Attribution-Guided Exit Policy for Reliable Early-Exit Inference

Haseena Rahmath P*
School of CSE&T
Bennett University,
Greater Noida, India
haseenarahmath@gmail.com

Muhammed Ashrah
Dept. of Computer Engineering
Jamia Millia Islamia
New Delhi, India
mohd.ashrah@gmail.com

Ajith Abraham
School of AI
Sai University
Chennai, India
abraham.ajith@gmail.com

Abstract

Early-exit neural networks improve inference efficiency by enabling predictions at intermediate layers, but their effectiveness critically depends on the exit policy that decides whether to stop or continue computation. Existing policies, often based on confidence thresholds, enable early termination but offer limited insight into the reliability of such decisions. To address this, we introduce an explainability-aware framework that operationalizes reliability in terms of the semantic consistency of early-exit decisions, using attribution-derived evidence. We propose Progressive Feature Attribution Maps (PFAM), which capture how feature importance evolves across exits, and the Interpretability-Based Early-Exit Score (IEES), a unified decision metric that integrates prediction confidence with attribution strength and cross-exit attribution stability. To enable efficient deployment, we further introduce a lightweight proxy predictor that estimates IEES using inexpensive, attribution-free features, allowing attribution-informed decisions without computing attributions at inference time. Experiments on convolutional architectures (ResNet, MobileNet, MSDNet) show that our approach improves accuracy–efficiency trade-offs, achieving up to 2.9 percentage points of accuracy gain while reducing inference time by up to 24% compared to full-depth inference. Similar trends are observed on a vision transformer and a BERT-based language model, suggesting applicability beyond convolutional architectures.

1 Introduction

Deep neural networks (DNNs) achieve state-of-the-art performance across many domains [1, 16, 14], but their high computational cost limits deployment on resource-constrained platforms. Early-exit DNNs address this by introducing intermediate classifiers, allowing inference to stop once predictions meet a confidence criterion [47], thereby reducing latency while preserving accuracy. However, in safety-critical settings such as healthcare or autonomous driving, efficiency alone is insufficient–decisions must also be reliable and grounded in semantically meaningful evidence to support appropriate trust [39]. Here, reliability refers specifically to the semantic consistency of an early-exit decision, assessed jointly through prediction confidence, attribution strength and attribution stability across exits; it does not imply model trustworthiness or system-level safety guarantees.

Existing exit policies, often based on confidence thresholds, determine when to stop inference but provide limited insight into the reliability of these decisions; early exits may be triggered despite weak or inconsistent supporting evidence. In parallel, most explainable AI (XAI) methods target single-exit models, producing explanations only at the final layer [3]. Early-exit networks, however, introduce multiple decision points, requiring explanations that reflect how model focus evolves across

*Work performed at Bennett University; currently affiliated with Vivish Technologies Pvt. Ltd. (MyGate), Kochi, India.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

depth rather than isolated per-exit interpretations—an aspect largely overlooked in conventional XAI methods [18].

To address these challenges, we propose an explainability-aware early-exit framework that models decision reliability using attribution signals. *Progressive Feature Attribution Maps* (PFAM) capture the evolution of feature importance across exits through localized and cumulative attributions. The *Interpretability-Based Early-Exit Score* (IEES) integrates prediction confidence, attribution strength, and cross-exit stability to assess intermediate exits. A lightweight proxy predictor estimates IEES from inexpensive, attribution-free features, enabling efficient inference without requiring attribution at test time. Together, these components enable reliability-aware exit decisions in settings where unreliable early exits can undermine trust.

We evaluate our framework on ResNet-18 (CIFAR-10), MSDNet (CIFAR-100, ImageNet), and MobileNetV3 (ImageNet). Results show consistent accuracy–efficiency improvements over confidence-based policies and, when only the exit policy of pretrained frameworks is replaced (Section 6), over their native policies. Qualitative examples and a small user study indicate that the resulting exits align better with stable, class-relevant attribution patterns. Similar trends are observed on a vision transformer (ImageNet) and a BERT-based model (SST-2), suggesting applicability beyond convolutional architectures.

Contributions. This work makes the following key contributions:

1. **PFAM:** Progressive attribution maps that capture both exit-wise and cumulative explanations, revealing how model focus evolves across inference stages.
2. **IEES:** An explainability-aware scoring mechanism that integrates confidence, attribution strength, and cross-exit stability to assess intermediate exits. This oracle IEES is computed offline; for deployment, it is approximated by a lightweight proxy predictor using attribution-free features, so no attribution is computed at inference.
3. **Evaluation:** Comparisons with existing exit policies and attribution methods demonstrate improved accuracy–efficiency trade-offs and stronger semantic consistency across exits.

Paper Organization. Section 2 reviews related work, Section 3 presents the methodology, and Sections 4–5 report experiments and results. Section 6 evaluates applicability across early-exit frameworks, Sections 7–8 provide discussion and limitations, and Section 9 concludes.

2 Related Work

Early-Exit DNNs. Early-exit architectures improve efficiency by terminating computation at intermediate layers once an exit criterion is met [36, 34]. Easy inputs exit early, while harder ones propagate deeper, making these models suitable for latency-constrained settings. BranchyNet [47] introduced intermediate classifiers in AlexNet, LeNet, and ResNet, achieving speedups with minimal accuracy loss. This idea has since been extended to CNNs such as MobileNets [23], VGG [15], MSDNet [24], and EfficientNet [46], as well as to Transformers [25, 53], LSTMs [30], and GNNs [37]. Recent extensions span multiple modalities. In vision, Demir et al. [12] enhance CNN-based early exits using learnable confidence branches and a unified accuracy–cost objective. In NLP, DeeBERT [52] and CEEBERT [5] show that intermediate Transformer layers contain sufficient discriminative information for entropy-based exiting, while FREE (Bae et al.) [4] accelerates early exits in autoregressive models. Early exiting has also been extended to large language and vision-language models (EE-LLM [10], CE-CoLLM [26], and FREE (Bajpai and Hanawal) [7])

Exit Policies. Exit policies determine when to stop inference. Classical approaches use fixed confidence or entropy thresholds [34, 55, 50], while adaptive variants employ reinforcement learning [48] or learned controllers [11]. Budget-aware scheduling has also been explored [25]. ZTW [51] recycles predictions, BEEM [6] improves ensemble margins, and E3 [31] uses Integrated Gradients offline to guide allocation, though runtime exits remain scheduler-driven. Recent work improves decision quality through consistency and uncertainty modeling. PABEE [56] exits once predictions remain unchanged across a fixed number of consecutive exits, CALM [41] enforces sequence-level consistency, and FREE (Bae et al.) [4] adapts thresholds via a Beta-mixture strategy. Uncertainty-based methods further refine exit behavior: Meronen et al. [33] address overconfidence via Laplace-based estimates, and CEEBERT [5] learns thresholds online under domain shift. However, most policies focus on

accuracy–efficiency trade-offs, with limited attention to the semantic reliability of early predictions. Our work addresses this gap by incorporating attribution stability into the exit signal.

Explainable AI. XAI methods provide insight into model predictions in critical settings [18]. Saliency maps [43] and Integrated Gradients [45] highlight feature importance, while Grad-CAM [42] and Grad-CAM++ [9] produce class-discriminative visualizations. FullGrad [44] aggregates attribution across all intermediate feature maps of a single-exit network. Model-agnostic approaches such as LIME [38] and SHAP [32] rely on surrogate models or Shapley values. These methods, largely designed for single-exit models, provide isolated explanations without capturing how relevance evolves across intermediate exits. PFAM and IEES extend this paradigm: unlike FullGrad, which aggregates across layers of one prediction, PFAM aggregates attribution across explicit exit points, producing exit-wise and cumulative attributions, and IEES incorporates the resulting stability into the exit signal, supporting stage-wise and semantically grounded decision-making.

3 Methodology

We propose an explainability-aware framework for early-exit inference with three components: (i) PFAM, which tracks how model focus evolves across exits, (ii) IEES, which integrates confidence and attribution signals into a reliability score, and (iii) a lightweight *proxy* that estimates IEES from attribution-free features. Unlike prior approaches that use attribution mainly for training-time adjustments, offline analysis, or scheduling [31], we use attribution to define a reliability score that is learned offline and deployed via a lightweight proxy, avoiding attribution overhead at inference.

3.1 Framework Overview

The pipeline operates in three phases. (a) **Training:** A backbone (e.g., ResNet, MobileNet) is augmented with intermediate classifiers and trained jointly using standard early-exit supervision (details in Section 4). (b) **Calibration:** PFAMs characterize attribution focus across exits; stability is derived from these maps and attribution strength from the channel-weighted activation within the top-$k\%$ most salient locations of the cumulative attribution map. Together with confidence, these signals define the *oracle* IEES, whose attribution components provide complementary information that improves separation between reliable and unreliable exit decisions (Appendix I). The weights (w_1, w_2, w_3), which balance confidence, attribution strength, and stability, and the per-exit thresholds τ_i^* are optimized. The resulting oracle IEES scores then supervise proxy training on attribution-free features, yielding a lightweight predictor ready for deployment. (c) **Inference:** The trained proxy predicts IEES at each exit; no attribution maps are computed. If $\widehat{\text{IEES}}_i(x) \geq \tau_i^*$, inference halts; otherwise it proceeds deeper. PFAM maps are optional at test time and used only for visualization or auditing.

Terminology. We distinguish *oracle* IEES, $\text{IEES}_i(x)$ (Eq. 5), computed offline with attribution maps to supervise proxy training, from *proxy* IEES, $\widehat{\text{IEES}}_i(x)$ (Eq. 9), which is the only score used at inference. Unless otherwise specified, "IEES" in the text, tables and figures refers to the proxy-based policy.

3.2 Progressive Feature Attribution Maps (PFAM)

PFAM extends attribution from single predictions to progressive trajectories, capturing how model focus evolves across exits. For exit i, an attribution map is defined as

$$A_i(x) = \text{AttributionMethod}(f_i, x), \tag{1}$$

where f_i denotes the backbone up to exit i, instantiated with any attribution technique (e.g., Grad-CAM, IG, LayerCAM). In our experiments, we adopt Grad-CAM, yielding

$$A_i(x) = \text{ReLU}\left(\sum_{\kappa=1}^{K} \alpha_{i,\kappa}\, Z_{i,\kappa}(x)\right), \tag{2}$$

where $Z_{i,\kappa}(x) \in \mathbb{R}^{H\times W}$ is the activation map of channel κ at exit i, $\alpha_{i,\kappa}$ its gradient-derived weight, and K the number of channels (see Appendix O for transformer and sequence model extensions). To

capture progression, PFAM computes cumulative attribution

$$C_i(x) = \frac{1}{i} \sum_{j=1}^{i} A_j(x), \tag{3}$$

where maps $A_j(x)$ are bilinearly upsampled to a common resolution $H_0 \times W_0$ prior to aggregation, ensuring spatial compatibility across exits. Stability is measured via

$$Q_i(x) = \begin{cases} 0, & i = 1, \\ \dfrac{1}{|\Omega|} \displaystyle\sum_{p \in \Omega} \big| C_i(x)[p] - C_{i-1}(x)[p] \big|, & i \geq 2, \end{cases} \tag{4}$$

where Ω denotes the spatial domain ($Q_1(x) = 0$, as there is no previous exit). Smaller $Q_i(x)$ indicates more stable attribution focus. Since $C_i - C_{i-1} = (A_i - C_{i-1})/i$, the difference between consecutive cumulative maps is scaled by $1/i$ through the running-average construction. This aggregation is intended to suppress exit-specific noise and stabilize attribution focus progressively, although Q_i is not necessarily monotonically decreasing with depth. For $i \geq 2$, the resulting Q_i values are min-max normalized per exit before scoring (Appendix M).

Unlike conventional explainability methods that generate independent explanations for individual predictions, PFAM models the progressive evolution of attribution across successive exits. By incorporating this attribution-level stability alongside prediction confidence, IEES evaluates not only how confident a prediction is, but also whether the supporting evidence has stabilized, providing an additional reliability signal for early-exit decisions. PFAM does not propose a new attribution algorithm and is attribution-agnostic by design: it can be instantiated with different attribution methods (e.g., Integrated Gradients, with comparable results; Appendix F), so advances in attribution methods can be adopted without changing the IEES formulation.

3.3 Interpretability-Based Early-Exit Score (IEES)

IEES is guided by *semantic reliability*: an exit is considered reliable when its prediction is (i) confident, (ii) supported by class-relevant attribution evidence, and (iii) stable across exits. IEES estimates the reliability of exit i by combining confidence, attribution strength, and attribution stability:

$$\mathrm{IEES}_i(x) = w_1 S_i(x) + w_2 F_i(x) - w_3 Q_i(x), \tag{5}$$

where $S_i(x) = \max_c \, p_i(c \mid x)$, with $p_i(c \mid x)$ the softmax probability at exit i; $Q_i(x)$ is defined in Eq. 4; and $F_i(x)$ (defined below in Eq. 7) captures class-relevant evidence. The weights (w_1, w_2, w_3) are non-negative and tuned on a held-out validation set. $F_i(x)$ and $Q_i(x)$ (for $i \geq 2$) are min-max normalized to $[0, 1]$ per exit, using validation-set statistics at that exit, prior to scoring (Appendix M). Let $M_i(x)$ denote the channel-weighted activation map at exit i, bilinearly upsampled to $H_0 \times W_0$ as for $A_j(x)$:

$$M_i(x)[h, w] = \sum_{\kappa=1}^{K} \alpha_{i,\kappa} \, Z_{i,\kappa}(x)[h, w]. \tag{6}$$

Attribution strength is then computed as the average activation over the top-$k\%$ most discriminative spatial locations:

$$F_i(x) = \frac{1}{|\Omega_k^i(x)|} \sum_{p \, \in \, \Omega_k^i(x)} M_i(x)[p], \tag{7}$$

where the salient mask $\Omega_k^i(x)$ retains positions whose cumulative attribution attains at least the $(1 - k/100)$-quantile of $C_i(x)$:

$$\Omega_k^i(x) = \Big\{ \, p \in \Omega \;\Big|\; C_i(x)[p] \geq q_k^i(x) \Big\}, \tag{8}$$

and $q_k^i(x)$ is the $(1 - k/100)$-quantile of $C_i(x)$, computed per sample at exit i. We use k=20 across all experiments (Appendix P). Before normalization, positive values of $F_i(x)$ indicate strongly activated, spatially concentrated class evidence, while near-zero or negative values reflect weak or diffuse support, reducing the exit score.

The signs in Eq. (5) follow the semantic reliability principle: confidence $(+S)$ and relevance $(+F)$ are rewarded, while instability $(-Q)$ is penalized. IEES integrates these signals into a unified score that can serve as a drop-in replacement for confidence in standard exit policies, consistent with prior work highlighting stability as a desirable property of explanations [18]. Empirical support is provided by Appendices H and I and by the ablations in Appendix J (Tables 10–12); calibration details are in Appendix M.

3.4 Proxy Exit Predictor

Since computing attributions at inference is costly, we train a lightweight regressor g_θ at each exit to approximate the oracle IEES (Eq. 5):

$$\widehat{\mathrm{IEES}}_i(x) = g_\theta(\phi_i(x)), \tag{9}$$

where $\phi_i(x) \in \mathbb{R}^{10}$ comprises five single-exit statistics (confidence, logit margin, entropy, mean/max activation) and five cross-exit signals (confidence delta $|S_i - S_{i-1}|$, class stability $\mathbf{1}[\hat{y}_i = \hat{y}_{i-1}]$, logit cosine similarity Q_i^{logit}, entropy delta, activation delta); see Appendix B. Cross-exit features are zero for $i{=}1$ and require only storing the previous exit's logits and scalar statistics, incurring negligible memory overhead. Training minimizes the per-exit loss

$$\mathcal{L}_{\text{proxy},i} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{x \in \mathcal{D}_{\text{train}}} \left(\widehat{\mathrm{IEES}}_i(x) - \mathrm{IEES}_i(x)\right)^2. \tag{10}$$

The proxy is trained on the training set $\mathcal{D}_{\text{train}}$, with 10% held out for validation. Among Random Forest, XGBoost, and a shallow MLP, Random Forest provides a favorable accuracy–latency trade-off and is used by default. Proxy fidelity is reported in Appendix B, where predictors achieve high fidelity, with $R^2 \geq 0.80$ and Spearman $\rho \geq 0.85$, at sub-millisecond latency. Computing attributions incurs ≈0.4–0.5 ms per sample, whereas the proxy requires ≈0.1 ms, a roughly 4–5× lower overhead. To isolate the effect of the supervision target from the proxy architecture, we compare proxies trained with alternative targets (confidence, recalibrated confidence, and correctness) in Section 5 (Table 2); on CIFAR-10, IEES supervision performs best.

Inference. At runtime, if $\widehat{\mathrm{IEES}}_i(x) \geq \tau_i^*$, inference halts; otherwise computation proceeds deeper. In the worst case, runtime approaches backbone latency plus the overhead of the intermediate exit heads and of the proxy evaluated at each exit.

4 Experimental Setup

Datasets and Architectures. We evaluate on CIFAR-10 [29], CIFAR-100[29], and ImageNet[13] using ResNet-18 [19], MSDNet [24], and MobileNetV3[22] backbones, respectively. CIFAR-10/100 contain 50k training and 10k test images (32×32); 5k training samples are held out for validation. ImageNet has 1.2M training and 25k testing images (224×224), with 25k held out for validation. All backbones are augmented with three side exits (ResNet-18: stages 2/3/4; MSDNet: layers 3/5/9; MobileNetV3: blocks 3/5/7), following BranchyNet and MSDNet conventions. Each exit uses a 3×3 convolution, adaptive pooling, and a linear classifier.

Training. Backbone and exit classifiers are trained jointly using standard early-exit supervision [47, 24], with a weighted sum of exit-specific cross-entropy losses and branch weights decreasing from early to late exits. Optimization follows standard settings for each backbone; full details are provided in Appendix N.

Calibration and Proxy. Calibration tunes weights (w_1, w_2, w_3) and thresholds τ_i^* via grid search on validation data ($\tau \in [0.5, 0.95]$, using a coarse grid followed by local refinement; details in Appendix M). A Random Forest proxy regressor is trained on attribution-free features to approximate IEES, with details and extended evaluation reported in Appendix B. Proxy training uses the training set with 10% held out for validation to avoid test leakage.

Evaluation Protocols. We report top-1 accuracy, exit distribution, latency, and AUC (area under the accuracy–FLOPs curve) using the Budgeted Batch Classification (BBC) protocol [24]. BBC applies a global threshold τ^* to match FLOPs budgets, while inference uses per-exit thresholds τ_i^*. All thresholds are tuned on validation data and fixed at test time. Robustness is evaluated under synthetic

corruptions (CIFAR-10-C [20]) and by perturbing τ_i^* within $\pm 10\%$. Unless otherwise specified, all results use proxy IEES at inference.

Explainability Benchmark. PFAM is compared with Grad-CAM, Integrated Gradients (IG), SmoothGrad, and Occlusion. Since these baselines target single-exit models, they are evaluated on truncated sub-networks up to each exit E_i (Appendix C), while PFAM operates across all exits. Attribution quality is assessed using: (i) *semantic progression*—Monotonic Feature Alignment (MFA ↑) and convergence Intersection-over-Union (IoU ↑); (ii) *faithfulness*—deletion AUC (↓) and insertion AUC (↑); and (iii) human preference via a user study. Metric definitions and study details are provided in Appendices D and K.

Baselines. We primarily benchmark against confidence-based exits, the de facto standard, in three variants: (i) *Confidence (Optimized)* with thresholds tuned on validation; (ii) *IEES* calibrated in the same manner for fairness; and (iii) *Confidence (IEES-Matched)*, where thresholds are adjusted to reproduce IEES's exit distribution. We further compare with recent exit policies on pretrained models (Section 6).

Hardware. Experiments use an NVIDIA RTX A6000 (48GB) and a 3.91 GHz Intel Xeon CPU. Inference latency is measured as average wall-clock time per test sample with batch size 1. Each result is averaged over five independent runs and reported as mean ± standard deviation.

5 Results and Analysis

We evaluate the proposed framework through four research questions addressing reliability, efficiency, robustness, and attribution-based explainability in early-exit models:

RQ1: Does IEES provide reliable exit decisions guided by attribution without sacrificing accuracy?
RQ2: Does IEES improve accuracy–efficiency trade-offs under compute budgets?
RQ3: How robust is IEES to threshold variation and distribution shift?
RQ4: How does PFAM compare to standard attribution methods in explaining exit decisions?

The evaluation uses complementary protocols: Section 5.1 compares policies at optimized thresholds and matched exit distributions (isolating the decision rule), Section 5.2 uses BBC FLOPs–accuracy curves, Section 5.3 tests threshold perturbation and distribution shift, Section 5.4 compares PFAM with standard attribution methods, and Section 6 evaluates IEES in pretrained frameworks at matched inference budgets. Unless otherwise specified, inference-time results use proxy IEES; oracle IEES, which requires attribution maps, is used only offline (calibration and proxy supervision) and in the qualitative figures.

5.1 RQ1: Does IEES provide reliable exit decisions guided by attribution without sacrificing accuracy?

We compare IEES with confidence-based exit policies under two settings: (i) *optimized thresholds*, tuned for the best accuracy–efficiency trade-off, and (ii) *matched exit rates*, where confidence thresholds reproduce IEES's exit distribution to isolate the effect of the decision rule. Table 1 shows that IEES consistently improves overall accuracy under both settings.

On CIFAR-10, IEES exits approximately 76% of inputs by Exit 2 (cumulative across Exits 1–2), achieving 89.77% accuracy at 3.14 ms, compared with 87.35% at 3.05 ms for optimized confidence and 86.91% at 4.12 ms for full inference. When confidence thresholds are adjusted to match IEES's exit distribution, accuracy drops to 86.13%, indicating that the confidence-based decision rule is less effective under identical exit conditions. Notably, IEES slightly exceeds full-depth accuracy, which may be attributed to regularization effects from multi-exit supervision, as observed in prior work [47].

On CIFAR-100, IEES achieves 77.89% accuracy at 27.59 ms, exceeding optimized confidence (76.95%, 25.30 ms) and matched confidence (75.78%, 26.94 ms). On ImageNet, IEES resolves approximately 70% of inputs by Exit 3, with only 30% requiring full-depth inference, achieving 76.91% accuracy at 29.49 ms, compared with 75.04% at 27.56 ms for optimized confidence and 74.06% at 36.63 ms for full-depth inference. Matched confidence again underperforms (74.31%). Across these datasets, IEES achieves higher accuracy with a modest latency increase over optimized confidence, while reducing latency relative to full-depth inference. This trade-off reflects IEES routing a larger proportion of harder samples to deeper exits. The deepest exit processes the residual samples that were not resolved earlier, resulting in lower per-exit accuracy than the overall accuracy.

Table 1: Exit-wise accuracy and exit distribution. E1–E4 accuracies are computed over samples exiting at that exit. Conf. (Opt.) and IEES (Opt.) use thresholds tuned on validation (IEES: proxy-based policy); Conf. (Matched) matches IEES's exit distribution. Mean ± std over 5 runs.

Dataset / Model	Exit Policy	Exit Rate (%)				Accuracy (%)				Overall Acc.	Latency (ms)
		E1	E2	E3	E4	E1	E2	E3	E4		
CIFAR-10 ResNet-18	No early exit	–	–	–	100.0	–	–	–	86.91	86.91 ± 0.00	4.12 ± 0.00
	Conf. (Opt.)	48.02	30.24	9.35	12.39	88.91	89.57	89.71	74.12	87.35 ± 0.03	3.05 ± 0.02
	IEES (Opt.)	**20.18**	**55.92**	**17.26**	**6.64**	**90.01**	**91.12**	**90.04**	**77.01**	**89.77 ± 0.02**	**3.14 ± 0.02**
	Conf. (Matched)	20.18	55.92	17.26	6.64	89.03	86.37	87.06	72.89	86.13 ± 0.04	3.00 ± 0.03
CIFAR-100 MSDNet	No early exit	–	–	–	100.0	–	–	–	76.67	76.67 ± 0.00	32.04 ± 0.00
	Conf. (Opt.)	29.30	31.94	16.10	22.66	78.63	79.06	80.33	69.44	76.95 ± 0.05	25.30 ± 0.03
	IEES (Opt.)	**20.08**	**23.92**	**28.89**	**27.11**	**81.71**	**83.19**	**81.13**	**66.94**	**77.89 ± 0.03**	**27.59 ± 0.02**
	Conf. (Matched)	20.08	23.92	28.89	27.11	77.89	79.96	79.98	66.04	75.78 ± 0.07	26.94 ± 0.02
ImageNet MobileNetV3	No early exit	–	–	–	100.0	–	–	–	74.06	74.06 ± 0.00	36.63 ± 0.00
	Conf. (Opt.)	18.56	35.95	25.42	20.07	77.01	78.02	78.10	64.02	75.04 ± 0.03	27.56 ± 0.02
	IEES (Opt.)	**16.83**	**25.82**	**26.94**	**30.41**	**81.12**	**80.34**	**81.01**	**68.05**	**76.91 ± 0.01**	**29.49 ± 0.03**
	Conf. (Matched)	16.83	25.82	26.94	30.41	76.92	77.59	77.85	66.96	74.31 ± 0.04	28.53 ± 0.01

This is consistent with prior work [47, 24]. An ablation study in Appendix J further confirms that confidence, attribution strength, and stability contribute complementary, non-redundant signals in IEES.

Supervision target. To isolate the effect of the supervision signal from the proxy architecture, Table 2 trains the same Random Forest on the same attribution-free features with different targets (CIFAR-10/ResNet-18). Supervising with recalibrated confidence or correctness gives 87.97% and 88.02% accuracy, compared with 87.35% for confidence-only exits, whereas IEES supervision gives 89.77%. The gain is therefore not due to the proxy architecture alone. Per-exit results are in Appendix B.1.

Attribution-Guided Exit Behavior. Figure 1 illustrates how IEES combines confidence, attribution strength, and stability to guide exit decisions. In Figure 1(a), improving confidence and attribution quality enable a correct, well-supported exit at Exit 3. In Figure 1(b), confidence-only reasoning exits prematurely despite unstable attribution, while attribution-only reasoning may stop at Exit 3; IEES combines both signals and defers to Exit 4 for a more reliable outcome. In Figure 1(c), strong attribution enables a correct early exit at Exit 1 despite moderate confidence. In Figures 1(d)–1(f), high confidence precedes diffuse or unstable attribution, prompting IEES to defer until maps align with class-relevant regions, trading some latency for a more consistent exit. Figure 1(e) further illustrates progressive attribution refinement, leading to a later but more reliable exit, while Figure 1(g) shows a persistently ambiguous sample deferred to the deepest exit. Overall, these cases demonstrate how combining confidence with attribution consistency reduces premature exits and promotes more reliable decisions, particularly in trust-sensitive settings.

5.2 RQ2: Does IEES improve accuracy–efficiency trade-offs under compute budgets?

We evaluate deployment-time efficiency under the BBC paradigm, where a global FLOPs budget governs early-exit decisions (Section 4, Evaluation Protocols). Figure 2 presents FLOPs–accuracy curves for ResNet-18 on CIFAR-10, MSDNet on CIFAR-100, and MobileNetV3 on ImageNet, with the red marker indicating the optimal budget threshold and the black marker full-depth inference. Across all settings, IEES achieves higher accuracy at equal or lower computational cost than confidence-based exits. This trend is reflected in higher AUC values: 0.644 vs. 0.540 (CIFAR-10), 0.554 vs. 0.495 (CIFAR-100), and 0.600 vs. 0.520 (ImageNet). Relative to full-depth inference, IEES also reduces latency by enabling earlier exits for samples with stable attribution signals, yielding reductions of 24% on CIFAR-10 and 20% on ImageNet (Table 1). In contrast, confidence-based policies show limited gains under stricter budget constraints. Overall, these results indicate that incorporating

Table 2: Proxy supervision targets on CIFAR-10/ResNet-18 (same Random Forest and attribution-free input features; only supervision targets differ). Confidence-only uses no proxy.

Supervision target	Overall Acc. (%)	Time (ms)
Confidence only (no proxy)	87.35 ± 0.03	3.05
Conf. Recalib. RF	87.97 ± 0.04	3.08
Correctness RF	88.02 ± 0.01	3.13
IEES RF (ours)	**89.77 ± 0.02**	3.14

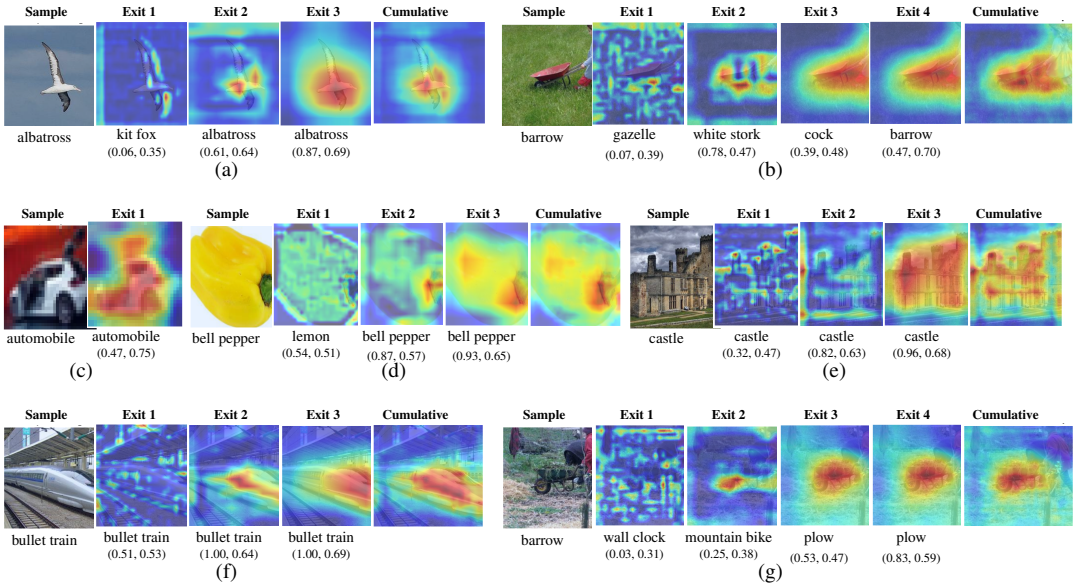


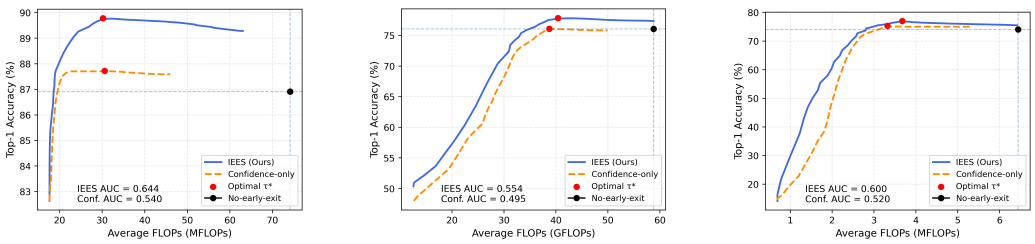
Figure 1: Qualitative examples of attribution maps with ground-truth and predicted labels, along with prediction confidence (P) and IEES score (I). In each row, the label below the input sample is the ground truth, while labels below exit maps show the predicted class with its corresponding (P, I). (a) Correct exit at Exit 3 with strong confidence and attribution. (b) IEES defers to Exit 4 due to unstable attribution, correcting a premature confidence-only exit. (c) Early exit at Exit 1 based on stable attribution; the cumulative map is omitted as it is identical to Exit 1. (d) IEES avoids premature exit under high but unreliable confidence. (e) Attribution quality refines progressively across exits. (f) Exit is delayed until semantic focus emerges. (g) IEES defers fully for persistently ambiguous inputs.

attribution-based stability leads to improved accuracy–efficiency trade-offs under fixed compute budgets. Additional threshold–accuracy and threshold–exit analyses are provided in Appendix G.

5.3 RQ3: How robust is IEES to threshold variation and distribution shift?

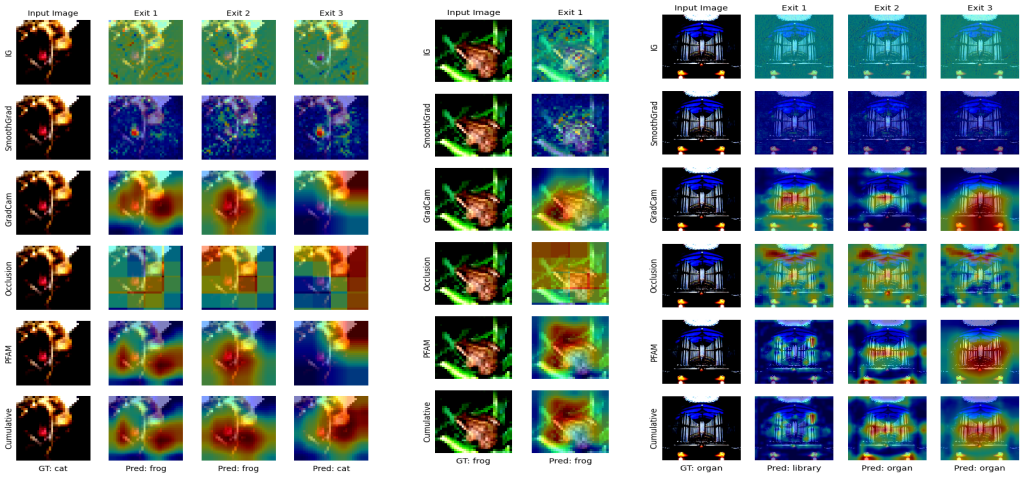
Threshold Sensitivity. To evaluate calibration robustness, we perturb decision thresholds at each exit by $\pm 10\%$ around the optimal values τ_i^*. As summarized in Table 5 (Appendix A), accuracy changes by at most 1.8 points across datasets and remains above full-depth inference in all settings, while higher thresholds increase latency (by up to 26% on CIFAR-10). These results suggest that IEES is reasonably robust to mild threshold miscalibration.

Distribution Shift Evaluation. We further assess robustness under distribution shift using CIFAR-10-C [20], with thresholds calibrated on clean data and fixed at test time. Table 4 (Appendix A) reports accuracy and average exit index under common corruptions (severity level 3). Under increasing corruption, IEES progressively routes more samples to deeper exits; for example, the average exit index increases from 2.10 on clean data to 2.91 under motion blur and 3.02 under Gaussian



(a) CIFAR-10 (ResNet-18) (b) CIFAR-100 (MSDNet) (c) ImageNet (MobileNetV3)

Figure 2: Accuracy–FLOPs trade-off under BBC. IEES consistently achieves higher accuracy at similar or lower computational cost than confidence-based policies. Red and black markers denote the optimal threshold and full-model inference, respectively. Curves are smoothed for visualization.



(a) (b) (c)

Figure 3: PFAM vs. standard attribution methods: (a) for a misclassified Exit 1 sample, PFAM isolates the error more clearly; (b) for a correct Exit 1 case, it stays focused while others diffuse; (c) across exits, it maintains consistency, reinforcing model confidence.

noise. Although accuracy decreases under these conditions (73.34% and 68.55%, respectively), the degradation is gradual and occurs without retraining or recalibration. This behavior is associated with reduced confidence and attribution stability under corruption, prompting IEES to defer decisions to later exits. Overall, these results indicate that IEES degrades gracefully and adjusts its exit behavior under distribution shift without additional supervision.

5.4 RQ4: How does PFAM compare to standard attribution methods in explaining exit decisions?

We compare PFAM with standard post-hoc attribution methods—Integrated Gradients (IG), SmoothGrad, Grad-CAM, and Occlusion—adapted to operate at individual early exits. While these baselines generate independent single-level explanations at each exit, PFAM produces both exit-wise and cumulative heatmaps, capturing how attribution focus evolves across inference stages. Figure 3 illustrates qualitative differences. In Figure 3(a), baseline methods produce diffuse or less stable saliency for a misclassified *cat* sample, whereas PFAM progressively sharpens focus on the object across exits. Figure 3(b) shows PFAM maintaining more consistent localization as predictions mature, while cumulative maps aggregate evidence into a coherent attribution trajectory. In Figure 3(c), PFAM remains more stable across exits, whereas baseline explanations vary, highlighting the advantage of progressive attribution for interpreting exit decisions. Additional visual comparisons are provided in Appendix L. Quantitatively, PFAM achieves substantially higher IoU and competitive MFA across datasets, with higher MFA than the baselines on CIFAR-100 and ImageNet. It also achieves higher insertion AUC across settings; deletion AUC is competitive, with IG and Occlusion performing strongly on CIFAR-10 and ImageNet. Full results are reported in Appendix E (Table 8). A small-scale user study (Appendix K) further supports these findings: participants generally preferred PFAM visualizations and judged IEES-guided exits as more reasonable. Overall, PFAM enhances transparency in early-exit inference by providing progressive, stage-aware explanations, whereas standard post-hoc methods offer isolated snapshots of model behavior.

6 Applicability Across Early-Exit Frameworks

IEES operates at the exit-decision level, replacing the exit scoring rule without modifying the backbone or exit heads. We assess its architectural generalizability across CNN, ViT, and NLP Transformer families using four representative frameworks: EENet [25], MSDNet [24], DVT [49], and DeeBERT [52]. EENet is evaluated in both vision and NLP settings, with an additional comparison against a patience-based consistency policy (PABEE-style [56]) on DeeBERT. Using publicly available pretrained checkpoints, we first evaluate each framework's original exit policy, then replace it with IEES while keeping the architecture unchanged. For fairness, we match the original policy's

Table 3: IEES applied across early-exit frameworks with matched inference budgets. Only the decision rule is replaced on pretrained models. Speed (×) denotes relative speedup w.r.t. full-depth inference and defines the matched inference budget. Acc. (Orig.) and Acc. (IEES) report accuracy of the original and IEES policies, respectively. PABEE-style denotes a patience-based consistency exit applied to the DeeBERT checkpoint.

Method	Backbone / Dataset	Modality	Exit Policy	Speed (×)	Acc. (Orig.)	Acc. (IEES)	△ Acc.
EENet [25]	ResNet-56 / CIFAR-10	Vision	Scheduler	1.55	92.79 ± 0.03	93.82 ± 0.02	1.03
MSDNet [24]	MSDNet / ImageNet	Vision	Max Conf.	1.58	73.51 ± 0.04	74.63 ± 0.02	1.12
DVT [49]	T2T-ViT-12 / ImageNet	Vision	Max Conf.	1.31	77.72 ± 0.02	79.49 ± 0.03	1.77
DeeBERT [52]	BERT-base / SST-2	NLP	Entropy	1.40	92.92 ± 0.02	93.56 ± 0.01	0.64
PABEE-style [56]	BERT-base / SST-2 (DeeBERT ckpt.)	NLP	Patience	1.55	92.90 ± 0.03	93.47 ± 0.02	0.57
EENet [25]	BERT-base / SST-2	NLP	Scheduler	1.62	91.82 ± 0.03	92.45 ± 0.02	0.63

inference budget and calibrate IEES to achieve comparable cost. Each row in Table 3 represents an independent evaluation at its corresponding matched budget. IEES consistently improves accuracy by 0.57–1.77 percentage points at comparable inference costs, demonstrating the benefits of attribution-aware reliability scoring across diverse exit policies and architectures.

7 Discussion

This work demonstrates that early-exit decisions benefit from signals beyond confidence. By combining confidence, attribution strength, and cross-exit stability, IEES improves exit quality while preserving efficiency through proxy-based deployment. It reduces premature exits, improving accuracy under compute constraints and robustness to threshold perturbations and distribution shifts. Additional analyses (Appendices H, I) confirm the complementary roles of these signals in guiding semantically consistent decisions. The proxy enables low-overhead deployment, while PFAM reveals how evidence evolves across exits, supporting attribution-informed decision-making.

Broader Impact. This work improves the efficiency and transparency of early-exit neural networks, enabling deployment in resource-constrained settings. However, premature exits may introduce risks in high-stakes applications, requiring safeguards such as fallback or abstention mechanisms; the notion of reliability used here concerns the semantic consistency of exit decisions and does not constitute a safety guarantee.

8 Limitations and Future Work

IEES relies on a proxy predictor to approximate attribution-informed scores at inference. While efficient, proxy fidelity bounds decision quality and may lead to suboptimal exits; improving robustness via online or continual adaptation is an important direction. PFAM is currently instantiated with gradient-based methods (e.g., Grad-CAM, IG); extending to gradient-free or black-box attributions and evaluating their impact on stability remains future work. The attribution-strength term assumes spatially concentrated evidence, potentially limiting its applicability to distributed-evidence tasks (e.g., lung CT and histopathology) and well-calibrated models on simple datasets, which remain untested. Finally, evaluation is limited to image classification and sentiment analysis; extensions to structured prediction, generation, and multimodal settings, as well as broader comparison with modern backbones and emerging exit policies, are left for future work.

9 Conclusion

We introduced IEES, an exit policy for early-exit inference that combines confidence, attribution strength, and cross-exit stability to improve decision quality. It establishes an attribution-grounded reliability signal based on the semantic consistency of exit decisions, efficiently approximated at inference using a lightweight proxy predictor that avoids attribution computation. PFAM captures the evolution of feature importance across exits, providing insights into the decision process. Experiments on CIFAR-10/100 and ImageNet demonstrate consistent accuracy–efficiency gains over confidence-based policies, with robustness to threshold variation and distribution shift. Further evaluation by integrating IEES into pretrained frameworks confirms consistent gains across architectures and settings.

Acknowledgments and Disclosure of Funding

The authors thank the reviewers for their constructive feedback, which helped improve the clarity and presentation of this work. This work was carried out using computing resources provided by Bennett University. The authors received no external research funding. The authors declare no competing interests.

References

[1] B. Alshemali and J. Kalita. Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 191:105210, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.105210.

[2] D. Alvarez Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[3] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

[4] S. Bae, J. Ko, H. Song, and S.-Y. Yun. Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5910–5924. EMNLP, 2023.

[5] D. J. Bajpai and M. K. Hanawal. CeeBERT: Cross-domain inference in early exit BERT. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1736–1748, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.101. URL https://aclanthology.org/2024.findings-acl.101/.

[6] D. J. Bajpai and M. K. Hanawal. BEEM: Boosting performance of early exit DNNs using multi-exit classifiers as experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EzrZX9bd4G.

[7] D. J. Bajpai and M. K. Hanawal. Free: fast and robust vision language models with early exits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23599–23615, 2025.

[8] U. Bhalla, S. Srinivas, and H. Lakkaraju. Discriminative feature attributions: Bridging post hoc explainability and inherent interpretability. *Advances in Neural Information Processing Systems*, 36:44105–44122, 2023.

[9] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[10] Y. Chen, X. Pan, Y. Li, B. Ding, and J. Zhou. Ee-llm: large-scale training and inference of early-exit large language models with 3d parallelism. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[11] X. Dai, X. Kong, and T. Guo. Epnet: Learning to exit with flexible multi-branch network. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 235–244, Virtual Event, Ireland, 2020. Association for Computing Machinery (ACM). CIKM 2020.

[12] E. Demir and E. Akbas. Early-exit convolutional neural networks. *arXiv preprint arXiv:2409.05336*, 2024.

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

[14] A. S. Dhanjal and W. Singh. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83(8):23367–23412, 2024.

[15] R. Dong, Y. Mao, and J. Zhang. Resource-constrained edge ai with early exit prediction. *Journal of Communications and Information Networks*, 7(2):122–134, 2022.

[16] P. Fraternali, F. Milani, R. N. Torres, and N. Zangrando. Black-box error diagnosis in deep neural networks for computer vision: a survey of tools. *Neural Computing and Applications*, 35 (4):3041–3062, 2023.

[17] J. Gildenblat and contributors. pytorch-grad-cam: Advanced ai explainability for PyTorch, 2021. URL `https://github.com/jacobgil/pytorch-grad-cam`.

[18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

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

[20] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

[21] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[22] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[24] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Hk2aImxAb`.

[25] F. Ilhan, K.-H. Chow, S. Hu, T. Huang, S. Tekin, W. Wei, Y. Wu, M. Lee, R. Kompella, H. Latapie, G. Liu, and L. Liu. Adaptive deep neural network inference optimization with eenet. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1362–1371, 2024. doi: 10.1109/WACV57701.2024.00140.

[26] H. Jin and Y. Wu. Ce-collm: Efficient and adaptive large language models through cloud-edge collaboration. In *2025 IEEE International Conference on Web Services (ICWS)*, pages 316–323. IEEE, 2025.

[27] J. Kim, S. Kim, J. Kong, and S. Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

[28] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 09 2020. doi: 10.48550/arXiv.2009.07896.

[29] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[30] E. Lattanzi, C. Contoli, and V. Freschi. Do we need early exit networks in human activity recognition? *Engineering Applications of Artificial Intelligence*, 121:106035, 2023.

[31] C. Lin, Z. Chen, Z. Zhang, and J. Liu. E3: Early exiting with explainable AI for real-time and accurate DNN inference in edge-cloud systems. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 385–397, 2025.

[32] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[33] L. Meronen, M. Trapp, A. Pilzer, L. Yang, and A. Solin. Fixing overconfidence in dynamic neural networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2680–2690, 2024.

[34] R. G. Pacheco, K. Bochie, M. S. Gilbert, R. S. Couto, and M. E. M. Campista. Towards edge computing using early-exit convolutional neural networks. *Information*, 12(10):431, 2021.

[35] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck. Online and linear-time attention by enforcing monotonic alignments. In *International conference on machine learning*, pages 2837–2846. PMLR, 2017.

[36] H. Rahmath P, V. Srivastava, and K. Chaurasia. A strategy to accelerate the inference of a complex deep neural network. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 57–68. Springer, online, 2023.

[37] H. Rahmath P, K. Chaurasia, A. Gupta, and V. Srivastava. Hypergcn – a multi-layer multi-exit graph neural network to enhance hyperspectral image classification. *International Journal of Remote Sensing*, 45(14):4848–4882, 2024. doi: 10.1080/01431161.2024.2370501.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[39] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[40] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[41] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Tran, Y. Tay, and D. Metzler. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[44] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[45] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[46] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[47] S. Teerapittayanon, B. McDanel, and H.-T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE, 2016.

[48] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 409–424, Munich, Germany, 2018. Springer.

[49] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Advances in neural information processing systems*, 34:11960–11973, 2021.

[50] W. Wenjian, X. Qian, X. Jun, and H. Zhikun. DynamicSleepNet: A multi-exit neural network with adaptive inference time for sleep stage classification. *Frontiers in Physiology*, 14:1171467, 2023.

[51] M. Wołczyk, B. Wójcik, K. Bałazy, I. T. Podolak, J. Tabor, M. Śmieja, and T. Trzcinski. Zero time waste: Recycling predictions in early exit neural networks. *Advances in Neural Information Processing Systems*, 34:2516–2528, 2021.

[52] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.204. URL `https://aclanthology.org/2020.acl-main.204/`.

[53] J. Xin, R. Tang, Y. Yu, and J. Lin. Berxit: Early exiting for BERT with better fine-tuning and extension to regression. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume*, pages 91–104, Online, 2021. ACL.

[54] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

[55] H. Yu, D. Liu, Z. Zhang, and J. Wang. A dynamic transformer network with early exit mechanism for fast detection of multiscale surface defects. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10, 2023.

[56] W. Zhou, C. Xu, T. Ge, J. McAuley, K. Xu, and F. Wei. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems*, 33:18330–18341, 2020.

Appendix

This appendix provides extended technical details, derivations, and additional results supporting the main paper. Appendix A reports supporting tables for RQ3, including threshold sensitivity and robustness under distribution shift. Appendix B details the proxy predictor, including implementation, comparisons of Random Forest, XGBoost, and MLP regressors, and a proxy fidelity analysis. Implementation details for Grad-CAM, Integrated Gradients, SmoothGrad, and Occlusion baselines are given in Appendix C, while formal definitions of additional attribution evaluation metrics (MFA, convergence IoU, and deletion/insertion AUC) are provided in Appendix D. Appendix E extends the attribution stability analysis for RQ4, while Appendix F validates the attribution-agnostic nature of PFAM, and Appendix G examines the effect of threshold selection on accuracy–latency trade-offs.

Further analyses include Appendix H (correlation analysis of IEES components), Appendix I (distributional statistics of exits under different datasets and policies), and Appendix J (ablations of IEES design choices). Appendix K reports a small-scale user study evaluating PFAM visualizations and IEES-guided exits, while Appendix L presents additional qualitative comparisons. Finally, Appendix M provides calibration details and Appendix N describes training settings. Appendix B.1 gives per-exit results for the proxy supervision targets, Appendix O describes attribution for transformer and sequence models, and Appendix P reports sensitivity to the top-k mask. Together, these sections provide the full technical foundation for the results reported in the main paper.

A Threshold Sensitivity and Robustness (RQ3)

We provide the full tables supporting RQ3, including threshold perturbation (Table 5) and robustness under distribution shift (Table 4). These results are presented here for completeness, with detailed interpretation in Section 5.3.

Table 4: IEES performance under distribution shift (CIFAR-10-C; severity level 3). Thresholds are calibrated on clean data and held fixed. Accuracy decreases under all corruptions while the average exit index increases. See Section 5.3 for interpretation.

Corruption Type	**Accuracy (%)**	**Avg. Exit Index**
None (Clean)	89.77 ± 0.02	2.10 ± 0.02
Motion Blur	73.34 ± 0.06	2.91 ± 0.03
Brightness	81.22 ± 0.05	2.63 ± 0.02
Gaussian Noise	68.55 ± 0.07	3.02 ± 0.04

Table 5: Accuracy (%) and average inference time under threshold perturbation ($\tau_i^* \pm 10\%$). Results are shown for ResNet-18 on CIFAR-10, MSDNet on CIFAR-100, and MobileNetV3 on ImageNet. See Section 5.3 for interpretation.

Policy (Threshold)	**Exit Rate (%)**				**Per-Exit Accuracy (%)**				**Overall Acc.**	**Avg. Time (ms)**
	E1	E2	E3	E4	E1	E2	E3	E4		
CIFAR-10 (ResNet-18)										
IEES ($\tau_i^* - 10\%$)	30.72	51.44	13.15	4.69	88.02	88.96	89.44	73.55	88.01 ± 0.04	3.00 ± 0.03
IEES (τ_i^*)	20.18	55.92	17.26	6.64	90.01	91.12	90.04	77.01	89.77 ± 0.02	3.14 ± 0.02
IEES ($\tau_i^* + 10\%$)	13.96	53.97	11.79	20.28	91.49	90.94	91.15	77.87	88.39 ± 0.03	3.95 ± 0.01
CIFAR-100 (MSDNet)										
IEES ($\tau_i^* - 10\%$)	25.91	25.73	30.07	18.29	78.89	79.94	79.51	64.46	76.71 ± 0.04	26.00 ± 0.05
IEES (τ_i^*)	20.08	23.92	28.89	27.11	81.71	83.19	81.13	66.94	77.89 ± 0.03	27.59 ± 0.02
IEES ($\tau_i^* + 10\%$)	18.78	22.93	26.23	32.06	81.58	82.41	81.04	66.92	76.93 ± 0.04	29.81 ± 0.01
ImageNet (MobileNetV3)										
IEES ($\tau_i^* - 10\%$)	21.91	28.73	26.07	23.29	78.89	78.94	79.51	65.46	75.94 ± 0.04	27.90 ± 0.05
IEES (τ_i^*)	16.83	25.82	26.94	30.41	81.12	80.34	81.01	68.05	76.91 ± 0.01	29.49 ± 0.02
IEES ($\tau_i^* + 10\%$)	14.78	21.93	22.03	41.26	82.67	79.41	81.44	67.92	75.60 ± 0.04	31.81 ± 0.01

Table 6: Proxy fidelity on held-out test data (averaged across exits). Higher R^2 and ρ indicate stronger approximation.

Dataset / Backbone	R^2	Spearman ρ	MAE	RMSE
CIFAR-10 / ResNet-18	0.87	0.91	0.069	0.077
CIFAR-100 / MSDNet	0.84	0.88	0.081	0.092
ImageNet / MobileNetV3	0.80	0.85	0.072	0.084

B Proxy Predictor Details

We evaluated lightweight regressors as candidate proxies to approximate the oracle IEES score at inference. The input feature vector comprises ten attribution-free statistics organized in two groups.

Single-exit features (5): softmax confidence, logit margin, prediction entropy, and the mean and maximum activations from the exit branch. These features correlate with semantic confidence and are inexpensive to compute from the forward pass.

Cross-exit features (5): confidence delta $|S_i(x) - S_{i-1}(x)|$, predicted class stability $\mathbf{1}[\hat{y}_i = \hat{y}_{i-1}]$, logit cosine similarity $Q_i^{\text{logit}}(x)$ (as defined in Appendix J.1, used here as an input feature rather than a score component), entropy delta $|H_i(x) - H_{i-1}(x)|$, and mean activation delta $|\bar{a}_i - \bar{a}_{i-1}|$. These features provide the proxy with direct signals about cross-exit prediction stability, bridging the gap between single-exit inputs and the cross-exit stability term $Q_i(x)$. For i=1, cross-exit features are set to zero.

All features are computable from the forward pass; cross-exit features additionally require storing the previous exit's logits and scalar activation statistics, incurring negligible memory overhead. The target output is the calibrated oracle IEES score computed offline at each exit. Proxies were trained on the training set and validated on a held-out 10% split, with fidelity reported on disjoint test data to avoid leakage.

We compared Random Forest (RF), shallow multi-layer perceptrons (MLPs), and XGBoost regressors. Across datasets, RF provided a favorable fidelity–latency trade-off: an RF with 100 trees (maximum depth 8, minimum leaf size 5) achieved high fidelity with low additional overhead. MLPs exhibited weaker generalization in low-data regimes, and XGBoost required stronger regularization without outperforming RF. We therefore adopt RF as the default proxy.

Table 6 reports fidelity on CIFAR-10, CIFAR-100, and ImageNet, averaged across exits. In all cases, proxies achieve high agreement with oracle IEES, with $R^2 \geq 0.80$, Spearman $\rho \geq 0.85$, and MAE between 0.069 and 0.081. These results show strong agreement with the oracle IEES score, enabling accurate attribution-free approximation at inference. When substituted for the oracle, the proxy incurred accuracy differences below 0.3% across datasets, and exit allocations deviated by less than 2% even under distribution shifts (e.g., CIFAR-10-C).

In terms of runtime, RF proxy inference adds approximately 0.1 ms per sample, compared to 0.4–0.5 ms for attribution-based IEES computation. This represents a substantial reduction in overhead, making proxies practical for deployment in early-exit systems where efficiency and reliability must coexist.

B.1 Proxy Supervision Baselines

To isolate the contribution of IEES as a supervision target, we evaluate four variants (one baseline without a proxy and three Random Forest proxies) using identical input features but different supervision signals on CIFAR-10/ResNet-18. All proxy variants use the same RF architecture and attribution-free feature vector described above.

- **Confidence Only**: standard max-probability threshold; no proxy is trained.
- **Conf. Recalib RF**: RF trained to predict softmax confidence, evaluating whether recalibrating confidence alone improves routing.
- **Correctness RF**: RF trained to predict binary correctness, evaluating whether a learned uncertainty estimate suffices.

Table 7: Proxy supervision baselines on CIFAR-10/ResNet-18. All variants use identical RF architecture and attribution-free input features; only the supervision target differs. Exit rate and per-exit accuracy (%) are reported alongside overall accuracy and average inference time.

Supervision Target	Exit Rate (%)				Accuracy (%)				Overall Acc.	Time (ms)
	E1	E2	E3	E4	E1	E2	E3	E4		
Confidence only	48.02	30.24	9.35	12.39	88.91	89.57	89.71	74.12	87.35 ± 0.03	3.05
Conf. Recalib RF	37.72	38.82	12.70	10.76	89.03	90.43	89.01	74.23	87.97 ± 0.04	3.08
Correctness RF	22.29	53.55	16.10	8.06	89.14	89.02	89.12	76.11	88.02 ± 0.01	3.13
IEES RF (Ours)	**20.18**	**55.92**	**17.26**	**6.64**	**90.01**	**91.12**	**90.04**	**77.01**	$\mathbf{89.77 \pm 0.02}$	**3.14**

- **IEES RF (Ours)**: RF trained to predict the oracle IEES score, incorporating attribution strength $F_i(x)$ and stability $Q_i(x)$.

Results are reported in Table 7. Recalibrated-confidence and correctness supervision both yield modest improvements over raw confidence (87.97% and 88.02% vs. 87.35%), whereas IEES supervision achieves 89.77%, the highest accuracy among the variants. This suggests that the proxy is not merely learning a better uncertainty estimate; performance improves as the supervision signal becomes more informative.

Three factors likely contribute to this improvement. First, continuous IEES regression preserves reliability ordering across exits — binary correctness collapses distinct reliability levels into a single label, discarding gradient information. Second, confidence-based supervision reflects information already present in the input features, providing limited additional signal beyond what the proxy could infer directly. Third, attribution-based stability captures complementary signals beyond prediction-space metrics, as supported by Table 11 (Appendix J.1), where a logit-stability variant achieves lower performance (88.05% vs. 89.77% on CIFAR-10). Appendix H shows that F and Q exhibit complementary correlations with accuracy (0.38–0.51 and -0.28 to -0.37, respectively), alongside confidence (0.58–0.69), while Appendix I confirms clearer exit/continue separation when attribution signals are incorporated, particularly in ambiguous confidence regimes.

C Baseline Attribution Methods

To enable fair comparison with PFAM, we adapt four widely used attribution techniques—Integrated Gradients (IG), SmoothGrad, Grad-CAM, and Occlusion—to operate on early exits. For each exit E_i, we construct a sub-model by truncating the backbone up to that exit's classifier and apply the attribution method independently, targeting the predicted class logit. This yields a separate attribution map for each exit and input.

All baselines are implemented with the Captum library [28] using default parameters unless otherwise specified. IG uses 50 interpolation steps; SmoothGrad samples 25 noisy variants with Gaussian noise ($\sigma = 0.2$); Occlusion employs 15×15 sliding windows with stride 8; and Grad-CAM targets the last convolutional layer before each exit. All heatmaps are resized to 224×224 and normalized per sample for consistency.

In contrast, PFAM operates natively within the early-exit framework. It reuses intermediate activations computed during the forward pass and requires only a lightweight backward step at selected exits, implemented via registered hooks. Unlike baseline methods, PFAM does not require constructing truncated sub-models and produces both exit-wise and cumulative maps within a unified procedure. Since IEES (or its proxy) governs exit decisions, PFAM visualizations are optional at test time and can be invoked for auditing or analysis.

D Additional Attribution Evaluation Metrics

We quantify attribution quality across early exits using three complementary metrics: (i) *semantic progression*, measured by Monotonic Feature Alignment (MFA, ↑); (ii) *convergence*, measured by IoU with the final attribution (↑); and (iii) *faithfulness*, assessed using normalized deletion and insertion AUC (↓/↑).

Our design follows established principles of robustness and consistency in explanations [2, 54] and leverages standard perturbation-based faithfulness protocols [40, 21]. Classical pixel-level change metrics, including cumulative Δ-stability [2], stepwise L1 differences, and Earth Mover's Distance (EMD), can yield degenerate values for PFAM (e.g., 0 or nan), since PFAM produces progressively normalized maps with stable mass distribution. We therefore do not adopt these metrics.

D.1 Monotonic Feature Alignment (MFA)

Monotonic Feature Alignment (MFA) quantifies whether intermediate exit attributions progressively align with the final attribution. Let A_i denote the attribution map at exit i, and A_L the final attribution. Similarity is measured via cosine similarity:

$$s_i = \cos(A_i, A_L), \tag{11}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity computed between the flattened attribution maps. MFA is then defined as:

$$\mathrm{MFA} = \mathbb{1}\big[s_1 \leq s_2 \leq \cdots \leq s_{L-1}\big], \tag{12}$$

which equals 1 if similarity increases monotonically across exits, and 0 otherwise. This reflects the intuition that explanations should progressively align with the final attribution map. The formulation is conceptually related to monotonic alignment mechanisms in sequence modeling [35, 27], adapted here to progressive attribution refinement.

D.2 Convergence IoU to Final

We further measure semantic alignment by computing the overlap between each exit attribution and the final attribution. Let $\mathrm{TopK}(A_i)$ denote the set of top-$k\%$ salient pixels in A_i, selected by thresholding at the $(1 - k/100)$ quantile. The convergence IoU at exit i is:

$$\mathrm{IoU}(A_i, A_L) = \frac{|\mathrm{TopK}(A_i) \cap \mathrm{TopK}(A_L)|}{|\mathrm{TopK}(A_i) \cup \mathrm{TopK}(A_L)|}. \tag{13}$$

We report the average IoU across all non-final exits. Higher values indicate stronger alignment with the semantic focus of the final attribution. IoU-based measures are commonly used to evaluate saliency map localization quality [9, 8].

MFA and IoU capture complementary aspects of attribution quality. MFA reflects whether similarity to the final attribution increases monotonically (a directional measure), whereas IoU measures spatial overlap of salient regions. As a result, high MFA does not necessarily imply high IoU. For example, an attribution method may exhibit consistent directional alignment with the final map while remaining spatially diffuse, leading to low IoU despite perfect MFA.

D.3 Faithfulness: Deletion and Insertion AUCs

Faithfulness evaluates whether perturbing features ranked highly by the attribution impacts the prediction in the expected direction [40, 21]. For input x, class y, and attribution map A, pixels (or superpixels) are ranked by attribution score and progressively perturbed. At fraction $f \in [0, 1]$, we denote the perturbed input as x_f and the predicted probability as $p_y(x_f)$. The normalized trajectory is:

$$\hat{p}_y(f) = \frac{p_y(x_f) - p_y(x_{\text{base}})}{p_y(x) - p_y(x_{\text{base}}) + \epsilon}, \tag{14}$$

where x_{base} is the mean image across the dataset and $\epsilon = 10^{-8}$ avoids numerical instability when the denominator approaches zero. The area under this curve (AUC) is:

$$\mathrm{AUC} = \int_0^1 \hat{p}_y(f)\, df. \tag{15}$$

Lower AUC in deletion (removing salient pixels should reduce confidence) and higher AUC in insertion (adding salient pixels should increase confidence) indicate greater faithfulness of the attribution.

Table 8: Attribution stability and faithfulness metrics across datasets. Higher MFA and IoU, lower deletion AUC, and higher insertion AUC indicate improved attribution quality. Grad-CAM, IG, SmoothGrad, and Occlusion are single-level maps computed independently at each exit (Appendix C), whereas PFAM produces cumulative cross-exit maps. Values are mean ± std over 100 samples. For MFA, which is a binary indicator function, std reflects sample-level outcome variance $\sqrt{p(1-p)}$. Best values in each row are bolded.

Dataset / Backbone	Metric	Grad-CAM	IG	SmoothGrad	Occlusion	PFAM
CIFAR-10 / ResNet-18	MFA (↑)	**1.00 ±0.00**	0.75 ±0.44	0.74 ±0.44	0.91 ±0.29	0.96 ±0.20
	IoU (↑)	0.20 ±0.00	0.65 ±0.07	0.36 ±0.06	0.71 ±0.15	**0.93 ±0.02**
	Deletion AUC (↓)	0.35 ±0.11	**0.19 ±0.19**	0.31 ±0.15	0.25 ±0.14	0.28 ±0.14
	Insertion AUC (↑)	0.55 ±0.13	0.58 ±0.26	0.56 ±0.17	0.62 ±0.16	**0.63 ±0.18**
CIFAR-100 / MSDNet	MFA (↑)	0.53 ±0.50	0.42 ±0.50	0.48 ±0.50	0.62 ±0.49	**0.95 ±0.22**
	IoU (↑)	0.30 ±0.13	0.50 ±0.12	0.33 ±0.11	0.58 ±0.23	**0.92 ±0.03**
	Deletion AUC (↓)	0.16 ±0.15	0.15 ±0.17	0.17 ±0.14	0.17 ±0.14	**0.14 ±0.15**
	Insertion AUC (↑)	0.46 ±0.20	0.27 ±0.23	0.43 ±0.19	0.45 ±0.17	**0.50 ±0.21**
ImageNet / MobileNetV3	MFA (↑)	0.63 ±0.49	0.51 ±0.50	0.88 ±0.33	0.37 ±0.49	**0.97 ±0.17**
	IoU (↑)	0.21 ±0.09	0.31 ±0.06	0.32 ±0.05	0.27 ±0.09	**0.91 ±0.02**
	Deletion AUC (↓)	0.17 ±0.09	**0.11 ±0.06**	0.14 ±0.09	0.12 ±0.08	0.14 ±0.12
	Insertion AUC (↑)	0.29 ±0.11	0.16 ±0.10	0.29 ±0.13	0.31 ±0.13	**0.40 ±0.15**

E Attribution Stability Analysis (RQ4)

We present the full quantitative evaluation corresponding to RQ4 (Section 5.4), focusing on attribution stability and progression across exits. Our analysis uses the semantic progression metrics defined in Appendix D—Monotonic Feature Alignment (MFA, ↑) and convergence IoU to the final attribution (↑)—to assess how attributions evolve and align across exits. In addition, we report normalized deletion and insertion AUCs to evaluate attribution faithfulness, using a mean image as the baseline x_{base} in Eq. 14. Together, these metrics quantify whether attributions evolve progressively, align with the final attribution map, and highlight features important for model predictions.

Results. Across CIFAR-10, CIFAR-100, and ImageNet, PFAM achieves higher convergence IoU than all baseline methods and competitive MFA in most settings. On CIFAR-100 with MSDNet, PFAM reaches MFA = 0.95 and IoU = 0.92, whereas baselines range between 0.3–0.6. On ImageNet, PFAM similarly attains MFA = 0.97 and IoU = 0.91, exceeding all alternatives. On CIFAR-10, Grad-CAM achieves a perfect MFA of 1.00, while PFAM attains 0.96; however, PFAM substantially outperforms Grad-CAM on IoU (0.93 vs. 0.20), indicating that while Grad-CAM's exit-wise maps align directionally with the final attribution, they lack spatial precision. These results show that PFAM produces attribution maps that evolve smoothly and align spatially with the final attribution, whereas baseline methods often exhibit less stable or less localized progression.

Faithfulness metrics further indicate that PFAM achieves competitive or improved scores across settings. For instance, on CIFAR-10 with ResNet-18, PFAM achieves an insertion AUC of 0.63 compared to 0.55 for Grad-CAM and 0.56 for SmoothGrad. On deletion AUC, IG and Occlusion achieve stronger scores on CIFAR-10 (0.19 and 0.25 vs. 0.28) and ImageNet (0.11 and 0.12 vs. 0.14), while PFAM remains competitive across all three datasets. Overall, these findings quantitatively support PFAM's property of progressive refinement across exits. Attribution stability analysis shows that PFAM provides progressive, stable, and faithful explanations across exits, supporting its use in explanation-aware early-exit systems.

F Attribution-Agnostic Validation of PFAM

PFAM is designed to be attribution-agnostic (Section 3.2), accepting any differentiable attribution method as its instantiation. While our primary experiments adopt Grad-CAM for efficiency, we validate this property empirically by evaluating PFAM instantiated with Integrated Gradients (IG, 50 interpolation steps) [45] on CIFAR-10/ResNet-18. All other components of IEES remain unchanged.

As shown in Table 9, both instantiations produce similar exit distributions and outperform confidence-only exits (87.35%, Table 1 of the main paper), supporting PFAM's attribution-agnostic design. PFAM (Grad-CAM) achieves slightly higher overall accuracy (89.77% vs. 89.00%). This difference may reflect the effect of attribution method choice on the quality of offline IEES supervision labels,

as IG and Grad-CAM produce different attribution maps that yield different calibrated IEES scores used to train the proxy. The similar exit distributions further suggest that the cumulative aggregation structure of PFAM, rather than the specific attribution method, plays a key role in guiding exit decisions.

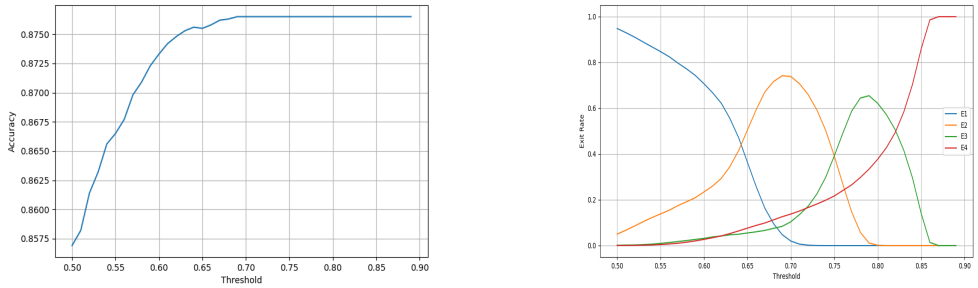
Table 9: Attribution-agnostic evaluation of PFAM on CIFAR-10/ResNet-18. PFAM is instantiated with Integrated Gradients [45] and Grad-CAM [42] respectively. Exit rate and per-exit accuracy (%) are reported alongside overall accuracy (mean ± std over 5 runs). Both outperform confidence-only exits (87.35 ± 0.03, Table 1).

PFAM Instantiation	**Exit Rate (%)**				**Accuracy (%)**				**Overall Acc.**
	E1	E2	E3	E4	E1	E2	E3	E4	
PFAM (IG)	19.95	56.06	15.27	8.72	89.36	90.55	89.54	77.32	89.00 ± 0.02
PFAM (Grad-CAM)	**20.18**	**55.92**	**17.26**	**6.64**	**90.01**	**91.12**	**90.04**	**77.01**	$\mathbf{89.77 \pm 0.02}$

G Effect of Threshold Selection on Accuracy and Exit Distribution

To complement the quantitative sensitivity results in Section 5.3, we visualize how varying the decision threshold affects IEES accuracy and exit allocation. Figure 4(a) shows the threshold–accuracy curve on CIFAR-10 (ResNet-18): increasing thresholds delays early exits, improving accuracy up to an optimal point ($\tau^* = 0.72$), beyond which gains plateau. Figure 4(b) shows the corresponding change in exit behavior: lower thresholds trigger earlier exits, while higher thresholds defer decisions to deeper branches.

These trends illustrate that IEES enables controllable trade-offs between efficiency and decision quality. In practice, calibration allows balancing predictive accuracy and computational cost by selecting thresholds that align with deployment requirements.



(a) Threshold vs. Accuracy

(b) Threshold vs. Exit Rate

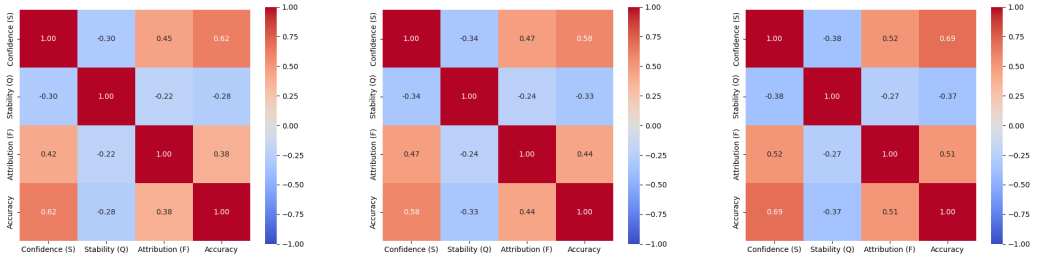
Figure 4: Effect of threshold variation on IEES performance for ResNet-18 on CIFAR-10. (a) Accuracy increases with stricter thresholds until saturation near the optimal point. (b) Exit distribution shifts progressively deeper as thresholds rise, reflecting increased decision confidence.

H Correlation Analysis of IEES Components

IEES integrates three complementary signals—confidence (S), attribution strength (F), and stability (Q)—to support accurate and feature-grounded early-exit decisions. Correlation matrices (Figure 5) across CIFAR-10, CIFAR-100, and ImageNet characterize the relationships of these components with prediction accuracy and with each other.

Key Observations:

- **Confidence (S):** Confidence shows a strong positive correlation with accuracy across all datasets (e.g., 0.62 on CIFAR-10, 0.58 on CIFAR-100, and 0.69 on ImageNet), indicating its importance in



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet

Figure 5: Correlation matrices of IEES components across CIFAR-10, CIFAR-100, and ImageNet. Confidence and attribution strength are positively correlated with accuracy, while stability shows the expected negative correlation, confirming their complementary roles in guiding early-exit decisions.

guiding exit decisions. As expected, S is positively associated with F (0.42–0.52) and negatively with Q (-0.30 to -0.38), suggesting that confident predictions tend to coincide with stronger activations and more stable attributions.

- **Attribution Strength (F):** Attribution strength also correlates positively with accuracy (0.38 on CIFAR-10, 0.44 on CIFAR-100, 0.51 on ImageNet). While its correlation is lower than that of S, it complements confidence by reflecting whether predictions are supported by class-relevant activations. Its moderate negative correlation with Q (-0.22 to -0.27) suggests that stronger activations are often associated with more stable attribution patterns.
- **Stability (Q):** Stability correlates negatively with accuracy (-0.28, -0.33, and -0.37), consistent with its role in penalizing unstable attribution patterns. This behavior indicates that exits associated with greater attribution variability tend to be less reliable.

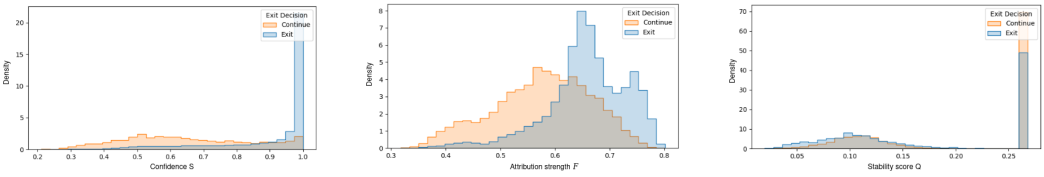
Dataset-Specific Trends: The strength of correlations varies across datasets. On CIFAR-10, correlations are moderate, reflecting the relatively lower complexity of the dataset. CIFAR-100 shows clearer separation between signals, while ImageNet exhibits stronger alignment between S, F, and accuracy, with Q acting as a consistent negative indicator. This progression suggests that dataset-dependent calibration can help balance accuracy and computational efficiency.

Conclusion: The three IEES components exhibit distinct yet complementary relationships with prediction accuracy: confidence correlates strongly with correctness, attribution strength reflects feature alignment, and stability penalizes variability across exits. Their combination provides a structured basis for dynamic exit decisions across datasets of varying complexity.

I Distribution Analysis of IEES Components

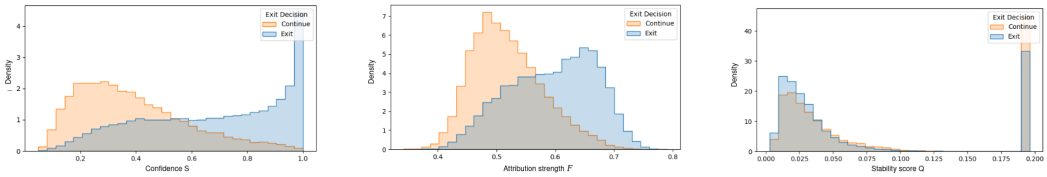
We analyze the distributions of confidence, attribution strength, and attribution stability across "Exit" and "Continue" decisions (Figures 6–8). This analysis highlights how each component contributes to guiding early exits and complements the others in shaping decision behavior.

Key Observations:



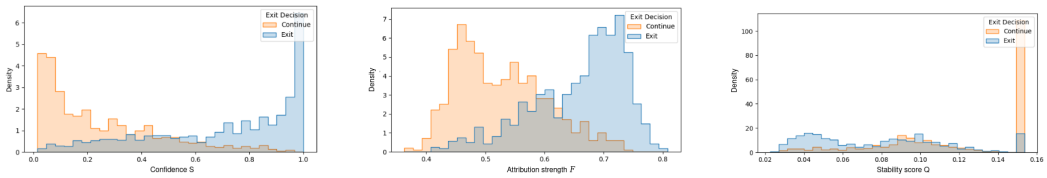
(a) Confidence (b) Attribution Strength (c) Attribution Stability

Figure 6: Distribution analysis of IEES components on CIFAR-10 using ResNet-18 (confidence, attribution strength, and attribution stability) across exit decisions (Exit vs. Continue).



(a) Confidence (b) Attribution Strength (c) Attribution Stability

Figure 7: Distribution analysis of IEES components on CIFAR-100 using MSDNet (confidence, attribution strength, and attribution stability) across exit decisions (Exit vs. Continue).



(a) Confidence (b) Attribution Strength (c) Attribution Stability

Figure 8: Distribution analysis of IEES components on ImageNet using MobileNetV3 (confidence, attribution strength, and attribution stability) across exit decisions (Exit vs. Continue).

- **Confidence as an Initial Signal:** Confidence distributions show clear separation at high values, indicating that highly confident samples are more likely to exit early. However, overlap in the mid-range (0.4–0.7) creates an ambiguity region, where confidence alone may lead to premature or unstable exits. This motivates the inclusion of additional signals.

- **Attribution Strength as a Refinement Signal:** Attribution strength (see Figures 6(b), 7(b), 8(b)) refines decision behavior by favoring exits when features exhibit stronger class-relevant activation. Its effect varies across datasets: sharper separation in CIFAR-10, broader distributions in CIFAR-100 (reflecting higher task complexity), and higher effective thresholds in ImageNet.

- **Attribution Stability as a Reliability Filter:** Attribution stability (see Figures 6(c), 7(c), 8(c)) acts as a safeguard against unstable exits. Lower values (stable attributions) are more frequently associated with Exit decisions, whereas higher values (unstable attributions) are more often associated with Continue decisions. This pattern suggests that stability helps defer decisions when attribution patterns fluctuate.

Interactions and IEES Justification: The joint behavior of these components helps explain the effectiveness of IEES. Confidence encourages early exits but becomes less reliable in ambiguous regions. Attribution strength complements it by favoring exits supported by class-relevant activations, while attribution stability discourages exits when attribution patterns vary across depths. Together, these signals form a composite score that is more robust than confidence-only baselines.

Dataset-Specific Thresholding: Exit behavior varies across datasets. In CIFAR-10, clearer separation in confidence and attribution strength supports earlier exits. In CIFAR-100, broader distributions reflect increased difficulty, requiring stronger activation signals before exiting. On ImageNet, distributions shift upward across all components, consistent with higher semantic complexity and deeper architectures. These trends suggest the need for dataset-dependent threshold selection rather than fixed exit criteria.

Conclusion: Distribution analysis shows that confidence alone is insufficient to characterize reliable early exits. Attribution strength provides additional evidence of class-relevant activation, while attribution stability helps identify uncertain cases. The observed dataset-dependent patterns further indicate that IEES adapts across architectures and datasets more effectively than fixed confidence-based policies.

J Ablation of IEES Components

IEES integrates three complementary signals: (i) Confidence (S), capturing exit-wise certainty; (ii) Attribution strength (F), reflecting class-relevant activation patterns; and (iii) Attribution stability (Q), capturing variability of attributions across exits. To assess their individual contributions, we ablate components on CIFAR-10 (ResNet-18). Thresholds for each variant are tuned on a held-out validation set, and results are averaged over 5 runs (mean $\pm$ std).

Table 10: Ablation of IEES components on CIFAR-10 (ResNet-18). Accuracy and latency are averaged over 5 runs. Latency differences reflect earlier or deferred exit allocations depending on the available signals.

Exit Signal	**Accuracy (%)**	**Latency (ms)**
Confidence only (S)	87.35 ± 0.03	3.05 ± 0.02
Confidence + Stability ($S + Q$)	88.90 ± 0.04	3.20 ± 0.03
Confidence + Attribution ($S + F$)	88.10 ± 0.03	2.98 ± 0.02
Full IEES ($S + F + Q$)	$\mathbf{89.77 \pm 0.02}$	$\mathbf{3.14 \pm 0.02}$

Results show that confidence alone yields the lowest performance. Adding stability ($S + Q$) improves accuracy by reducing unstable exit decisions, with a modest increase in latency. Adding attribution strength ($S + F$) also improves accuracy and slightly reduces latency, suggesting that stronger activation patterns enable earlier confident exits. The full combination ($S + F + Q$) achieves the best overall performance, indicating that the three signals provide complementary information for improving accuracy and efficiency relative to confidence-only policies.

J.1 Ablation: Logit-Based Stability vs. Attribution-Based Stability

Stability is a central component of the IEES formulation. To isolate the contribution of attribution-level stability, we evaluate an alternative that replaces attribution-based stability with consistency measured directly in logit space. This logit-stability variant retains the overall score structure (confidence and attribution-strength terms), modifying only the stability component:

$$Q_i^{\text{logit}}(x) = \frac{\mathbf{z}_i(x)^\top \mathbf{z}_{i-1}(x)}{\|\mathbf{z}_i(x)\|_2 \, \|\mathbf{z}_{i-1}(x)\|_2}.$$

Since Q_i^{logit} is a similarity measure (larger values indicate greater stability), it enters Eq. 5 with a positive sign to reward stability, in contrast to the attribution-based $Q_i(x)$ which is subtracted as an instability penalty:

$$\text{IEES}_i^{\text{logit}}(x) = w_1 S_i(x) + w_2 F_i(x) + w_3 Q_i^{\text{logit}}(x). \tag{16}$$

Both variants are calibrated using the same validation-based grid search to select exit thresholds. Results across CIFAR-10, CIFAR-100, and ImageNet are summarized in Table 11. Logit-based stability provides moderate improvements over confidence-only routing on CIFAR-10 and CIFAR-100, indicating that prediction-space smoothness is a useful signal. However, it consistently underperforms IEES, which incorporates stability in attribution patterns rather than only in logits. These results suggest that attribution-level stability offers a more informative signal for early-exit decisions.

Table 11: Ablation comparing IEES with attribution-based stability to a logit-stability variant. All components except the stability term remain identical.

Dataset / Backbone	**Method**	E1	E2	E3	E4	**Overall Acc.**	**Time (ms)**
CIFAR-10 / ResNet-18	Logit-Stability	89.30	89.85	89.60	75.00	88.05	3.07
	IEES (Attrib-Stability)	90.01	91.12	90.04	77.01	**89.77**	3.14
CIFAR-100 / MSDNet	Logit-Stability	80.00	80.80	80.70	64.10	77.45	25.55
	IEES (Attrib-Stability)	81.71	83.19	81.13	66.94	**77.89**	27.59
ImageNet / MobileNetV3	Logit-Stability	77.10	77.50	77.60	65.70	75.00	27.65
	IEES (Attrib-Stability)	81.12	80.34	81.01	68.05	**76.91**	29.49

Table 12: Ablation of PFAM vs. raw per-exit Grad-CAM within IEES on CIFAR-10/ResNet-18. All other IEES components remain identical. Exit rate and per-exit accuracy (%) are reported alongside overall accuracy.

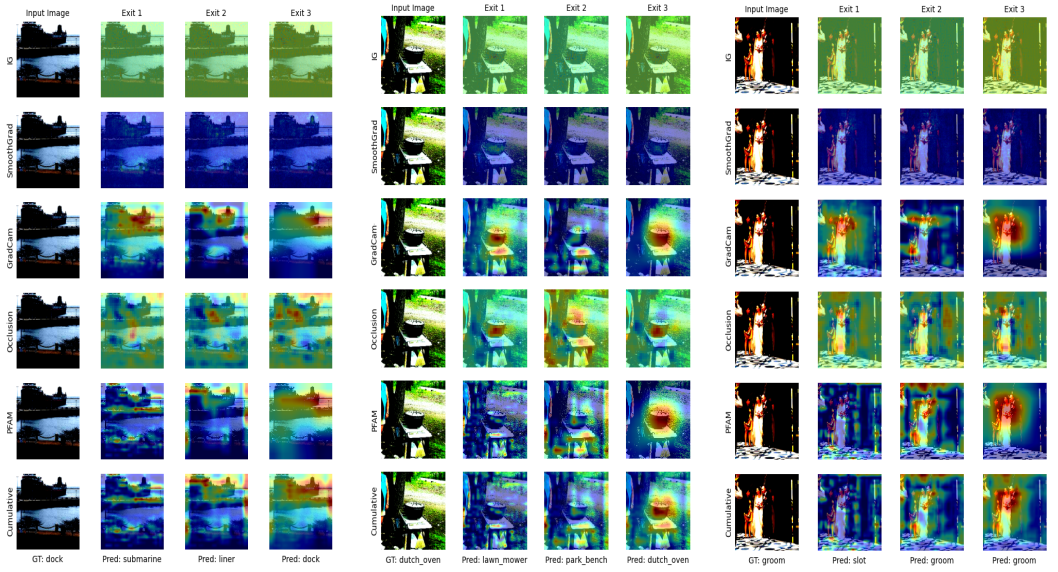
Method	**Exit Rate (%)**				**Accuracy (%)**				**Overall Acc.**
	E1	E2	E3	E4	E1	E2	E3	E4	
Confidence only	48.02	30.24	9.35	12.39	88.91	89.57	89.71	74.12	87.35 ± 0.03
IEES (Raw Grad-CAM)	29.72	38.82	19.45	12.01	89.81	90.04	86.26	75.12	87.44 ± 0.01
IEES (PFAM)	**20.18**	**55.92**	**17.26**	**6.64**	**90.01**	**91.12**	**90.04**	**77.01**	$\mathbf{89.77 \pm 0.02}$

J.2 Ablation: PFAM vs. Raw Per-Exit Attribution in IEES

PFAM's cumulative formulation is designed to stabilize both attribution strength $F_i(x)$ and stability $Q_i(x)$ across exits. To isolate its contribution, we replace PFAM's cumulative attribution maps with raw per-exit Grad-CAM within IEES on CIFAR-10/ResNet-18, keeping all other components identical. Since raw Grad-CAM lacks cumulative structure, we approximate stability using consecutive per-exit maps, isolating the effect of PFAM's progressive aggregation.

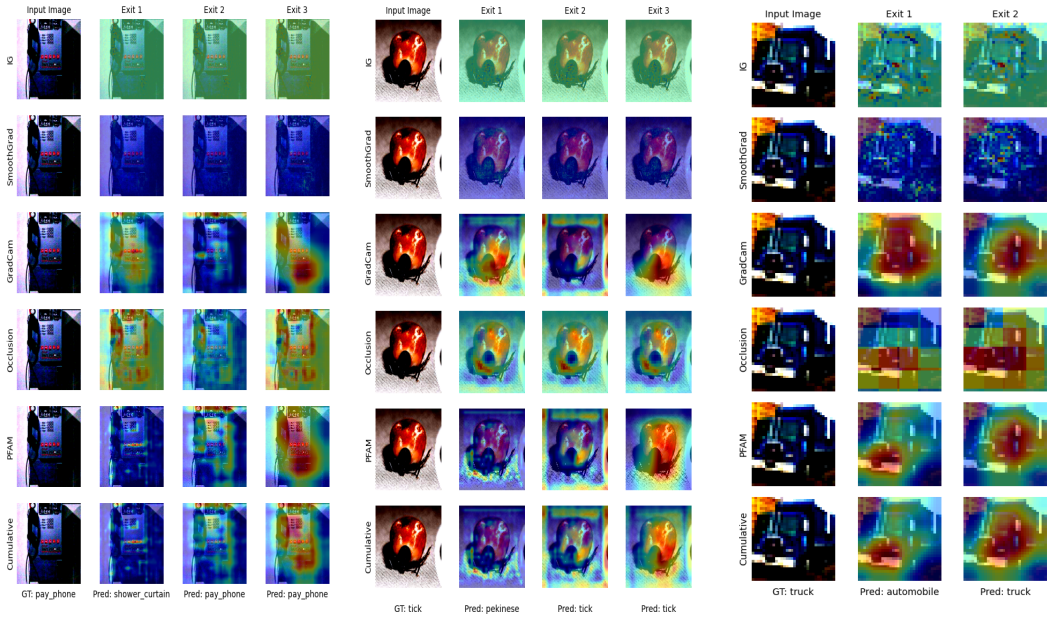
Results are reported in Table 12. Raw Grad-CAM achieves 87.44%, close to confidence-only routing (87.35%), while IEES with PFAM achieves 89.77% (+2.33%). Raw Grad-CAM lacks cumulative structure, which affects both components of IEES: the spatial mask $\Omega_k^i(x)$ used to compute $F_i(x)$ is derived from noisier per-exit maps rather than stable cumulative attributions, and the stability estimate $Q_i(x)$ exhibits higher variance as consecutive per-exit maps fluctuate more than cumulative ones. This leads to reduced performance particularly at deeper exits (E3: 86.26% vs. 90.04% for PFAM).

Table 8 (Appendix E) provides additional supporting evidence: standalone Grad-CAM achieves IoU = 0.20 compared to PFAM's 0.93 on CIFAR-10/ResNet-18, highlighting a substantial gap in spatial alignment that explains why raw per-exit attribution yields noisier estimates of both $F_i(x)$ and $Q_i(x)$. These results indicate that PFAM serves as a functional component within IEES, contributing to improved decision behavior compared to using raw per-exit attributions.



(a) (b) (c)

Figure 9: Extended comparison of PFAM with standard attribution methods (IG, SmoothGrad, Grad-CAM, Occlusion). PFAM consistently captures class-relevant regions across exits and provides more coherent visual explanations than baselines.



(a) (b) (c)

Figure 10: Comparison of attribution behavior across exits. PFAM produces stable and progressively refined attributions aligned with model depth, compared to baseline methods.

K User Study on Attribution Quality

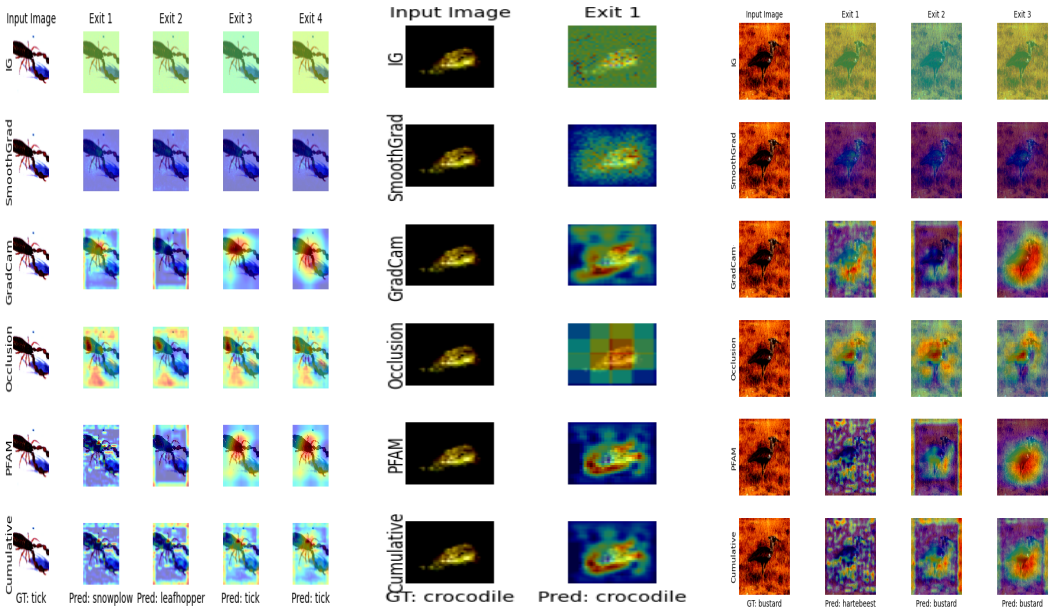
To complement quantitative metrics, we conducted a small-scale user study with 12 participants who had prior experience with visual recognition tasks (e.g., computer vision coursework or research). The study aimed to assess whether PFAM heatmaps were perceived as more informative and better aligned with predicted classes than standard attribution methods, and whether IEES-guided early exits appeared reasonable.

Participants were presented with samples from CIFAR-10, CIFAR-100, and ImageNet that were randomly selected in a class-stratified manner. Each sample included the model prediction, PFAM visualizations (exit-wise and cumulative), and matched baselines (Grad-CAM, IG, SmoothGrad, Occlusion). Participants were asked to select the heatmap that best localized class-relevant regions and to evaluate whether the corresponding exit decision appeared reasonable. Each participant evaluated 20 image comparisons, yielding 240 total evaluations (N=240) across the three datasets, which forms the basis of the reported binomial confidence interval.

Results. PFAM was preferred in 70–80% of comparisons across datasets (binomial 95% CI: 58–88%), and IEES exit decisions were judged reasonable in approximately 90% of cases. These findings provide supporting human-centered evidence that PFAM visualizations are perceived as clearer, and that IEES-guided decisions are generally aligned with user expectations.

All participants provided informed consent, no personal data were collected, and only non-sensitive tasks were involved. Participation was voluntary and uncompensated.

Participant Instructions. Participants were instructed as follows: “You will be shown model predictions along with attribution maps. Please select the visualization that best highlights the regions relevant to the predicted class, and indicate whether the corresponding early-exit decision appears reasonable.”



(a) (b) (c)

Figure 11: Examples under challenging or ambiguous predictions. PFAM maintains consistent attribution focus across exits, whereas baseline methods may exhibit greater variability.

L Extended Visual Comparison: PFAM vs. Traditional Attribution Methods

Figures 9–11 provide extended visual comparisons of PFAM against standard attribution methods (IG, SmoothGrad, Grad-CAM, Occlusion) across different exits. These examples expand upon Figure 3 in Section 5.4, illustrating attribution behavior, prediction correctness, and consistency across depth.

Misclassification and Attribution Behavior. Examples in Figure 9 illustrate PFAM's behavior under misclassification. In Figure 9(a), the model misclassifies at Exit 1; while IG and Grad-CAM produce diffuse or misplaced attention, PFAM highlights regions associated with the predicted class, making the failure more interpretable. In Figure 9(b), the model classifies correctly at Exit 1, and PFAM maintains relatively consistent attribution across exits, whereas SmoothGrad produces more fragmented heatmaps. Figure 9(c) further shows PFAM preserving attention alignment across exits.

Spatial Precision and Progressive Refinement. As shown in Figure 10, PFAM provides comparable or improved spatial precision relative to Grad-CAM, particularly at deeper exits where baseline methods may lose focus. In Figure 10(a), PFAM captures fine-grained attention aligned with salient features. In Figure 10(b), it reflects progressive refinement across exits, whereas IG may exhibit drift. Figure 10(c) illustrates an Exit 2 case where PFAM produces clearer and more coherent maps relative to Occlusion and other baselines.

Robustness in Challenging Inputs. Figure 11 highlights PFAM's behavior under challenging inputs. In Figure 11(a), PFAM retains focus on relevant regions while other methods introduce more diffuse patterns. In fine-grained or ambiguous cases such as Figure 11(b) and Figure 11(c), PFAM produces sharper and more localized attributions, maintaining consistency across exits.

Cumulative Attribution Mapping. PFAM supports cumulative attribution maps that aggregate information across exits. These maps illustrate how model focus evolves across depth, providing a view of attribution progression not available in standard post-hoc methods.

Summary. Overall, these examples show that PFAM provides stable and progressively refined attribution patterns across exits. Compared to baseline methods, PFAM often yields more consistent and coherent visualizations, particularly in deeper layers and challenging cases. Its cumulative maps further illustrate how attribution evolves with depth, supporting analysis of both correct and incorrect predictions.

M Additional Calibration Details

IEES calibration is performed post-training using a held-out validation split; no attribution is computed at test time. Prior to calibration, $F_i(x)$ and $Q_i(x)$ are min-max normalized to $[0, 1]$ per exit (for $i \geq 2$ in the case of Q_i), using the minimum and maximum of each quantity over the validation set at that exit; these statistics are then fixed, ensuring commensurability with $S_i(x)$ and consistent mapping onto the threshold grid $[0.5, 0.95]$. The process then proceeds in two stages. First, the component weights (w_1, w_2, w_3) are selected through a coarse grid search to balance confidence, attribution strength, and attribution stability. The final weights used in the experiments are $(0.8, 0.1, 0.1)$ for CIFAR-10 (ResNet-18), $(0.7, 0.2, 0.1)$ for CIFAR-100 (MSDNet), and $(0.5, 0.4, 0.1)$ for ImageNet (MobileNetV3). Once these weights are fixed, exit thresholds τ_i are tuned per exit by sweeping values in $[0.5, 0.95]$, using a coarse grid followed by local refinement around promising regions, and selecting values that maximize validation accuracy. Separating weight selection and threshold tuning helps avoid instability that can arise when thresholds are optimized over uncalibrated scores. All tuning is performed strictly on the validation split, with no test set involvement.

N Training Details

We follow standard early-exit training protocols as introduced in BranchyNet [47] and extended in MSDNet [24]. Backbone networks are augmented with intermediate classifiers and trained jointly using a weighted sum of exit-specific cross-entropy losses, enabling deep supervision across exits. This joint optimization improves gradient flow to earlier layers and acts as a regularizer for intermediate representations.

All models are trained using SGD with Nesterov momentum 0.9, following commonly adopted early-exit training configurations [47, 24]. ResNet-18 on CIFAR-10 uses an initial learning rate of 0.1 with step decay and weight decay 5×10^{-4}, trained for 200 epochs. MSDNet on CIFAR-100 follows its original training configuration [24]: initial learning rate 0.1, step decay, weight decay 10^{-4}, and 300 epochs. MobileNetV3 on ImageNet is fine-tuned from a publicly available pretrained checkpoint [22] for 150 epochs with cosine decay and weight decay 4×10^{-5}. A batch size of 64 is used for all models.

Branch weights decrease from early to late exits following the convention of BranchyNet [47], encouraging progressively refined representations at deeper exits. Standard data augmentation is applied: random crop with 4-pixel padding and horizontal flip for CIFAR-10/100, and random resized crop (224×224) with horizontal flip for ImageNet. All models and baselines are trained using identical settings per architecture, and no additional tuning is performed for IEES beyond standard calibration on validation data.

Compute budget. All experiments were conducted on a single NVIDIA RTX A6000 (48 GB) GPU with an Intel Xeon CPU at 3.91 GHz. Approximate wall-clock training times are: ResNet-18 on CIFAR-10 ($\sim$2 GPU-hours), MSDNet on CIFAR-100 ($\sim$10 GPU-hours), and MobileNetV3 on ImageNet ($\sim$12 GPU-hours, reduced by pretrained initialization). Calibration and proxy training add approximately 1 GPU-hour per dataset. Total compute across experiments, including multiple runs and ablations, is approximately on the order of 80–100 GPU-hours.

O Attribution Implementation Across Architectures

To support the cross-architecture evaluations in Section 6, we briefly describe how gradient-based attribution is implemented for transformer-based and sequence models. PFAM requires attribution maps at intermediate exits, but is not restricted to convolutional feature maps; it applies to architectures with structured internal representations and differentiable outputs.

Vision Transformers. For transformer-based vision models, intermediate representations at exit i take the form $Z_i(x) \in \mathbb{R}^{N \times D}$, where N is the number of tokens and D is the embedding dimension. Following standard practice [17], tokens are reshaped into a spatial grid ($N = H \times W$), yielding $\tilde{Z}_i(x) \in \mathbb{R}^{D \times H \times W}$. Gradients are computed with respect to these representations, treating embedding dimensions as channels, which produces spatial attribution maps compatible with PFAM.

Language Models (BERT). For BERT-based models, representations at exit i are $Z_i(x) \in \mathbb{R}^{L \times D}$, where L is the sequence length. Token positions define the attribution domain, and gradients are computed with respect to hidden states at each exit, following gradient-based attribution methods used in NLP [45].

We then adapt the Grad-CAM formulation to token representations by treating hidden dimensions as channels:

$$A_i(x) = \text{ReLU}\left(\sum_{d=1}^{D} \bar{\alpha}_{i,d}\, Z_i(x)[:, d]\right), \quad \bar{\alpha}_{i,d} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{\partial s}{\partial Z_i(x)[\ell, d]}, \tag{17}$$

yielding token-level attribution maps. This extension is consistent with prior adaptations of gradient-based attribution to non-convolutional architectures. The resulting maps allow PFAM and IEES to operate consistently across CNN, transformer, and sequence-based models without modifying the underlying architectures.

P Sensitivity to Top-k Spatial Mask

Table 13 reports the effect of varying the spatial masking threshold k in Eq. 8 on CIFAR-10 (ResNet-18). All other components of IEES remain identical.

Table 13: Accuracy (%) under varying top-k spatial mask. Results are mean $\pm$ std over 5 runs.

k (%)	Overall Acc. (%)	Avg. Latency (ms)
10	88.21 ± 0.04	3.09 ± 0.03
20	$\mathbf{89.77 \pm 0.02}$	$\mathbf{3.14 \pm 0.02}$
30	89.51 ± 0.03	3.12 ± 0.02
50	88.83 ± 0.03	3.07 ± 0.02
100	88.10 ± 0.03	2.98 ± 0.02

Performance is stable across $k \in \{20, 30\}$, with degradation at k=10 due to excessive spatial sparsity and at k=100, where the mask covers the full spatial domain and attribution strength reduces to the global average of M_i. We adopt k=20 as the default.

NeurIPS Paper Checklist

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 clearly state the contributions of PFAM, IEES, and proxy-based deployment. These claims are supported by experimental results in Sections 5 and 6, with additional supporting analyses in the Appendix, including component analyses (H, I, J) and extended evaluations of PFAM and the proxy predictor (F, B).

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the proposed approach are explicitly discussed in Section 8.

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: [N/A]

   Justification: The paper does not include formal theoretical results, assumptions, or proofs; contributions are empirical and methodological.

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: The paper provides sufficient details on datasets, model architectures, training procedures, calibration, and evaluation protocols in Section 4, with additional implementation, ablation, and extended evaluation details in the Appendix (Appendices N, M, B, and J), enabling reproduction of the reported 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: All datasets used are publicly available. The source code implementing IEES, PFAM, and the proxy predictor is publicly available at `https://github.com/haseenarahmath/pfam_iees`

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: The paper specifies datasets, model architectures, training procedures, calibration, and evaluation protocols in Section 4, with full implementation and hyperparameter details provided in the Appendix (Appendices N and M).

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: The paper reports mean ± standard deviation over 5 independent runs with different random seeds.

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: Experiments were conducted on an NVIDIA RTX A6000 (48 GB) GPU with an Intel Xeon CPU (3.91 GHz), as described in Section 4. Approximate training times and total compute are reported in Appendix N.

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: The research uses publicly available datasets and standard evaluation practices, including a limited user study without collection of personal or sensitive data (Appendix K), and complies with the NeurIPS Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses both benefits and potential risks in Section 7 (Broader Impact subsection).

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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve releasing models or datasets that pose a high risk of misuse.

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: The paper uses publicly available datasets and models and properly credits original sources through citations.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces new methods and provides implementation details in the paper and supplementary material.

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: [Yes]

    Justification: The paper includes a limited user study, with study details and participant instructions provided in Appendix K.

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: [No]

    Justification: The study involved participants performing non-sensitive visual perception tasks with no personal data collected. Informed consent was obtained from all participants. No IRB approval was required under our institution's policy for low-risk perception studies of this type. Study details are provided in Appendix K.

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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used in the core methodology or experiments.