When Do Models Actually Decide? Mapping the Layer-Wise Decision Timeline in Pretrained Neural Networks

Supplementary Material

9. Overview

This supplementary material provides comprehensive additional details, extended experimental results, and in-depth analyses that complement the main paper. We organize the content as follows:

- **Section 10**: Detailed implementation specifications including probe training protocols and computational infrastructure.
- **Section 11**: Mathematical foundations with formal derivations of stability criteria and statistical measures.
- **Section 12**: Comprehensive metric definitions including stable-but-wrong rates, earliest-correct-stable layers, and flip statistics.
- **Section 13**: Extended experimental results across all three ResNet architectures (18/50/101).
- **Section 14**: Deep-dive analyses with calibration studies and computational profiling.

10. Implementation Details

10.1. Network Architecture and Layer Selection

For each ResNet variant, we extract representations from 12 strategically selected layers spanning the network hierarchy. Table 2 provides complete specifications including architectural positions, output dimensionalities, and spatial resolutions.

The selection strategy ensures representative coverage across all four residual stages plus the stem and classification head. For ResNet-18, which uses BasicBlock rather than Bottleneck blocks, we use matched stage-level anchors at the same L0–L11 indices but do not reuse the bottleneck-specific internal block names listed in Table 2.

10.2. Probe Training Protocol

Linear probes are trained independently for each layer using a carefully tuned protocol designed to maximize probe performance without overfitting. We partition the ImageNet validation set (50,000 samples) into 70% probe training (35,000 samples) and 30% probe validation (15,000 samples) using a fixed random seed for reproducibility.

Optimization Configuration. Each probe $\mathcal{P}_l : \mathbb{R}^{C_l} \to \mathbb{R}^{1000}$ is parameterized as a single affine transformation: $\mathcal{P}_l(\bar{\mathbf{h}}^{(l)}) = \mathbf{W}_l\bar{\mathbf{h}}^{(l)} + \mathbf{b}_l$ where $\mathbf{W}_l \in \mathbb{R}^{1000\times C_l}$ and $\mathbf{b}_l \in \mathbb{R}^{1000}$. We optimize via stochastic gradient descent with the following hyperparameters:

Table 2. **Layer extraction specifications for bottleneck ResNets (ResNet-50/101).** Each anchor is indexed sequentially (L0–L11) and mapped to its architectural position. ResNet-18 uses matched stage-level anchors with the same indices but different internal block names and channel dimensions. Spatial resolution decreases progressively through stages while channel dimensionality increases, creating the characteristic pyramidal structure of residual networks.

Index	Layer Name	Stage	Channels	Resolution
L0	conv1	stem	64	112×112
L1	maxpool	stem	64	56×56
L2	layer1.0	1	256	56×56
L3	layer1.end	1	256	56×56
L4	layer2.0	2	512	28×28
L5	layer2.end	2	512	28×28
L6	layer3.0	3	1024	14×14
L7	layer3.mid	3	1024	14×14
L8	layer3.end	3	1024	14×14
L9	layer4.0	4	2048	7×7
L10	layer4.mid	4	2048	7×7
L11	avgpool	head	2048	1×1

- **Learning rate**: $\eta_0 = 0.1$ (initial), reduced by $10\times$ upon validation plateau
- **Momentum**: $\mu = 0.9$ (Nesterov accelerated gradient)
- **Weight decay**: $\lambda = 10^{-4}$ (L2 regularization)
- **Batch size**: 256 samples per gradient step
- **Maximum epochs**: 100 with early stopping (patience 10)
- **Loss function**: Cross-entropy with label smoothing ($\epsilon = 0.1$)

Spatial Aggregation Strategy. For convolutional layers producing spatial feature maps $\mathbf{h}^{(l)} \in \mathbb{R}^{B\times C\times H\times W}$, we apply global average pooling (GAP):

$$\bar{\mathbf{h}}^{(l)} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{h}^{(l)}_{:,:,i,j} \in \mathbb{R}^{B\times C} \tag{5}$$

11. Mathematical Foundations

11.1. Formal Decision Layer Definition

We provide a rigorous mathematical formulation of decision layers with complete notation and edge case handling.

Notation. Let T denote the last probe index; in all experiments, $T = 11$, corresponding to 12 anchor points

$L0, \ldots, L11$. For an input $\mathbf{x}_i$ with ground truth label $y_i \in \{1, \ldots, C\}$, define

$$\hat{y}_i^{(l)} = \arg\max_c \mathcal{P}_l(\mathbf{h}_i^{(l)})_c, \qquad l \in \{0, \ldots, T\}. \quad (6)$$

Stability Window. For a persistence requirement $k \in \{1, 2, 3, 4\}$, define the k-stable predicate:

$$\text{Stable}_k(i, l) = \begin{cases} \text{True} & \text{if } \hat{y}_i^{(l)} = \hat{y}_i^{(l+1)} = \cdots = \hat{y}_i^{(l+k)} \\ \text{False} & \text{otherwise} \end{cases} \quad (7)$$

Decision Layer. The decision layer $l_i^*(k)$ for sample i with stability window k is

$$l_i^*(k) = \begin{cases} \min\{l \in \{0, \ldots, T-k\} : \text{Stable}_k(i, l)\} & \text{if such } l \text{ exists,} \\ T & \text{otherwise.} \end{cases} \quad (8)$$

Thus the terminal anchor T also acts as a fallback bucket for never-stable samples.

11.2. Statistical Measures and Distributions

For a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we compute distributional characteristics:

$$\mu_k = \frac{1}{N} \sum_{i=1}^N l_i^*(k) \qquad \text{(mean decision depth)} \quad (9)$$

$$\tilde{l}_k = \text{median}\{l_i^*(k)\}_{i=1}^N \qquad \text{(median decision depth)} \quad (10)$$

$$\sigma_k = \sqrt{\frac{1}{N} \sum_{i=1}^N (l_i^*(k) - \mu_k)^2} \qquad \text{(standard deviation)} \quad (11)$$

Category Partitioning. We partition samples into three disjoint categories:

$$\mathcal{E}_k = \{i : l_i^*(k) < 0.3T\} \qquad \text{(early deciders)} \quad (12)$$

$$\mathcal{M}_k = \{i : 0.3T \le l_i^*(k) < 0.7T\} \qquad \text{(mid deciders)} \quad (13)$$

$$\mathcal{L}_k = \{i : l_i^*(k) \ge 0.7T\} \qquad \text{(late deciders)} \quad (14)$$

Cumulative Distribution Function.

$$F_k(l) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[l_i^*(k) \le l] \quad \text{for } l \in \{0, \ldots, T\} \quad (15)$$

11.3. Probe Accuracy and Representation Quality

Linear probe accuracy at layer l:

$$\alpha^{(l)} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{y}_i^{(l)} = y_i] \quad (16)$$

11.4. Why Larger k Delays Decisions

A larger stability window is mechanically stricter: any sample that satisfies a $(k+1)$-step stability criterion also satisfies the k-step criterion, but not conversely. Therefore increasing k can only move decision layers rightward or leave them unchanged, and it can only increase the mass assigned to the terminal fallback bucket. We present this as a qualitative monotonicity argument rather than as a fitted probabilistic model of layerwise prediction changes.

12. Extended Metrics and Definitions

12.1. Stable-But-Wrong (SBW) Rate

The stable-but-wrong metric quantifies what fraction of samples that decide at layer l make incorrect predictions:

$$\text{SBW}_k^{(l)} = \frac{|\{i : l_i^*(k) = l \wedge \hat{y}_i^{(l)} \ne y_i\}|}{|\{i : l_i^*(k) = l\}|} \quad (17)$$

High SBW rates at early layers indicate stability does not imply correctness.

12.2. Earliest Correct and Stable Layer (ECSL)

The earliest correct and stable layer identifies when predictions both stabilize and match ground truth:

$$\text{ECSL}_i(k) = \begin{cases} \min\{\, l \in \{0, \ldots, T-k\} : \text{Stable}_k(i, l) \wedge \hat{y}_i^{(l)} = y_i \,\} & \text{if such } l \text{ exists} \\ T & \text{otherwise} \end{cases} \quad (18)$$

The ECSL provides an optimistic bound on early exit potential with a perfect oracle.

12.3. Always Wrong and Fixed (AWF) Depth

The always-wrong-and-fixed metric identifies samples that never produce correct predictions but achieve stable (incorrect) predictions:

$$\text{AWF}_i(k) = \begin{cases} \min\{\, l \in \{0, \ldots, T-k\} : \text{Stable}_k(i, l) \wedge \forall l' \in \{0, \ldots, T\},\ \hat{y}_i^{(l')} \ne y_i \,\} & \text{if such } l \text{ exists} \\ T & \text{otherwise} \end{cases} \quad (19)$$

For both ECSL and AWF, the terminal value T is a fallback convention and should not be interpreted as a genuinely late stabilization layer.

12.4. Prediction Flip Statistics

Flip Count. The number of times a sample's prediction changes:

$$\text{Flips}(i) = \sum_{l=1}^T \mathbb{1}[\hat{y}_i^{(l)} \ne \hat{y}_i^{(l-1)}] \quad (20)$$

First Flip Depth.

$$\text{FirstFlip}(i) = \begin{cases} \min\{l \in \{1, \ldots, T\} : \hat{y}_i^{(l)} \ne \hat{y}_i^{(0)}\} & \text{if such } l \text{ exists} \\ T & \text{otherwise} \end{cases} \quad (21)$$

Table 3. **Decision depth statistics for ResNet-18.** Mean, median, and standard deviation of decision layers across stability windows. Never-stabilize percentage indicates samples failing to achieve k-consecutive agreement.

Window	Mean	Median	Std Dev	Never Stabilize
$k = 1$	4.12	3.00	3.13	1.1%
$k = 2$	7.40	9.00	3.86	43.8%
$k = 3$	9.24	11.00	3.23	74.7%
$k = 4$	10.00	11.00	2.68	86.7%

Table 4. **Decision depth statistics for ResNet-50.**

Window	Mean	Median	Std Dev	Never Stabilize
$k = 1$	2.86	2.00	2.94	0.5%
$k = 2$	5.51	5.00	4.10	17.6%
$k = 3$	7.84	11.00	4.38	63.0%
$k = 4$	9.10	11.00	3.81	79.0%

Table 5. **Decision depth statistics for ResNet-101.**

Window	Mean	Median	Std Dev	Never Stabilize
$k = 1$	2.64	2.00	2.76	0.3%
$k = 2$	5.60	6.00	3.64	10.5%
$k = 3$	8.02	11.00	4.17	62.9%
$k = 4$	9.02	11.00	3.80	77.2%

Maximum Run Length.

$$\text{MaxRun}(i) = \max_{0 \le a \le b \le T} \left\{ b - a + 1 : \hat{y}_i^{(a)} = \hat{y}_i^{(a+1)} = \cdots = \hat{y}_i^{(b)} \right\} \tag{22}$$

13. Extended Experimental Results

13.1. Decision Depth Statistics Across Architectures

Tables 3, 4, and 5 present comprehensive decision depth statistics for all three ResNet architectures across stability windows $k \in \{1, 2, 3, 4\}$.

Key Observations. ResNet-18 shows systematically later decisions (mean 7.40 at $k = 2$) compared to ResNet-50/101 (mean 5.51/5.60), reflecting capacity constraints. The never-stabilize rates increase steeply with k across all architectures, with ResNet-18 showing the most dramatic increase (1.1% $\rightarrow$ 86.7%). ResNet-50 and ResNet-101 exhibit nearly identical patterns despite $2\times$ depth difference, suggesting capacity matters more than depth.

13.2. Stable-But-Wrong Rates Across Architectures

Tables 6, 7, and 8 present SBW rates at the decision layer for $k = 2$ across all three architectures.

Table 6. **Stable-but-wrong rates for ResNet-18 ($k = 2$).** For each anchor L0–L11, the fraction of samples assigned to that decision layer whose prediction is incorrect. Early anchors show SBW>90%.

Layer	SBW Rate
L0	96.3%
L1	95.1%
L2	97.0%
L3	97.2%
L4	92.8%
L5	93.0%
L6	83.4%
L7	84.8%
L8	8.3%
L9	9.4%
L10	—
L11	20.3%

Table 7. **Stable-but-wrong rates for ResNet-50 ($k = 2$).**

Layer	SBW Rate
L0	97.7%
L1	97.7%
L2	97.0%
L3	93.4%
L4	88.1%
L5	92.0%
L6	79.5%
L7	45.7%
L8	7.9%
L9	7.2%
L10	—
L11	23.5%

Table 8. **Stable-but-wrong rates for ResNet-101 ($k = 2$).**

Layer	SBW Rate
L0	97.3%
L1	97.8%
L2	97.3%
L3	97.2%
L4	93.6%
L5	92.8%
L6	84.4%
L7	20.4%
L8	7.3%
L9	7.6%
L10	—
L11	32.4%

Universal Early-Layer Incorrectness. All three architectures exhibit SBW rates above 90% through roughly L0–L5, with L6 still remaining high at 79–84%. Thus early

Table 9. **ECSL/AWF/DL distributions for ResNet-18 ($k = 2$).**

Layer	ECSL %	AWF %	DL %
L0	0.26	0.24	6.97
L1	0.12	0.12	2.40
L2	0.19	0.17	6.15
L3	0.22	0.22	7.01
L4	0.38	0.38	4.99
L5	0.31	0.32	4.00
L6	1.95	2.02	9.86
L7	1.00	1.04	4.69
L8	1.39	1.36	1.13
L9	12.84	12.42	8.94
L10	0.00	0.00	0.00
L11	81.35	81.70	43.85

Table 10. **ECSL/AWF/DL distributions for ResNet-50 ($k = 2$).**

Layer	ECSL %	AWF %	DL %
L0	0.41	0.43	18.08
L1	0.17	0.17	7.37
L2	0.28	0.29	9.29
L3	0.30	0.29	3.87
L4	0.62	0.64	4.07
L5	1.06	1.08	8.53
L6	1.75	1.82	5.60
L7	0.93	0.93	0.98
L8	2.00	1.97	1.04
L9	47.61	46.46	23.54
L10	0.00	0.00	0.00
L11	44.85	45.92	17.63

stable predictions are overwhelmingly incorrect. The sharp drop by L8–L9, where SBW falls to single digits for the late semantic anchors, coincides with the consolidation documented in the main paper.

13.3. ECSL, AWF, and DL Distributions

Tables 9, 10, and 11 compare distributions of earliest correct stable layer (ECSL), always wrong and fixed (AWF), and standard decision layer (DL) for $k = 2$.

Late Concentration Pattern. ECSL and AWF place large mass at L9 and at the terminal anchor L11 across all architectures. The L11 mass should be interpreted as a fallback bucket for samples with no earlier qualifying layer, not as evidence that these samples genuinely become stable only at avgpool. In contrast, DL retains substantial early mass, which quantifies the gap between descriptive stability and correctness-aware exit readiness.

Table 11. **ECSL/AWF/DL distributions for ResNet-101 ($k = 2$).**

Layer	ECSL %	AWF %	DL %
L0	0.32	0.34	11.60
L1	0.18	0.19	8.10
L2	0.24	0.24	8.79
L3	0.14	0.13	4.79
L4	0.33	0.36	4.60
L5	0.76	0.76	8.15
L6	3.01	3.02	14.25
L7	1.08	1.06	0.86
L8	2.50	2.43	1.57
L9	59.24	58.00	26.79
L10	0.00	0.00	0.00
L11	32.20	33.48	10.50

Table 12. **Gating performance for ResNet-18.** Baseline accuracy 69.76%.

Policy	Accuracy	Latency (ms)
Baseline	69.76%	15.81
Stability-only ($k = 2$)	37.49%	12.60

Table 13. **Gating performance for ResNet-50.** Baseline accuracy 76.15%.

Policy	Accuracy	Latency (ms)
Baseline	76.15%	49.46
Stability-only ($k = 2$)	34.68%	28.73
Conf $\geq$95% baseline	73.65%	49.13
Hybrid ($k = 2$) $\geq$95%	73.66%	49.18

Table 14. **Gating performance for ResNet-101.** Baseline accuracy 77.37%.

Policy	Accuracy	Latency (ms)
Baseline	77.37%	78.56
Stability-only ($k = 2$)	32.94%	40.91
Conf $\geq$95% baseline	75.15%	78.01
Hybrid ($k = 2$) $\geq$95%	75.01%	78.06

13.4. Gating Policy Performance

Tables 12, 13, and 14 summarize gating policy performance versus baseline.

Universal Gating Failure. Stability-only gating fails catastrophically across all architectures (33–37% accuracy), while confidence gating yields modest accuracy degradation (3–5% drop) with minimal speedup (99–100% baseline latency).

Table 15. **Temperature scaling for ResNet-18.** ECE and NLL before and after calibration.

Exit	T^*	ECE raw	ECE cal	NLL raw→cal
L1	0.232	0.85%	0.68%	6.803→6.687
L3	0.184	1.38%	0.47%	6.800→6.553
L9	3.187	18.55%	18.69%	6.257→6.698
L10	0.508	14.93%	5.28%	0.916→0.878

Table 16. **Temperature scaling for ResNet-50.**

Exit	T^*	ECE raw	ECE cal	NLL raw→cal
L1	0.248	0.91%	0.81%	6.784→6.646
L3	0.140	1.45%	1.92%	6.786→6.522
L9	0.366	42.16%	3.72%	2.978→1.956
L10	0.570	12.55%	3.16%	0.728→0.656

Table 17. **Temperature scaling for ResNet-101.**

Exit	T^*	ECE raw	ECE cal	NLL raw→cal
L1	0.235	0.81%	0.64%	6.801→6.680
L3	0.179	1.31%	1.33%	6.778→6.526
L9	0.397	40.51%	3.63%	1.974→1.281
L10	0.622	11.67%	2.23%	0.712→0.635

Table 18. **Robustness for ResNet-18.** Accuracy delta versus clean validation.

Corruption	Δ Accuracy
brightness_s1	−1.10%
brightness_s2	−4.93%
gauss_blur_s1	−4.77%
gauss_blur_s2	−4.77%
gauss_noise_s1	−0.17%
gauss_noise_s2	−1.73%

13.5. Calibration Analysis

Tables 15, 16, and 17 present temperature scaling calibration for selected exits.

Severe L9 Miscalibration. Layer L9 exhibits extreme miscalibration across all three architectures (18.55–42.16% ECE), with ResNet-50/101 showing the most severe L9 miscalibration (40.51–42.16% ECE) and temperature scaling providing dramatic improvements there. ResNet-18 shows anomalous behavior ($T^* = 3.187$, ECE slightly increases after scaling), reflecting lower accuracy at this layer.

13.6. Robustness Under Corruption

Tables 18, 19, and 20 present accuracy deltas under ImageNet-C corruptions.

Table 19. **Robustness for ResNet-50.**

Corruption	Δ Accuracy
brightness_s1	−1.13%
brightness_s2	−4.03%
gauss_blur_s1	−5.83%
gauss_blur_s2	−5.83%
gauss_noise_s1	−1.03%
gauss_noise_s2	−2.50%

Table 20. **Robustness for ResNet-101.**

Corruption	Δ Accuracy
brightness_s1	−0.63%
brightness_s2	−4.97%
gauss_blur_s1	−3.87%
gauss_blur_s2	−3.87%
gauss_noise_s1	−0.27%
gauss_noise_s2	−0.57%

Table 21. **Flip statistics for ResNet-18.**

Metric	Mean	Median
Flip count	7.58	8.0
First-flip depth	1.28	1.0
Maximum run length	3.08	3.0
Earliest equal to final	9.53	10.0
Earliest correct	9.48	10.0

Table 22. **Flip statistics for ResNet-50.**

Metric	Mean	Median
Flip count	6.56	7.0
First-flip depth	1.78	1.0
Maximum run length	3.64	3.0
Earliest equal to final	9.07	9.0
Earliest correct	9.05	9.0

Blur Sensitivity. Gaussian blur causes worst degradation across all architectures (3.87–5.83%), while noise effects are milder (0.17–2.50%).

13.7. Flip Statistics

Tables 21, 22, and 23 summarize prediction flip statistics from complete per-layer sequences.

High Flip Counts. Mean flip counts of 6.22–7.58 indicate predictions change at 56–69% of layer transitions, explaining why strict stability criteria ($k \geq 3$) yield high never-stabilize rates.

Table 23. **Flip statistics for ResNet-101.**

Metric	Mean	Median
Flip count	6.22	6.0
First-flip depth	1.65	1.0
Maximum run length	3.79	3.0
Earliest equal to final	9.02	9.0
Earliest correct	9.00	9.0

14. Detailed Analysis and Visualizations

14.1. Figure S1: SBW and Distribution Analysis (ResNet-50)

Figure 4 presents three complementary analyses for ResNet-50 exposing the gap between stability and correctness.

14.2. Figure S2: Calibration Across Architectures

Figure 5 compares calibration metrics (ECE and NLL) across all three architectures at selected exit points.

14.3. Figure S3: Flip Analysis Across Architectures

Figure 6 presents flip count distributions revealing pervasive prediction instability.

14.4. Figure S4: CDF Across Stability Windows

Figure 7 compares cumulative decision curves across all stability windows and architectures.

15. Discussion and Practical Implications

15.1. Cross-Architecture Patterns

Our comprehensive analysis across ResNet-18/50/101 reveals several universal patterns that transcend architectural specifics:

Capacity Dominates Over Depth. ResNet-50 and ResNet-101 exhibit nearly identical decision timelines (mean 5.51 vs 5.60 layers at $k = 2$, Figure 7b,c) despite $2\times$ depth difference (50 vs 101 layers). Both architectures use identical stage structures (64-256-512-1024-2048 channels) and bottleneck blocks, differing only in blocks per stage. This overlap indicates decision formation is determined primarily by representational capacity (channel dimensions, stage transitions) rather than mere layer count.

Limited Capacity Delays Decisions. ResNet-18 shows systematically later decisions (mean 7.40 layers at $k = 2$) and a much stronger late-assigned bias (53.92% vs. 42.21% for ResNet-50), together with a substantially higher never-stabilize rate (43.8% vs. 17.6%). Using BasicBlock instead of Bottleneck and lower channel counts (64-128-256-512), ResNet-18's limited capacity forces deferred decisions until sufficient semantic information accumulates. This 34% later mean decision depth (7.40 vs. 5.51) directly reduces early exit opportunities.

Universal SBW Pattern. All three architectures exhibit SBW rates exceeding 90% for layers L0–L6 (Tables 6, 7, 8), confirming early stable predictions are overwhelmingly incorrect regardless of architectural capacity. The sharp drop at L8–L9 (45–93% $\rightarrow$ 7–9%) coincides with semantic consolidation, validating that correctness requires deep processing across architectures.

Consistent Flip Behavior. Mean flip counts show modest architectural variation: 7.58 (R-18) $\rightarrow$ 6.56 (R-50) $\rightarrow$ 6.22 (R-101), representing 69%, 60%, and 56% change rates respectively (Tables 21, 22, 23). Despite different capacities, all exhibit pervasive instability explaining why strict criteria ($k \geq 3$) yield 63–87% never-stabilize rates. The slight decrease with capacity suggests deeper networks explore prediction space more efficiently.

15.2. Practical Early Exit Guidance

Target $k = 2$ Stability. Our analysis reveals $k = 1$ accepts too many transient agreements (0.3–1.1% never-stabilize, high SBW), while $k \geq 3$ imposes excessive conservatism (62–87% never-stabilize). The $k = 2$ regime balances reliability and coverage: 10.5–43.8% never-stabilize rates indicate most samples can theoretically stabilize, while 2-consecutive agreement filters many transient flips (mean 6.2–7.6 flips).

Architecture-Aware Design. For early exit systems, ResNet-50 is often a more practical target than simply scaling depth to ResNet-101. Both achieve similar peak accuracy (76.15% vs 77.37%) and very similar decision timelines, but ResNet-50 offers a 37% lower baseline latency (49.46 ms vs. 78.56 ms). Avoid shallow low-capacity architectures (ResNet-18) for early exit despite lower baseline latency, as late-heavy bias (43.85% at L11) severely limits exit opportunities.

Two-Tier Exit Strategy. Bimodal distributions (Figure 4b) suggest two-tier exits at layers 1–3 (early deciders) and 9–11 (late deciders after semantic consolidation). Mid-range exits (L4–L8) capture minimal mass (0.98–8.53% per layer) and waste architectural capacity. However, high SBW rates at early exits (Table 7) require additional gating beyond stability.

Combine Multiple Signals. Stability-only gating fails catastrophically (33–37% accuracy, Tables 12, 13, 14),

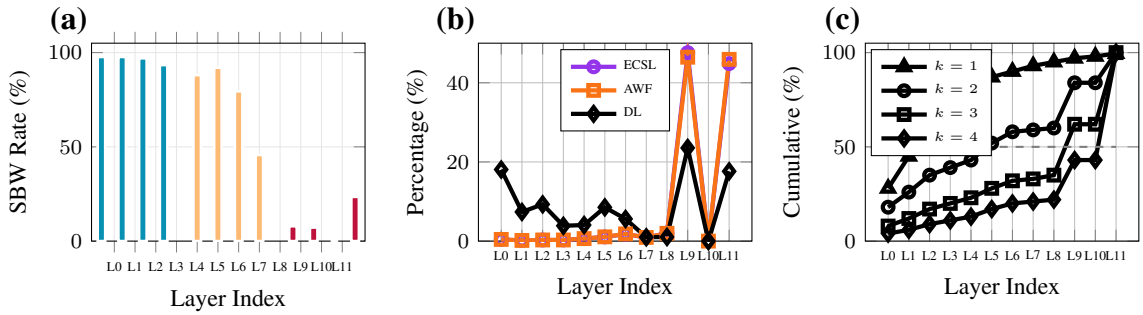

Figure 4. **Stable-but-wrong and distribution analysis (ResNet-50).** (a) SBW rates per layer show early layers >90% incorrect, with sharp drop at L8–L9. Colors indicate categories: cyan=early (L0–L3), orange=mid (L4–L7), pink=late (L8–L11). (b) ECSL/AWF concentrate at L9/L11 (47–48% and 45–46%), while DL shows bimodal pattern with 38.6% early mass. (c) Cumulative curves across k show systematic rightward shifts, with the 50% milestone moving from L2 ($k = 1$) to L5 ($k = 2$), and to L9 or later under stricter criteria.

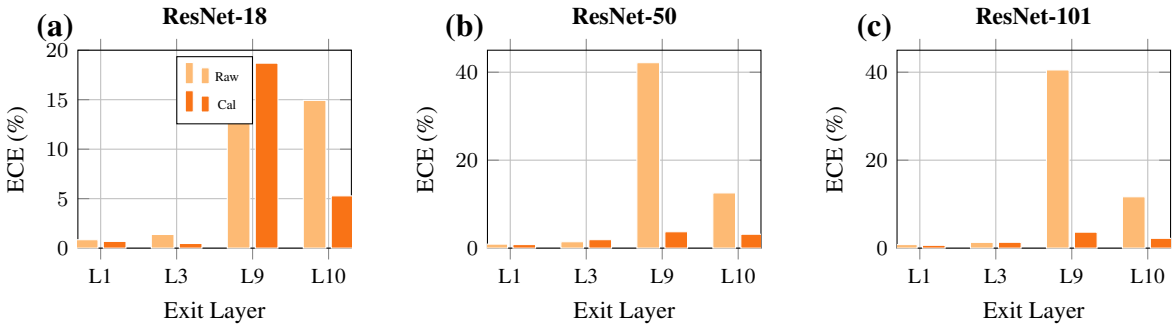

Figure 5. **Calibration analysis across architectures.** Expected calibration error (ECE) before (raw, light) and after (cal, dark) temperature scaling for four representative exits (L1, L3, L9, L10). All architectures show severe L9 miscalibration, but the benefit of temperature scaling is architecture-dependent: ResNet-50/101 improve dramatically at L9, whereas ResNet-18 does not. Early exits (L1, L3) are already well-calibrated even without scaling (<2% ECE).

while confidence-only gating achieves modest accuracy degradation (3–5% drop from baseline for ResNet-50/101) but minimal speedup (99–100% baseline latency). Future systems must explore learned gating networks trained to predict exit readiness from early-layer features, combining stability, confidence, representation distance metrics, and potentially sample-specific visual properties.

15.3. Calibration Insights

Temperature scaling reveals severe miscalibration at semantic transition (L9: 18–42% raw ECE, Tables 15, 16, 17). Calibration dramatically improves ECE (3–4% after scaling) but does not resolve the efficiency gap: high-confidence correct predictions still require semantic consolidation regardless of confidence reliability. The fundamental barrier is representational (insufficient discriminative power in early layers) rather than calibrational.

15.4. Robustness Patterns

Gaussian blur causes worst degradation (3.87–5.83%, Tables 18, 19, 20), likely removing high-frequency details discriminating fine-grained classes. Noise effects are milder (0.17–2.50%), consistent with spatial aggregation (GAP) averaging noise across locations. ResNet-101 shows best blur robustness (-3.87% vs -4.77% for R-18, -5.83% for R-50), suggesting deeper processing improves robustness to certain corruptions.

16. Qualitative Examples of Early-Stable Errors and Late-Correct Stabilization

To complement the aggregate statistics in the main paper, Figure 8 shows qualitative examples from two particularly diagnostic quadrants of the decision-timeline space: *early-stable but wrong* and *late-stable and correct*. We focus on these two groups because they most directly illustrate the central point of our analysis, namely that early stabilization should not be conflated with correctness.

The top row contains samples whose predicted class becomes stable at very shallow anchors but remains incorrect at the decision layer. In many such cases, the confidence trajectory stays low for most of the network and rises only

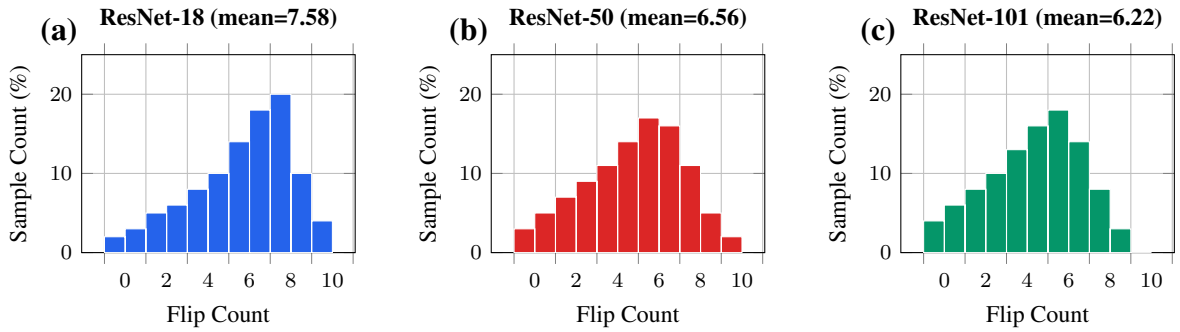

Figure 6. **Flip analysis reveals pervasive instability.** Histogram of flip counts (number of prediction changes across 11 layer transitions) for ResNet-18 (left), ResNet-50 (center), and ResNet-101 (right). Modal flip counts are 6–8 flips, indicating predictions change at 55–73% of transitions. Very few samples (<5%) maintain stable predictions (0–2 flips). Mean flip counts decrease with model depth/capacity: 7.58 (R-18), 6.56 (R-50), and 6.22 (R-101), with shallower networks showing more exploration.

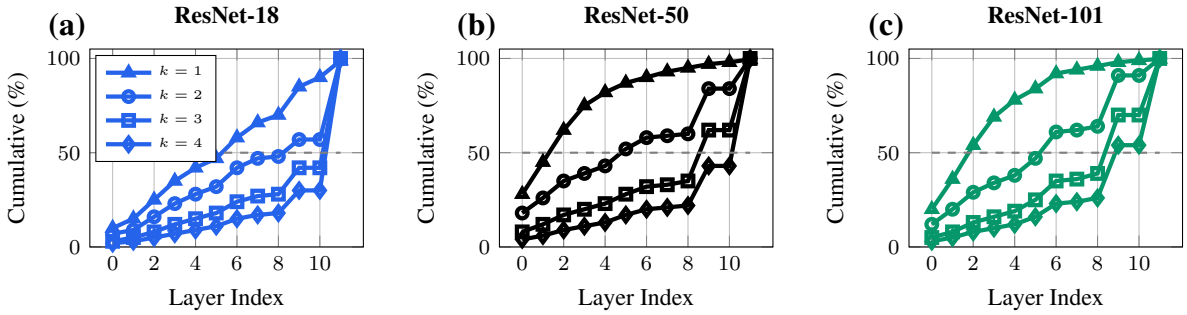

Figure 7. **Cumulative decision curves across stability windows.** For ResNet-18 (left), ResNet-50 (center), and ResNet-101 (right), cumulative percentage of samples achieving stability by each layer for $k \in \{1, 2, 3, 4\}$. All architectures show systematic rightward shifts with increasing k, confirming universal stability sensitivity. S-curve shapes persist across all combinations. ResNet-18 shows rightmost curves (mean 7.40 at $k = 2$), while ResNet-50/101 nearly overlap (mean 5.51/5.60), indicating capacity matters more than depth. The 50% milestone (dashed line) moves steadily rightward as k increases, from early anchors at $k = 1$ to substantially later anchors at $k = 2$, and to the final anchors for $k \geq 3$.

near the end, showing that early prediction stability does not imply early semantic certainty. By contrast, the bottom row contains samples that only stabilize near the final anchors but do so correctly, highlighting that many correct predictions require late semantic consolidation.

Together, these examples provide a qualitative view of the stability–correctness gap quantified in the main text. They also clarify why stability alone is not a sufficient early-exit criterion: a sample may appear prediction-stable long before the network has formed a confident and correct semantic representation. The remaining two quadrants (*early-stable and correct* and *late-stable but wrong*) were also inspected, but we omit them here for brevity because they are less central to the specific failure mode emphasized in this paper.

17. Conclusions

This supplementary material has provided comprehensive details across all three ResNet architectures, establishing universal patterns in decision formation:

First, the stable-but-wrong analysis quantifies why stability-based gating fails: early stable predictions are 90–98% incorrect across all architectures, creating a fundamental correctness-stability gap that simple gating cannot bridge.

Second, ECSL/AWF distributions reveal both correct and incorrect predictions concentrate at late layers after semantic consolidation, limiting early exit potential even with perfect oracles.

Third, flip analysis demonstrates pervasive instability (6–8 flips across 11 transitions), explaining why strict criteria ($k \geq 3$) yield prohibitive never-stabilize rates (63–87%).

Fourth, cross-architecture comparison reveals capacity dominates depth: ResNet-50/101 show identical timelines despite 2× depth difference, while ResNet-18's limited capacity delays decisions 34%.

Fifth, calibration improves confidence reliability but not efficiency—the barrier is representational, requiring archi-

Figure 8. **Qualitative examples illustrating the stability–correctness gap in ResNet-50 under k=2 stability.** The top row shows *early-stable but wrong* examples, while the bottom row shows *late-stable and correct* examples. For each sample, the image thumbnail is followed by a compact stability visualization and a confidence trajectory across the 12 anchor layers. In the stability strip, the top lane indicates whether the prediction at each anchor matches the decision-layer class, and the bottom lane indicates whether it matches the final-layer class. The dashed vertical line marks the decision layer, and the curve below shows the max-softmax confidence across anchors. These examples illustrate that prediction stability can arise well before confidence rises sharply, and that many early-stable cases remain incorrect, whereas many correct cases stabilize only near the final anchors.

tectural innovations beyond post-hoc calibration.

These findings provide empirical foundations for early exit system design while revealing fundamental challenges. Future work must develop learned gating mechanisms, confidence calibration techniques, or architectural innovations that bridge the correctness-stability gap we have comprehensively characterized.