-scale source dataset. It contains 202,752 gain spectrum measurements collected from 16 commercial-grade ROADM booster and pre-amplifier EDFAs. The measurements cover a wide range of gain settings and channel-loading conditions, with each loading configuration measured repeatedly under different input power levels. In this work, we selected the subset with a target gain of 18 dB to align the source-domain gain range with that of the target dataset and to represent a commonly used EDFA gain setting in practical optical systems. This selection resulted in 41,440 samples, which were further divided into training and validation sets with a 95% to 5% ratio for pretraining.

OFC/Kaggle Dataset. The OFC/Kaggle dataset [26] was used as the target dataset. It contains gain spectrum measurements from multiple EDFAs under different channel-loading scenarios, where the measured responses include practical nonidealities such as aging-induced drift and SHB-related spectral variations. The dataset is designed to reflect data-scarce deployment conditions, where only a small fraction of labeled samples is available for model training. Specifically, the train-to-test ratios range from 2:13 to 1:32. Several operating configurations are present only in the test set, which introduces unseen cases and further increases the difficulty of generalization. After data cleaning and removal of abnormal samples, only 473 valid samples were retained for fine-tuning and then divided into training and validation sets with a 95% to 5% ratio. The final test set contains 21,056 samples covering aging-related effects, SHB-related effects, and unseen operating configurations.

B. Experimental Design

The experiments consist of one main comparison and three ablation studies. Unless otherwise specified, all learned models follow the two-stage pretraining and fine-tuning workflow described in Section II. They are trained with the masked objective in Eq. (18) using the physics-derived baseline gain spectrum in Eq. (14).

Main experiments. The proposed model is evaluated against two reference methods under the same preprocessing and evaluation protocol. The first reference is a deep neural network (DNN) transfer-learning baseline based on the established dense EDFA gain model reported in [18]. In the original study, a component-level fully connected network is trained for each EDFA, and transfer learning is performed separately from a source EDFA model to each target EDFA model using limited target-domain measurements. The network employs batch normalization and ELU activations, and directly predicts the absolute gain spectrum. For a fair comparison in our global-model setting, we adapt the dense architecture to the same input schema and two-stage training workflow as the proposed model, while preserving its absolute-gain prediction target. This design allows the proposed FourierKAN residual model to be compared with a representative DNN transfer-learning approach for EDFA gain spectrum modeling. The second reference is the physics-derived baseline, which directly outputs Eq. (14) without any learned residual correction. This baseline corresponds to the zero-residual case of the proposed formulation and quantifies the gain obtained from learning the residual ripple.

Ablation studies. The ablation studies quantify the contribution of the main components of the proposed framework. Each variant is derived from the full configuration by modifying one design factor at a time. First, the physics-guided formulation ablation removes the physics-derived baseline and directly regresses the absolute gain spectrum, thereby evaluating the role of residual-gain learning. Second, the architecture ablation replaces the FourierKAN backbone with same-capacity multilayer perceptron (MLP), one-dimensional convolutional neural network (CNN), and Transformer encoder backbones while keeping the residual-learning formulation unchanged. Third, the transfer-learning ablation compares the full two-stage workflow with three alternatives: direct transfer without target-domain fine-tuning, target-only training from scratch, and single-stage training on the combined source and target datasets.

C. Implementation Details

The proposed FourierKAN residual-ripple model uses five hidden layers with widths $(256, 256, 128, 128, 128)$, $K = 4$ Fourier frequencies, dropout of 0.2, and residual projections between consecutive blocks. Unless otherwise stated, all learned models use the same feature preprocessing, activated-channel mask, masked mean-squared-error (MSE) objective, and 5% validation split. The learning rate is scheduled by ReduceLROnPlateau with a reduction factor of 0.5, patience of 10 validation checks, and minimum learning rate of 10^{-7}. Early stopping is applied according to the validation loss. The detailed hyperparameters used in the pretraining and fine-tuning stages are summarized in Table I.

TABLE I: Training Hyperparameters.

Setting	**Pretraining**	**Fine-tuning**
Dataset	COSMOS-EDFA	OFC/Kaggle
Optimizer	Adam	AdamW
Learning rate	1×10^{-3}	2×10^{-4}
Weight decay	1×10^{-4}	5×10^{-5}
Batch size	256	64
Maximum epochs	500	500
Early-stopping patience	80 epochs	60 epochs
Validation interval	2 epochs	1 epoch
Gradient clipping	1.0	1.0
Loss	masked MSE	masked MSE

The DNN baseline uses hidden widths $(256, 128, 128, 128)$ with batch normalization and ELU activation. It directly predicts the absolute gain spectrum and follows the same two-stage training protocol. Since this baseline contains batch-normalization layers, the fine-tuning stage freezes both the batch-normalization statistics and affine parameters to avoid unstable updates from the small target-domain dataset. In addition, the trailing incomplete mini-batch is dropped in both stages to avoid unreliable batch-statistic estimates. Its gradient clipping threshold is set to 3.0, following the training setting in [18].

For the architecture ablation, the FourierKAN backbone is replaced with three alternative architectures while keeping the remaining training pipeline unchanged. The MLP baseline uses the same hidden widths as the proposed model. The CNN baseline uses channel widths $(64, 96, 128, 160)$ with a kernel size of 7. The Transformer baseline uses a model dimension of $d_{\text{model}} = 96$, four attention heads, four encoder layers, and a feed-forward dimension of 192.

IV. Results and Discussion

A. Training Process

Fig. 2 shows the pretraining and fine-tuning loss curves for a representative run with a random seed of 48. In both stages, the training and validation losses decrease rapidly and remain stable, indicating effective optimization under the adopted settings. Pretraining reaches the maximum number of 500 epochs, whereas fine-tuning converges after approximately 150 epochs.

During pretraining, the validation loss is slightly lower than the training loss. This behavior is mainly attributed to dropout, which is applied only during training to mitigate overfitting. Since dropout randomly deactivates a fraction of neurons during training but is disabled during validation, the validation loss can be lower because the full network is used for prediction. During fine-tuning, the validation loss in Fig. 2(b) is slightly higher than the training loss. This difference is mainly attributed to the limited size of the fine-tuning dataset, from which an even smaller validation subset is separated. Consequently, the validation loss is more sensitive to the random data split. Runs with other random seeds also show cases where the validation loss is slightly lower than the training loss, confirming that these small differences are consistent with the stochastic nature of the split and training process.

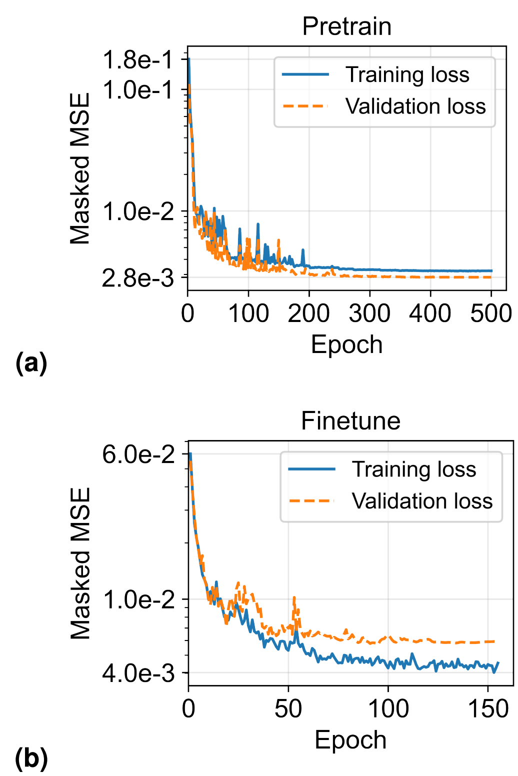

Fig. 2: Loss curves during the network training process: (**a**) pretraining stage and (**b**) fine-tuning stage (with random seed 48).

B. Modeling Performance Comparison

To evaluate the modeling accuracy, the proposed method is compared with the DNN baseline and the physics-derived baseline. As described above, both the proposed model and the DNN baseline are trained using the same two-stage workflow, including source-domain pretraining and target-domain fine-tuning, and are then tested on the target-domain test subsets. All learning-based methods are independently trained with four different random seeds. The resulting test errors are summarized in Fig. 3. The aging, SHB, and unseen subsets are shown with different colors, while the overall results are computed over all target-domain test samples. Each bar represents the mean test error over four independent runs, and the error bar denotes the corresponding standard deviation. Since the physics-derived baseline is deterministically calculated from the target gain and target tilt, it does not involve random initialization or stochastic training, resulting in zero standard deviation over repeated runs.

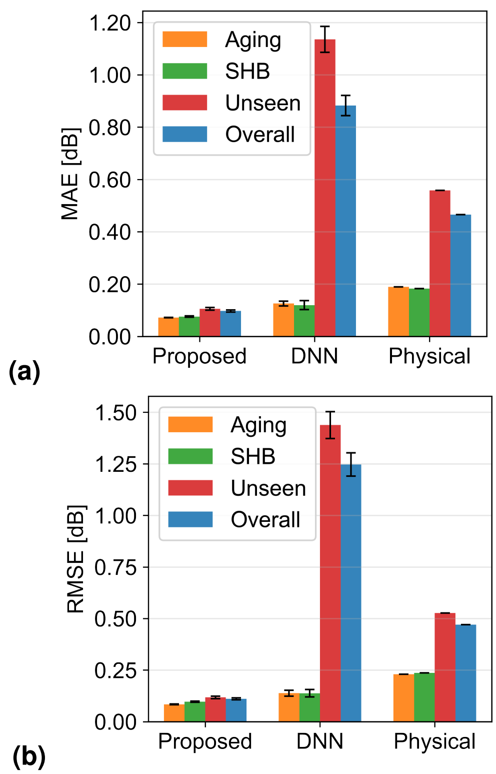

Fig. 3: Modeling performance across test subsets. (**a**) MAE and (**b**) RMSE for the proposed model, DNN baseline, and physics-derived baseline.

As shown in Fig. 3, the proposed method achieves the best overall accuracy among the three methods, with an MAE of 0.097 dB and an RMSE of 0.111 dB. Compared with the DNN baseline, which obtains an overall MAE of 0.883 dB and an RMSE of 1.247 dB, the proposed method reduces the MAE and RMSE by 89% and 91%, respectively. Compared with the physics-derived baseline, which obtains an overall MAE of 0.465 dB and an RMSE of 0.471 dB, the proposed method further reduces the MAE and RMSE by 79% and 76%, respectively. These results demonstrate that the proposed strategy provides substantially higher accuracy for EDFA gain spectrum prediction.

It is also worth noting that the DNN baseline achieves accuracy comparable to the proposed method on the aging and SHB subsets. However, its error increases significantly on the unseen subset, where both the MAE and RMSE exceed 1 dB. This degradation dominates its overall performance and indicates limited generalization to EDFA configurations that are not covered by the target-domain fine-tuning data. This behavior can be attributed to the fact that the DNN baseline directly regresses the absolute gain spectrum in a purely data-driven manner, without using a physics-based prior to constrain the low-order gain trend. In contrast, the proposed method incorporates the analytical gain and tilt baseline as prior information, such that the neural network mainly learns the residual wavelength-dependent correction. As a result, the proposed method shows only a slight error increase on the unseen subset, with an MAE of 0.105 dB and an RMSE of 0.118 dB. Compared with the DNN baseline, these results correspond to error reductions of 91% and 92% in MAE and RMSE, respectively. This demonstrates that the proposed model achieves high modeling accuracy while substantially improving the generalization capability for EDFA gain spectrum prediction.

C. Per-Channel Modeling Performance

The per-channel error distribution is further analyzed to assess whether the predicted gain spectra remain reliable across the 95 WDM channels. Fig. 4 compares the proposed model with the DNN baseline in terms of the channel-wise median absolute error and the 95th-percentile absolute error. The former characterizes the typical prediction accuracy at each wavelength channel, while the latter corresponds to the boundary of the shaded region and reflects the error distribution.

For the aging and SHB subsets in Fig. 4(a) and Fig. 4(b), both models achieve relatively small median errors over most active channels. Nevertheless, the proposed model shows a lower 95th-percentile upper boundary for most wavelength channels, indicating that large-error events are effectively suppressed at individual channels. The shaded region of the proposed model is also generally more concentrated, which suggests a more stable error distribution across different samples. This improvement is consistent with the proposed residual formulation, where the physics-derived baseline captures the smooth gain and tilt trend, while the FourierKAN branch learns the remaining wavelength-dependent correction associated with device condition, SHB, and channel-loading effects.

The advantage of the proposed model becomes more evident on the unseen subset in Fig. 4(c). The average channel-wise median error is reduced from 1.063 dB for the DNN baseline to 0.073 dB for the proposed model, while the mean upper-envelope error, measured by the 95th percentile, decreases from 2.124 dB to 0.234 dB. The larger high-percentile error of the DNN baseline indicates that direct absolute-gain regression becomes poorly calibrated when the operating configuration

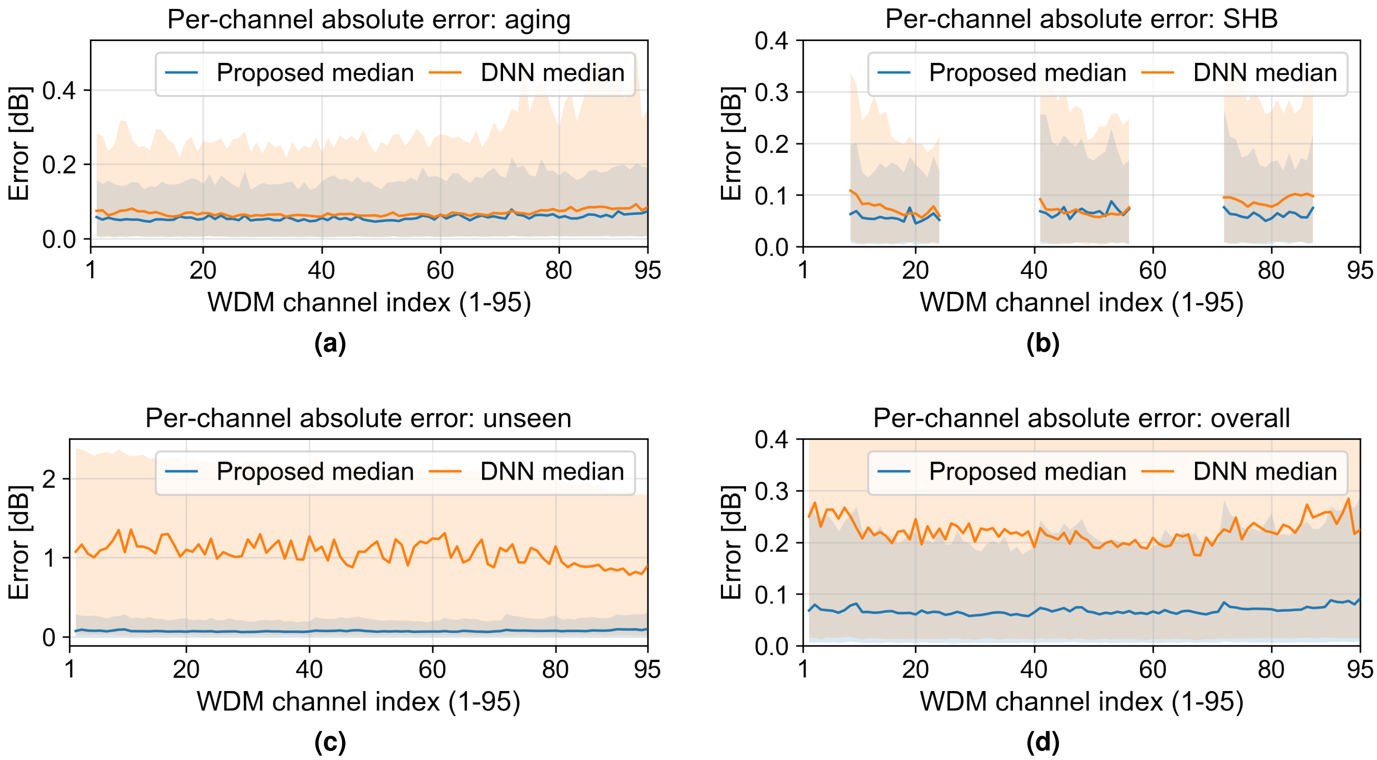

Fig. 4: Per-channel absolute error to the measured gain spectra for (**a**) aging, (**b**) SHB, (**c**) unseen, and (**d**) overall test subsets. Solid curves denote the channel-wise median absolute error, and the shaded regions denote the 5th–95th percentile error envelope.

deviates from the fine-tuning distribution. In contrast, the proposed model is anchored by the target gain and tilt baseline, which constrains the low-order spectral trend and leaves only the residual ripple to be learned. This formulation improves robustness under unseen channel-loading and device conditions. Over the overall test set, the proposed model reduces the average 95th-percentile error from 2.099 dB to 0.221 dB. This result indicates that the proposed strategy improves the overall modeling accuracy. More importantly, it suppresses the upper tail of the per-channel error distribution, thereby reducing the occurrence of large errors at individual wavelength channels.

D. Ablation Studies

The preceding results demonstrate that the proposed model improves the overall prediction accuracy, generalizes well to unseen operating configurations, and provides reliable channel-wise performance. To further verify the effectiveness of each key component in the proposed framework, ablation studies are conducted from three perspectives. First, the physics-based residual formulation is examined to determine whether it reduces the extrapolation burden under unseen configurations. Second, the FourierKAN backbone is evaluated to verify its ability to represent wavelength-dependent spectral ripple. Third, the two-stage transfer-learning workflow is analyzed to assess whether it effectively separates broad EDFA spectral learning from target-domain calibration.

1) Physics-based formulation: Fig. 5 evaluates the contribution of the physics-derived baseline and the residual-ripple

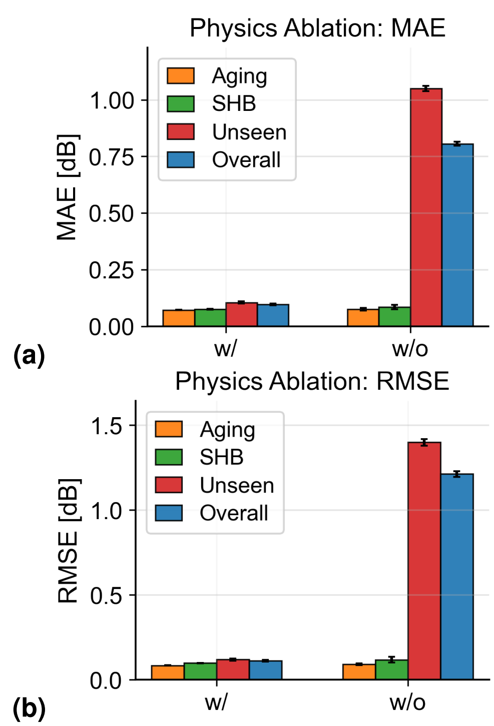

Fig. 5: Ablation of the physics-based residual formulation grouped by test subset. (**a**) MAE and (**b**) RMSE obtained with (w/) and without (w/o) the physics-based residual formulation.

learning formulation. When the same FourierKAN backbone is trained to directly regress the absolute gain spectrum, the overall MAE increases from 0.097 dB to 0.806 dB, and the overall RMSE increases from 0.111 dB to 1.212 dB. The degradation is most pronounced on the unseen subset, where the MAE increases from 0.105 dB to 1.049 dB. By contrast, the aging and SHB subsets show a smaller performance gap between the two target formulations, suggesting that direct absolute-gain regression can still interpolate partially observed device conditions but becomes much less reliable under unseen operating configurations.

These results confirm that the physics-based residual formulation is the main source of the improved generalization capability. Without baseline subtraction, the neural network must jointly learn the commanded gain level, the gain-tilt slope, and the wavelength-dependent residual ripple from limited target-domain data. This creates a difficult extrapolation problem when the input power profile, channel-loading pattern, or device condition deviates from the fine-tuning distribution. With residual-ripple learning, the deterministic gain and tilt trend is removed from the learning target, so the model only needs to estimate a smaller correction around the physics-derived baseline. The severe degradation on the unseen subset after removing this formulation indicates that the baseline is not merely a numerical offset. Instead, it provides a physical reference that stabilizes prediction outside the observed target-domain configurations.

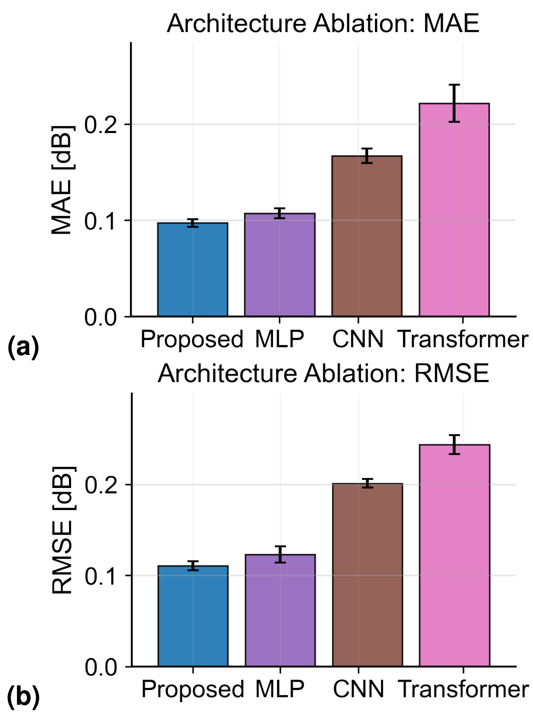

Fig. 6: Architecture ablation results: (**a**) MAE and (**b**) RMSE for the proposed FourierKAN residual-ripple model, MLP, CNN, and Transformer backbones.

2) Architecture comparison: To evaluate the effect of the neural network architecture, the proposed FourierKAN residual-ripple model is further compared with MLP, CNN, and Transformer variants under the same physics-based residual formulation and two-stage training protocol. The resulting prediction accuracy is summarized in Fig. 6. The proposed model achieves the lowest overall MAE and RMSE of 0.097 dB and 0.111 dB, respectively. The same-width MLP is the closest alternative, with an overall MAE of 0.107 dB and an RMSE of 0.123 dB, corresponding to moderate reductions of 9% in MAE and 10% in RMSE. The CNN and Transformer baselines show larger errors, with overall MAEs of 0.167 dB and 0.222 dB, respectively. Compared with the ablation of the physics-based formulation, the architecture-dependent gap is smaller. This suggests that the target formulation primarily determines whether the learning problem is well conditioned, while the backbone further affects how accurately the residual ripple can be represented.

The relatively small gap between FourierKAN and the same-width MLP indicates that the physics-based residual formulation has already simplified the learning problem. FourierKAN should therefore be viewed as a complementary backbone rather than the primary source of the performance gain. Its role is to provide a moderate but consistent improvement in residual-spectrum approximation under the physics-based formulation.

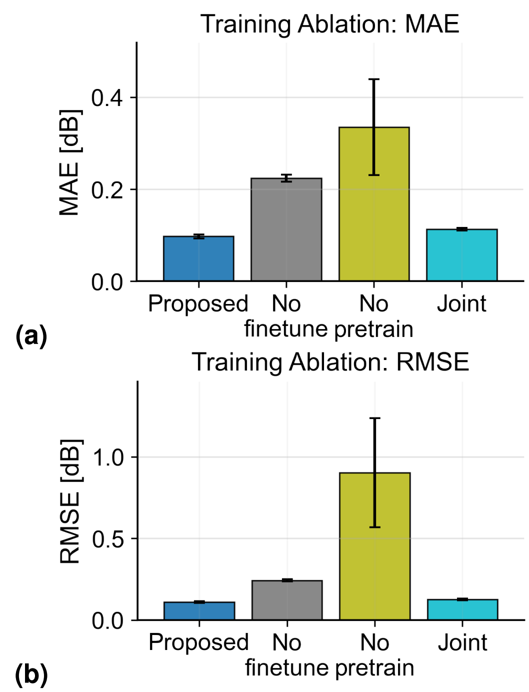

Fig. 7: Ablation results of different training strategies. (**a**) MAE and (**b**) RMSE obtained with the full two-stage workflow, source-only pretraining without fine-tuning, target-only training without pretraining, and joint training on the combined source and target datasets.

3) Training strategy: To assess the contribution of the transfer-learning workflow, the full two-stage strategy is compared with three alternative training strategies. The corresponding prediction accuracy is summarized in Fig. 7. When target-domain fine-tuning is removed, the overall MAE increases to 0.224 dB, indicating that source-domain pretraining alone cannot fully compensate for the mismatch between COSMOS-EDFA and the OFC/Kaggle target domain. When the model is trained only on the target dataset without COSMOS pretraining, the overall MAE and RMSE further increase to 0.335 dB and 0.903 dB, respectively. This degradation reflects the instability caused by the limited amount of target-domain training data. Joint training on the union of the two datasets is more competitive, with an overall MAE of 0.113 dB and an RMSE of 0.127 dB, but it still underperforms the proposed sequential pretraining and fine-tuning workflow.

These results show that the two stages provide complementary benefits. Pretraining on the large source-domain dataset enables the model to learn general EDFA spectral characteristics from a broad distribution of operating conditions, including the relationships among input power spectra, channel loading, and residual gain ripple. However, this source-domain representation still requires calibration using the small target-domain dataset, since the target domain contains device-specific responses and operating-condition-dependent effects that are not fully covered by the source distribution. In contrast, target-only training does not provide sufficient labeled samples to learn a stable residual mapping, while joint training may be biased toward the larger source-domain distribution and does not explicitly prioritize target-domain adaptation. The sequential workflow therefore provides a more effective transfer-learning strategy. Source-domain pretraining establishes a broad spectral representation, while target-domain fine-tuning calibrates this representation to the target devices and operating regimes. This combination enables more accurate EDFA gain modeling under limited target-domain data.

V. Conclusion

In this paper, we proposed a physics-guided FourierKAN framework for accurate and generalizable EDFA gain spectrum modeling. Based on the mathematical derivation of the EDFA gain response under AGC and tilt-control mechanisms, the measured gain spectrum is decomposed into a physically meaningful baseline and a residual spectral component. The baseline is determined by the target gain and tilt settings, and represents the dominant gain profile commanded by the amplifier control system. The residual term is centered around zero and captures complex physical effects and practical nonidealities that cannot be represented by the baseline alone. A FourierKAN-based residual-ripple model is then trained with a two-stage transfer learning strategy to achieve accurate and data-efficient residual modeling.

Experiments on a real-world EDFA dataset show that the proposed method achieves 0.097 dB overall MAE and maintains stable performance under unseen operating configurations, outperforming both data-driven and physics-derived baselines. Ablation studies further verify the effectiveness of the physics-based residual formulation, FourierKAN modeling, and transfer learning strategy. These results demonstrate the potential of the proposed framework for accurate EDFA DT construction in QoT estimation and autonomous optical network management. They also confirm the feasibility of incorporating physical knowledge into data-driven modeling methods to improve both accuracy and generalization.

References

[1] Cisco, "Cisco Research: A major infrastructure shift is underway. AI could double the strain or solve it," 2025, [Online]. Available: https://investor.cisco.com/news/news-details/2025/Cisco-Research-A-Major-Infrastructure-Shift-Is-Underway--AI-Could-Double-the-Strain-or-Solve-It/default.aspx.

[2] Y. Song, M. Zhang, Y. Zhang, Y. Shi, S. Shen, X. Tang, S. Huang, and D. Wang, "Lifecycle management of optical networks with dynamic-updating digital twin: A hybrid data-driven and physics-informed approach," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 5, pp. 1767–1779, 2025.

[3] Y. Zhang, Q. Qiu, X. Liu, X. Yu, D. Fu, X. Liu, Z. Wang, H. Lin, Y. Chen, L. Yi, W. Hu, and Q. Zhuge, "AI agent for autonomous optical networks: architectures, technologies, and prospects [invited tutorial]," *Journal of Optical Communications and Networking*, vol. 18, no. 2, pp. A159–A178, 2026. [Online]. Available: https://doi.org/10.1364/JOCN.576017

[4] X. Liu, Y. Zhang, Q. Qiu, Y. Cheng, W. Hu, and Q. Zhuge, "Field trial of an LLM-powered AI agent for autonomous optical networks: a full-lifecycle demonstration," *Journal of Optical Communications and Networking*, vol. 18, no. 5, pp. A179–A190, 2026. [Online]. Available: https://doi.org/10.1364/JOCN.575402

[5] Q. Qiu, Y. Zhang, X. Liu, S. Wu, L. Yi, W. Hu, and Q. Zhuge, "Expertise-guided LLM agent for autonomous optical power optimization in field-deployed optical networks," *Journal of Optical Communications and Networking*, vol. 18, no. 5, pp. 481–491, 2026. [Online]. Available: https://doi.org/10.1364/JOCN.588873

[6] Q. Zhuge, X. Liu, Y. Zhang, M. Cai, Y. Liu, Q. Qiu, X. Zhong, J. Wu, R. Gao, L. Yi, and W. Hu, "Building a digital twin for intelligent optical networks [Invited Tutorial]," *Journal of Optical Communications and Networking*, vol. 15, no. 8, pp. C242–C262, 2023. [Online]. Available: https://doi.org/10.1364/JOCN.483600

[7] D. Wang, Z. Zhang, M. Zhang, M. Fu, J. Li, S. Cai, C. Zhang, and X. Chen, "The role of digital twin in optical communication: Fault management, hardware configuration, and transmission simulation," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 133–139, 2021.

[8] A. Richard, "Autonomous networks: Empowering digital transformation for the telecoms industry," TM Forum White Paper, 2019, https://www.tmforum.org/resources/standard/autonomous-networks-empowering-digital-transformation-telecoms-industry/.

[9] P. Poggiolini, "The GN model of non-linear propagation in uncompensated coherent optical systems," *Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3857–3879, 2012.

[10] A. Ferrari, A. Napoli, J. K. Fischer, N. Costa, A. D'Amico, J. Pedro, W. Forysiak, E. Pincemin, A. Lord, A. Stavdas, J. P. F.-P. Gimenez, G. Roelkens, N. Calabretta, S. Abrate, B. Sommerkorn-Krombholz, and V. Curri, "Assessment on the achievable throughput of multi-band ITU-T G.652.D fiber transmission systems," *Journal of Lightwave Technology*, vol. 38, no. 16, pp. 4279–4291, 2020.

[11] Y. Pointurier, "Design of low-margin optical networks," *Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A9–A17, 2017. [Online]. Available: https://doi.org/10.1364/JOCN.9.0000A9

[12] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An overview on application of machine learning techniques in optical networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383–1408, 2019.

[13] Y. Liu, X. Liu, L. Liu, Y. Zhang, M. Cai, L. Yi, W. Hu, and Q. Zhuge, "Modeling EDFA gain: Approaches and challenges," *Photonics*, vol. 8, no. 10, p. 417, 2021. [Online]. Available: https://doi.org/10.3390/photonics8100417

[14] C. Giles and E. Desurvire, "Modeling erbium-doped fiber amplifiers," *Journal of Lightwave Technology*, vol. 9, no. 2, pp. 271–283, 1991.

[15] M. Bolshtyansky, "Spectral hole burning in erbium-doped fiber amplifiers," *Journal of Lightwave Technology*, vol. 21, no. 4, pp. 1032–1038, 2003.

[16] S. Zhu, C. L. Gutterman, W. Mo, Y. Li, G. Zussman, and D. C. Kilper, "Machine learning based prediction of erbium-doped fiber wdm line amplifier gain spectra," in *2018 European Conference on Optical Communication (ECOC)*, 2018, pp. 1–3.

[17] F. D. Ros *et al.*, "Machine learning-based EDFA gain model generalizable to multiple physical devices," in *European Conference on Optical Communication (ECOC)*, 2020. [Online]. Available: https://doi.org/10.1109/ECOC48923.2020.9333297

[18] Z. Wang, D. C. Kilper, and T. Chen, "Open EDFA gain spectrum dataset and its applications in data-driven EDFA gain modeling," *Journal of*

Optical Communications and Networking, vol. 15, no. 9, pp. 588–599, 2023. [Online]. Available: https://doi.org/10.1364/JOCN.491901

[19] A. Raj, Z. Wang, T. Chen, D. C. Kilper, and M. Ruffini, "Generalized few-shot transfer learning architecture for modeling the EDFA gain spectrum," *Journal of Optical Communications and Networking*, vol. 17, no. 9, pp. D106–D117, 2025. [Online]. Available: https://doi.org/10.1364/JOCN.560987

[20] Z. Wang, Y.-K. Huang, S. Han, D. Kilper, and T. Chen, "Multi-span optical power spectrum prediction using cascaded learning with one-shot end-to-end measurement," *Journal of Optical Communications and Networking*, vol. 17, no. 1, pp. A23–A33, 2025. [Online]. Available: https://doi.org/10.1364/JOCN.533634

[21] Y. Liu, X. Liu, Y. Zhang, M. Cai, M. Fu, X. Zhong, L. Yi, W. Hu, and Q. Zhuge, "Building a digital twin of an EDFA for optical networks: a gray-box modeling approach," *Journal of Optical Communications and Networking*, vol. 15, no. 11, pp. 830–838, 2023. [Online]. Available: https://doi.org/10.1364/JOCN.499530

[22] Y. Chen, X. Liu, Y. Liu, Y. Zhang, L. Yi, W. Hu, and Q. Zhuge, "A comprehensive analysis on machine learning based EDFA gain model," in *Optica Advanced Photonics Congress 2022*, ser. Technical Digest Series. Maastricht, Limburg, Netherlands: Optica Publishing Group, 2022, p. NeTh1C.4, photonic Networks and Devices. [Online]. Available: https://doi.org/10.1364/NETWORKS.2022.NeTh1C.4

[23] J. Yu, S. Zhu, C. L. Gutterman, G. Zussman, and D. C. Kilper, "Machine-learning-based EDFA gain estimation," *Journal of Optical Communications and Networking*, vol. 13, no. 4, pp. B83–B91, 2021, invited. [Online]. Available: https://doi.org/10.1364/JOCN.417584

[24] A. Mehrabian, P. M. Adi, M. Heidari, and I. Hacihaliloglu, "Implicit neural representations with fourier kolmogorov-arnold networks," 2025. [Online]. Available: https://arxiv.org/abs/2409.09323

[25] J. Xu, Z. Chen, J. Li, S. Yang, W. Wang, X. Hu, and E. C.-H. Ngai, "FourierKAN-GCF: Fourier kolmogorov-arnold network – an effective and efficient feature transformation for graph collaborative filtering," *arXiv preprint arXiv:2406.01034*, 2024. [Online]. Available: https://arxiv.org/abs/2406.01034

[26] M. Han and Z. Wang, "OFC 2026 ML challenge," https://kaggle.com/competitions/ofc-2026-ml-challenge, 2025.

[27] M. Hashimoto, M. Yoshida, and H. Tanaka, "The characteristics of WDM systems with hybrid AGC EDFA in the photonics network," in *Optical Fiber Communications Conference*, ser. OSA Trends in Optics and Photonics, A. Sawchuk, Ed., vol. 70. Anaheim, California, United States: Optica Publishing Group, 2002, p. ThR5.

[28] R. Sommer, R. M. Fortenberry, B. Flintham, and P. C. Johnson, "Multiple filter functions integrated into multi-port GFF components," in *Optical Fiber Communication Conference and Exposition and The National Fiber Optic Engineers Conference*, ser. OSA Technical Digest Series. Anaheim, California, United States: Optica Publishing Group, 2007, p. JWA24.

[29] A. Saleh, R. Jopson, J. Evankow, and J. Aspell, "Modeling of gain in erbium-doped fiber amplifiers," *IEEE Photonics Technology Letters*, vol. 2, no. 10, pp. 714–717, 1990.

[30] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 7537–7547. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf

[31] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 7462–7473. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf