RISK-CONTROLLED POST-COMPRESSION VERIFICATION FOR VISUAL TOKEN REDUCTION WITH EXACT DENSE FALLBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

When a system trusts a compressed visual prediction on a particular input, how often does that choice alter the prediction of a frozen dense teacher? Aggregate accuracy and analytical compute do not answer this deployment-contract question. We study post-compression verification: a frozen vision transformer runs a token-merged continuation, accepts its output only when a calibrated confidence rule passes, and otherwise restores an exact dense continuation from an uncompressed activation saved before the first merge. Six merge ratios, one hundred thresholds, and four targets are calibrated jointly as one finite family of $2{,}400$ candidates with per-candidate error probability 2.08×10^{-5}. Across four dataset–backbone combinations and three seeds, the 1% target accepts 79.5%–95.4% of examples; the largest per-seed calibration upper bound is 0.991%, while observed conditional dense-teacher disagreement is 0.42%–0.55%. A separately implemented program that does not import the production selector recomputes every reported policy and result. The selected policies use 0.74–0.82 of the analytical transformer compute. On the measured hardware platform, wall-clock gains appear only at batch one and reverse at larger batches, underscoring that analytical compute reduction is not a hardware-speedup guarantee. The study therefore reports four non-interchangeable layers of evidence: certified conditional disagreement, analytical transformer cost, compute saved by exact prefix reuse, and measured end-to-end latency.

1 INTRODUCTION

Vision transformers process an image as a sequence of patch tokens; their attention and feed-forward blocks repeatedly transform this sequence, so reducing the number of tokens can substantially lower the multiply–add count of later blocks. Token pruning, token merging, and related dynamic-computation mechanisms exploit this structure by discarding or consolidating patches that appear redundant (Bolya et al., 2023; Rao et al., 2021; Liang et al., 2022). These approaches are normally evaluated through aggregate accuracy and an analytical cost proxy. Those averages are informative for model comparison, but they do not answer a deployment question that is local to the routing decision: when the system chooses to trust a compressed execution on a particular input, how frequently does that choice alter the prediction that a frozen dense model would have made?

This question differs from ordinary predictive accuracy. A compressed branch can preserve average accuracy while changing a non-negligible subset of individual decisions, and disagreement with the dense teacher is not always an error relative to the ground-truth label. The relevant event depends on the deployment contract. In this work the contract is fidelity to a fixed dense teacher: among examples for which the compressed output is accepted, the conditional probability of disagreeing with the dense prediction should not exceed a user-specified target. This definition isolates the incremental behavioral risk introduced by compression and makes no claim that the dense teacher itself is correct.

Risk-control methods provide a natural language for this deployment contract. Distribution-free risk-controlling prediction sets and learn-then-test show how a held-out calibration sample can select from a pre-specified family while controlling a target risk under exchangeability (Bates et al.,

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.

2021; Angelopoulos et al., 2025). Risk-controlled early-exit systems apply related ideas when deciding whether an intermediate prediction can replace the full model (Jazbec et al., 2024; Bajpai & Hanawal, 2025). Post-compression verification poses a different systems problem: the candidate prediction is produced after the token sequence has already been changed, and a rejection must recover the dense computation without silently changing its output. Three gaps follow. The risk object must be defined on accepted compressed predictions, a fallback should avoid recomputing the dense prefix, and calibration must account for jointly searching compression budgets and acceptance thresholds.

Our system addresses these gaps as a deployment contract rather than as another claim of universally faster token pruning. It saves the uncompressed attention output immediately before the first token merge. A rejected example resumes the dense multilayer perceptron and all later dense blocks from that tensor, whereas an accepted example returns the compressed prediction. The fallback logits therefore equal a standalone dense forward pass without restarting the shared prefix. We calibrate six merge ratios, one hundred thresholds, and four risk targets as one finite family per seed, then select among certified candidates by architecture-derived expected compute. A separately implemented recomputation program that does not import the production selector rebuilds the policies and audits every reported field.

The paper also treats evaluation discipline as part of the method. Certified conditional disagreement answers whether an accepted output respects the contract; analytical transformer cost estimates arithmetic saved by routing; shared-prefix accounting isolates work avoided on rejection; and measured latency reveals device- and batch-dependent overhead. None of these quantities substitutes for another. Accordingly, hardware timing appears as a boundary analysis, not as the premise of the contribution.

We make three contributions.

Contribution 1. A deployment contract for visual token reduction that controls dense-teacher disagreement conditional on accepting a compressed prediction. The contract is instantiated by post-compression verification and an exact shared-prefix dense fallback that reuses the pre-merge attention tensor. ToMe serves only as the deterministic compression branch; no new merging primitive is claimed.

Contribution 2. A pre-specified finite-family calibration and separately implemented audit protocol that makes the cost of model selection explicit. All 48 policies select $r = 0.30$; a fixed-budget diagnostic quantifies the resulting multiplicity penalty and motivates pre-registering a deployment budget when that choice is available.

Contribution 3. An evaluation methodology that reports four non-interchangeable layers: certified conditional disagreement, analytical transformer cost, compute saved by exact prefix reuse, and measured end-to-end latency. Across four controlled dataset–backbone combinations and three seeds, we compare the contract with a license-safe risk-controlled early-exit adaptation and report hardware slowdowns wherever dynamic execution loses to dense batching.

2 Related Work

We position our system along three structural difference axes.

Risk-controlled adaptive inference. Selective classification formalizes abstention as a risk–coverage trade-off (Geifman & El-Yaniv, 2017); early-exit and adaptive-depth systems shorten the executed network (Huang et al., 2018; Dehghani et al., 2019; Elbayad et al., 2020; Banino et al., 2021). Risk-controlled early-exit systems further decide whether an intermediate prediction can replace the full model (Schuster et al., 2021; Jazbec et al., 2024; Bajpai & Hanawal, 2025). CATs calibrates intermediate prediction heads to preserve consistency with the original Transformer (Schuster et al., 2021). These decisions occur before the remaining dense layers are evaluated and leave the token sequence unchanged. Our event instead arises after the token representation has been compressed, and rejection must restore an exact uncompressed continuation from a saved state.

Verification with shared computation. SpecVLM uses a verifier to guide visual-token pruning for video LLM speculative decoding (Ji et al., 2025). LayerSkip performs self-speculative language-

model decoding by drafting from early layers and verifying generated tokens with the remaining model, sharing activations between stages (Elhoushi et al., 2024). These works establish the broader value of verification and shared computation, but their verification object is a generated token sequence rather than the conditional dense-teacher disagreement of an accepted image-classification prediction. Our fallback also restores the exact dense classifier state instead of verifying a draft sequence.

Risk control and calibrated routing. Distribution-free risk-controlling prediction sets and learn-then-test select from a predictor family under exchangeability (Bates et al., 2021; Angelopoulos et al., 2025); risk control has been applied to early exiting (Schuster et al., 2021; Jazbec et al., 2024; Bajpai & Hanawal, 2025). These are our closest statistical neighbors. We control conditional disagreement of accepted compressed outputs relative to a frozen dense teacher, not ground-truth error of a prediction set. We instantiate the finite-family binomial guarantee for this risk object and implement a license-safe early-exit adaptation under its own calibration protocol for matched-target comparison (§5).

Token reduction as an accuracy–compute frontier. ToMe is our compressor (Bolya et al., 2023); DynamicViT and EViT are fixed compression baselines (Rao et al., 2021; Liang et al., 2022). They optimize an accuracy–compute frontier and do not offer a finite-sample guarantee on the accepted subset relative to a dense teacher. Conversely, a contract that can fall back may use more expected compute than an always-compress method. The two settings answer different deployment questions and are reported separately.

3 METHOD

3.1 PROBLEM FORMULATION

We use ToMe (Bolya et al., 2023) as the deterministic compression branch and apply a standard finite-family union-bound correction to the dense-teacher disagreement risk defined below. Let $f_0 : \mathcal{X} \to \mathbb{R}^K$ denote the logits of a fixed dense classifier and let f_r denote a compressed classifier parameterized by a merge ratio $r \in \mathcal{R}$. The dense and compressed predictions are

$$y_0(x) = \arg\max_k f_0(x)_k, \qquad y_r(x) = \arg\max_k f_r(x)_k, \tag{1}$$

where y_0 and y_r are the dense and compressed predicted classes, and $\arg\max_k$ returns the class $k \in \{1, \dots, K\}$ with the largest logit. The compressed branch also produces an uncertainty score

$$s_r(x) = 1 - \max_k \mathrm{softmax}(f_r(x))_k, \tag{2}$$

where $\mathrm{softmax}(f_r(x))_k$ is the k-th compressed class probability, so $s_r(x) = 0$ on a fully overconfident compressed prediction. We use maximum softmax confidence as a simple, deterministic preset score; we do not claim it is optimal and we leave a score-function ablation to future work (§7). For a threshold τ, the policy accepts the compressed output when $s_r(x) \le \tau$ and otherwise returns the dense output. Its acceptance indicator is $A_{r,\tau}(x) = \mathbb{1}\{s_r(x) \le \tau\}$. The disagreement event is $D_r(x) = \mathbb{1}\{y_r(x) \neq y_0(x)\}$. The target quantity is the accepted-subpopulation conditional risk

$$R(r, \tau) = \Pr(D_r(X) = 1 \mid A_{r,\tau}(X) = 1), \tag{3}$$

where X is a deployment example drawn exchangeably with the calibration data, $A_{r,\tau}$ is the acceptance indicator, and D_r is the dense–compressed disagreement event, so $R(r, \tau)$ conditions disagreement on the accepted subset only.

For a user-specified level α, the calibration objective selects a policy with $R(r, \tau) \le \alpha$ while minimizing expected computation. This guarantee is relative to the dense teacher, not to the true class; ground-truth test labels are used only to report end-to-end accuracy after the policy is frozen. The merge-ratio set is $\mathcal{R} = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, the threshold set is $\mathcal{T} = \{0.00, 0.01, \dots, 0.99\}$, and the reported targets are $\mathcal{A} = \{0.01, 0.025, 0.05, 0.10\}$. The simultaneously corrected candidate family therefore contains $|\mathcal{R}| \times |\mathcal{T}| \times |\mathcal{A}| = 6 \times 100 \times 4 = 2400$ policy–target pairs per seed. All ratios, thresholds, targets, tie-breaking rules, and cost formulas are fixed before calibration disagreement labels are inspected.

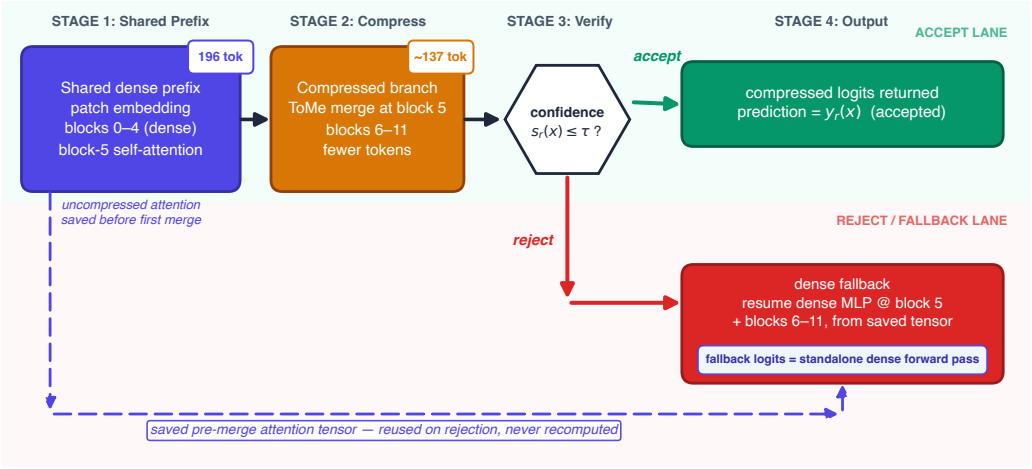

Figure 1: Shared-prefix post-compression execution. The uncompressed attention output of block 5 is saved before the first merge; the merged branch continues through blocks 6–11, and a rejected prediction resumes dense inference from the saved tensor (dashed arrow) without recomputing the shared prefix. Green denotes an accepted compressed output, red a dense fallback.

3.2 Shared-prefix post-compression execution

The dense backbone contains twelve transformer blocks. The compressed path merges tokens in zero-based blocks 5, 7, and 9. At the first merge block, self-attention is evaluated on the full token sequence and the uncompressed attention output is retained; tokens are then merged before the block's multilayer perceptron and continue through later blocks under the fixed merge schedule. If the compressed output is accepted, its logits are returned.

If the compressed output is rejected, the saved uncompressed attention output enters the dense multilayer perceptron of block 5, followed by the remaining unmodified dense blocks and classifier head. Patch embedding, blocks 0 through 4, and block-5 attention are not recomputed. Because the saved tensor is exactly the tensor used by an ordinary dense forward pass at that boundary, the fallback output is numerically equivalent to dense inference. A restart-on-reject variant serves as an ablation: it recomputes the full dense network after rejection and produces identical logits, confidence scores, acceptance masks, and predictions, differing only in rejected-example cost. This is post-compression verification: the accept-or-fallback decision is made after the efficient branch has produced its output, because pre-compression features could not reliably identify the rare low-risk failures in development experiments. Figure 1 summarizes the accepted and fallback execution paths.

3.3 Finite-family calibration

For a fixed candidate $j = (r, \tau, \alpha)$ and a calibration set of size n, let

$$N_j = \sum_{i=1}^{n} A_{r,\tau}(x_i), \qquad E_j = \sum_{i=1}^{n} A_{r,\tau}(x_i) D_r(x_i), \tag{4}$$

where n is the calibration-set size, N_j is the number of accepted calibration examples for candidate j, and E_j is the number of disagreements among those accepted examples. When $N_j > 0$, E_j is the number of disagreements among accepted calibration examples. Let $U(E_j, N_j; \delta_j)$ be the one-sided Clopper–Pearson upper confidence bound for a binomial proportion at error probability δ_j. Each seed uses family-wise error probability $\delta = 0.05$ over all 2,400 candidates and per-candidate

$$\delta_j = \frac{0.05}{2400} = 2.08333 \times 10^{-5}, \tag{5}$$

where 0.05 is the family-wise error budget shared across the 2,400 candidates and the denominator is the family size per seed. A candidate is certified when $N_j > 0$ and $U(E_j, N_j; \delta_j) \le \alpha$. Among

certified candidates for a target, the selector minimizes architecture-derived expected computation; ties are resolved deterministically by the locked candidate ordering. If no candidate is certified, the policy accepts no compressed outputs and uses dense fallback for all examples, leaving conditional risk undefined rather than reporting it as zero.

The simultaneous statement follows from the union bound. For any fixed candidate, the one-sided exact interval fails to upper-bound its conditional Bernoulli disagreement probability with probability at most δ_j. The probability that any interval fails is at most $\sum_j \delta_j = 0.05$, so on the complementary event every reported bound is valid simultaneously and any selected candidate whose upper bound is at most its target has true conditional disagreement at most that target. The claim applies independently within each seed and requires exchangeability of calibration and deployment examples for the disagreement event; the three seeds are experimental replicates and are not pooled.

Empirically, every one of the 48 selected policies (4 combinations $\times$ 3 seeds $\times$ 4 targets) uses the largest pre-specified merge ratio, $r = 0.30$; Appendix C reports all selected thresholds. Thus the budget coordinate is not empirically active within this grid, although it was part of the selection surface inspected on calibration data. The $2{,}400$-way correction therefore remains the valid correction for this protocol. Replacing it post hoc by the 400 candidates at $r = 0.30$ would not preserve the declared simultaneous guarantee; a tighter correction would require fixing that ratio before inspecting calibration outcomes and rerunning calibration as a new protocol.

Analytical cost is based on attention and multilayer-perceptron multiply–adds using the exact token count before attention and after merging; the cost model is detailed in Appendix A. This is an analytical proxy, not a measured speedup.

4 Experimental Setup

4.1 Data and training

The benchmark uses CIFAR-100 and Tiny-ImageNet with DeiT-Tiny and DeiT-Small (patch size 16, resolution 224), giving four controlled dataset–backbone combinations. We deliberately use this compact matrix to isolate the risk-control and fallback mechanisms from large-model capacity and infrastructure confounds; the evidence therefore supports this controlled setting rather than an ImageNet-scale generalization claim. The two datasets use different official sizes, so the splits are stated separately rather than pooled: for CIFAR-100 the official $50{,}000$ training examples are stratified without replacement into a $45{,}000$-example gradient-training subset and a $5{,}000$-example risk-calibration subset; for Tiny-ImageNet the official $100{,}000$ training examples are stratified into a $90{,}000$-example gradient-training subset and a $10{,}000$-example risk-calibration subset. On both datasets calibration examples never contribute gradients or checkpoint selection, and every baseline uses the same split. CIFAR-100 uses its official test set for blind evaluation; Tiny-ImageNet uses the official labeled validation set because the challenge test labels are unavailable, and the unlabeled test set is not used. Splits and training are repeated for seeds 2027, 2028, and 2029. The dense model uses ImageNet-pretrained parameters with a reinitialized task head; training lasts 20 epochs with AdamW, and we fix the final epoch for every seed so that neither calibration nor test accuracy selects a checkpoint. Full hyperparameters are listed in Appendix B.

4.2 Compute environment

Final training and blind-test evaluation of the dense teachers, the compressed branches, calibration, and the separate recomputation audit ran on cloud instances with NVIDIA RTX 5090 GPUs under CUDA. The main-method compute ratios in Tables 1 and 2 are derived from token-dependent attention and multilayer-perceptron multiply–adds (Appendix A); the early-exit proxy in Appendix 5 instead counts executed equal-token blocks. Neither is a measured wall-clock speedup. Measured end-to-end latency was collected separately on an Apple M3 Pro using the PyTorch MPS backend (§5), and that platform is used only for the hardware boundary, not for training or risk evaluation. Results are therefore not reported as if they were produced on a single device.

Table 1: Jointly calibrated policies at the 1% target. Coverage, observed disagreement, and compute are cross-seed means; Max cert. UB is the largest per-seed calibration upper bound and is the certificate-bearing quantity.

Dataset	Backbone	Cov.	Obs. disagree.	Max cert. UB	Compute
CIFAR-100	DeiT-Tiny	89.31%	0.420%	0.973%	0.7699
CIFAR-100	DeiT-Small	95.39%	0.450%	0.979%	0.7421
Tiny-ImageNet	DeiT-Tiny	79.47%	0.550%	0.991%	0.8238
Tiny-ImageNet	DeiT-Small	89.88%	0.500%	0.990%	0.7724

4.3 Calibration freeze, blind test, and separate recomputation audit

Evaluation splits into two stages. The calibration stage saves per-example dense and compressed logits for all six ratios, enumerates the fixed family, selects one policy per target, serializes the policies, and records a cryptographic digest. The blind-test stage builds the test set only after the policy file exists and verifies the digest is unchanged. A separately implemented recomputation program consumes only saved logits, labels, and serialized outputs (not the production selector) and rebuilds every stage; every checked field agrees across the three seeds. Audit counts are in Appendix D.

4.4 Baselines and hardware measurement

We train DynamicViT (Rao et al., 2021) and EViT (Liang et al., 2022), each at three keep rates $(0.7, 0.8, 0.9)$, across the four combinations and three seeds, yielding 72 fixed-baseline runs that are independently audited. Each fixed baseline compresses every test example, so its disagreement is unconditional over all test examples and it runs a *different deployment contract* from our risk-controlled policy (§5.3). The license-safe early-exit adaptation uses a single intermediate exit-layer head accepted under a calibrated threshold, selected from a $100 \times 4 = 400$-candidate family with $\delta = 0.05/400 = 1.25 \times 10^{-4}$; this is an adaptation of the early-exit idea, not an exact reproduction of the original paper. Hardware timing uses an Apple M3 Pro with the PyTorch MPS backend, DeiT-Small checkpoints, batch sizes $1, 16, 64$, and 10 warmup plus 50 timed iterations with synchronization; speedup is dense mean latency divided by the comparison mean latency. Across the paper, variability is reported as cross-seed sample standard deviation (Table 7), and the shared-prefix reuse is ablated against a restart-on-reject variant in §5.4.

5 Results

5.1 RQ1: Does verification buy a controlled fidelity–compute trade-off?

Table 1 reports the jointly calibrated policies at the 1% target. Across all four combinations, they accept 79.5%–95.4% of blind-set examples, use 0.74–0.82 analytical transformer compute, and have observed conditional disagreement of 0.42%–0.55%. More importantly, the largest per-seed calibration upper bound in each combination is 0.973%–0.991%, so every selected policy satisfies the declared certificate; the blind-test estimate is reported separately. Mixed accuracy remains within 0.10 percentage points of the dense teacher.

The strongest self-comparison is available on CIFAR-100/DeiT-Tiny (Table 2). Always accepting the same $r = 0.30$ ToMe branch uses 0.7114 compute but disagrees with the dense teacher on $2.98 \pm 0.23\%$ of examples. Verification lowers observed disagreement among returned compressed predictions to $0.418 \pm 0.056\%$, with $89.31 \pm 0.53\%$ coverage and 0.7699 ± 0.0029 compute. Thus, on this combination, approximately 0.059 additional analytical compute buys a 2.56-point reduction in disagreement among outputs trusted as compressed, with exact dense fallback for the remainder.

5.2 RQ2: What is the multiplicity cost of searching compression budgets?

All 48 selected policies reach the upper ratio boundary $r = 0.30$. The budget coordinate is therefore inactive in the observed matrix, while its six values still enlarge the simultaneous family from 400 to $2{,}400$. On the three CIFAR-100/DeiT-Tiny seeds, a counterfactual 400-candidate family restricted

Table 2: CIFAR-100/DeiT-Tiny ablations at the 1% target, mean±sample-std over three seeds. The fixed-budget row is a counterfactual protocol valid only if $r = 0.30$ is specified before calibration; Max UB is the worst seed.

Policy	Coverage	Obs. disagree.	Max UB	Compute
Joint family (2,400)	89.31±0.53%	0.418±0.056%	0.973%	0.7699±0.0029
Fixed $r = 0.30$ (400; diagnostic)	90.04±0.03%	0.481±0.053%	0.999%	0.7659±0.0002
Always accept $r = 0.30$	100.00%	2.983±0.231%	—	0.7114

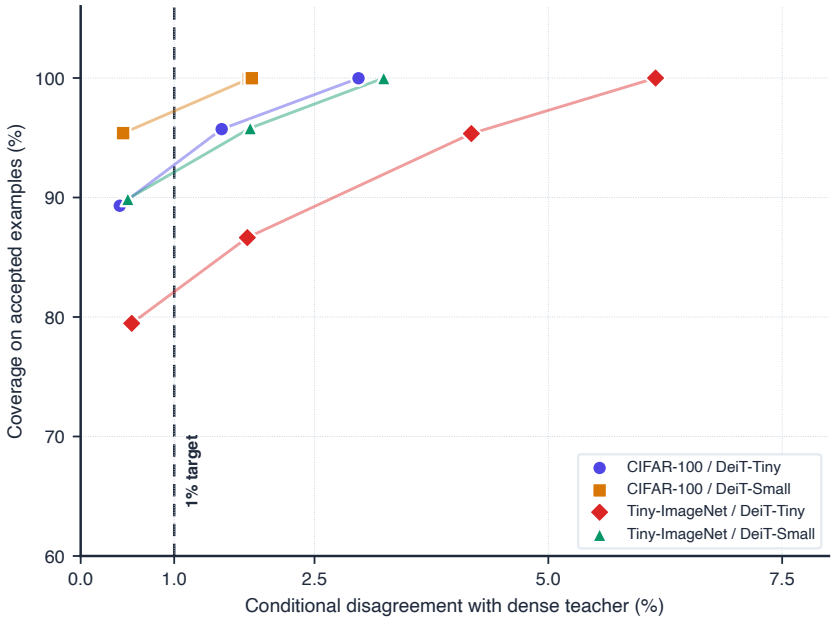

Figure 2: Observed blind-set coverage–disagreement frontier. The dashed line marks 1%; overlapping markers denote targets selecting the same threshold.

to $r = 0.30$ raises mean coverage from 89.31% to 90.04% and reduces compute from 0.7699 to 0.7659; its observed disagreement rises from 0.418% to 0.481%, and its worst calibration upper bound remains below 1%. This quantifies a small but strictly unfavorable multiplicity cost for the full grid. Because $r = 0.30$ was identified after observing the joint calibration outcomes, this row is valid only as the result of a protocol that would have fixed the ratio in advance; it is a diagnostic and does not replace the main certificate.

Figure 2 shows the observed blind-set frontier across targets; certification uses calibration upper bounds, not the plotted test observations. The 5% and 10% policies coincide when they select the same threshold, so the number of distinct policies is smaller than 48.

5.3 RQ3: How does the matched-risk early-exit path behave?

The license-safe early-exit adaptation controls the same conditional-disagreement target, but accepts fewer examples: 44.6%–74.4% versus 79.5%–95.4% for post-compression verification (full results in Appendix 5). Their analytical proxies are not comparable because ours counts token-dependent attention and MLP multiply–adds whereas the early-exit proxy counts equal-token blocks. The common MPS measurement provides a shared system axis on CIFAR-100/DeiT-Small: our path is faster at batch one (1.093× versus 0.614×), while early exit is faster at batches 16 and 64 (0.954× versus 0.738×, and 1.082× versus 0.762×). Among the two adaptive paths, only early exit exceeds dense throughput at batch 64. This reversal shows that the preferred adaptive path depends on batch regime, not only on risk and coverage.

Table 3: End-to-end Apple MPS speedups over dense DeiT-Small. Main, early-exit, and fixed-pruning rows use CIFAR-100; $1.0\times$ denotes dense latency.

Regime	Batch 1	Batch 16	Batch 64
Post-compression vs. dense	$1.093\times$	$0.738\times$	$0.762\times$
Post-compression vs. restart	$1.023\times$	$1.029\times$	$1.029\times$
Early exit vs. dense	$0.614\times$	$0.954\times$	$1.082\times$
DynamicViT keep 0.7 vs. dense	$1.209\times$	$1.287\times$	$1.242\times$
EViT keep 0.7 vs. dense	$1.437\times$	$1.187\times$	$1.162\times$

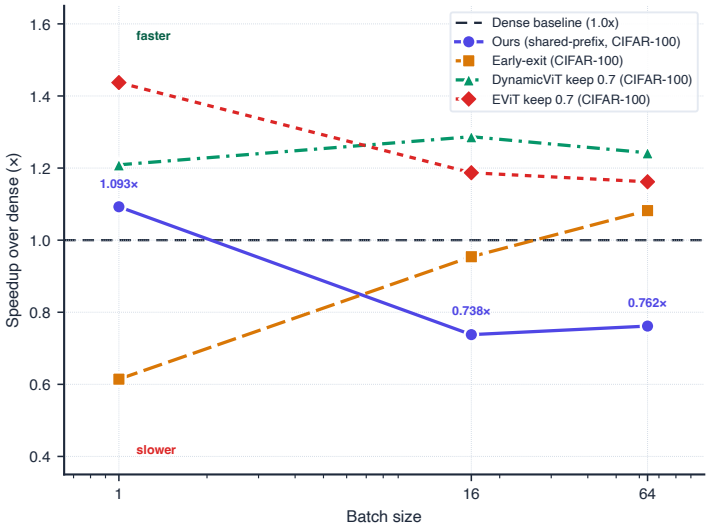

Figure 3: Measured MPS speedup over dense DeiT-Small on CIFAR-100. The dashed line is dense inference ($1.0\times$).

Fixed-pruning context. Across $72/72$ audited DynamicViT/EViT runs, unconditional dense-teacher disagreement ranges from 4.37% to 18.17% (mean 10.01%). These models always compress and never fall back, so this is a different deployment contract, reported as context rather than a superiority comparison; details are in Appendix F.

5.4 Shared-prefix ablation

Because shared-prefix and restart-on-reject execution are prediction-equivalent, their cost difference isolates prefix reuse. On CIFAR-100/DeiT-Tiny, shared-prefix execution saves 4.84 ± 0.24 compute percentage points over restart at the 1% target, consistent with $C_{\text{restart}} - C_{\text{shared}} = (1-q)c_p$. The saving falls to 1.94 ± 0.12 points at 2.5% and approaches zero at 5% and 10% as acceptance nears 100%. Shared-prefix execution is also 1.02–$1.03\times$ faster than restart on hardware across batch sizes. These results justify prefix reuse without claiming it is the only possible fallback.

6 Discussion

The experiments expose when the contract is useful and what it costs. On CIFAR-100/DeiT-Tiny, verification changes a 2.98% always-accept disagreement rate into 0.418% disagreement on accepted outputs for about 0.059 additional analytical compute. The ratio grid saturates at $r = 0.30$, and the fixed-budget diagnostic shows that searching six budgets costs 0.73 coverage points in this setting. At the system level, post-compression verification wins batch-one latency, whereas early exit wins at larger batches. These results connect the statistical contract to both its multiplicity cost and its hardware regime.

The guarantee is deliberately narrower than correctness: it controls agreement with a frozen teacher under exchangeability. Statistical certification, analytical compute, shared-prefix savings, and measured latency remain distinct evidence layers and should not substitute for one another.

7 LIMITATIONS

The fixed DynamicViT/EViT baselines lack a calibration-and-fallback wrapper, so they are a different contract rather than a superiority test. All 48 selected policies reach the ratio-grid boundary $r = 0.30$; this matrix therefore does not demonstrate adaptive budget selection. A smaller correction would require pre-registering that ratio rather than choosing it after calibration. CATs is a close consistency-controlled early-exit neighbor, but we do not implement its vision adaptation; our comparison applies only to the stated license-safe adaptation. Max-softmax confidence is preset without score ablation. The study covers two datasets, two backbones, three seeds, and one MPS latency platform; it does not establish ImageNet-scale, larger-backbone, GPU, or CPU latency behavior. The guarantee assumes exchangeability, not distribution-shift robustness, and controls teacher agreement rather than correctness.

We make no claim of being the first to apply risk control or visual token reduction. Our contribution is this post-compression instance, exact shared-prefix fallback, and audit.

8 CONCLUSION

We presented risk-controlled post-compression verification with exact dense fallback. Across four combinations and three seeds, the 1% policies certify upper bounds no larger than 0.991%, accept 79%–95% of examples, and use 0.74–0.82 analytical compute. On CIFAR-100/DeiT-Tiny, verification reduces always-accept ToMe disagreement from 2.98% to 0.418% on accepted outputs for about 0.059 additional compute. The saturated ratio grid reveals a measurable multiplicity cost, while MPS results show that post-compression and early-exit routing dominate in different batch regimes. The result is a verifiable deployment contract whose certificate, compute, prefix reuse, and latency are reported separately.

REPRODUCIBILITY STATEMENT

All protocol choices (splits, seeds, merge ratios, threshold grid, risk targets, tie-breaking rules, and cost formulas) are fixed before any calibration disagreement label is inspected. Per-example logits are rebuilt by a separately implemented recomputation program that does not import the production selector, and every checked field agrees across the three seeds. The anonymous supplementary bundle contains both main-method audit files (covering all 12 runs), the fixed- and early-exit audits, the shared-prefix ablation results, protocol locks, code, and a SHA-256 manifest. The appendices document compute accounting, training hyperparameters, audit coverage, and per-seed results.

AI USE STATEMENT

Generative AI assisted with literature search and language editing. All cited works and the final manuscript were checked by the authors.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):700–727, 2025. doi: 10.1214/24-AOAS1998.

Divya Jyoti Bajpai and Manjesh Kumar Hanawal. Beyond greedy exits: Improved early exit decisions for risk control and reliability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-5684.

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to ponder. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=1EuxRTe0WN.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Daniel Bolya, Chen-Yi Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SJg7KhVKPH.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed A. 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. URL https://aclanthology.org/2024.acl-long.681/.

Yossi Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. doi: 10.48550/arXiv.1705.08500.

Gao Huang, Danlu Chen, Tianyuan Li, Feng 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. URL https://openreview.net/forum?id=ciTxLiWUJl.

Metod Jazbec, Adam Timans, Tanja Hadži Veljković, Klas Sakmann, Difan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 129825–129854, 2024. doi: 10.52202/079017-4124.

Yicheng Ji, Jiale Zhang, Huijian Xia, Junyang Chen, Mike Zheng Shou, Guandao Chen, and Hang Li. SpecVLM: Enhancing speculative decoding of video LLMs via verifier-guided token pruning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7205–7219, 2025. doi: 10.18653/v1/2025.emnlp-main.366.

Youwei Liang, Chong Ge, Zhongzhe Tong, Yujun Song, Jian Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn˙.

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 34 (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Consistent accelerated inference via confident adaptive transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4962–4979. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.406. URL https://aclanthology.org/2021.emnlp-main.406/.

A Compute accounting

Analytical cost is based on attention and multilayer-perceptron multiply–adds using the exact token count before attention and after merging. For the six ratios, the compressed-branch transformer cost relative to dense execution is $0.9430, 0.8865, 0.8346, 0.7890, 0.7484$, and 0.7114. The shared decision prefix through block-5 attention has a relative cost $c_p = 0.4526$, reflecting that the first merge occurs after attention so the decision block's attention remains dense while its multilayer perceptron can benefit from fewer tokens.

Let c_r denote compressed-branch cost and q the acceptance probability. The shared-prefix expected ratio is

$$C_{\text{shared}}(r, \tau) = c_r + (1 - q)(1 - c_p), \tag{6}$$

because the compressed branch always runs and a rejected input adds only the dense suffix not already shared. Restart-on-reject instead has

$$C_{\text{restart}}(r, \tau) = c_r + (1 - q). \tag{7}$$

The analytical proxy excludes patch embedding, normalization, classification heads, confidence computation, token-matching kernels, indexing, memory allocation, synchronization, and device scheduling. It is therefore reported as a compute ratio rather than a measured speedup; end-to-end timing is measured separately on Apple MPS.

B Training hyperparameters

The dense model uses ImageNet-pretrained parameters with a reinitialized task head. Training lasts 20 epochs with batch size 64, AdamW, learning rate 10^{-4}, weight decay 0.05, three warmup epochs followed by cosine decay, and gradient clipping at 1.0. The final epoch is always used; neither calibration accuracy nor test accuracy selects a checkpoint, and the same fixed-final checkpoint supplies the dense teacher and all compressed branches within a seed. The compressor uses key-based bipartite soft matching, token-size-weighted feature aggregation, and proportional attention. Ratios are applied under the same layer schedule (blocks $5, 7, 9$) for calibration and test. Zero-compression equivalence, full-accept equivalence to an independently evaluated compressed branch, full-reject equivalence to dense inference, and agreement with the fixed upstream patch under identical schedules are tested before formal evaluation.

C SELECTED POLICY PARAMETERS

Table 4 reports the selected policy coordinates. All 48 policies select the upper grid boundary $r = 0.30$; thresholds vary with the risk target, dataset, backbone, and seed. This is an empirical outcome of the locked $2{,}400$-candidate search, not a justification for reducing the family after selection.

Table 4: Selected thresholds τ for seeds 2027/2028/2029. The selected merge ratio is $r = 0.30$ in every cell.

Dataset	Backbone	1%	2.5%	5%	10%
CIFAR-100	DeiT-Tiny	0.32/0.29/0.30	0.50/0.50/0.48	0.91/0.89/0.89	0.91/0.89/0.89
CIFAR-100	DeiT-Small	0.33/0.40/0.37	0.79/0.79/0.80	0.86/0.86/0.88	0.86/0.86/0.88
Tiny-ImageNet	DeiT-Tiny	0.29/0.28/0.31	0.44/0.45/0.46	0.67/0.67/0.65	0.94/0.93/0.95
Tiny-ImageNet	DeiT-Small	0.29/0.30/0.28	0.52/0.55/0.52	0.93/0.94/0.93	0.93/0.94/0.93

D INDEPENDENT RECOMPUTATION AUDIT

The main-method audit covers $12/12$ runs: $3/3$ CIFAR-100/DeiT-Tiny runs ($252/252$ checked fields) in the formal audit and the remaining $9/9$ runs ($2556/2556$ checked fields) in the extension audit, for $2808/2808$ checked fields overall. Both independent programs rebuild policy selection and blind-set metrics without importing the production selector or evaluator. The fixed-baseline audit covers $72/72$ runs (2 methods $\times$ 3 keep rates $\times$ 4 combinations $\times$ 3 seeds) and $4392/4392$ checked fields, including top-1 accuracy, same-pass dense-teacher accuracy, and dense-teacher disagreement recomputed from hashed inputs. The early-exit adaptation audit covers $12/12$ runs (4 combinations $\times$ 3 seeds) and $4116/4116$ checked fields. Both audits use an independent implementation that does not import the production selector, routing module, or evaluator, and both report zero blind-test target violations.

E EARLY-EXIT MATCHED-TARGET RESULTS

Table 5: Post-compression verification and the license-safe early-exit adaptation at the 1% target, cross-seed mean. The compute columns use different proxies and must not be compared numerically across methods.

Dataset	Backbone	Post-compression			Early exit		
		Cov.	Risk	Compute	Cov.	Risk	Compute
CIFAR-100	DeiT-Tiny	89.31%	0.420%	0.7699	62.71%	0.410%	0.8916
CIFAR-100	DeiT-Small	95.39%	0.450%	0.7421	74.37%	0.340%	0.8440
Tiny-ImageNet	DeiT-Tiny	79.47%	0.550%	0.8238	44.57%	0.390%	0.9246
Tiny-ImageNet	DeiT-Small	89.88%	0.500%	0.7724	66.35%	0.570%	0.8687

F FULL FIXED-BASELINE DISAGREEMENT TABLE

Table 6 summarizes the 72 audited fixed baselines, mean$\pm$std over three seeds, dense-teacher disagreement over all test examples (100% coverage). These are always-compress official models under their own deployment contract; they compress every example, never fall back, and are not wrapped with the calibration-and-fallback protocol used by our method. Disagreement grows as fewer tokens are kept and is larger on Tiny-ImageNet.

Table 6: Fixed official baselines, dense-teacher disagreement over all test examples. All $72/72$ runs are independently audited. These always-compress numbers run a different deployment contract from our risk-controlled policy.

Method	Dataset	Backbone	Keep	Disagreement
DynamicViT	CIFAR-100	DeiT-Small	0.70–0.90	4.41–5.69%
DynamicViT	CIFAR-100	DeiT-Tiny	0.70–0.90	4.83–7.95%
DynamicViT	Tiny-ImageNet	DeiT-Small	0.70–0.90	8.46–10.60%
DynamicViT	Tiny-ImageNet	DeiT-Tiny	0.70–0.90	9.15–14.29%
EViT	CIFAR-100	DeiT-Small	0.70–0.90	7.66–8.26%
EViT	CIFAR-100	DeiT-Tiny	0.70–0.90	9.45–10.23%
EViT	Tiny-ImageNet	DeiT-Small	0.70–0.90	12.68–13.98%
EViT	Tiny-ImageNet	DeiT-Tiny	0.70–0.90	15.53–17.81%

G LICENSING NOTE

The submitted supplementary implementation is released under the MIT License. Its vendored ToMe component retains the upstream CC BY-NC 4.0 license and is therefore restricted to non-commercial use; third-party notices are included in the bundle.

H PER-SEED DETAILED RESULTS

Table 7 reports the main method at the 1% target per seed and cross-seed mean$\pm$sample-std. Dense teacher accuracies are 85.33% (CIFAR-100/DeiT-Tiny), 88.75% (CIFAR-100/DeiT-Small), 76.81% (Tiny-ImageNet/DeiT-Tiny), and 84.15% (Tiny-ImageNet/DeiT-Small). Coverage and conditional disagreement are stable across the three independently trained seeds within each combination.

Table 7: Main method at the 1% target, cross-seed mean$\pm$sample-std. Conditional disagreement is defined only on accepted examples; the compute ratio is an analytical proxy, not a measured speedup.

Dataset	Backbone	Coverage	Cond. disagree.	Mixed acc.	Compute
CIFAR-100	DeiT-Tiny	89.31±0.53%	0.420±0.060%	85.26±0.23%	0.7699±0.0029
CIFAR-100	DeiT-Small	95.39±0.69%	0.450±0.160%	88.76±0.22%	0.7421±0.0038
Tiny-ImageNet	DeiT-Tiny	79.47±0.54%	0.550±0.090%	76.76±0.21%	0.8238±0.0030
Tiny-ImageNet	DeiT-Small	89.88±0.18%	0.500±0.080%	84.14±0.15%	0.7724±0.0010