Advancing Fluorescence Detection and Ranging in Scattering Media with Mixture-of-Experts and Evidential Critics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Attention-based models dominate sequence transduction, yet in medical time-series datasets they often misallocate focus to irrelevant regions while missing critical context. We present EvidenceMoE, a Mixture-of-Experts architecture that assigns experts based on prior physics knowledge and refines their outputs through an Evidential Dirichlet feedback mechanism providing per-expert reliability scores. In our work on fluorescence lifetime-guided cancer surgery, we assigned expert models to relevant time-series segments encoding tumor depth and microenvironment based tumor delineation knowledge from physics (i.e., the radiative transport equation for photon propagation in tissue), rather than learned only from data. Unlike other prior models that address either depth (e.g., Fluorescence LiDAR) or fluorescence decay (fluorescence lifetime or FLI for drug–target binding), EvidenceMoE jointly captures both within a unified framework, achieving errors as low as 0.030 NRMSE for depth and 0.074 NRMSE for FLI on simulated and experimental datasets, closely matching ground-truth measurements.

1 Introduction

Fluorescence-guided surgery (FGS) enhances intraoperative tumor visualization, enabling surgeons to achieve more precise resections while sparing healthy tissue (1; 2). Despite its promise, current FGS systems face two fundamental limitations. First, intensity-based imaging alone often fails to delineate tumor boundaries accurately, as fluorescence accumulation can occur passively and generate misleading signals (3; 4). Second, it provides only 2D surface fluorescence intensity map, leaving surgeons without information about tumor depth (tumor location beneath the tissue surface) (5).

A natural solution to the first limitation is fluorescence lifetime imaging (FLI), which leverages fluorescence decay information rather then intensity alone to robustly distinguish specific molecular interactions from nonspecific probe accumulation (6; 7; 8). However, FLI being indirect computational imaging method requires time-resolved fluorescence images acquisition following solving ill-posed inverse problems, resulting in computationally expensive pipelines that limit it's clinical translation (9; 10). To address the second limitation, fluorescence detection and ranging (FLiDAR) has been proposed for depth estimation (11; 12). In ideal conditions, where scattering and absorption are absent or known, FLiDAR can localize fluorescence probes with high accuracy. Yet, in real surgical environments, biological tissue introduces highly variable scattering and absorption, severely degrading performance unless prior knowledge of optical properties is provided (13; 5; 11; 8).

In this work, we demonstrate that time-resolved fluorescence sequences alone contain sufficient information to jointly infer both depth and lifetime without requiring any auxiliary measurements of tissue optical properties. Our approach builds on the physics of photon transport in scattering media (biological tissues), where different temporal regions of the fluorescence decay sequence encode distinct aspects of the underlying tissue–fluorophore interaction (14; 12). By learning to

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

attend selectively to these physics-relevant regions, our model achieves accurate depth localization and robust lifetime estimation in complex tissue environments, paving the way toward clinically viable FGS systems with depth-resolved molecular contrast.

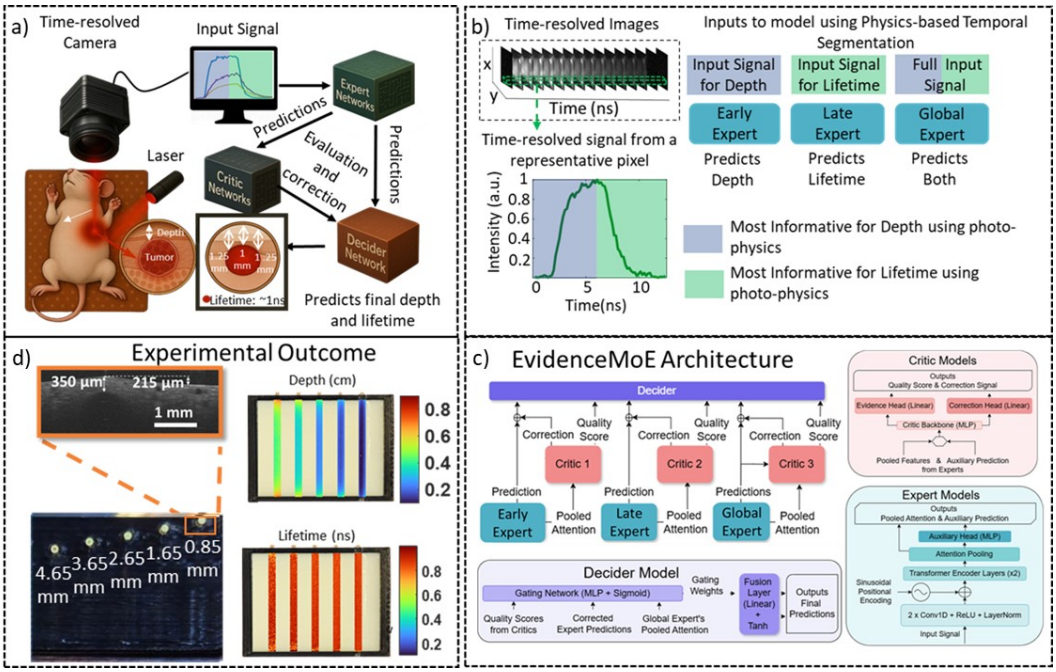

Figure 1: **End-to-end EvidenceMoE Workflow for FLiDAR-based Tumor Lifetime and Depth Estimation.** (a) A laser excites the fluorescent tumor target, and the resulting photons are captured by a time-resolved camera to generate temporal decay images. (b) These images are processed by physics-guided experts that specialize on early photon arrivals for depth, late decay dynamics for lifetime, and the full signal for global context. (c) Within the EvidenceMoE architecture, evidential critics assess expert reliability and provide corrections, while the decider fuses these pathways into final robust estimates of tumor depth and fluorescence lifetime. (d) The framework is validated on a tissue-mimicking phantom with inclusions at varying depths.

Herein, we introduce a Physics-guided Mixture-of-Experts (MoE) architecture (15; 16), where expert roles are pre-defined using knowledge from photo-physics rather than learned from data. Each expert generates a prediction, which is evaluated and corrected by an evidence-based critic serving as an internal quality assessor. This reduces expert domination (17) and allows a Decider Network to intelligently fuse these reviewed contributions into a final, reliable estimate of depth and lifetime (Figure 1). Note that "physics-guided" here is distinct from physics-informed approaches.

The principal contributions of this work are:

1. A direct framework for jointly estimating fluorescence depth and lifetime from raw time-resolved fluorescence image sequences without requiring tissue optical properties.
2. An uncertainty-aware inference scheme using evidential Dirichlet correction to quantify model reliability.
3. A Mixture-of-Experts architecture leveraging the knowledge from photon transport equation to attend to temporal segments, with adaptive fusion for FLiDAR inference.

2 Model Architecture

Our proposed EvidenceMoE framework is designed to jointly estimate the fluorescence probe depth and fluorescence lifetime directly from raw time-resolved fluorescence image sequences. The framework targets biological samples in which the fluorescent inclusion is located at depths of 1–5 mm beneath the tissue surface. Subsequent subsections will provide descriptions of the specialized expert networks, the Evidence-Based Dirichlet Critics (EDCs), and the Decider network.

2.1 High-Level Overview and Signal Flow

EvidenceMoE model, illustrated in Figure 1 (c), processes an input time-resolved fluorescence images through a sequence of specialized, interconnected modules. Each input signal is represented as a vector $\mathbf{x} \in \mathbb{R}^L$, where L denotes the total number of discrete time bins capturing the temporal distribution of detected photons for the respective pixel.

Expert pathways The framework utilizes three parallel expert pathways, each tailored to different aspects of the signal: (1) an Early Expert E_e, (2) a Late Expert E_l, and (3) a Global Expert E_g. Each pathway is specialized to extract information from distinct temporal regions of the fluorescence decay and to produce auxiliary predictions.

Critic pathways Each expert is paired with a dedicated critic that evaluates the reliability of the expert's auxiliary prediction, quantifies uncertainty, and generates a corrective residual, to account for the stochasticity inherent in photon transport and signal noise in scattering media.

Decider network The final module adaptively fuses corrected expert predictions, critic-derived quality scores, and global contextual features of entire signal into a single robust estimate of fluorescence depth and lifetime.

2.2 Physics-Guided Mixture-of-Experts: Temporal Specialization

Our proposed MoE architecture leverages the understanding that early-arriving photons in the FLiDAR images are predominantly correlated with the target depth, whereas the decay characteristics of the later portion of the time-resolved images are more significantly influenced by the material's intrinsic fluorescence lifetime. Consequently, each expert network processes a designated portion of the input (cf. Figure 1(b)).

Architecture details Each expert shares a common backbone, as detailed in Appendix C: 1D convolutional layers to capture local temporal features, a transformer encoder to model long-range dependencies, an attention pooling mechanism to extract a compact feature vector $\phi_k \in \mathbb{R}^H$, and multi-layer perceptron (MLP) head generates the final prediction $y_{\text{aux},k}$.

2.3 Evidence-Based Dirichlet Critics (EDCs)

Each EDC is paired with a specific expert and receives a rich input representation that includes both the expert's pooled internal feature vector and its auxiliary prediction. These are concatenated into a single input vector, allowing the critic to consider both the latent activations and initial predictions when evaluating reliability.

Architecture details. Given an expert's pooled features and auxiliary prediction, concatenated as $z_k = \text{concat}(\phi_k, y_{\text{aux},k})$ where $z_k \in \mathbb{R}^{H+D_k}$, the critic applies a shared MLP backbone followed by two heads. The first, an *evidence head*, that estimates the parameters $\{\alpha_{k,d}, \beta_{k,d}\}$ of independent Beta distributions, modeling the uncertainty associated with each output dimension. The second, a *correction head*, produces a residual vector $\Delta_k \in \mathbb{R}^{D_k}$, which is used to refine the expert's auxiliary output.

Critic score. In the absence of direct supervision, training the evidence head requires a proxy for an expert's prediction quality. For each dimension d, the target score is given by

$$q_{k,d}^{\text{target}} = \frac{1}{1 + \kappa \cdot \text{MAE}_{k,d}} \quad \text{with} \quad \text{MAE}_{k,d} = \frac{1}{N} \sum_{i=1}^{N} \left| y_{aux,k,d}^{(i)} - y_{true,k',d}^{(i)} \right|, \tag{1}$$

where N represents the batch size and κ is a scaling hyperparameter controlling sensitivity. This formulation enables the critic to learn a continuous notion of reliability directly grounded in the expert's observed performance.

3 Decider Network: Adaptive and Informed Fusion

The Decider architecture (F, parameterized by θ_F), illustrated in Figure 1 (c), consists of a gating mechanism and a fusion layer.

Architecture details. The network employs a gating mechanism that assigns dynamic weights to the early, late, and global experts through a two-layer MLP with sigmoid activation: $w = \sigma(W_{g2} \cdot \text{ReLU}(W_{g1} u_{\text{gate}} + b_{g1}) + b_{g2})$, where $(W_{g1}, b_{g1}, W_{g2}, b_{g2})$ are learnable parameters within F that determine the relative influence of each expert branch. The gated expert contributions are computed as $y_{\text{gated}} = [y_{\text{aux},e} \cdot w_e, y_{\text{aux},l} \cdot w_l, y_{\text{aux},g,d} \cdot w_g, y_{\text{aux},g,l} \cdot w_g] \in \mathbb{R}^{D_{\text{experts}}}$ and concatenated with the decider feature ϕ_g. This combined vector is passed through a linear fusion layer to produce the raw 2D output: $y_{raw} = H_{\text{fus}}(\text{concat}(y_{\text{gated}}, \phi_g); \theta_F)$. Finally, the raw output y_{raw} is transformed via a $\tanh$ activation to yield the final predictions for depth y_d and fluorescence lifetime y_l.

4 Empirical Studies

This section details empirical validation of the EvidenceMoE framework, highlighting the critical role of physics-guided temporal segmentation for accurate depth and lifetime measurement while managing photon stochasticity using EDCs. We evaluate the framework using both simulated datasets and experimental phantom data. For simulation studies, datasets were partitioned into 80% training, 20% validation, and 100 test samples. Data generation is explained in detail in Appendix D, briefly to generate realistic FLiDAR data, we leveraged Monte Carlo (MC) simulation, a robust methodology for simulating photon transport in scattering biological tissues (18; 19), using the Monte Carlo eXtreme (MCX) tool (19) following established workflows (20; 21; 22).

Performance of the Full EvidenceMoE Framework. In the evaluation of our proposed EvidenceMoE model, we set the hyperparameter $\kappa = 2$ (see Equation 1) while training the EDCs. This parameter influences how prediction errors are translated into the target quality scores that the EDCs learn to predict. Our choice of $\kappa = 2$ was made to ensure a balanced and interpretable relationship between error and quality, where, for example, a 20% prediction error yields around a 70% quality score. As reported in Table 2 (row *Full model* ($\kappa = 2$)), this framework demonstrates strong performance. The accuracy of these estimations is further illustrated in Figure 2, demonstrating lifetime predictions (Figure 2 b) exhibit high precision, closely aligning with the ground truth. While depth predictions (Figure 2 a) also show strong agreement, they display a slightly larger spread compared to lifetime; yet, the maximum depth errors remain small, around 0.07 cm (0.7 mm), indicating a high degree of accuracy. Notably, the predicted depth quality scores (Figure 2 c) average around 95%, while the lifetime quality scores (Figure 2 d) average around 96.5%. This observation suggests the slightly higher quality scores for lifetime correspond with its marginally better predictive precision compared to depth, underscoring the utility of the EDC-generated quality scores in reflecting prediction reliability. Beyond overall performance metrics, we conduct targeted evaluations and ablation studies (see subsection D.3) to validate that each architectural component performs its intended function according to our design principles.

Leveraging Prior Knowledge from Physics for Expert Assignment The physics-guided design assigns each expert to distinct temporal segments of the FLiDAR signal: the Early expert focuses on initial photon arrivals (depth), while the Late Expert targets later decay characteristics (lifetime). Attention maps from each expert, visualized in Figure 3, confirm this specialization: the depth focused expert emphasizes the signal's rising edge, while the fluorescence lifetime focused expert concentrates on the falling edge, validating that each expert captures specific features relevant to its task.

The Challenge of Stochasticity. Accurate parameter estimation from FLiDAR signals in scattering media is complicated by the inherent stochasticity of photon transport, which introduces significant noise and variability. To explicitly account for the varying uncertainty in the expert output, heteroscedastic loss was also used in isolated experts, our results (Table 2, *Heteroscedastic experts only*) show that this approach converges more slowly, requiring ∼500 epochs versus 70 for EvidenceMoE, and yields lower performance on depth estimation (D.NRMSE: 0.036 vs. 0.030). This demonstrates that, without dynamic expert assessment and fusion, basic uncertainty prediction alone is insufficient to address the intricacies of FLiDAR signal analysis, motivating our more comprehensive EvidenceMoE architecture.

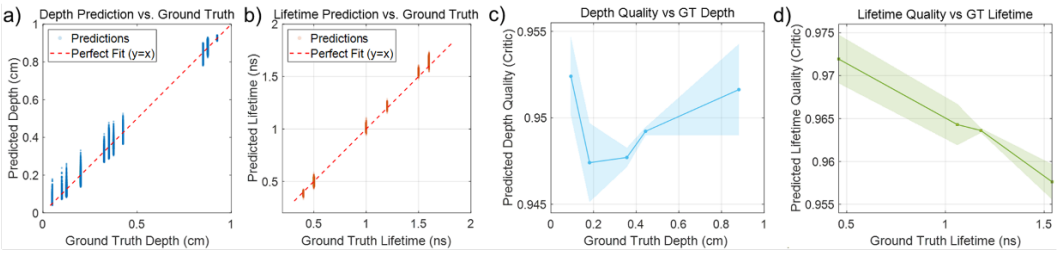

Figure 2: Performance results of EvidenceMoE model

Experimental Validation. To bridge the gap between simulation and real-world application, we conducted experimental validation using a tissue-mimicking phantom The phantom was fabricated

from 1% agar to replicate tissue structure, with intralipid incorporated to induce scattering properties analogous to biological tissue. Five capillary tubes, filled with a 10 μM concentration of the near-infrared fluorophore Alexa Fluor 700 (AF700), were embedded within the phantom. The tubes were arranged in a stepwise depth configuration, with each subsequent tube placed $1\ mm$ deeper than the last, creating a controlled ground truth for depth assessment as shown in Figure 1 (d). Our model achieved depth estimation errors ≤ 0.09 cm and lifetime errors ≤ 0.13 ns across all inclusions (Table 1). Using optical coherence tomography (OCT) as a validation benchmark for depth, we confirmed that our method could localize all five buried inclusions, whereas OCT itself visualized only the shallowest target within it's limit of 2mm resolution in highly scattering media.

Tube Number	Predicted Depth	Actual Depth	Lifetime*
1	0.11 cm $\pm$ 0.05	0.09 cm	0.87 ns $\pm$ 0.03
2	0.18 cm $\pm$ 0.06	0.17 cm	0.87 ns $\pm$ 0.02
3	0.26 cm $\pm$ 0.08	0.27 cm	0.87 ns $\pm$ 0.02
4	0.37 cm $\pm$ 0.09	0.37 cm	0.88 ns $\pm$ 0.03
5	0.53 cm $\pm$ 0.11	0.47 cm	0.88 ns $\pm$ 0.06

*While the manufacturer's reported lifetime for Alexa Fluor 700 dye is ≈ 1 ns in solution (23), in scattering media, with a traditional least squares solver estimating it at 0.9 ns.

Table 1: Experimental validation results of EvidenceMoE framework on tissue-mimicking phantom with stepwise depth configuration.

Implementation for Real-Time Inference. To evaluate its clinical feasibility, the EvidenceMoE framework was benchmarked on an NVIDIA H100 GPU with 80GB of HBM3 memory. The model processed a full 500×250 pixel frame in 1.375 seconds. Crucially, practical applications like tumor imaging often focus on a smaller region of interest (approximately 1/4 of the total field of view), which reduces latency to a sub-second timeframe suitable for near real-time feedback. This performance confirms a viable pathway for deploying the framework on dedicated hardware to support low-latency, uncertainty-aware surgical guidance.

5 Conclusion and Discussion

In this work, we introduced EvidenceMoE, a Mixture-of-Experts architecture that integrates physics-based prior knowledge with sequential data learning for time-resolved fluorescence imaging. Through ablation studies and validation on both Monte Carlo simulations and tissue-mimicking phantom experiments, we showed that EvidenceMoE achieves accurate depth and lifetime estimation. Notably, it could locate fluorescence probes at depths up to 5 mm in experimental scattering media, surpassing the $\approx 2mm$ limit of established optical methods such as Optical Coherence Tomography. These results highlight the clinical translation potential of EvidenceMoE for fluorescence-guided surgery, where precise depth and molecular information are essential for reliable tumor delineation and resection. In a nutshell, EvidenceMoE combines photophysics knowledge with Mixture-of-Experts learning to deliver accurate depth and lifetime estimates in time-resolved fluorescence imaging, enabling clinically translatable guidance. Moving forward, we will extend our study to *ex vivo* and *in vivo* samples to capture the complexity of biological tissues, and pursue embedded hardware implementations to satisfy real-time surgical constraints. This work represents a step toward depth-resolved, lifetime-enabled fluorescence imaging in the operating room, bridging physics, machine learning, and clinical needs.

References

[1] H. L. Stewart and D. J. Birch, "Fluorescence guided surgery," *Methods and Applications in Fluorescence*, vol. 9, no. 4, p. 042002, 2021.

[2] P. A. Sutton, M. A. van Dam, R. A. Cahill, S. Mieog, K. Polom, A. L. Vahrmeijer, and J. van der Vorst, "Fluorescence-guided surgery: comprehensive review," *BJS open*, vol. 7, no. 3, p. zrad049, 2023.

[3] J. T. Liu and N. Sanai, "Trends and challenges for the clinical adoption of fluorescence-guided surgery," *Journal of Nuclear Medicine*, vol. 60, no. 6, pp. 756–757, 2019.

[4] J. S. D. Mieog, F. B. Achterberg, A. Zlitni, M. Hutteman, J. Burggraaf, R.-J. Swijnenburg, S. Gioux, and A. L. Vahrmeijer, "Fundamentals and developments in fluorescence-guided cancer surgery," *Nature reviews Clinical oncology*, vol. 19, no. 1, pp. 9–22, 2022.

[5] J. T. Smith, E. Aguénounon, S. Gioux, and X. Intes, "Macroscopic fluorescence lifetime topography enhanced via spatial frequency domain imaging," *Optics letters*, vol. 45, no. 15, pp. 4232–4235, 2020.

[6] R. I. Dmitriev, X. Intes, and M. M. Barroso, "Luminescence lifetime imaging of three-dimensional biological objects," *Journal of Cell Science*, vol. 134, no. 9, pp. 1–17, 2021.

[7] A. Verma, V. Pandey, C. Sherry, T. Humphrey, C. James, K. Matteson, J. T. Smith, A. Rudkouskaya, X. Intes, and M. Barroso, "Fluorescence lifetime imaging for quantification of targeted drug delivery in varying tumor microenvironments," *Advanced Science*, vol. 12, no. 3, p. 2403253, 2025.

[8] A. F. Petusseau, S. S. Streeter, A. Ulku, Y. Feng, K. S. Samkoe, C. Bruschini, E. Charbon, B. W. Pogue, and P. Bruza, "Subsurface fluorescence time-of-flight imaging using a large-format single-photon avalanche diode sensor for tumor depth assessment," *Journal of Biomedical Optics*, vol. 29, no. 1, pp. 016 004–016 004, 2024.

[9] W. Becker, "Fluorescence lifetime imaging–techniques and applications," *Journal of microscopy*, vol. 247, no. 2, pp. 119–136, 2012.

[10] J. T. Smith, R. Yao, N. Sinsuebphon, A. Rudkouskaya, N. Un, J. Mazurkiewicz, M. Barroso, P. Yan, and X. Intes, "Fast fit-free analysis of fluorescence lifetime imaging via deep learning," *Proceedings of the national academy of sciences*, vol. 116, no. 48, pp. 24 019–24 030, 2019.

[11] S.-H. Han, S. Farshchi-Heydari, and D. J. Hall, "Analytical method for the fast time-domain reconstruction of fluorescent inclusions in vitro and in vivo," *Biophysical journal*, vol. 98, no. 2, pp. 350–357, 2010.

[12] J. Wu, L. Perelman, R. R. Dasari, and M. S. Feld, "Fluorescence tomographic imaging in turbid media using early-arriving photons and laplace transforms," *Proceedings of the National Academy of Sciences*, vol. 94, no. 16, pp. 8783–8788, 1997.

[13] P. A. Valdés, F. Leblond, V. L. Jacobs, B. C. Wilson, K. D. Paulsen, and D. W. Roberts, "Quantitative, spectrally-resolved intraoperative fluorescence imaging," *Scientific reports*, vol. 2, no. 1, p. 798, 2012.

[14] L. Zhao, H. Yang, W. Cong, G. Wang, and X. Intes, "L p regularization for early gate fluorescence molecular tomography," *Optics letters*, vol. 39, no. 14, pp. 4156–4159, 2014.

[15] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[16] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: a literature survey," *Artificial Intelligence Review*, vol. 42, pp. 275–293, 2014.

[17] Z. Chi, L. Dong, S. Huang, D. Dai, S. Ma, B. Patra, S. Singhal, P. Bajaj, X. Song, X.-L. Mao *et al.*, "On the representation collapse of sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 600–34 613, 2022.

[18] J. Chen, *Optical tomography in small animals with time-resolved Monte Carlo methods*. Rensselaer Polytechnic Institute, 2012.

[19] Q. Fang and D. A. Boas, "Monte carlo simulation of photon migration in 3d turbid media accelerated by graphics processing units," *Optics express*, vol. 17, no. 22, pp. 20 178–20 190, 2009.

[20] N. I. Nizam, V. Pandey, I. Erbas, J. T. Smith, and X. Intes, "A novel technique for fluorescence lifetime tomography," *bioRxiv*, 2024.

[21] N. I. Nizam, I. Erbas, V. Pandey, and X. Intes, "Monte-carlo based data generator for fluorescence lifetime applications," in *Optical Tomography and Spectroscopy*. Optica Publishing Group, 2024, pp. JS4A–27.

[22] V. Pandey, I. Erbas, X. Michalet, A. Ulku, C. Bruschini, E. Charbon, M. Barroso, and X. Intes, "Deep learning-based temporal deconvolution<? pag\break?> for photon time-of-flight distribution retrieval," *Optics letters*, vol. 49, no. 22, pp. 6457–6460, 2024.

[23] Thermo Fisher Scientific, "Fluorescence quantum yields (QY) and lifetimes (τ) for Alexa Fluor dyes-Table 1.5," https://www.thermofisher.com/us/en/home/references/molecular-probes-the-handbook/tables/fluorescence-quantum-yields-and-lifetimes-for-alexa-fluor-dyes.html, accessed August 29, 2025.

[24] N. Yuan, V. Pandey, A. Verma, J. C. Williams, X. Intes, and M. Barroso, "Antibody-target binding quantification in living tumors using macroscopy fluorescence lifetime forster resonance energy transfer imaging (mfli fret)," in *Visualizing and Quantifying Drug Distribution in Tissue VIII*, vol. 12821. SPIE, 2024, pp. 17–20.

[25] J. McGinty, N. P. Galletly, C. Dunsby, I. Munro, D. S. Elson, J. Requejo-Isidro, P. Cohen, R. Ahmad, A. Forsyth, A. V. Thillainayagam *et al.*, "Wide-field fluorescence lifetime imaging of cancer," *Biomedical optics express*, vol. 1, no. 2, pp. 627–640, 2010.

[26] M. R. Karim, M. N. Reza, H. Jin, M. A. Haque, K.-H. Lee, J. Sung, and S.-O. Chung, "Application of lidar sensors for crop and working environment recognition in agriculture: A review," *Remote Sensing*, vol. 16, no. 24, p. 4623, 2024.

[27] L. A. Silva, F. S. Ferreira, G. S. Oliveira, A. L. Moura, R. A. de Oliveira, and A. S. Reyna, "Exploring disordered light transport in scattering media to optimize random lasers," *The Journal of Physical Chemistry C*, vol. 128, no. 12, pp. 5321–5329, 2024.

[28] J. Ma, S. Zhuo, L. Qiu, Y. Gao, Y. Wu, M. Zhong, R. Bai, M. Sun, and P. Y. Chiang, "A review of tof-based lidar," *Journal of Semiconductors*, vol. 45, no. 10, p. 101201, 2024.

[29] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," *Proceedings of the Royal Society A*, vol. 478, no. 2266, p. 20210068, 2022.

[30] T. M. Mitchell, "The need for biases in learning generalizations," 1980.

[31] X. Xu, L. Kong, H. Shuai, L. Pan, Z. Liu, and Q. Liu, "Limoe: Mixture of lidar representation learners from automotive scenes," *arXiv preprint arXiv:2501.04004*, 2025.

[32] D. Katkoria, J. Sreevalsan-Nair, M. Sati, and S. Karunakaran, "Me-odal: Mixture-of-experts ensemble of cnn models for 3d object detection from automotive lidar point clouds," in *International Conference on Deep Learning Theory and Applications*. Springer, 2024, pp. 279–300.

[33] S. Mu and S. Lin, "A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications," *arXiv preprint arXiv:2503.07137*, 2025.

[34] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[36] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International conference on machine learning*. PMLR, 2015, pp. 1613–1622.

[37] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[38] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on bayesian neural networks—a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022.

[39] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[40] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[41] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[42] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in neural information processing systems*, vol. 33, pp. 14 927–14 937, 2020.

[43] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[44] A. C. Ulku, C. Bruschini, I. M. Antolović, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512× 512 spad image sensor with integrated gating for widefield flim," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, 2018.

[45] J. T. Smith, A. Rudkouskaya, S. Gao, J. M. Gupta, A. Ulku, C. Bruschini, E. Charbon, S. Weiss, M. Barroso, X. Intes *et al.*, "In vitro and in vivo nir fluorescence lifetime imaging with a time-gated spad camera," *Optica*, vol. 9, no. 5, pp. 532–544, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the problem of FLiDAR signal analysis in scattering media. These claims match the experimental results provided later in the paper, in the empirical results section, specifically the ablation studies sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[Yes]**

   Justification: Future work needed for ex vivo/in vivo validation; current validation limited to simulated data and phantom experiments.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The mathematical equations provided describe the model architecture, loss functions, and evaluation metrics rather than theoretical propositions.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 provides experimental setup; Sections 8-9 detail model architecture and training methodology; Monte Carlo data generation fully described; GitHub implementation provided.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation of EvidenceMoE is publicly available on GitHub. (https://anonymous.4open.science/r/EvidenceMoE-4728/)

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail our experimental setup in Section 3.4 and Appendix, including data splits , optimizer, learning rates, batch sizes, epoch schedules, hyperparameter selection criteria, and evaluation protocols for both training and testing .

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report each result accompanied by its standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: NVIDIA H100 GPU with 80GB HBM3 memory; processing time 1.375 seconds for 500×250 pixel frame (Section 4).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 discusses clinical translation potential for fluorescence-guided surgery; addresses improved tumor delineation while acknowledging need for *in vivo* and *ex vivo* validation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The model and data, as described, do not fall into the category of high-risk assets like large pretrained language models or image generators, nor does it involve scraped datasets that might pose inherent safety or privacy risks requiring specific release safeguards. The primary concern would be the scientific validity and interpretation of the model's outputs, which is addressed through the general research methodology and validation.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We credit the original creators of all external assets and provide full reference details for each.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We accompany our new assets with full documentation: Appendix details the code modules and usage instructions for the EvidenceMoE framework, while the simulated dataset format, experimental data parameters, generation parameters, and data splits were specified; additionally, a structured README in the released repository covers installation and licensing.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research does not involve crowdsourcing experiments or direct research with human subjects for data collection or experimentation. The FLiDAR data used for validation was generated through Monte Carlo simulations, as described in Section 4.1.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research focuses on the EvidenceMoE framework, a novel deep learning architecture for FLiDAR signal analysis. This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Principles of Time-Resolved FLiDAR, Photon Scattering, and Fluorescence Lifetime

Fluorescence LiDAR (FLiDAR) is a specialized technique that extends LiDAR capabilities by analyzing light-induced fluorescence in scattering media, where LiDAR performance degrades by photon scattering. When photons from the laser pulse are absorbed by specific molecules (fluorophores) within the target material, these molecules transition to an excited electronic state. They subsequently relax to their ground state, partly by emitting photons, a process known as *fluorescence* (24; 7; 6). Time-resolved camera systems are designed to detect and temporally resolve this fluorescence emission. The characteristic rate at which the fluorescence intensity declines after excitation follows an exponential curve, and is termed the *fluorescence lifetime* (τ). This lifetime is an intrinsic property of the fluorophore and is sensitive to its local chemical and structural environment (7; 24). Consequently, fluorescence lifetime serves as a useful contrast parameter with applications in biomedical diagnostics to differentiate tissue states (such as healthy or cancerous tissue (25)) or in environmental science for vegetation analysis (26).

Accurately retrieving depth and fluorescence lifetime using FLiDAR in scattering media presents distinct challenges. Firstly, the laser pulse for excitation and basic ranging experiences scattering en route to the target (27; 28). Secondly, photons that are emitted by the target fluorophores will also propagate through the scattering medium to reach the time-resolved camera. During this transit, these fluorescence photons are subject to similar scattering processes. This additional scattering means that the temporal profile of the fluorescence decay, as recorded by the detector, is not solely governed by the intrinsic fluorescence lifetime of the molecule. Instead, the observed decay profile becomes a convolution of the intrinsic exponential decay with the temporal dispersion effects introduced by photon scattering within the medium.

Therefore, the signal acquired by FLiDAR in such conditions is a composite, reflecting both the target's range and its fluorescence decay properties, each distorted by scattering. Disentangling these convolved effects to estimate the true depth and the intrinsic fluorescence lifetime accurately requires sophisticated signal processing. Despite these complexities, specific temporal characteristics of the signal provide differential information:

- **Early photons:** Early-arriving photons to the time-resolved camera, having statistically undergone fewer scattering events, correlate more strongly with the shortest path length to the target, thus primarily encoding depth information (14).
- **Late photons:** The decay characteristics of the later portion of the signal are more significantly influenced by the fluorescence lifetime.

B Background and Related Work

Our proposed framework integrates concepts from Mixture-of-Experts, uncertainty quantification, and evidential reasoning, all applied to the specific challenges of interpreting time-resolved FLiDAR images from scattering media. We briefly review these foundational areas and position our contributions within their context.

B.1 Physics-Guided Inductive Biases for FLiDAR Temporal Segmentation

With a finite training set, a model's ability to generalize to new inputs depends on the preferences or assumptions it encodes, its inductive biases, which narrow the set of solutions consistent with the observed data(29). These inductive biases are the inherent assumptions within a model architecture that guide its learning process and ability to generalize from finite data (30). For complex data such as FLiDAR images from scattering media, where distinct temporal segments like early and late photon arrivals convey different physical information (14), the choice of inductive bias substantially impacts learning outcomes. An effective inductive bias guides the learning algorithm by incorporating domain knowledge about the signal, such as by structuring the model to process these physically meaningful segments differentially. This targeted approach constrains the hypothesis space, directing the model to focus on relevant features and relationships that align with known physical principles. Consequently, such guidance can lead to more efficient model training, including faster convergence and improved sample efficiency, as the model is steered away from learning spurious correlations, such as noise, ultimately enhancing its ability to interpret the underlying signal characteristics.

B.2 Mixture-of-Experts (MoE)

Mixture-of-Experts (MoE) models (15; 16) implement a divide-and-conquer strategy by routing each input to a small subset of expert subnetworks via a gating network. In MoE, the gate selects experts through sparse routing to balance computation across the model. This paradigm has been applied to both multi-representation sensing and temporal forecasting. In LiMoE, features from range images, voxels, and point clouds are fused through an MoE layer for LiDAR perception in air (31). ME-ODAL applies MoE routing to 3D object detection in point clouds (32). Across these applications, expert specialization, where each expert learns a distinct function or handles a particular data subset, underpins model behavior (16).

Existing temporal MoE methods process each sequence holistically, relying on auxiliary objectives (e.g., load balancing) to promote expert diversity and expecting the gate to infer segment relevance from raw inputs (33). Such approaches do not exploit known signal physics. Our Physics-guided MoE instead assigns one expert to each of the predefined temporal segments, reflecting the physics of photon transport. This segment-based assignment embeds an inductive bias: each expert models the dynamics specific to its interval.

B.3 Limitations of Standard Uncertainty Quantification in Deep Learning

The challenging nature of FLiDAR-based parameter estimation in scattering media, stemming from distorted and convolved signals, elevates the importance of providing reliable uncertainty estimates alongside deep learning predictions. Common approaches of uncertainty estimation include Monte Carlo Dropout (34), Deep Ensembles (35), and Bayesian Neural Networks (BNNs) (36). While effective, these methods can suffer from limitations (37): MC Dropout's estimates may not always be well-calibrated; Deep Ensembles incur significant computational overhead during training and inference (35); and approximate inference in BNNs can be complex to implement and tune (38). Furthermore, these methods primarily capture aleatoric (data) uncertainty or epistemic (model) uncertainty (39; 40), but do not explicitly model the quality of a specific prediction based on input evidence and do not provide correction signals based on the uncertainty.

B.4 Evidential Deep Learning

Evidential Deep Learning (EDL) trains a deterministic network, for example, a stack of standard MLP layers, to output the parameters of a higher-order evidential distribution, rather than placing priors on weights as in BNNs with Bayesian layers (41; 42). Training of EDL models generally relies on loss functions that combine a negative log-likelihood term with regularizers (e.g. KL divergence to a noninformative prior or evidence penalties). Our Evidence-Based Dirichlet Critic (EDC) module acts as an evaluator for each expert. Inspired by actor-critic frameworks in reinforcement learning (43), though adapted for supervised regression, the critic assesses the actor's (expert's) output.

C Detailed Model Architecture for Reproducibility

C.1 Physics-Guided Mixture-of-Experts

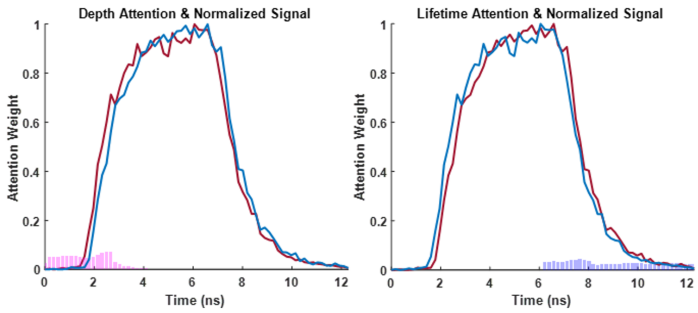

Figure 3: Visualization of Pooled Attention Weights Over Time for Both Depth and Lifetime Experts, Computed Over a Batch Size of 512.

Each expert (E_k) $k \in \{e, l, g\}$, parameterized by $\theta_E = \{\theta_{E_e}, \theta_{E_l}, \theta_{E_g}\}$, shares a common internal architecture optimized for feature extraction, consisting of three main stages:

1. **Hybrid CNN-Transformer Encoder:** The input segment $x_k \in^{L_k}$ is initially processed by a sequence of 1D convolutional layers (kernel size K, with residual connections, Layer Normalization, and GELU activation). This allows the network to explicitly capture localized characteristics of the time resolved signal, such as:
   - The sharpness and timing of the initial rising edge related to photon arrival time (depth).
   - Local decay rates or changes in slope within short segments of the fluorescence tail (lifetime).
   - The presence and shape of small secondary peaks or instrumental response function artifacts within the window.

   This extraction of local waveform motifs precedes the subsequent components. The output of the CNN block is then augmented with sinusoidal Positional Encoding before being passed to a multi-layer Transformer encoder. The Transformer utilizes self-attention mechanisms to model long-range temporal dependencies and contextual relationships across the entire input segment $\mathbf{x}_k$. The encoder outputs a sequence of refined feature vectors $h_k^{enc} \in^{L_k \times H}$, where H is the hidden dimension size.
2. **Pooling Layer (f_k):** To obtain a fixed-size representation from the variable-length output sequence of the encoder, an Attention Pooling mechanism is employed. This layer learns attention weights over the time steps of h_k^{enc} and computes a weighted sum, producing a single feature vector $\phi_k \in^H$. This vector summarizes the relevant information from the expert's input segment.
$$\phi_k = \text{AttentionPool}(h_k^{enc}; \theta_{E_k}) \tag{2}$$
3. **Auxiliary Prediction Head (h_k):** A small Multi-Layer Perceptron (MLP) with non-linear activation (ReLU) maps the pooled feature vector ϕ_k to the expert's specific auxiliary prediction $y_{aux,k}$ (dimension $D_k = 1$ for early/late, $D_k = 2$ for global).
$$y_{aux,k} = h_k(\phi_k; \theta_{E_k}) \tag{3}$$

Role of Global Expert Features: Beyond its auxiliary prediction $y_{aux,g}$, the pooled feature vector ϕ_g from the global expert serves a dual purpose. It acts as the primary source of global context for the downstream fusion mechanism within the Final Decider Head, informing the gating decisions.

C.2 Evidence-Based Dirichlet Critics (EDC)

To assess the reliability of each expert's auxiliary prediction and provide a mechanism for refinement, we employ a dedicated critic network C_k for each expert $k \in \{e, l, g\}$, parameterized by $\theta_C = \{\theta_{C_e}, \theta_{C_l}, \theta_{C_g}\}$. We adopt the Evidential Deep Learning (EDL) framework (42) to enable the critics to quantify uncertainty in their quality assessments.

Critic Input Features: To enable an informed reliability assessment by the critic C_k, we provide it with an input representation z_k. This input concatenates the pooled feature vector ϕ_k from the expert's encoder (Equation 2) with the expert's auxiliary prediction $y_{aux,k}$ (Equation 3):

$$z_k = \text{concat}(\phi_k, y_{aux,k}) \tag{4}$$

Critic Architecture: Each critic C_k utilizes an identical architecture comprising a shared MLP backbone (b_k) followed by two separate linear heads:

1. **Shared Backbone (b_k):** The concatenated input z_k is processed by a shared MLP backbone, b_k. A relatively shallow architecture (two layers) with moderate hidden dimensions [32, 16] is employed to keep the critic computationally lightweight while providing sufficient capacity to extract relevant features from the input z_k. The backbone outputs a shared latent feature representation $h_k \in^{16}$.
$$h_k = b_k(z_k; \theta_{C_k}) \tag{5}$$
2. **Evidence Head (evi_k):** A dedicated linear layer, evi_k, maps the shared features h_k to the raw evidence outputs $e_k \in^{2 \times D_k}$ (one positive and one negative evidence value per output dimension D_k of the corresponding expert).
$$e_k = evi_k(h_k; \theta_{C_k}) \tag{6}$$

   To ensure the parameters of the resulting Beta/Dirichlet distribution are strictly positive and greater than one (required for a well-defined distribution and stable calculation of

variance and KL divergence terms), the raw evidence is transformed using the $\mathrm{softplus}(x) = \log(1 + e^x)$ function followed by adding 1:

$$\alpha_{k,d} = \mathrm{softplus}(e_{k,pos,d}) + 1 \tag{7}$$

$$\beta_{k,d} = \mathrm{softplus}(e_{k,neg,d}) + 1 \tag{8}$$

For each dimension $d = 1..D_k$, where $e_{k,pos}$ and $e_{k,neg}$ are the corresponding slices of e_k. The resulting $\alpha_{k,d}, \beta_{k,d} > 1$ parameterize D_k independent Beta distributions. The total evidence $S_{k,d} = \alpha_{k,d} + \beta_{k,d}$ represents the precision or concentration parameter of the distribution, quantifying the amount of evidence gathered from the data supporting the quality prediction; a higher $S_{k,d}$ indicates greater confidence (lower variance) in the quality estimate.

3. **Correction Head ($corr_k$):** A separate linear layer maps h_k to the correction signal $\Delta_k \in^{D_k}$.

$$\Delta_k = corr_k(h_k; \theta_{C_k}) \tag{9}$$

The critic's forward pass thus yields $(\alpha_k, \beta_k, \Delta_k)$. The mean of the predicted Beta distribution serves as the point estimate for quality score q_k:

$$q_{k,d} = \frac{\alpha_{k,d}}{\alpha_{k,d} + \beta_{k,d}} \tag{10}$$

These mean quality scores (specifically q_e, q_l, and the two components of q_g, forming $q_{full} \in^4$) are utilized by the downstream decider head.

C.3 Decider Network: Adaptive and Informed Fusion

To synthesize the multiple, quality-assessed predictions from the expert pathways into a unified estimation of depth and lifetime, the EvidenceMoE architecture incorporates a Decider network. The Decider architecture F, illustrated in Figure 1 (c) consists of a gating mechanism and a fusion layer.

Gating mechanism for dynamic expert weighting A core component of the Decider is a learned gating network G. The gate's decision w is informed by a concatenation of the corrected auxiliary predictions, $u_{gate} = \mathrm{concat}(y_{aux,k}, \phi_g, q_{full})$, the global decider feature, and the full quality scores q_{full}, a vector containing the mean quality scores derived from the EDCs for all relevant output dimensions of the experts. The gating network G outputs gating weights, w_e, w_l, and w_g for early, late, and global experts respectively.

$$w = \sigma(W_{g2} \cdot \mathrm{ReLU}(W_{g1} u_{\text{gate}} + b_{g1}) + b_{g2}) \tag{11}$$

where $(W_{g1}, b_{g1}, W_{g2}, b_{g2})$ are learnable parameters within F. These weights determine the relative influence of each expert branch.

Fusion and final prediction

The gated expert contributions are concatenated with the decider feature ϕ_g:

$$y_{\text{gated}} = [y_{\text{aux},e} \cdot w_e, y_{\text{aux},l} \cdot w_l, y_{\text{aux},g,d} \cdot w_g, y_{\text{aux},g,l} \cdot w_g] \in \mathbb{R}^{D_{\text{experts}}} \tag{12}$$

This vector of gated expert contributions, y_{gated}, is then concatenated with the ϕ_g and passed through a final fusion layer (H_{fus}), a linear layer, to produce the raw 2D output $y_{raw} \in \mathbb{R}^2$:

$$y_{raw} = H_{fus}(\mathrm{concat}(y_{gated}, \phi_g); \theta_F) \tag{13}$$

Finally, this raw output y_{raw} is transformed via an $\tanh$ activation to yield the final predictions for depth y_d and fluorescence lifetime y_l.

C.4 Decider Head and Fusion Mechanism

The final stage of the model is the Decider Head (F, parameterized by θ_F), which performs a learned, context-aware fusion of the information streams originating from the three expert branches. This head utilizes a gating network (G) that processes the potentially corrected auxiliary predictions ($y_{aux,k}$), the critics' reliability estimates (q_{full}), and global contextual features (ϕ_g) to compute dynamic, input-dependent weights (w) for each expert branch. These weighted expert contributions are then combined with the global context in a subsequent fusion layer (H_{fus}) to generate the final, refined 2D prediction y_d, y_l.

Inputs to the Decider Head: The head receives three primary inputs per sample:

1. **Corrected Auxiliary Predictions (y_{aux}):** This is the set of auxiliary predictions from the experts, adjusted by the correction signals provided by the critics (Equation 9), $y_{aux} = \{y_{aux,e} \in^1, y_{aux,l} \in^1, y_{aux,g} \in^2\}$.

2. **Decider Feature (ϕ_g):** This is the pooled feature vector $\phi_g \in^H$ from the global expert's encoder (Equation 2).

3. **Full Quality Scores (q_{full}):** A vector containing the mean quality estimates derived from the Evidence Critics for all four quality dimensions: $q_{full} = [q_e, q_l, q_{g,d}, q_{g,l}] \in^4$ (Equation 10).

Gating Network (G): The core of the fusion mechanism is a learned gating network G, designed to adaptively weight the contributions of the three expert branches based on the specific input characteristics. The gate's decision is informed by the corrected predictions y_{aux}, the global context ϕ_g, and the critics' full quality assessments q_{full}. The combined input u_{gate} is defined as:

$$u_{gate} = \text{concat}(y_{aux,k}, \phi_g, q_{full}) \tag{14}$$

The gating network architecture consists of a two-layer MLP with a ReLU activation after the first layer and a final Sigmoid activation applied after the second layer (mapping to the 3 expert weights):

$$w = \sigma(W_{g2} \cdot \text{ReLU}(W_{g1} u_{gate} + b_{g1}) + b_{g2}) \tag{15}$$

where $\sigma(\cdot)$ is the Sigmoid function, and $(W_{g1}, b_{g1}, W_{g2}, b_{g2})$ are learnable parameters within θ_F. Additionally, gating dropout may apply dropout to w_g during training as a regularization technique.

Applying Gates: The gating weights w modulate the corrected auxiliary predictions $y_{aux,k}$:

$$y_{gated,e} = y_{aux,e} \cdot w_e \tag{16}$$

$$y_{gated,l} = y_{aux,l} \cdot w_l \tag{17}$$

$$y_{gated,g} = y_{aux,g} \cdot w_g \tag{18}$$

Fusion Layer (H_{fus}): The gated expert contributions are concatenated with the decider feature ϕ_g and passed through a final linear fusion layer H_{fus} to produce the raw 2D output y_{raw}:

$$y_{gated\_concat} = [y_{gated,e}, y_{gated,l}, y_{gated,g}] \tag{19}$$

$$u_{fus} = [y_{gated\_concat}, \phi_{decider\_fus}] \tag{20}$$

$$y_{raw} = H_{fus}(u_{fus}; \theta_F) \tag{21}$$

D Training Methodology

D.1 Realistic Data Generation

For an AI model to perform reliably in real-world scenarios, especially in complex fields like medical imaging, it must be trained on data that closely mirrors experimental conditions. This section explains how we generated such realistic training data for our study on Fluorescence LiDAR (FLiDAR) in biological tissue. Our goal was to generate representative data that captures the nuances of how light travels through tissue and how our specific camera system detects it. The realistic data generation allows us to incorporate its experimentally determined Instrument Response Function (IRF), representing the system's intrinsic temporal response to a laser pulse, and its documented noise profile into our data generation process (44; 45).

This experimentally determined IRF represents the system's intrinsic temporal response to a laser pulse, and we incorporate system's documented noise profile into data generation flow. Using directly measured IRF is critical as it ensures that the temporal dispersion characteristics embedded in our simulated data faithfully replicate those of the imaging apparatus. Similarly, integrating the camera's documented noise profile is essential for ensuring that our simulations' stochastic variations and statistical fidelity of photon counts accurately reflect those of the physical sensor.

Furthermore, to realistically model the complex interactions of light within the tissue, we utilized Monte Carlo (MC) simulations, a well-established and robust methodology in biomedical optics for simulating photon transport in scattering media (18; 19). Our simulations were performed using Monte Carlo eXtreme (MCX) (19), adopting a two-stage approach to generate time-resolved fluorescence signals. The MC data generation workflow is further detailed in (20; 21). Moreover, given that the noise profile of the specific time-resolved camera was also characterized and its parameters defined in the camera's documentation, we introduced these system-specific noises (photon (shot) noise, dark counts, afterpulsing, and read noise) into the `MCXLab`-generated signals. By applying these characterized noise sources to the initial MCX simulation outputs and utilizing the noise-free and noise-integrated datasets, our training and validation procedures are grounded in data that realistically reflect the detector's behavior.

D.2 Loss Function

The model parameters, partitioned into expert (θ_E), critic (θ_C), and final head (θ_F) groups, are concurrently trained by minimizing a composite, multi-component loss function $\mathcal{L}_{total}$. This loss function is designed to achieve several simultaneous objectives: optimizing the final prediction accuracy (y_{pred}), encouraging the experts to learn informative intermediate representations (ϕ_k) and generate reasonable auxiliary predictions ($y_{aux,k}$), training the Evidence Critics to produce quality estimates (represented by α_k, β_k) and correction signals (Δ_k) that approximate the experts' scaled residual errors ($\Delta_k \approx (y_{true,k'} - y_{aux,k})/\lambda_{damp}$), and applying specific regularization terms (such as KL divergence for evidence and an evidence-weighted penalty for corrections). The overall loss is a weighted sum of these individual components:

$$\mathcal{L}_{total} = \lambda_{pri}\mathcal{L}_{primary} + \lambda_{aux}\mathcal{L}_{aux} + \lambda_{crit\_q}\mathcal{L}_{quality} + \lambda_{corr}\mathcal{L}_{corr} + \lambda_{pen}\mathcal{L}_{penalty} \tag{22}$$

where the weights $\lambda_{(\cdot)}$ are scalar hyperparameters controlling the relative importance of each term, as defined in the configuration. Each loss component is detailed below.

Primary Loss ($\mathcal{L}_{primary}$): This is the main objective function driving the overall prediction task. It measures the discrepancy between the final fused prediction $y_{pred} \in^2$ and the ground truth $y_{true} \in^2$ (containing depth and lifetime). We use the Mean Absolute Error (L1 loss):

$$\mathcal{L}_{primary} = \text{MAE}(y_{pred}, y_{true}) = \frac{1}{D_{out}} \sum_{j=1}^{D_{out}} \mathbb{E}_{\text{batch}} \left[|y_{pred,j} - y_{true,j}| \right] \tag{23}$$

where $D_{out} = 2$ (depth and lifetime dimensions) and $\mathbb{E}_{\text{batch}}$ denotes the expectation (mean) over the batch samples.

Auxiliary Loss ($\mathcal{L}_{aux}$): To foster expert specialization and enhance training stability, an auxiliary loss term, $\mathcal{L}_{aux}$, provides direct supervision to each individual expert network E_k. This loss computes the L1 distance between the expert's auxiliary prediction ($y_{aux,e}$ for depth, $y_{aux,l}$ for lifetime, $y_{aux,g}$ for both) and the corresponding ground truth targets ($y_{true,k'}$). Enforcing this intermediate accuracy encourages each expert to learn representations directly relevant to its specific task domain (temporal segment and target variable).

$$\mathcal{L}_{aux} = \frac{1}{N_{exp}} \sum_{k \in \{e,l,g\}} \mathbb{E}_{\text{batch}}[\text{L1}(y_{aux,k}, y_{true,k'})] \tag{24}$$

where $N_{exp} = 3$.

Critic Quality Loss ($\mathcal{L}_{quality}$): This loss component is responsible for training the evidence head parameters (α, β) of each Evidence Critic C_k. It employs the Evidential Deep Learning (EDL) formulation specifically to learn a calibrated predictive distribution (parameterized by α_k, β_k) over the quality score associated with the corresponding expert's auxiliary prediction, $y_{aux,k}$. The training objective aims to align the mean of this predicted distribution, $q_k = \alpha_k/(\alpha_k + \beta_k)$, with a target quality score $q_{gt,k}$ derived from the expert's actual error (Equation 26), while simultaneously using the evidential variance and KL divergence terms (Equations 27, 28) to ensure the distribution's concentration (total evidence $S_k = \alpha_k + \beta_k$) reflects the true uncertainty or reliability. The target quality $q_{gt,k}$ is calculated using the $y_{aux,k}$:

$$MAE_{k,d} = \frac{1}{N} \sum_{i=1}^{N} \left| y_{aux,k,d}^{(i)} - y_{true,k',d}^{(i)} \right| \tag{25}$$

$$q_{gt,k,d} = (1 + \kappa \cdot \text{MAE}_{k,d} + \epsilon)^{-1} \tag{26}$$

where κ is the hyperparameter and ϵ prevents division by zero.

The quality loss combines the Evidential Regression loss ($\mathcal{L}_{evi}$) and a KL divergence regularizer ($\mathcal{L}_{KL}$), summed over the four quality dimensions (k, d) corresponding to (early, depth), (late, lifetime),

(global, depth), (global, lifetime):

$$\mathcal{L}_{evi}(\alpha, \beta, q_{gt}) = \underbrace{(q_{gt} - \frac{\alpha}{\alpha+\beta+\epsilon})^2}_{\text{MSE Term}} + \underbrace{\frac{\alpha\beta}{(\alpha+\beta+\epsilon)^2(\alpha+\beta+1+\epsilon)}}_{\text{Variance Term}} \tag{27}$$

$$\begin{aligned}\mathcal{L}_{KL}(\alpha, \beta) &= \mathrm{KL}(\mathrm{Beta}(\alpha, \beta) || \,\mathrm{Beta}(1, 1)) \\ &= \mathrm{lgamma}(\alpha + \beta) - \mathrm{lgamma}(\alpha) - \mathrm{lgamma}(\beta) \\ &\quad + (\alpha - 1)(\digamma(\alpha) - \digamma(\alpha + \beta)) \\ &\quad + (\beta - 1)(\digamma(\beta) - \digamma(\alpha + \beta))\end{aligned} \tag{28}$$

$$\mathcal{L}_{quality} = \mathbb{E}_{\text{batch}}\left[\sum_{k,d}\left(\mathcal{L}_{evi}(\alpha^{loss}_{k,d}, \beta^{loss}_{k,d}, q_{gt,k,d}) + \lambda_{KL}\max(0, \mathcal{L}_{KL}(\alpha^{loss}_{k,d}, \beta^{loss}_{k,d}))\right)\right] \tag{29}$$

The $\mathcal{L}_{evi}$ term drives the mean predicted quality towards the target while penalizing high confidence for incorrect predictions via the variance term. The $\mathcal{L}_{KL}$ term regularizes the learned distribution, preventing collapse and encouraging uncertainty quantification. Gradients from $\mathcal{L}_{quality}$ only update critic parameters θ_C.

Correction Loss ($\mathcal{L}_{corr}$): This loss component trains the correction head of each critic C_k to output a signal Δ_k (Equation 9) that aims to reduce the error in the expert's auxiliary prediction $y_{aux,k}$. Specifically, it minimizes the Huber loss between the damped corrected prediction ($y_{aux,k} + \lambda_{damp}\Delta_k$) and the ground truth target $y_{true,k'}$.

$$\mathcal{L}_{corr} = \mathbb{E}_{\text{batch}}\left[\sum_{k\in\{e,l,g\}} \mathrm{Huber}(y_{aux,k} + \lambda_{damp}\Delta_k, y_{true,k'})\right] \tag{30}$$

This loss is calculated using the auxiliary predictions $y_{aux,k}$ and correction signals Δ_k generated during the standard forward pass. Consequently, the gradients derived from $\mathcal{L}_{corr}$ backpropagate to update both the critic parameters θ_{C_k} (through Δ_k) and the expert parameters θ_{E_k}.

$$\mathcal{L}_{corr} = \frac{1}{N_{exp}}\sum_{k\in\{e,l,g\}} \mathbb{E}_{\text{batch}}\left[\mathrm{Huber}(y_{aux,k} + \lambda_{damp}\Delta_k, y_{true,k'})\right] \tag{31}$$

Evidence Penalty Loss ($\mathcal{L}_{penalty}$): To further regulate the correction mechanism and promote robustness, we use an evidence-weighted penalty term, $\mathcal{L}_{penalty}$. This loss component links the confidence of the critic's quality assessment (as measured by the total evidence $S^{loss}_{k,d} = \alpha^{loss}_{k,d} + \beta^{loss}_{k,d}$) to the magnitude of the correction signal Δ^{loss}_k it proposes. The rationale is to discourage the critic from making large adjustments to the expert's prediction when its own assessment of the expert's quality is uncertain (i.e., when the evidence S is low). This is achieved by penalizing the squared L2 norm of the correction signal, inversely weighted by the total evidence:

$$\mathcal{L}_{penalty} = \mathbb{E}_{\text{batch}}\left[\sum_{k,d}\gamma\frac{(\Delta^{loss}_{k,d})^2}{S^{loss}_{k,d} + \epsilon}\right] \tag{32}$$

where γ is the hyperparameter, the sum is over the relevant output dimensions (k, d), and ϵ ensures numerical stability. This loss is calculated using the parameters $\alpha^{loss}, \beta^{loss}, \Delta^{loss}$ derived from the critic inputs (z^{loss}_k), ensuring that its gradients only update the critic parameters θ_C. This evidence-aware regularization encourages more cautious corrections under uncertainty compared to standard L2 regularization on Δ_k alone.

D.2.1 Phased Training Strategy

A three-phased training strategy was utilized to stabilize training.

1. **Phase 1 (Expert Pretraining):** Initially, only the expert parameters (E) are trained for a set number of epochs (N_1). Here, the primary learning signal comes from the auxiliary loss $\mathcal{L}_{\text{aux}}$ (detailed in Appendix D) associated with each expert, while the critic (C) and decider (F) components remain frozen.
2. **Phase 2 (Critic and Decider Integration):** In the second phase, the expert parameters (E) are frozen. The critic (C) and decider (F) parameters are then trained for N_2 epochs, utilizing the full

Table 2: Ablation studies for proposed model.

Configuration	D.NRMSE ↓	D.AbsRel↓	D.RMSE$_{\log}$ ↓	L.NRMSE(f) ↓	Q.Depth ↑	Q.Life ↑
Damping factor $d = 0.5$	0.032 ± 0.009	0.164 ± 0.160	0.170 ± 0.150	0.063 ± 0.023	0.954 ± 0.001	0.966 ± 0.007
Damping factor $d = 1.0$	0.034 ± 0.010	0.202 ± 0.240	0.189 ± 0.180	0.058 ± 0.018	0.952 ± 0.004	0.966 ± 0.006
Critic quality-loss $\lambda_{cq} = 4$	0.034 ± 0.011	0.150 ± 0.120	0.169 ± 0.130	0.071 ± 0.022	0.949 ± 0.003	0.962 ± 0.009
Phased training (1 / 6)	0.033 ± 0.009	0.171 ± 0.180	0.175 ± 0.150	0.064 ± 0.014	0.951 ± 0.002	0.964 ± 0.008
Phased training (5 / 15)	0.032 ± 0.008	0.153 ± 0.130	0.174 ± 0.140	$\mathbf{0.058 \pm 0.010}$	0.951 ± 0.003	0.962 ± 0.008
Phased training (3 / 8)	0.034 ± 0.009	0.153 ± 0.130	0.166 ± 0.130	0.059 ± 0.017	0.952 ± 0.002	0.968 ± 0.005
Phased training (10 / 10)	0.036 ± 0.013	0.198 ± 0.200	0.383 ± 0.520	0.083 ± 0.026	0.949 ± 0.000	0.963 ± 0.006
No evidential correction	0.035 ± 0.009	0.206 ± 0.240	0.193 ± 0.180	0.063 ± 0.018	0.951 ± 0.001	0.964 ± 0.007
No quality gating	0.032 ± 0.009	0.143 ± 0.110	0.167 ± 0.120	0.074 ± 0.028	0.950 ± 0.002	0.963 ± 0.009
No decider features	0.032 ± 0.006	0.160 ± 0.150	0.172 ± 0.140	0.077 ± 0.030	0.947 ± 0.001	0.959 ± 0.008
No decider fusion	0.034 ± 0.011	0.181 ± 0.180	0.179 ± 0.150	0.059 ± 0.015	0.953 ± 0.004	0.965 ± 0.006
No gating dropout	0.033 ± 0.007	0.169 ± 0.170	0.171 ± 0.150	0.072 ± 0.017	0.951 ± 0.001	0.961 ± 0.007
No phased training	0.031 ± 0.010	0.145 ± 0.120	0.159 ± 0.130	0.063 ± 0.023	0.948 ± 0.002	0.964 ± 0.008
Mean pooling	0.036 ± 0.010	0.170 ± 0.150	0.174 ± 0.140	0.069 ± 0.024	0.949 ± 0.006	0.968 ± 0.007
Heteroscedastic experts only	0.036 ± 0.011	0.139 ± 0.093	0.164 ± 0.110	0.063 ± 0.051	–	–
Full model ($\kappa = 2$)	$\mathbf{0.030 \pm 0.007}$	$\mathbf{0.140 \pm 0.120}$	$\mathbf{0.155 \pm 0.120}$	0.074 ± 0.022	0.950 ± 0.003	0.965 ± 0.006

composite loss $\mathcal{L}_{total}$ (detailed in Appendix D) to learn how to evaluate the pre-trained experts and fuse their outputs.

3. **Phase 3 (Joint Training):** Finally, all model parameters (E, C, F) are jointly fine-tuned for the remaining N_3 epochs using the complete loss function $\mathcal{L}_{total}$. This phase incorporates a gradual unfreezing schedule for the expert learning rates.

The specific number of epochs allocated to each phase (N_1, N_2, N_3) was subject to variation in some experiments, as further detailed in the ablation studies (Table 2).

D.3 Ablation Studies

To evaluate the distinct contributions of its architectural elements and design choices, we conducted a series of ablation studies on the EvidenceMoE framework. These studies systematically assessed the impact of individual mechanisms on depth and lifetime estimation by creating model variants with specific components removed or altered. Key aspects investigated included the necessity of evidential correction, quality gating, decider features, and decider fusion. Additionally, we explored the influence of hyperparameter settings, phased training schedules, and pooling mechanisms. Each ablated configuration was evaluated on the same test dataset. Table 2 presents comprehensive ablation results for our primary model configuration with $\kappa = 2$, while Table 3 shows additional ablation studies conducted with $\kappa = 8$ to evaluate the sensitivity of our approach to this hyperparameter. The implementation of EvidenceMoE is publicly available on GitHub. [1].

Table 3: Ablation study for $\kappa = 8$.

Configuration	D.NRMSE ↓	D.AbsRel↓	D.RMSE$_{\log}$ ↓	L.NRMSE(f) ↓	Q.Depth ↑	Q.Life ↑
Critic quality-loss $\lambda_{cq} = 4$	0.033 ± 0.011	0.146 ± 0.110	0.181 ± 0.130	0.058 ± 0.012	0.856 ± 0.025	0.900 ± 0.034
No evidential correction	0.036 ± 0.015	0.148 ± 0.100	0.182 ± 0.130	0.060 ± 0.011	0.859 ± 0.012	0.869 ± 0.044
No quality gating	0.033 ± 0.010	0.146 ± 0.110	0.176 ± 0.130	0.056 ± 0.012	0.868 ± 0.024	0.886 ± 0.031
No decider features	0.032 ± 0.011	0.146 ± 0.110	0.184 ± 0.130	0.067 ± 0.025	0.865 ± 0.014	0.892 ± 0.037
No decider fusion	0.035 ± 0.013	0.167 ± 0.160	0.182 ± 0.140	0.073 ± 0.015	0.828 ± 0.020	0.879 ± 0.022
Uniform gating	0.034 ± 0.010	0.157 ± 0.120	0.191 ± 0.150	0.066 ± 0.018	0.861 ± 0.021	0.879 ± 0.034
No gating dropout	0.041 ± 0.014	0.215 ± 0.180	0.373 ± 0.400	0.073 ± 0.022	0.865 ± 0.015	0.866 ± 0.029
No phased training	0.037 ± 0.015	0.181 ± 0.150	0.269 ± 0.260	0.076 ± 0.021	0.862 ± 0.007	0.890 ± 0.037
Mean pooling	0.034 ± 0.009	0.155 ± 0.130	0.173 ± 0.130	0.075 ± 0.027	0.818 ± 0.069	0.866 ± 0.058
No auxiliary MAE	0.037 ± 0.014	0.172 ± 0.130	0.234 ± 0.180	0.063 ± 0.018	0.850 ± 0.009	0.850 ± 0.034
Model trained with ($\kappa = 8$)	0.040 ± 0.013	0.216 ± 0.230	0.202 ± 0.180	$\mathbf{0.055 \pm 0.011}$	0.857 ± 0.023	0.896 ± 0.023
Full model ($\kappa = 2$)	$\mathbf{0.030 \pm 0.007}$	$\mathbf{0.140 \pm 0.120}$	$\mathbf{0.155 \pm 0.120}$	0.074 ± 0.022	0.950 ± 0.003	0.965 ± 0.006

[1] `https://anonymous.4open.science/r/EvidenceMoE-4728/`