DEEPER IS NOT BETTER FOR QUANTIZED CLIP: EARLY READOUT RECOVERS ACCURACY WHILE CUTTING COMPUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Readout depth is a decision variable. CLIP predicts a label from the direction of its final-layer image embedding, on the premise that every transformer block moves that direction closer to the correct class. We show that this premise fails for INT8-quantized CLIP and is already unreliable in full precision. In FP32 ViT-B/32, the cosine agreement between intermediate and final representations rises with depth but regresses at three of the eight candidate layers; under INT8, the deviation introduced by quantization grows from 0.09% of the native inter-layer change at Layer 1 to 75.64% at Layer 11. We model transformer depth as a noisy readout channel in which each block adds a semantic increment and a perturbation, and show that under this model a per-sample readout gated on the top-1/top-2 margin, with depth-dependent thresholds, is expected to beat full-depth readout once perturbation growth outpaces semantic gain. LRA-EE implements this rule on a frozen INT8 backbone: a patch-token average replaces the immature shallow [CLS] token, a learned gate trained without ground-truth labels decides when to stop, and exit layers are selected by an FP32-only maturity criterion applied identically to every backbone. Across four INT8 CLIP backbones on ImageNet-1K, accuracy never falls below the full-depth baseline while FLOPs fall by 7.9–14.5%: $+3.27$%p on ViT-B/32, $+4.27$%p on ViT-B/16, $+3.16$%p on EVA-CLIP, and $+0.66$%p on ViT-L/14. In other words, the quantized model is more accurate when it computes less. A four-quadrant decomposition attributes the gain to a Rescue Effect—8.1% of samples are correct at an early exit but wrong at full depth, against 4.9% the reverse—and an exact-route control splits it into a precision-independent component ($+2.42$%p) and a quantization-specific increment ($+0.86 \pm 0.03$%p over three seeds). The evidence is empirical: the backbone is never updated and the exit heads are calibrated on 2.3% of the ImageNet training images, so the result is post-training readout calibration of a compressed model rather than zero-shot deployment.

1 INTRODUCTION

CLIP classifies an image by the direction of a single vector (Radford et al., 2021). Its vision transformer (Dosovitskiy et al., 2021) moves this vector once per block; the standard recipe reads the vector out only after the last block, compares it by cosine similarity with text-prompt embeddings, and returns the nearest class. Deployment on integer accelerators adds a second convention on top of the first: post-training INT8 quantization (Jacob et al., 2018; Nagel et al., 2020; Xiao et al., 2023) is applied block by block, and the error it introduces is treated as a per-layer approximation that the final readout absorbs. Both conventions rest on one premise—that every additional block moves the embedding closer to the correct direction.

We show that this premise fails for INT8-quantized CLIP and is already unreliable at full precision. Two depth profiles, both measured on a 1,000-image calibration subset (Section 3), carry the argument. The first is *representational maturity* $M(L)$, the cosine agreement between the patch-token average at block L and at the final block of the FP32 model. In ViT-B/32, $M(L)$ rises from -0.10 at Layer 4 to 0.40 at Layer 11, but three of the eight candidate layers (5, 7, and 9) score lower than

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

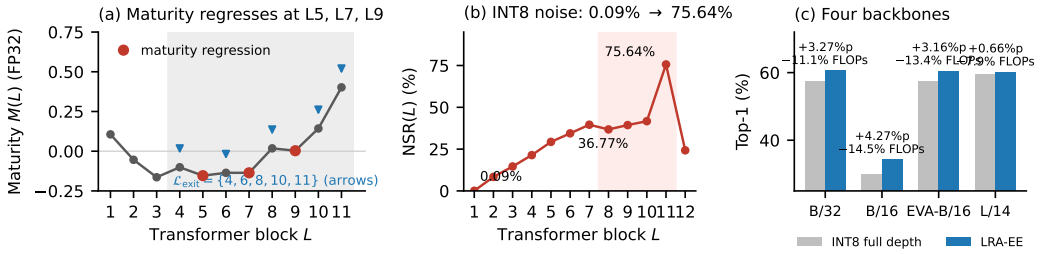


Figure 1: **Depth is a noisy readout channel in INT8 CLIP ViT-B/32.** (a) Representational maturity $M(L)$, the FP32 cosine agreement between the patch-token average at block L and at the final block, rises from -0.10 at Layer 4 to 0.40 at Layer 11 but regresses at Layers 5, 7, and 9 (red); the maturity rule of Section 4.3 excludes exactly these layers, leaving $\mathcal{L}_{\text{exit}} = \{4, 6, 8, 10, 11\}$. (b) The INT8 noise-to-signal ratio $\text{NSR}(L)$ grows from 0.09% (Layer 1) to 36.77% (Layer 8) and 75.64% (Layer 11): by the penultimate block, the INT8 quantization MSE reaches 75.64% of the FP32 inter-layer-change MSE. (c) Reading each sample out at its gated depth yields 60.65% Top-1 against 57.38% at full depth (+3.27%p) with 11.1% fewer FLOPs; 8.1% of samples are rescued and 4.9% are lost. The same procedure gives $+4.27$, $+3.16$, and $+0.66$%p on ViT-B/16, EVA-CLIP, and ViT-L/14 at 14.5%, 13.4%, and 7.9% FLOPs saving.

the layer before them; we call such a step a *maturity regression*. The second is the *noise-to-signal ratio* $\text{NSR}(L)$, the mean squared deviation that INT8 quantization introduces at block L divided by the mean squared change that the FP32 model itself makes between blocks $L-1$ and L. $\text{NSR}(L)$ grows from 0.09% at Layer 1 to 75.64% at Layer 11 ($r = 0.76$ with depth): by the penultimate block, quantization moves the representation three quarters as far as the block's own computation does. The consequence at the output is direct. Of 10,000 ImageNet validation images, 813 are classified correctly by an intermediate-layer readout and incorrectly by the full-depth INT8 model, whereas only 489 show the reverse pattern (Figure 1).

Existing tools address neither profile. Post-training quantization methods (Li et al., 2021; Yuan et al., 2022; Li et al., 2023; Dettmers et al., 2022; Wei et al., 2022) reduce the per-block error at a fixed depth and leave the accumulation across depth untouched. Early-exit methods (Teerapittayanon et al., 2016; Huang et al., 2018; Xin et al., 2020; Zhou et al., 2020; Schuster et al., 2022) treat depth as a resource to be traded for accuracy, gate on one confidence score with one threshold, and, in vision transformers, read the [CLS] token (Wang et al., 2021; Bakhtiarnia et al., 2021), which at Layer 10 of INT8 ViT-B/32 reaches 25.12% accuracy against 55.72% for the patch-token average (Section 6.2). Kaya et al. (2019) observed that a full-precision CNN can *overthink*—answer correctly at an intermediate layer and then change its mind—and Skean et al. (2025) showed that intermediate embeddings of language models can outperform final ones. Neither line quantifies how quantization changes the balance, and neither gives a per-sample rule for when to stop.

The question this paper answers is therefore: at which depth should a quantized transformer be read out, and for which inputs? We treat depth as a *noisy readout channel*. Each residual block writes a semantic increment $f_L(z_L)$ and, under quantization, a perturbation ϵ_L into the residual stream, $z_{L+1} = z_L + f_L(z_L) + \epsilon_L$ (He et al., 2016; Veit et al., 2016). The prediction at any depth is fixed by the margin between the top-1 and top-2 cosine scores, so a correct prediction survives further computation only while the accumulated perturbation stays below that margin, and a wrong one is corrected only if the remaining semantic increments are large enough to cross it. Reading these two quantities off the measured profiles—$M(L)$ for the semantic increment and $\text{NSR}(L)$ for the perturbation—yields a readout rule (Section 4.2): stop at the first candidate layer whose margin clears a depth-dependent threshold, skip subsequent layers at which maturity regresses, and loosen the threshold as the remaining semantic gain shrinks. In simpler terms, a sample should keep computing only while the next block is more likely to help than to hurt.

LRA-EE (Layer-wise Representation-Aware Early Exit) implements this rule on a frozen INT8 backbone with three components. *Spatio-Semantic Aggregation* (SSA) reads out the mean of

the patch tokens in place of the [CLS] token, which raises intermediate-layer accuracy by 7.11–30.60%p on ViT-B/32. A learned gate on three scalars—confidence, top-1/top-2 margin, and feature-activation variance—decides whether to stop; it is trained on agreement with the full-depth prediction and never sees a ground-truth label. *Layer-adaptive Confidence Thresholding* (LCT) applies a depth-dependent threshold schedule selected on a 5,000-image calibration split. The candidate exit layers come from one rule for every backbone: a depth-relative window from one third of the depth to the penultimate block, with the window's first depth retained as an anchor and every subsequent layer at which $M(L) < M(L-1)$ removed. The rule uses FP32 forward passes only and needs no labels, no quantized model, and no trained head; it excludes three, four, zero, and two layers on ViT-B/32, ViT-B/16, EVA-CLIP, and ViT-L/14 respectively, tracking each backbone's own profile.

The evidence, all on ImageNet-1K (Deng et al., 2009; Russakovsky et al., 2015), is as follows. On four INT8 CLIP backbones, accuracy never falls below the full-depth baseline while FLOPs fall by 7.9–14.5%: $+3.27$%p on ViT-B/32 (57.38%→60.65%), $+4.27$%p on ViT-B/16, $+3.16$%p on EVA-CLIP (Fang et al., 2023; Sun et al., 2023), and $+0.66$%p on ViT-L/14 (Table 1). A four-quadrant decomposition attributes the gain to the *Rescue Effect*, a net margin of $+3.24$%p between samples rescued by early readout (8.1%) and samples lost to it (4.9%) (Section 7.1). Freezing the per-sample routes and replaying them through the FP32 model splits this margin into a precision-independent part ($+2.42 \pm 0.06$%p) and a quantization-specific increment ($+0.86 \pm 0.03$%p over three seeds), whose sign is what the noisy-channel model predicts from a rising $\mathrm{NSR}(L)$ (Section 7.2). A factorial ablation shows that routing is necessary: a distilled SSA head at the final layer, trained on four times more data, reaches 53.47%, below the baseline, whereas routing with hard labels and no distillation reaches 57.88% and routing with distillation reaches 60.65% (Section 6.1). A single confidence threshold shared across layers peaks at 57.84% with 8.8% FLOPs saving, so the multi-feature gate and the depth-dependent thresholds each contribute (Section 6.2); the cost of heads and gate is measured, not assumed (Section 5.3).

Two points of scope belong next to these numbers. The backbone is never updated; the SSA heads are trained on 30,000 labeled ImageNet training images (2.3% of the training split), and the evaluation split is not used for model, gate, or threshold selection. The result is post-training readout calibration of a compressed model, not zero-shot deployment, and the accuracy figures are in-domain ImageNet figures. Within that scope the paper makes three claims. First, quantized CLIP has an interior optimal readout depth, and the two measurable profiles $M(L)$ and $\mathrm{NSR}(L)$ locate it (Section 3). Second, a margin-gated, depth-calibrated readout converts this into an accuracy gain at lower compute, with one procedure and no per-backbone tuning, on four backbones (Sections 4–5). Third, the gain decomposes into a precision-independent overthinking component and a quantization-specific component, and the two are measured separately (Section 7). Nothing in the procedure is specific to CLIP beyond a residual encoder and a cosine readout, and its two profiles cost a few forward passes, so the question it answers—at which depth a compressed encoder should be read out—can be asked of any such model before deployment.

2 Related Work

Where a transformer can be read out. Probes and similarity analyses show that intermediate layers carry usable, sometimes better, features (Alain & Bengio, 2017; Kornblith et al., 2019; Raghu et al., 2021; Tenney et al., 2019; Skean et al., 2025), and the residual-stream view explains why late blocks can be dropped (Veit et al., 2016; Huang et al., 2016; Fan et al., 2020; Gromov et al., 2025; Dong et al., 2021). These are population statements about full-precision models; we add the quantized case and a per-sample rule.

Early exit and adaptive inference. From the reject option (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) through BranchyNet, MSDNet, and cascades (Teerapittayanon et al., 2016; Huang et al., 2018; Bolukbasi et al., 2017), BERT-family exits (Xin et al., 2020; Liu et al., 2020; Zhou et al., 2020; Xin et al., 2021; Wang et al., 2022), depth-adaptive decoding (Elbayad et al., 2020; Schuster et al., 2022; Elhoushi et al., 2024), and vision and vision–language variants (Wang et al., 2021; Bakhtiarnia et al., 2021; Rao et al., 2021; Yin et al., 2022; Meng et al., 2022; Kong et al., 2022; Kim et al., 2022; Tang et al., 2023; Han et al., 2022), every method targets full precision, reads a [CLS]-type token, and gates on one confidence with one threshold. Kaya et al. (2019) named overthinking and Schwartz et al. (2020) tied depth to instance difficulty; neither treats

quantization. We measure the [CLS] gap (7.11–30.60%p), the single-threshold gap (57.84% vs. 60.65%), and the quantization increment (+0.86%p) directly.

Post-training quantization and CLIP adaptation. PTQ work minimizes per-block error at fixed depth (Jacob et al., 2018; Nagel et al., 2020; Li et al., 2021; Esser et al., 2020; Bondarenko et al., 2021; Dettmers et al., 2022; Wei et al., 2022; Xiao et al., 2023; Liu et al., 2021; Yuan et al., 2022; Lin et al., 2022; Li et al., 2023; 2022; Frantar et al., 2023; Lin et al., 2024; Yao et al., 2022; Wang et al., 2024; Li et al., 2025), and accumulation across layers has been analyzed since fixed-point CNNs (Lin et al., 2016; Hubara et al., 2018; Banner et al., 2019); LRA-EE holds the quantizer fixed and changes the depth, so the two compose. Prompt, adapter, and interpolation methods (Zhou et al., 2022; Shu et al., 2022; Gao et al., 2024; Wortsman et al., 2022) change what the final embedding is compared against; we change where it is taken from, on a frozen encoder. CLIP's cosine geometry (Wang & Isola, 2020; Liang et al., 2022) makes the margin, not the maximum probability (Hendrycks & Gimpel, 2017; Guo et al., 2017), the natural gating statistic (Schapire et al., 1998); CLIP compression (Wu et al., 2023; Vasu et al., 2024) is orthogonal.

3 Two Depth Profiles of a Quantized CLIP Encoder

Let $Z_L(x) \in \mathbb{R}^{(N+1)\times d}$ be the residual stream after block L, with patch tokens $z_{L,1}, \ldots, z_{L,N}$ and [CLS] token $z_{L,0}$; the *Spatio-Semantic Aggregation* (SSA) is

$$F_{\mathrm{SSA}}(Z_L) = \mathrm{LayerNorm}\Big(\tfrac{1}{N}\textstyle\sum_{i=1}^{N} z_{L,i}\Big). \tag{1}$$

Two functions of depth, computed on 1,000 calibration-split images (Section 5.1), carry the argument.

Representational maturity measures how far block L has moved the representation toward its final form, in FP32 and without labels or heads:

$$M(L) = \mathbb{E}_x\Big[\cos\big(F_{\mathrm{SSA}}(Z_L^{\mathrm{FP32}}(x)),\, F_{\mathrm{SSA}}(Z_{L_{\max}}^{\mathrm{FP32}}(x))\big)\Big]. \tag{2}$$

On ViT-B/32, $M(L)$ rises from -0.10 at Layer 4 to 0.40 at Layer 11 but Layers 5, 7, and 9 satisfy $M(L) < M(L-1)$, a *maturity regression* (Figure 1a). Each of the other three backbones has its own pattern (Appendix B).

Noise-to-signal ratio measures the INT8 perturbation at block L relative to the work the block itself does, with h_L the raw hidden state:

$$\mathrm{NSR}(L) = \frac{\mathbb{E}_x\big[\mathrm{MSE}\big(h_L^{\mathrm{INT8}}(x),\, h_L^{\mathrm{FP32}}(x)\big)\big]}{\mathbb{E}_x\big[\mathrm{MSE}\big(h_L^{\mathrm{FP32}}(x),\, h_{L-1}^{\mathrm{FP32}}(x)\big)\big]}. \tag{3}$$

On ViT-B/32 it grows from 0.09% at Layer 1 to 75.64% at Layer 11 across 9 of 11 transitions ($r = 0.76$ with depth) and doubles inside the exit region from 36.77% at Layer 8 (Figure 1b; the Layer-12 dip is a jump in the FP32 denominator, Table 7). Only $M(L)$ is used for selection; $\mathrm{NSR}(L)$ is independent evidence.

4 Depth as a Noisy Readout Channel

The mathematics below is elementary. Its use is to fix which quantity a gate must estimate, which observable is closest to it, and into which two effects the gain must decompose; Sections 5–7 then measure each. Proofs are in Appendix A.

4.1 Readout policies and the Rescue identity

Each block acts on the residual stream as

$$Z_{L+1} = Z_L + f_L(Z_L) + E_L, \qquad E_L = 0 \text{ in FP32}, \tag{4}$$

where E_L is the perturbation INT8 quantization of block L injects. A *readout* at depth L is $\rho_L(x) = \psi_L\big(F_{\mathrm{SSA}}(Z_L(x))\big) \in \mathbb{S}^{p-1}$ with ψ_L a head followed by ℓ_2 normalization; $\rho_{L_{\max}}$ is the native CLIP projection of the [CLS] token, i.e., the full-depth model. With unit text anchors $t_1, \dots, t_K$ and true class y,

$$s_{L,k}(x) = \langle \rho_L(x), t_k \rangle, \qquad \hat{y}_L(x) = \arg\max_k s_{L,k}(x),$$
$$m_L(x) = s_{L,(1)} - s_{L,(2)} \ \text{(unsigned margin)}, \qquad \mu_L(x) = s_{L,y} - \max_{k \neq y} s_{L,k} \ \text{(signed margin)}, \tag{5}$$

and the readout is correct, $a_L(x) = 1$, iff $\mu_L(x) > 0$. A *readout policy* $\pi : x \mapsto \pi(x) \in \mathcal{L}_{\mathrm{exit}} \cup \{L_{\max}\}$ has accuracy $A(\pi) = \mathbb{E}_x[a_{\pi(x)}(x)]$ and compute $C(\pi) = \mathbb{E}_x[\pi(x)]/L_{\max}$; FLOPs saving is $1 - C(\pi)$. Its gain over full depth is an exact identity:

$$A(\pi) - A(L_{\max}) = \underbrace{\Pr\left[a_\pi = 1,\, a_{L_{\max}} = 0\right]}_{\text{Rescue}} - \underbrace{\Pr\left[a_\pi = 0,\, a_{L_{\max}} = 1\right]}_{\text{Loss}} =: \Delta_{\mathrm{Rescue}}(\pi). \tag{6}$$

In other words, a policy beats full depth exactly when it rescues more samples than it loses; the diagonal of the four-quadrant table of Section 7.1 is irrelevant by construction.

4.2 A TWO-TERM MODEL OF CONTINUING COMPUTATION

For depths $L < L'$ we model the signed margin as

$$\mu_{L'}(x) = \mu_L(x) + g_{L \to L'}(x) + \xi_{L \to L'}(x), \tag{7}$$

where g is the *semantic drift* the FP32 model would produce and ξ is the *perturbation* from $E_L, \dots, E_{L'-1}$ projected on the decision direction, with $\mathbb{E}[\xi \mid x] = 0$, $\mathrm{Var}[\xi \mid x] = \sigma^2_{L \to L'}$, symmetric continuous cdf $F_\sigma(u) = F(u/\sigma)$, and $\xi \equiv 0$ in FP32. The two terms are the two profiles: $M(L)$ is the population signature of accumulated drift, $\mathrm{NSR}(L)$ the per-block growth of σ^2 relative to it. A readout with margin μ at L is correct at full depth with probability $F_\sigma(\mu + g)$ (Appendix A.2). The model is local: it follows the boundary between the top two classes and is meaningful while perturbations are small against the gaps to other classes; g and σ are never estimated per sample, and $M(L)$ and $\mathrm{NSR}(L)$ serve as their population proxies.

At a candidate depth the gate observes features $\phi_L(x)$ and chooses between stopping and continuing. Write

$$p_L = \Pr[\mu_L > 0 \mid \phi_L], \quad q_L^+ = \Pr[\mu_{L_{\max}} > 0 \mid \mu_L > 0, \phi_L], \quad q_L^- = \Pr[\mu_{L_{\max}} > 0 \mid \mu_L \leq 0, \phi_L], \tag{8}$$

for posterior correctness, survival of a correct readout, and correction of a wrong one.

Proposition 1 (Depth-dependent Chow rule) *Stopping at L has expected accuracy at least that of continuing to $L_{\max}$ iff*

$$p_L \ \geq \ \tau_L^\star \ := \ \frac{q_L^-}{q_L^- + 1 - q_L^+} \qquad (q_L^- + 1 - q_L^+ > 0). \tag{9}$$

The optimal one-step rule is a threshold on posterior correctness, and the threshold depends on depth only through the residual channel $L \to L_{\max}$. The multi-exit policy of equation 11 applies this rule greedily at each candidate; optimality over the sequence is not claimed.

Corollary 1 (Precision regimes) *Under equation 7: (i) if $\sigma_{L \to L_{\max}} = 0$ and $g_{L \to L_{\max}} \geq 0$ a.s., then $\tau_L^\star = 1$ and no sample should stop; (ii) if $\sigma_{L \to L_{\max}} = 0$ but $g_{L \to L_{\max}} < 0$ with positive probability (a maturity regression on the remaining path), then $\tau_L^\star < 1$ and early stopping helps some samples in full precision; (iii) if $\sigma_{L \to L_{\max}} > 0$ then $q_L^+ < 1$ and $\tau_L^\star < 1$, and $\tau_L^\star$ falls toward $\frac{1}{2}$ as σ grows, $\frac{1}{2}$ being the limit of the local model rather than a statement about the network. Quantization lowers the optimal threshold and enlarges the set of samples for which stopping is optimal.*

Corollary 2 (Direction of the schedule) *$\tau_L^\star$ decreases with L whenever the remaining loss probability $1 - q_L^+$ shrinks more slowly than the remaining correction probability q_L^-, i.e., when the residual channel is perturbation-dominated, which a rising $\mathrm{NSR}(L)$ makes the deep layers.*

Table 1: INT8 CLIP, ImageNet-1K evaluation split, mean ± std over three gate seeds. Exit sets come from equation 12 without per-backbone tuning.

Backbone	$\mathcal{L}_{\text{exit}}$	Baseline (%)	LRA-EE (%)	Δ (%p)	FLOPs saved (%)
ViT-B/32	$\{4, 6, 8, 10, 11\}$	57.38	60.65±0.03	+3.27	11.1±0.3
ViT-B/16	$\{4, 8, 9, 11\}$	30.06	34.33±0.08	+4.27	14.5±0.2
EVA-CLIP	$\{4, \ldots, 11\}$	57.41	60.57±0.04	+3.16	13.4±0.1
ViT-L/14	$[8, 23] \setminus \{17, 19\}$	59.59	60.25±0.05	+0.66	7.9±0.0

Table 2: Early-exit and adaptive-inference baselines on the same frozen INT8 ViT-B/32 and evaluation split (details in Appendix D.1).

Method	FLOPs saved (%)	Top-1 (%)	Δ (%p)
Full-depth INT8 baseline	0.00	57.38	–
CLS-EE (Teerapittayanon et al., 2016)	0.00	57.38	+0.00
PABEE (Zhou et al., 2020)	3.64	60.48	+3.10
AdaViT (Meng et al., 2022)	13.50	58.50	+1.12
LRA-EE (ours)	11.1	60.65	+3.27

Proposition 2 (Variance reduction of spatial aggregation) *If per-token perturbations on the decision direction have variance s^2 and pairwise correlation $r \in [0, 1]$, the mean of N tokens has perturbation variance s^2/N_{eff} with $N_{\text{eff}} = N/(1 + (N-1)r) \in [1, N]$, against s^2 for a single token.*

Corollary 1(ii)–(iii) are the two components measured in Section 7.2; Proposition 2 is why a shallow readout is viable: [CLS] is one draw from the channel, SSA averages N correlated draws. For any candidate depth after the window anchor, if $M(L) < M(L-1)$, the drift $g_{L-1\to L}$ is negative while σ^2 can only grow, so a readout at L is dominated by the one at $L-1$ as a stopping point and is excluded below.

4.3 Components and calibration-only selection

Readout and gate. Each $L \in \mathcal{L}_{\text{exit}}$ carries a two-layer MLP ψ_L on equation 1. The gate is a logistic estimate of p_L from three scalars,

$$G_L(x) = \sigma\big(w_1^{(L)} c_L + w_2^{(L)} m_L + w_3^{(L)} \text{FAV}_L + b^{(L)}\big), \qquad c_L = s_{L,(1)}, \qquad \text{FAV}_L = \text{Var}_{j \le d}\Big(\tfrac{1}{N}\textstyle\sum_{i=1}^{N} z_{L,i,j}\Big), \tag{10}$$

trained by binary cross-entropy on a held-out gating split with label $\mathbf{1}\{\hat{y}_L(x) = \hat{y}_{L_{\max}}(x)\}$: agreement with the full-depth prediction, never a class label. Agreement is a label-free surrogate for p_L that equals correctness on the samples full depth gets right and is conservative on the rest, since it discourages every exit that disagrees with full depth, rescues included; the Rescue Effect is obtained despite this label, not through it. The margin is the natural statistic since $|\mu_L| = m_L$ whenever the readout is correct; c_L and FAV_L carry calibration and scale.

Layer-adaptive thresholds. We approximate $\tau_L^\star$ by a one-parameter family and stop at the first layer that clears it:

$$\tau_L = \tau_0 \cdot \max\big(1 - \gamma(L-1),\, \tfrac{1}{2}\big), \qquad \pi_{\tau_0,\gamma}(x) = \min\big(\{L \in \mathcal{L}_{\text{exit}} : G_L(x) \ge \tau_L\} \cup \{L_{\max}\}\big), \tag{11}$$

with (τ_0, γ) chosen on this held-out split, γ restricted to $[0, 0.10]$; calibration selects $\gamma > 0$ on every backbone, as Corollary 2 predicts.

Exit set. One rule fixes the candidates for every backbone. Let $L_0 = \lceil L_{\max}/3 \rceil$ denote the anchor of the candidate window. We retain this anchor and apply maturity-regression filtering to subsequent depths:

$$\mathcal{L}_{\text{exit}} = \{L_0\} \cup \big\{L \in [L_0 + 1,\, L_{\max} - 1] : M(L) \ge M(L-1)\big\}. \tag{12}$$

which removes the shallowest third, the final block, and every maturity regression within the candidate window at depths after its anchor, at the cost of a few FP32 forward passes; it excludes three, four, zero, and two layers on ViT-B/32, ViT-B/16, EVA-CLIP, and ViT-L/14; Layer 9 of ViT-B/32 is excluded although its standalone accuracy exceeds Layer 8's (Appendix D.4). Heads are trained on the training split, while the gate and (τ_0, γ) are fitted on the same 5,000-image held-out split. The evaluation split is not used for model, gate, or threshold selection.

Predictions. (P1) $\Delta_{\text{Rescue}}(\pi) > 0$ under INT8; (P2) SSA beats [CLS] at every candidate depth; (P3) a depth-dependent gate beats one confidence threshold; (P4) Δ_{Rescue} is positive in FP32 and larger in INT8 for the same routes; (P5) no single-depth head recovers the gain (Sections 5.2–7).

5 Experiments

5.1 Setup

ImageNet-1K (Deng et al., 2009; Russakovsky et al., 2015), three disjoint subsets: 30,000 training images (2.3% of the split) for the SSA heads, a single 5,000-image held-out split used both to train the gate and to select (τ_0, γ), and the canonical 10,000-image evaluation subset of the validation set (Wang et al., 2021), which is not used for model, gate, or threshold selection. Backbones: CLIP ViT-B/32, ViT-B/16, ViT-L/14 (Radford et al., 2021) and EVA-CLIP ViT-B/16 (Fang et al., 2023; Sun et al., 2023), with a 5-template prompt ensemble and the full-precision text encoder. The vision encoder's linear weights are quantized by dynamic INT8 post-training quantization, activations stay floating point, and inference runs on CPU. Heads are two-layer MLPs trained for 3 epochs on cross-entropy plus a mean-squared term to the FP32 text-aligned embedding (*distilled supervision*, isolated in Section 6.1); (τ_0, γ) is selected on a two-dimensional grid by accuracy on this held-out split and the gate is the only trained component that sees no label. Backbones are never updated and hyperparameters are shared across backbones (Appendix C). We report Top-1 and FLOPs saving $1 - C(\pi)$ with block-uniform cost (exact for these blocks; excludes heads and gate), mean $\pm$ std over three gate seeds with heads and splits fixed (head-initialization variance is not included); wall-clock cost is measured in Section 5.3.

5.2 Main results: four backbones, one procedure (P1)

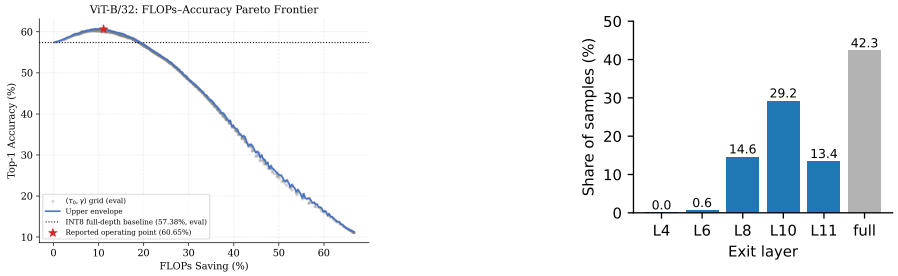
Every backbone has $\Delta_{\text{Rescue}}(\pi) > 0$ while saving 7.9–14.5% of block FLOPs (Table 1); the smallest gain, $+0.66$%p on ViT-L/14, is 13 seed standard deviations at -7.9% compute. ViT-B/16, whose INT8 baseline collapses to 30.06%, gains most; ViT-L/14, whose window is a long run of slowly rising maturity (Layers 12–23), gains least: the effect tracks the profiles, not capacity.

Baselines. CLS-EE, a confidence gate on [CLS] readouts, never exits; PABEE, patience over agreeing readouts, reaches 60.48% but saves 3.64%, since agreement across perturbed readouts is rare; AdaViT, token pruning retrained on the frozen INT8 backbone, reaches 58.50% at a saving within 2.4%p of ours (13.50% vs. 11.1%), since fewer tokens per block does not shorten the channel (Table 2).

Pareto frontier and exit distribution. Unlike the monotone trade-off of conventional early exit, the frontier of Figure 2a rises in both coordinates up to 11% saving; in Figure 2b, 99% of early exits fall at Layers 8–11 although nothing in equation 11 penalizes Layers 4 and 6, so the gate, not a prior, keeps the shallow readouts closed (Appendix D.2).

5.3 Wall-clock cost and a second quantization recipe

Block FLOPs omit the head, the gate, and per-block fixed costs, so we time the policy end to end on the same CPU, with exited samples removed from all later blocks and heads (Appendix F.1). The head-free full-depth model takes 30.65 ms per image at batch size 1; the LRA-EE path with every threshold disabled takes 36.21 ± 0.05 ms, so five head-and-gate evaluations cost 5.56 ms ($+18.1\%$); with exits enabled the same path takes 32.17 ± 0.02 ms. Exits remove 4.04 ms, an 11.2% reduction that matches the 11.1% block-FLOPs saving of Table 1: the FLOPs saving is a wall-clock saving of the same size. The policy is still 1.52 ms ($+5.0\%$) slower than the head-free model because the



(a) (b)

Figure 2: ViT-B/32. (a) Post-hoc visualization, on the evaluation split, of Top-1 accuracy against FLOPs saving over the full (τ_0, γ) grid (grey) and its upper envelope (blue). The reported operating point (star) was selected exclusively on the held-out split before evaluation; the evaluation sweep is shown only to visualize the accuracy–compute landscape and is not used for model or threshold selection. (b) Exit distribution of the selected policy (% of samples per exit layer).

Table 3: Head supervision × routing, ViT-B/32. Rows (ii)–(iii) attach an SSA head to the final layer only and receive 4.3× more training data.

#	Configuration	Supervision / routing	Train images	Top-1 (%)
(i)	INT8 baseline ([CLS], native head)	– / full depth	–	57.38
(ii)	SSA-Last	hard label / full depth	128k	53.23
(iii)	SSA-Last	distilled / full depth	128k	53.47
(iv)	LRA-EE	hard label / adaptive	30k	57.88±0.01
(v)	LRA-EE (ours)	distilled / adaptive	30k	60.65±0.03

current head, dominated by its 1,000-way cosine scores, costs more than the blocks it removes; a head below 0.8 ms turns the same exits into a net gain.

Nothing in Section 4 depends on the source of E_L, so we repeat the identical pipeline on ViT-B/32 under static per-tensor W8A8 (observers calibrated on the calibration split). The static recipe perturbs less (full-depth baseline 59.29% against 57.38%). LRA-EE still does not fall below it, reaching $59.72 \pm 0.01\%$ ($+0.43$%p) at 4.8% saving (Appendix F.2). This is the ordering Corollary 1(iii) implies: a smaller perturbation raises $\tau_L^\star$, so the Rescue Effect shrinks with the perturbation that causes it, without changing sign.

6 Ablations

6.1 Routing is necessary (P5)

With 128,117 training images, four times the budget of every other row, a final-layer SSA head trails the baseline by 4.15%p with hard labels and 3.91%p with distillation (Table 3). At the matched 30,000-image budget, routing with hard labels and no FP32 signal already exceeds the baseline ($+0.50$%p) and routing with distillation reaches $+3.27$%p: distillation amplifies routing ($+2.77$%p under identical routes) but cannot replace it.

6.2 Readout and gate ablations (P2, P3)

With backbone, quantization, and data fixed, SSA beats [CLS] at every candidate depth of ViT-B/32: 11.12% vs. 4.01% at Layer 4 ($+7.11$%p), 55.72% vs. 25.12% at Layer 10 ($+30.60$%p), and 56.90% vs. 42.81% at Layer 11 ($+14.09$%p). The gap is smallest where both are weak and largest where SSA is informative and [CLS] is not; Proposition 2 fixes its sign, the semantic term its depth dependence (Appendix D.3).

A single confidence threshold shared across layers, selected by the same leakage-safe procedure, peaks at 57.84% ($\tau = 0.33$, 8.8% saving) against 60.65% at 11.1% for LRA-EE and collapses to

Table 4: Four-quadrant decomposition, ViT-B/32, 10,000 evaluation images (count, share); $\Delta_{\text{Rescue}} = 813 - 489 = +324$ samples $= +3.24$%p.

	Full depth correct	Full depth wrong
Early exit correct	5,249 (52.5%)	**Rescue: 813 (8.1%)**
Early exit wrong	**Loss: 489 (4.9%)**	3,449 (34.5%)

11.12% at $\tau = 0.1$ (Appendix D.5): no shared threshold is both strict enough below $\tau = 0.3$ and permissive enough at Layers 8–11, which is the depth dependence of $\tau_L^\star$. In the per-feature ablation, margin alone reaches 60.47% at 12.0%, confidence alone 58.21%, and all three 60.65%, the ordering Proposition 1 implies.

7 Anatomy of the Rescue Effect

7.1 Four-quadrant decomposition (P1)

The diagonal, 87.0% of samples, is unaffected by the policy; among the remaining 13.0%, early readout rescues 813 and loses 489 (Table 4), matching the three-seed mean gain (Appendix D.8).

7.2 Precision-specific decomposition (P4)

Corollary 1 predicts a positive Δ_{Rescue} in full precision when maturity regresses on the remaining path, and a larger one under INT8 for the same policy. The test is an exact-route control: for each evaluation image we freeze the exit layer chosen by the INT8-calibrated policy and replay the same route through the FP32 model, so precision is the only variable and no FP32 policy is trained. Under identical routes, FP32 yields $\Delta_{\text{Rescue}} = +2.42 \pm 0.06$%p and INT8 $+3.27 \pm 0.03$%p; the per-seed difference is $+0.86 \pm 0.03$%p. Both hold: the precision-independent part is the larger, as regime (ii) implies for a window with three maturity regressions, and the increment is seed-stable with the predicted sign. The control isolates precision at fixed routes rather than comparing against an FP32-calibrated policy, whose gain would measure the saturation component under its own best routing; the claim rests on the sum, $+3.27$%p where compressed models are served (Appendix D.6).

7.3 Gate selectivity and interpretation

The 57.7% of samples the gate stops early are 76.18% correct; the 42.3% it sends to full depth are 39.47% correct, a 36.72%p gap that rules out random selection. The deferred group sits 18%p below the baseline because it is the set near the decision boundary, for which equation 9 declines to stop (Appendix D.7): the 813 rescued samples are fragile samples whose intact margin the remaining perturbation would have crossed, and the 489 lost are the reverse.

8 Limitations

The exit heads are calibrated on 2.3% of the target training split, so the accuracy figures are in-domain; the gate is label-free, and an earlier configuration recovered 60% and 40% of its FLOPs saving on CIFAR-100 and Food-101 from 256 unlabeled target images with no accuracy loss (Appendix E). The margin statistic presumes a fixed label set; retrieval needs a gallery-defined margin.

9 Conclusion

Readout depth is a decision variable, and for quantized CLIP the best readout is often not the last layer. Across four INT8 CLIP backbones, LRA-EE never falls below the full-depth baseline while saving 7.9–14.5% of block FLOPs, and exact-route analysis separates a precision-independent gain from a quantization-specific increment. The practical implication is to measure not only quantization error but also the depth at which encoder maturity stops improving.

Ethics statement

This work studies inference efficiency of public models on a public benchmark.

AI Use Statement

Large language models were used to edit grammar, wording, and LaTeX formatting. They were not used to design experiments, generate or analyze data, or write code that produced any reported number.

Reproducibility Statement

Section 5.1 and Appendix C specify the dataset splits, backbone variants, quantization settings, training objectives, optimization hyperparameters, gate seeds, and threshold-search procedure used in our experiments. Appendix A provides the proofs of the theoretical claims, while Appendices D–F report the extended ablations, precision-controlled analysis, and wall-clock and static-W8A8 measurements.

The supplementary material includes anonymized source code implementing the methodology used for the main LRA-EE experiments on all four backbones (SSA-head training, gate training, and layer-adaptive threshold selection), together with the principal ViT-B/32 ablations and control experiments: the head-supervision/routing ablation (factorial decomposition, Table 3), the per-feature gate ablation, the gate-selectivity analysis, the [CLS]-versus-SSA comparison, the exact-route FP32–INT8 precision control, true early-exit wall-clock measurement, and the static-W8A8 experiment. A README documents the run order and script dependencies. The code uses publicly available CLIP/EVA-CLIP model checkpoints and requires ImageNet-1K access; the dataset itself is not redistributed.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Multi-exit vision transformer for dynamic inference. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 527–536, 2017.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7947–7969, 2021.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2793–2804, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations (ICLR)*, 2020.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642, 2024. doi: 10.18653/v1/2024.acl-long.681.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations (ICLR)*, 2020.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations (ICLR)*, 2020.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teng Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132:581–595, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations (ICLR)*, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7436–7456, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *Computer Vision – ECCV 2016*, pp. 646–661. Springer, 2016.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations (ICLR)*, 2018.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18(187):1–30, 2018.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, 2018.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3301–3310, 2019.

Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 784–794, 2022. doi: 10.1145/3534678.3539260.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, Minghai Qin, and Yanzhi Wang. SPViT: Enabling faster vision transformers via latency-aware soft token pruning. In *Computer Vision – ECCV 2022*, pp. 620–640. Springer, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3519–3529, 2019.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. MBQ: Modality-balanced quantization for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4167–4177, 2025.

Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-ViT: Accurate and fully quantized low-bit vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021.

Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. RepQ-ViT: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17227–17236, 2023.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Darryl Lin, Sachin Talathi, and Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2849–2858, 2016.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1173–1179, 2022.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. FastBERT: A self-distilling BERT with adaptive inference time. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6035–6044, 2020.

Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28092–28103, 2021.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12309–12318, 2022.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 7197–7206, 2020.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12116–12128, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13937–13949, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The right tool for the job: Matching model and instance complexities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6640–6651, 2020.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 55854–55875, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Shengkun Tang, Yaqing Wang, Zhenglun Kong, Tianchi Zhang, Yao Li, Caiwen Ding, Yanzhi Wang, Yi Liang, and Dongkuan Xu. You need multiple exiting: Dynamic early exiting for accelerating unified vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10791, 2023.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. BranchyNet: Fast inference via early exiting from deep neural networks. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 2464–2469, 2016.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601, 2019.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. MobileCLIP: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15963–15974, 2024.

Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 550–558, 2016.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-VLM: Post-training quantization for large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jue Wang, Ke Chen, Gang Chen, Lidan Shou, and Julian McAuley. SkipBERT: Efficient inference with shallow layer skipping. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7287–7301, 2022. doi: 10.18653/v1/2022.acl-long.503.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 9929–9939, 2020.

Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11960–11973, 2021.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21970–21980, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 38087–38099, 2023.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2246–2251, 2020.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. BERxiT: Early exiting for BERT with better fine-tuning and extension to regression. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 91–104, 2021.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 27168–27183, 2022.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10809–10818, 2022.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. In *Computer Vision – ECCV 2022*, pp. 191–207. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. BERT loses patience: Fast and robust inference with early exit. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18330–18341, 2020.

A PROOFS

A.1 THE RESCUE IDENTITY EQUATION 6

For any policy π and any x, $a_{\pi(x)}(x) - a_{L_{\max}}(x) \in \{-1, 0, 1\}$ and

$$a_\pi(x) - a_{L_{\max}}(x) = \mathbf{1}\{a_\pi = 1, a_{L_{\max}} = 0\} - \mathbf{1}\{a_\pi = 0, a_{L_{\max}} = 1\}.$$

Taking expectations gives equation 6. The identity holds for every sample set, every backbone, and every precision; it is not a model assumption. □

A.2 SURVIVAL AND CORRECTION PROBABILITIES UNDER EQUATION 7

Fix x with signed margin $\mu = \mu_L(x)$ and drift $g = g_{L\to L_{\max}}(x)$, and let $\xi = \xi_{L\to L_{\max}}(x)$ have symmetric cdf F_σ. Then $\mu_{L_{\max}} > 0$ iff $\xi > -(\mu + g)$, so

$$\Pr[\mu_{L_{\max}} > 0 \mid x] = 1 - F_\sigma\big(-(\mu + g)\big) = F_\sigma(\mu + g).$$

For a correct readout ($\mu > 0$) this is the survival probability; for a wrong one ($\mu \le 0$) it is the correction probability. Conditioning on the gate features and integrating over the conditional law of (μ, g) gives $q_L^+ = \mathbb{E}[F_\sigma(\mu + g) \mid \mu > 0, \phi_L]$ and $q_L^- = \mathbb{E}[F_\sigma(\mu + g) \mid \mu \le 0, \phi_L]$. Because F_σ is increasing, q_L^+ and q_L^- are increasing in g; because $F_\sigma(u) \to \frac{1}{2}$ as $\sigma \to \infty$ for every fixed u, both tend to $\frac{1}{2}$ as the perturbation grows. □

A.3 PROOF OF PROPOSITION 1

Conditional on ϕ_L, stopping has expected accuracy $\Pr[\mu_L > 0 \mid \phi_L] = p_L$. Continuing has expected accuracy

$$\Pr[\mu_{L_{\max}} > 0 \mid \phi_L] = p_L\, q_L^+ + (1 - p_L)\, q_L^-.$$

Stopping is at least as good iff $p_L \ge p_L q_L^+ + (1 - p_L) q_L^-$, i.e., iff $p_L(1 - q_L^+) \ge (1 - p_L) q_L^-$, i.e., iff $p_L\big(1 - q_L^+ + q_L^-\big) \ge q_L^-$. When $1 - q_L^+ + q_L^- > 0$ this is $p_L \ge q_L^-/(q_L^- + 1 - q_L^+) = \tau_L^\star$. The threshold depends on L only through (q_L^+, q_L^-), which are functionals of the residual channel $L \to L_{\max}$. □

A.4 PROOF OF COROLLARY 1

(i) If $\sigma = 0$ and $g \ge 0$ a.s., then for $\mu > 0$ we have $F_0(\mu + g) = \mathbf{1}\{\mu + g > 0\} = 1$, so $q_L^+ = 1$ and $\tau_L^\star = q_L^-/q_L^- = 1$ whenever $q_L^- > 0$; if also $q_L^- = 0$ the two actions tie and continuing is weakly optimal. (ii) If $\sigma = 0$ and $\Pr[g < -\mu \mid \mu > 0, \phi_L] > 0$, then $q_L^+ < 1$, so the denominator of $\tau_L^\star$ exceeds the numerator and $\tau_L^\star < 1$; any sample with $p_L \in [\tau_L^\star, 1)$ should stop. (iii) If $\sigma > 0$ then $F_\sigma(\mu + g) < 1$ for every finite $\mu + g$, so $q_L^+ < 1$ and $\tau_L^\star < 1$; by Appendix A.2, $q_L^\pm \to \frac{1}{2}$ as $\sigma \to \infty$, hence $\tau_L^\star \to \frac{1}{2}$. The limit is a property of the scalar model, which reduces a K-way decision to the boundary between the top two classes; at perturbations comparable to the gaps to other classes the reduction no longer holds, and the model is used only in the local regime. □

A.5 PROOF OF COROLLARY 2

Write $u_L = q_L^-$ and $v_L = 1 - q_L^+$, so $\tau_L^\star = u_L/(u_L + v_L)$ is increasing in u_L and decreasing in v_L; equivalently $\tau_L^\star = 1/(1 + v_L/u_L)$ is decreasing in the ratio v_L/u_L. Along $L \to L_{\max}$ both u_L and v_L are functionals of the shrinking residual channel. If v_L/u_L is non-decreasing in L—the remaining loss probability shrinks no faster than the remaining correction probability—then $\tau_L^\star$ is non-increasing in L. Under equation 7, v_L is driven by $\sigma_{L\to L_{\max}}$ and u_L by $g_{L\to L_{\max}}$; a rising NSR means the perturbation share of the residual channel grows with L, which is the stated regime. □

A.6 PROOF OF PROPOSITION 2

Let $\eta_1, \dots, \eta_N$ be the per-token perturbations projected on the decision direction, with $\mathrm{Var}[\eta_i] = s^2$ and $\mathrm{Cov}[\eta_i, \eta_j] = r s^2$ for $i \ne j$. Then

$$\mathrm{Var}\Big[\tfrac{1}{N}\textstyle\sum_i \eta_i\Big] = \tfrac{1}{N^2}\big(N s^2 + N(N-1) r s^2\big) = s^2\, \tfrac{1 + (N-1)r}{N} = \tfrac{s^2}{N_{\mathrm{eff}}}.$$

$N_{\text{eff}} = N$ at $r = 0$ and $N_{\text{eff}} \to 1$ as $r \to 1$; for $r < 1$ the aggregate has strictly smaller variance than any single token. LayerNorm in equation 1 applies one per-sample affine map to signal and perturbation alike, so to first order it leaves the ratio unchanged. □

B Maturity Profiles of All Backbones

Tables 5 and 6 give $M(L)$ of equation 2 for every backbone, computed on the same 1,000 calibration images. The first depth L_0 of each candidate window is retained as an anchor. Bold entries mark subsequent maturity regressions inside the window and are exactly the layers equation 12 removes. The rule fires three, four, two, and zero times on ViT-B/32, ViT-B/16, ViT-L/14, and EVA-CLIP respectively; EVA-CLIP has a monotone profile and keeps all eight candidates.

Table 5: $M(L)$, layers 1–11. Bold: subsequent in-window depths with $M(L) < M(L-1)$, removed by equation 12; the window anchor L_0 is retained by definition. Window $[4, 11]$ (ViT-B/32, ViT-B/16, EVA-CLIP); ViT-L/14's window is $[8, 23]$, its first eleven layers shown for completeness only.

Layer	ViT-B/32	ViT-B/16	ViT-L/14	EVA-CLIP
1	0.106	0.204	−0.090	0.155
2	−0.054	0.032	−0.124	0.088
3	−0.164	0.053	−0.244	0.046
4	−0.101	0.001	−0.197	0.094
5	**−0.154**	**−0.095**	−0.169	0.164
6	−0.136	**−0.158**	−0.280	0.209
7	**−0.136**	**−0.173**	−0.098	0.320
8	0.018	−0.082	−0.051	0.377
9	**0.003**	−0.051	0.005	0.472
10	0.143	**−0.059**	0.029	0.618
11	0.402	0.218	0.113	0.807

Table 6: $M(L)$ for ViT-L/14 (24 blocks). Bold: regressions inside the window $[8, 23]$, giving $\mathcal{L}_{\text{exit}} = [8, 23] \setminus \{17, 19\}$.

Layer	$M(L)$	Layer	$M(L)$	Layer	$M(L)$
1	−0.090	9	0.005	17	**0.385**
2	−0.124	10	0.029	18	0.413
3	−0.244	11	0.113	19	**0.357**
4	−0.197	12	0.201	20	0.376
5	−0.169	13	0.263	21	0.416
6	−0.280	14	0.301	22	0.583
7	−0.098	15	0.314	23	0.638
8	−0.051	16	0.401	24	1.000

Noise-to-signal profile. Table 7 lists the two mean squared quantities behind equation 3 for ViT-B/32 on the same 1,000 calibration images. The numerator (INT8–FP32 deviation) increases at every layer; the ratio increases across 9 of the 11 transitions ($r = 0.76$ with depth) and falls at Layers 8 and 12, most sharply at Layer 12, where the FP32 inter-layer change itself jumps from 0.065 to 0.220.

Table 7: $\text{NSR}(L)$ and its two components, INT8 ViT-B/32, 1,000 calibration images.

Layer	1	2	3	4	5	6	7	8	9	10	11	12
INT8 deviation MSE ($\times 10^{-3}$)	1.06	1.84	2.03	2.27	2.87	3.46	4.04	5.42	11.29	30.61	49.05	53.36
FP32 inter-layer MSE ($\times 10^{-3}$)	1149.7	22.00	13.84	10.59	9.78	10.06	10.19	14.75	28.67	73.40	64.85	219.8
$\text{NSR}(L)$ (%)	0.09	8.38	14.63	21.41	29.32	34.43	39.59	36.77	39.37	41.70	75.64	24.28

Two observations connect the maturity tables to Section 5.2. First, the depth at which $M(L)$ becomes positive and stays positive through the rest of the window—Layer 8 on ViT-B/32, Layer 11

on ViT-B/16, Layer 9 on ViT-L/14, Layer 1 on EVA-CLIP—is where the selected policies begin to exit in volume, which is why EVA-CLIP and ViT-B/16 reach the largest FLOPs savings. Second, ViT-L/14's long run of slowly rising maturity (Layers 12–23) leaves little remaining drift for most of its window, so its residual channel is short in the sense of Corollary 2 and its gain is the smallest of the four.

C Experimental Details

Splits. Training subset: a fixed 30,000-image subset drawn from the first 10% of the ImageNet-1K training split (rows (ii)–(iii) of Table 3 use all 128,117 images of that 10%). A single 5,000-image held-out split, disjoint from the evaluation split, is used both to train the gate and to select (τ_0, γ). The evaluation split is not used for model, gate, or threshold selection. The profiles of Section 3 use 1,000 images of the held-out split.

Heads. ψ_L is $d_b \to d_b \to p_b$ with GELU, where $(N_b, d_b, p_b) = (49, 768, 512)$ for ViT-B/32, $(196, 768, 512)$ for ViT-B/16 and EVA-CLIP, and $(256, 1024, 768)$ for ViT-L/14. Loss: cross-entropy plus 15 times the mean-squared error between the ℓ_2-normalized output and the ℓ_2-normalized FP32 text-aligned embedding. Adam, $\mathrm{lr} = 5 \times 10^{-4}$, batch 32, 3 epochs, identical across backbones. On ViT-B/32, the heads at Layers 8, 10, and 11 were trained once in an earlier run with the same objective and reused; the heads at Layers 4 and 6 were trained fresh. All other backbones train every candidate head fresh.

Gate and thresholds. $3 \to 1$ logistic map per candidate layer, BCE, Adam, $\mathrm{lr} = 10^{-2}$, on the 5,000-image held-out split, labels $\mathbf{1}\{\hat{y}_L = \hat{y}_{L_{\max}}\}$. Seeds 42, 123, 2026 re-initialize the gate only. (τ_0, γ) is selected from the 99×11 grid by accuracy on the same held-out split. The resulting policy is then evaluated on the evaluation split, which is not used for selection.

Quantization. `torch.ao.quantization.quantize_dynamic` on every `nn.Linear` of the vision encoder (per-tensor, symmetric INT8 weights, floating-point activations). Text anchors: 5-template ensemble, full-precision text encoder, computed once.

Reproducibility. Splits, seeds (42, 123, 2026), hyperparameters, and rule equation 12 are given in Section 5.1 and Appendix C; the anonymized code reproduces every number on one CPU machine with ImageNet-1K access, and the evaluation split is not used for model, gate, or threshold selection.

Compute. All inference on CPU; head training on the same machine. No GPU is required to reproduce any number in the paper.

D Extended Results

D.1 Baseline configurations

CLS-EE attaches a confidence gate to the [CLS] readout at every candidate layer, with its threshold selected on the held-out split. PABEE stops when a fixed number of consecutive candidate readouts agree, with the patience selected on the same held-out split. AdaViT is retrained with the INT8 backbone held frozen and evaluated at a compute budget within 2.4%p of LRA-EE's FLOPs saving. All three share the frozen backbone and the evaluation split of Table 1.

D.2 Exit distribution and grid

Table 8 gives the exit distribution of the selected ViT-B/32 policy. The full (τ_0, γ) grid underlying Figure 2a is released with the code (`sweep_results.json`).

D.3 Standalone readout accuracy per layer

Table 9 gives the standalone (non-routed) accuracy of both readouts at the shallowest, mid-deep, and deepest candidate layers (Section 6.2).

Table 8: Exit distribution of the selected policy, ViT-B/32.

Exit at	Layer 4	Layer 6	Layer 8	Layer 10	Layer 11	Full depth
Share of samples (%)	0.0	0.6	14.6	29.2	13.4	42.3

Table 9: Standalone (non-routed) Top-1 of [CLS] and SSA heads, INT8 ViT-B/32.

Layer	[CLS] (%)	SSA (%)	Δ (%p)
4	4.01	11.12	+7.11
10	25.12	55.72	+30.60
11	42.81	56.90	+14.09

D.4 Maturity regression versus standalone accuracy

Layer 9 of ViT-B/32 has $M(9) = 0.003 < M(8) = 0.018$ and is excluded by equation 12 although its standalone SSA accuracy exceeds Layer 8's. Standalone accuracy is a property of the readout; the exclusion concerns its value as a stopping point, which also depends on the perturbation it carries. The decoupling recurs on ViT-B/16 and ViT-L/14; EVA-CLIP has no regression and loses no layer (Appendix B).

D.5 Gate: single-threshold sweep and per-feature ablation

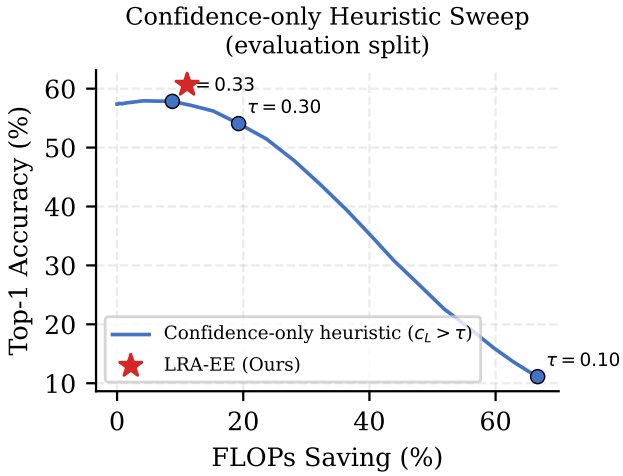


Figure 3: **ViT-B/32, one confidence threshold shared across layers (evaluation split).** Every $\tau \geq 0.5$ saves 0% FLOPs; the best point, $\tau = 0.33$, reaches 57.84% at 8.8% saving; $\tau = 0.30$ gives 54.07% at 19.3%; $\tau = 0.10$ collapses to 11.12% at 66% saving. LRA-EE (star) reaches 60.65% at 11.1%.

The single-threshold heuristic of Section 6.2, selected by the same leakage-safe procedure, has no operating point that is both cheaper and at least as accurate as full depth beyond 8.8% saving: its curve peaks at 57.84% and then falls monotonically, whereas the calibrated gate rises to 60.65% before falling (Figure 2a). The difference is the depth dependence of $\tau_L^\star$ in equation 9: a shared threshold strict enough to avoid the error-prone region below $\tau = 0.3$ is too strict to release the confidently correct samples at Layers 8–11.

Per-feature gate ablation. Table 10 retrains the $3 \to 1$ gate on every non-empty subset of its three features using the cached held-out-split features and reruns the (τ_0, γ) search on the same held-out split, three seeds each. Proposition 1 makes the posterior correctness p_L the decision statistic and equation 5 makes the margin its closest observable, so the margin should dominate. It does: the

margin alone recovers 60.47% at 12.0% saving, confidence alone 58.21%, and feature-activation variance alone 56.68%, below the full-depth baseline—FAV carries scale, not correctness, and is not a usable signal on its own. The full gate is the most accurate (60.65%, four seed standard deviations above confidence+margin) at a slightly lower saving than the margin alone; a margin-only gate is therefore a one-parameter deployment option that gives up 0.18%p for 0.9%p more saving. Two further readings: confidence alone with per-layer thresholds (58.21% at 10.1%) beats the same feature under one shared threshold (57.84% at 8.8%, Figure 3), which isolates the value of the layer-adaptive schedule; and no subset falls below the baseline except FAV alone, so the non-degradation property does not hinge on the feature choice.

Table 10: Gate feature ablation, INT8 ViT-B/32, three gate seeds; full-depth baseline 57.38%.

Gate features	Top-1 (%)	FLOPs saved (%)
Confidence only	58.21±0.07	10.1±0.4
Margin only	60.47±0.06	12.0±0.3
FAV only	56.68±0.15	8.3±0.4
Confidence + margin	60.53±0.04	11.3±0.2
Confidence + FAV	58.37±0.01	8.3±0.0
Margin + FAV	60.48±0.12	11.8±1.4
All three (LRA-EE)	60.65±0.03	11.1±0.3

D.6 Precision decomposition under identical routes

For each gate seed the INT8-calibrated policy fixes a route per evaluation image; the FP32 model is read out at the same layer for the same image, so no FP32 policy is ever trained or selected. Net Rescue is $+2.42 \pm 0.06$%p in FP32 and $+3.27 \pm 0.03$%p in INT8 over the three seeds; the per-seed difference is $+0.86 \pm 0.03$%p and is positive for every seed.

D.7 Gate selectivity stratification

Table 11: Stratification of the evaluation split by the gate's decision, INT8 ViT-B/32.

Group	Share (%)	Top-1 (%)
Early exit (any $L \in \mathcal{L}_{\text{exit}}$)	57.7	76.18
Full depth	42.3	39.47
Population (full-depth baseline)	100.0	57.38

The 36.72%p gap between the two groups is the empirical counterpart of $p_L \geq \tau_L^\star$ in equation 9: the gate stops the samples whose posterior correctness is high and defers the boundary cases. That the deferred group is 18%p below the population baseline confirms that deferral selects hard samples, and that the Rescue Effect does not come from stopping them.

D.8 Per-class rescue counts

Ranking the 1,000 classes by the number of rescued samples (correct at exit, wrong at full depth) concentrates the Rescue Effect on fine-grained natural categories—*sorrel*, *knot*, *admiral*—whose embeddings lie near top-1/top-2 boundaries, consistent with Section 7.3; no class contributes more than 11 of the 813 rescued samples, so the effect is broad rather than driven by a few classes. Figure 4 ranks the ten classes with the most rescued samples.

E Auxiliary Measurements from an Earlier Configuration

The measurements in this appendix were taken with an earlier configuration of the method (exit set $\{8, 10, 11\}$ on ViT-B/32, selected by a noise-based rule rather than equation 12). They are

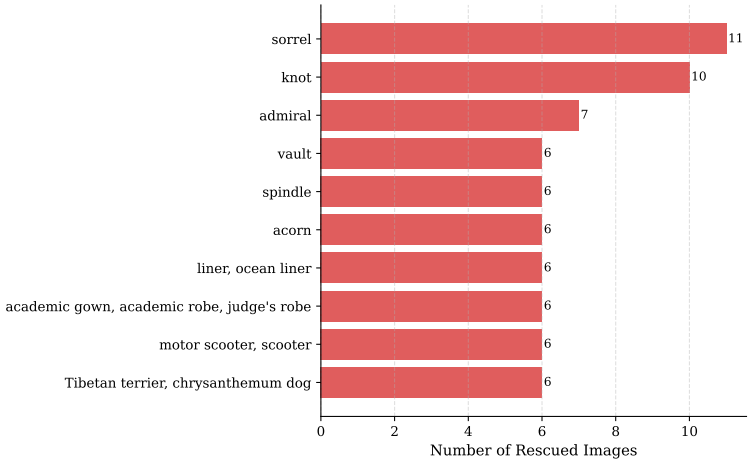


Figure 4: Top-10 ImageNet classes by rescue count (correct at early exit, wrong at full depth), ViT-B/32 evaluation split. The two most rescued classes, *sorrel* (11) and *knot* (10), are fine-grained natural categories; *admiral* (7) follows.

reported because the procedures they test—threshold recalibration and gate-feature redefinition—are unchanged, and because they establish the scope statements of Section 8. They are not part of the main evidence and their absolute numbers should not be compared with Table 1.

Cross-dataset gate transfer. Applying the ImageNet-calibrated gate and thresholds unchanged gave -2.24%p on CIFAR-100 and -9.70%p on Food-101, with FLOPs saving falling below 8%: CLIP's absolute confidence on a 101-way fine-grained label set is systematically lower, so the ImageNet thresholds are too strict. A one-parameter shift of the thresholds by the mean confidence of 256 unlabeled target images recovered about 60% (CIFAR-100) and 40% (Food-101) of the ImageNet-level FLOPs saving with no accuracy degradation. The failure is calibration-bound, as Proposition 1 implies for a threshold that depends on the label space.

Retrieval. Transferring the classification-calibrated gate to Flickr30k image-to-text retrieval unchanged reduced R@1 by 12.4%p (74.90 $\rightarrow$ 62.50) and threshold recalibration alone did not repair it (58.90): the confidence and margin features are defined on a 1,000-way label simplex that retrieval lacks. Redefining both features on the gallery-similarity distribution recovered a near-lossless operating point (R@1 74.30, -0.60%p) at 0.3% FLOPs saving, and the full threshold sweep traced a monotone trade-off with no region of simultaneous gain. Per-query ranking over a large gallery demands finer angular precision than 1,000-way classification, so retrieval is a conventional trade-off regime for this method; the backbone-side analysis of Section 3 is task-agnostic, the gate is not.

F Wall-clock Cost and a Second Quantization Recipe

Both measurements use the frozen backbones of Section 5.1 and are reproduced by the released code.

F.1 Wall-clock accounting

The benchmark implements true early exit: a sample that exits at a candidate layer is not computed by any later block or head. Two correctness checks precede the timings. The block-by-block forward pass used for early exit reproduces the reference full-depth features exactly (maximum absolute difference 0 over the evaluation split), and with every threshold set so that no sample exits, the LRA-EE path reproduces the full-depth predictions exactly. The 500-image timing subset has the same exit distribution as the 10,000-image evaluation split (Figure 2b). Timings are five-run means at batch size 1 on the CPU used for all other experiments, dynamic INT8 ViT-B/32.

Table 12: End-to-end CPU cost per image, INT8 ViT-B/32, batch size 1, five-run mean.

Path	Latency (ms)	Relative to the row above
Full depth, no heads or gate	30.65	–
LRA-EE path, all thresholds disabled	36.21±0.05	+5.56 ms (+18.1%): five heads and gates
LRA-EE path, exits enabled	32.17±0.02	−4.04 ms (−11.2%): skipped blocks and heads

The 11.2% wall-clock reduction from exits matches the 11.1% block-FLOPs saving of Table 1: on this platform block cost is uniform and the FLOPs proxy is exact. The 5.56 ms of head cost covers five SSA heads (two-layer MLP on the 768-dimensional aggregate, projection to 512, cosine scores against 1,000 text anchors) and five gates, 1.1 ms each, against 2.55 ms per block; the head's arithmetic is below 1% of a block's, so the measured cost is implementation overhead rather than a property of the method. The net position at batch size 1 is $+1.52$ ms ($+5.0\%$) relative to the head-free model; a head below $4.04/5 \approx 0.8$ ms makes the same policy a net wall-clock gain, and a head fused into the block that produces its input removes the cost almost entirely.

F.2 Static W8A8 on ViT-B/32

The vision encoder's six linear maps per block (query, key, value, output, and the two MLP layers) are quantized with static per-tensor INT8 weights and activations (symmetric weights, affine activations), observers calibrated on the held-out split; embeddings and LayerNorm stay in floating point. The identical pipeline follows: equation 12 is FP32-only, so the exit set $\{4, 6, 8, 10, 11\}$ is unchanged; the SSA heads are retrained on W8A8 hidden states (30,000 images, 3 epochs); the gate and (τ_0, γ) are fitted on the same held-out split. The evaluation split is used only after observer calibration, gate training, and threshold selection are fixed, and is not used for any of these choices. A second, independently written implementation reproduced every figure to the reported precision.

Table 13: Static W8A8 versus the main recipe, ViT-B/32.

Regime	Baseline (%)	LRA-EE (%)	Δ (%p)	FLOPs saved (%)
Dynamic W8, FP activations (main)	57.38	60.65±0.03	+3.27	11.1±0.3
Static W8A8	59.29	59.72±0.01	+0.43	4.8

The static recipe leaves the representation less perturbed (its full-depth baseline is 1.91%p above the dynamic one), and the Rescue Effect shrinks accordingly while keeping its sign. This is the ordering Corollary 1 predicts: with a smaller $\sigma_{L \to L_{\max}}$ the optimal threshold $\tau_L^\star$ moves back toward 1, fewer samples satisfy the stopping condition, and the policy saves less compute while still never degrading accuracy. The size of the effect is a property of how much a recipe perturbs the readout direction, not of the routing rule.

G Extended Related Work

Four bodies of work meet in this paper. We group them by the question each answers and state, for each, what LRA-EE adds.

Where a transformer can be read out. Linear probes (Alain & Bengio, 2017) and representation-similarity analyses (Kornblith et al., 2019; Raghu et al., 2021) established that intermediate layers of deep networks carry usable, and sometimes better, features; Tenney et al. (2019) located linguistic structure in the middle of BERT, and Skean et al. (2025) showed that mid-depth embeddings beat final-layer embeddings on 32 embedding tasks across transformers and state-space models. The residual-stream view of depth (He et al., 2016; Veit et al., 2016; Huang et al., 2016) explains why: blocks add increments to a shared stream, so late blocks can be dropped (Fan et al., 2020; Gromov et al., 2025) or lose rank (Dong et al., 2021; Noci et al., 2022) without destroying the representation. These are population-level statements about full-precision models. We add the quantized case, in which each block also writes a perturbation into the stream, measure the semantic and the perturba-

tion profile per layer (Section 3), and turn the two into a per-sample readout rule rather than a fixed truncation depth.

Early exit and adaptive inference. Reading out before the last layer descends from the reject option of Chow (1970) and from selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017): a classifier abstains, or here defers to deeper computation, when its confidence is low. BranchyNet (Teerapittayanon et al., 2016), MSDNet (Huang et al., 2018), and adaptive cascades (Bolukbasi et al., 2017) instantiated this for CNNs; DeeBERT (Xin et al., 2020), FastBERT (Liu et al., 2020), PABEE (Zhou et al., 2020), BERxiT (Xin et al., 2021), SkipBERT (Wang et al., 2022), depth-adaptive decoding (Elbayad et al., 2020; Schuster et al., 2022), and LayerSkip (Elhoushi et al., 2024) for language transformers; DVT (Wang et al., 2021), multi-exit ViTs (Bakhtiarnia et al., 2021), and token-level schemes (Rao et al., 2021; Yin et al., 2022; Meng et al., 2022; Kong et al., 2022; Kim et al., 2022) for vision; and MuE (Tang et al., 2023) for unified vision–language models (see Han et al. 2022 for a survey). Kaya et al. (2019) named the phenomenon these methods can exploit, overthinking, and Schwartz et al. (2020) showed that instance difficulty predicts the depth an input needs. Three properties separate LRA-EE from this line. Each method above targets a full-precision model, so its exit criterion is calibrated for a regime in which extra depth is at worst wasted; under INT8 we find that extra depth is harmful for more samples than it helps after Layer 8, with 8.1% rescued against 4.9% lost (Section 7.1). Each vision method above reads the [CLS] token, which we measure to be 7.11–30.60%p worse than the patch-token average at intermediate depth (Section 6.2). And each method above gates on one confidence signal with one threshold, whereas the noisy-channel model calls for a margin-based gate whose threshold depends on depth (Section 4.2); on ViT-B/32 the single-threshold heuristic reaches 57.84% at 8.8% FLOPs saving against 60.65% at 11.1% for LRA-EE (Section 6.2).

Post-training quantization of transformers. Integer-only inference (Jacob et al., 2018) and its post-training variants—adaptive rounding (Nagel et al., 2020), block reconstruction (Li et al., 2021), and learned step sizes (Esser et al., 2020)—minimize per-layer reconstruction error. Transformer-specific work traced the dominant error to activation outliers (Bondarenko et al., 2021; Dettmers et al., 2022; Wei et al., 2022; Xiao et al., 2023) and to the attention and LayerNorm statistics of ViTs (Liu et al., 2021; Yuan et al., 2022; Lin et al., 2022; Li et al., 2023; 2022); weight-only schemes (Frantar et al., 2023; Lin et al., 2024; Yao et al., 2022) carry the recipe to large models, and Q-VLM (Wang et al., 2024) and MBQ (Li et al., 2025) adapt it to large vision–language models through cross-layer dependency and modality-specific sensitivity. That quantization noise accumulates across layers has been analyzed since fixed-point CNNs (Lin et al., 2016; Hubara et al., 2018; Banner et al., 2019). Every method above holds the readout depth fixed and makes each block cheaper or more faithful. LRA-EE holds the quantizer fixed and changes the depth, so the two are complementary; our measurement that the INT8 deviation reaches 75.64% of the native inter-layer change at Layer 11 states how much error the best per-block quantizer would still have to remove at the final readout. Compression of CLIP itself (Wu et al., 2023; Vasu et al., 2024) yields a smaller full-depth model and is orthogonal to depth-adaptive readout.

Adapting a frozen CLIP, and why the margin is the right signal. Prompt learning (Zhou et al., 2022; Shu et al., 2022), feature adapters (Gao et al., 2024), and weight interpolation (Wortsman et al., 2022) improve a frozen or lightly tuned CLIP on a target domain by changing what the final embedding is compared against or how it is formed. LRA-EE changes where the embedding is taken from, uses 2.3% of the ImageNet training split for the exit heads and no labels for the gate, and leaves both encoders untouched, so it composes with these methods rather than replacing them. The geometry that makes readout depth matter—classification by cosine direction on a hypersphere (Wang & Isola, 2020) with a persistent gap between image and text embeddings (Liang et al., 2022)—is also what makes the margin, rather than the maximum softmax probability (Hendrycks & Gimpel, 2017; Guo et al., 2017), the natural gating statistic: the margin is the angular distance to the nearest decision boundary (Schapire et al., 1998), and the perturbation a block adds either clears that distance or does not.