SC-VTP: STATE-CONSISTENT VISUAL TOKEN PRUNING VIA DELETION-CONSISTENT TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual-token pruning requires selector training to reflect deployed computation. In hybrid decoders, attention masks leave discarded tokens active in convolutional history and recurrent memory. State-Consistent Visual Token Pruning (SC-VTP) coordinates key access, history advancement and memory decay/write with one differentiable retention gate. A compositional contract specifies when binary-gated execution recovers native deletion. With a frozen backbone and a 0.52M-parameter selector trained on 26,143 DocVQA questions, three paired seeds improve official-validation ANLS over attention-only training by 5.74 points on average. A separately trained Qwen3.8-27B pair improves ANLS by 10.38 points. Crossed soft/hard evaluation and trained component ablations connect execution fidelity to selector utility. The original seed reaches 93.67 ANLS versus 93.05 for native-deletion policy gradients and transfers to 84.18% TextVQA accuracy. Frozen scorers retain their attention-only advantage at 25% and 75% retention. On 32 hardware-test sources, full-input/SC-VTP geometric latency ratios are 1.16 and 1.04 for one-token and 32-token generation, including encoding and selection. SC-VTP connects deletion-consistent operators to measured pruning performance in hybrid decoders.

1 INTRODUCTION

Text-image understanding depends on small visual details: a number, name or qualifier can determine an answer. Visual-token pruning reduces the decoder sequence while retaining full-resolution encoding. We introduce *State-Consistent Visual Token Pruning* (SC-VTP), which trains selectors through a computation that represents deployed pruning.

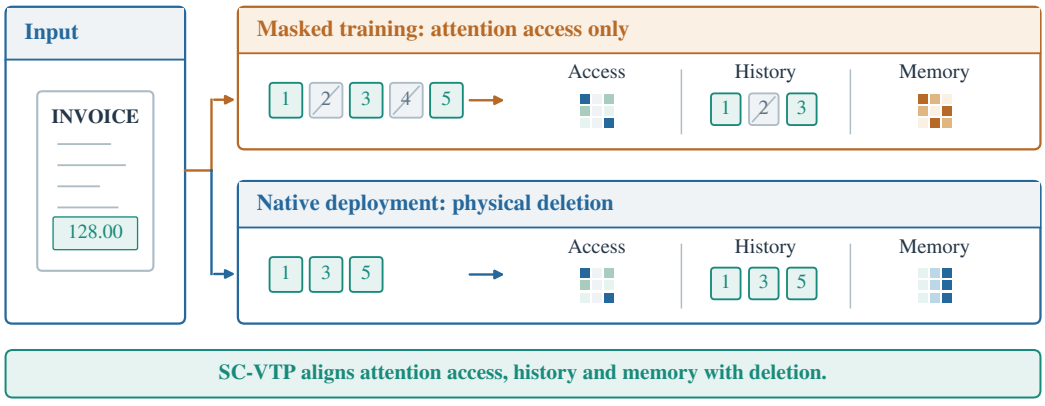

Figure 1: The execution problem behind visual-token pruning. The same selected positions can induce different histories and recurrent states under attention-masked training and physical deletion. SC-VTP controls key access, history advancement and memory updates together. The illustration is schematic; binary equivalence requires the conditions in Section 3.

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 shows the mismatch. Masking keys blocks one information route, while discarded tokens can still advance convolutional history and change recurrent memory. Native deletion skips those transitions. Sequence mismatch is established (Wu et al., 2025; He et al., 2026; Zhang et al., 2026); SC-VTP addresses the downstream training operator with a coordinated, differentiable realization for hybrid decoders.

Our contributions are: (i) a compositional deletion contract preserving logical coordinates while skipping discarded-state transitions; (ii) fixed-gate fidelity probes and trained history/memory ablations; and (iii) complete DocVQA evaluation with three 9B pairs, a 27B pair, task/budget transfer and pruning comparisons. Native-deletion policy gradients provide the direct-reward comparator, including its budget-dependent tradeoff.

2 Related work

Learned and recurrent token reduction. DynamicViT learns attention masks (Rao et al., 2021). Identity skipping, recurrent-state retention and state-space importance are established (Campos et al., 2018; Zhan et al., 2024; Jiang et al., 2026; Park et al., 2026). DyVM rearranges tokens to address Mamba's masked-training/removal mismatch (Wu et al., 2025); DiffPrune uses information throttling and hard top-K deployment (He et al., 2026); StepPrune combines sequential selection, noise-and-repair training and gathering (Zhang et al., 2026). SC-VTP develops a coordinated hybrid-decoder operator with conditional deletion endpoints. Our full-component DiffPrune comparison is a paper-based adaptation; DyVM and StepPrune supply conceptual comparisons.

Document-specific selection. LayoutLite uses policy gradients and layout supervision (Liu et al., 2026); S^2Prune uses spatial coverage and early responses (Jia et al., 2026); DocPrune combines background, question relevance and depth (Choi et al., 2026); ET-Prune adds evidence conditioning and adaptive budgets (Ding et al., 2026). Nüwa studies spatial integrity (Huang et al., 2026). These methods motivate preserving textual evidence. Our image-only scorer isolates the training-operator contribution; published cross-pipeline scores remain separate from matched comparisons.

Hybrid architecture and decoding. Gated Delta Networks provide the native decay and delta-rule updates (Yang et al., 2025); SC-VTP supplies their retention control. MoEMB ranks tokens by GDN write strength (Cui et al., 2026); our single-layer adaptation retains its early full-token computation. DSTP restores discarded evidence during generation (Kim et al., 2026). We evaluate fixed-support pruning for short-answer text-image QA.

3 State-Consistent Visual Token Pruning

In SC-VTP, state consistency coordinates attention access, convolutional history and recurrent memory with deletion. Figure 2 shows differentiable training and hard top-K deployment; the contract below specifies the binary endpoint.

3.1 A module-level deletion contract

Let G_m gather retained rows in order, carrying their original logical coordinates. A module is deletion-consistent if its binary-gated dense execution obeys

$$G_m \mathcal{M}_m(X, p) = \mathcal{M}(G_m X, G_m p). \tag{1}$$

This separates two clocks: a retained token keeps its spatial/rotary coordinate, while a discarded token must not advance a mixer's computational history. For attention, excluded keys must be inaccessible. For a causal state update $s_t = \Phi(s_{t-1}, x_t)$, a sufficient endpoint construction is

$$\Phi_z(s, x) = (1 - z)s + z\Phi(s, x), \qquad \Phi_0(s, x) = s, \quad \Phi_1(s, x) = \Phi(s, x). \tag{2}$$

The state includes every persistent component, such as both LSTM states or the full convolution history, and readout at retained positions matches the native operation. Identity skipping (Campos et al., 2018) supplies this continuous interface.

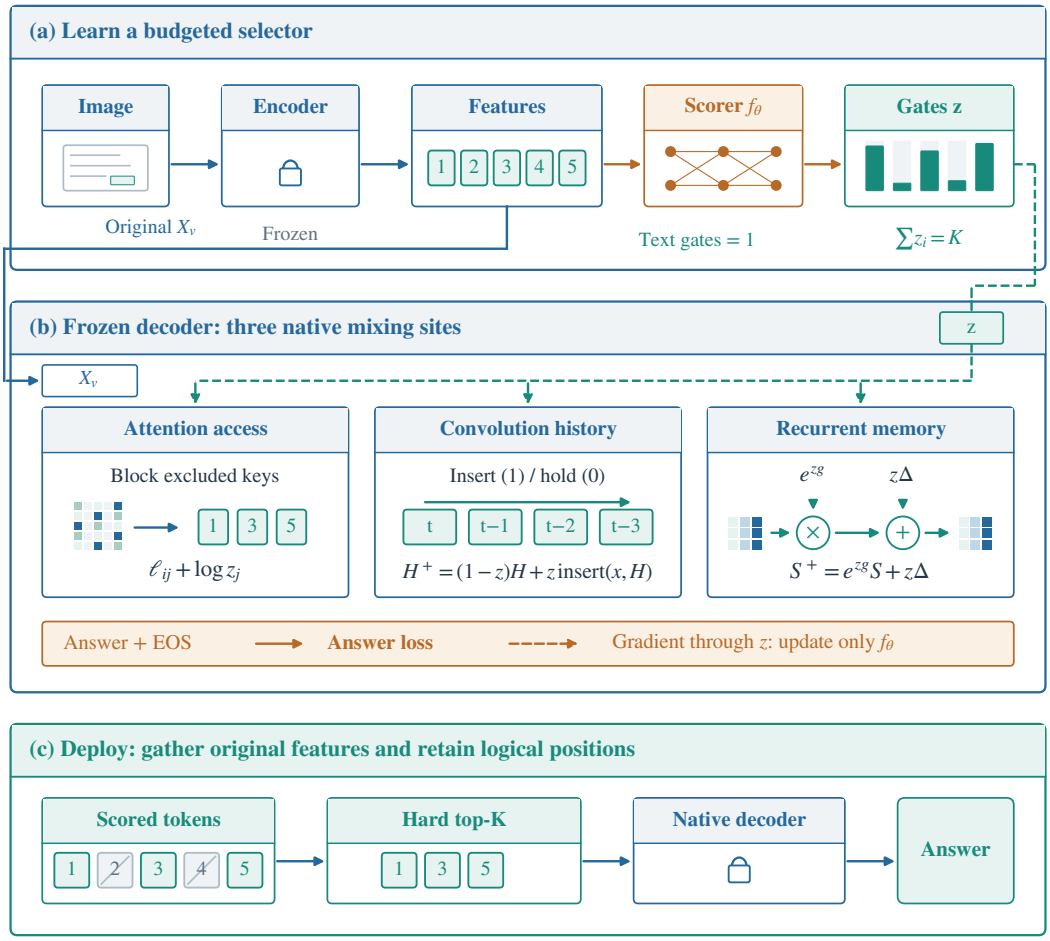

Figure 2: SC-VTP training and deployment. (a) The frozen encoder supplies original image features to the trainable scorer and decoder; question/prefix tokens are retained. (b) A shared gate controls attention columns, the complete history buffer and native memory decay/write at their respective sites, rather than adding a serial three-layer stack. Matrix and queue glyphs illustrate execution; at $z = 0$, both history and memory hold. The native delta write Δ uses the decayed state. Answer-token loss updates only the scorer. (c) Hard top-K gathers original features with their logical positions for native cached generation. Full image encoding is retained in both paths; finite gates are a continuous approximation.

Table 1: Architecture-specific realizations of one deletion contract. Logical positions are preserved in every row. The full pretrained-model study uses Gated DeltaNet/attention hybrid blocks; the other stateful rows have CPU reference qualification only.

Token mixer	Retention control	Binary interpretation
Causal attention	Add $\log z_j$ to key logits	Block excluded keys
Causal convolution	$H' = (1-z)H + z\,\mathrm{insert}(x, H)$	Freeze history
Affine state space	$A_z = (1-z)I + zA$, $b_z = zb$	Identity transition
Nonlinear recurrence	Equation 2 on complete state	Freeze full state
Gated DeltaNet	$g \leftarrow zg$, $\beta \leftarrow z\beta$, plus convolution control	Freeze all three

Composition. The contract composes when every mixer satisfies Equation 1, other operations are positionwise, initial states agree and discarded outputs enter no bypass. Substitution and induction

prove the result (Appendix M). For attention-only decoders with fixed coordinates, SC-VTP reduces to key gating; hybrid decoders require control of all mixing routes.

3.2 SETTING AND SELECTOR

Let a prepared prompt contain visual and text embeddings $h_1, \ldots, h_T$, with visual positions $\mathcal{V}$, $N = |\mathcal{V}|$, and preserved logical positions $p_1, \ldots, p_T$. A scorer predicts

$$s_i = w_2^\top \, \mathrm{GELU}(W_1 \, \mathrm{LN}(h_i) + b_1) + b_2, \quad i \in \mathcal{V}.$$

The frozen encoder supplies spatially contextualized image features. The question-agnostic scorer uses non-affine layer normalization, hidden width 128 and input width 4096/5120 for 9B/27B. The backbone remains frozen; this design isolates the training operator.

For retention budget K, training gates satisfy

$$z_i = \sigma((s_i - \lambda)/\tau), \qquad \sum_{i \in \mathcal{V}} z_i = K,$$

where λ is found by bisection and $\tau > 0$ is the temperature. Text gates are one. We anneal τ from 1.0 to 0.1 with a fixed cosine schedule and differentiate through the threshold's budget dependence (Appendix K). Budgeted soft top-K gates are established, including in DiffPrune (He et al., 2026). SC-VTP applies these gates to the downstream state/history operator.

Attention access. For a retained query, add $\log z_j$ to the causal key logit:

$$A_{ij} = \mathrm{softmax}_{j \le i} \left(q_i^\top k_j / \sqrt{d} + \log z_j \right).$$

Zero gates block keys; unit gates preserve native logits. A retained first text token ensures nonempty causal support. Attention-only training uses this gate while leaving convolutional and recurrent transitions unchanged.

Convolutional history. A depthwise causal convolution uses the current projected vector x_t and the previous $k - 1$ vectors. Let H_t store that history with the newest vector first. We compute the native convolution from x_t, H_t, then update

$$H_{t+1} = (1 - z_t) H_t + z_t [x_t, H_t^{1:k-2}].$$

At zero the entire history holds; at one the native insertion and shift occurs. Binary gates therefore make subsequent convolutions use retained predecessors.

The differentiable history implementation preserves native pre-activation rounding and uses quadratic sequence storage (Appendix A.1).

Gated DeltaNet state. Write the native state in the orientation $S_t \in \mathrm{R}^{d_k \times d_v}$, with column vectors $k_t \in \mathrm{R}^{d_k}$ and $v_t \in \mathrm{R}^{d_v}$. Let g_t be the learned log decay and β_t the native write strength. Omitting head indices, the gated transition is

$$\bar{S}_t = \exp(z_t g_t) S_{t-1}, \qquad S_t = \bar{S}_t + z_t \beta_t k_t (v_t - \bar{S}_t^\top k_t)^\top.$$

The native implementation receives $z_t g_t$ and $z_t \beta_t$: zero gates hold memory and unit gates recover native updates. Write-only gating would still decay memory. History and memory controls apply in every hybrid block.

3.3 BINARY ENDPOINT PROPERTY

The module contract composes to

$$\underbrace{G_m F_m^{\text{gate}}(X)}_{\text{full sequence, skipped transitions}} = \underbrace{F^{\text{native}}(G_m X)}_{\text{compact sequence, native transitions}} . \tag{3}$$

For deterministic, cache-free, unpadded batch-one prefill, this requires retained text, identical logical positions and initial states, positionwise other maps, and every mixing route satisfying the contract. Induction over retained positions and layers gives native execution. Appendix F proves the exact-arithmetic result with finite-activation, empty-history and no-dropout assumptions; BF16 runs and finite gates need not agree.

Finite-gate interpolation. Both e^{zg} and $(1-z)+ze^{g}$ recover the binary decay endpoints. For $g \leq 0$, their gap is at most $\frac{1}{2}g^2 z(1-z)$ (Appendix M); endpoint correctness does not establish better finite-gate gradients. All trained experiments use exponential decay. At $\tau = 0.1$, mean development gate-to-hard L_1 distances are 0.135/0.252 for state/attention scorers: the gates remain soft. Appendix P retains their distributions, without claiming a temperature or trained-path comparison.

3.4 OBJECTIVE AND DEPLOYMENT

Training optimizes only the scorer using teacher-forced cross entropy on the first accepted answer and EOS. Evaluation generation receives only image/question inputs; labels enter scoring after generation.

Inference gathers the top K original visual features in input order, preserving all text and logical spatial positions. The native decoder then generates with its compact cache. Full-resolution encoding remains; soft history operations and answer annotations are used only in training.

4 EXPERIMENTAL PROTOCOL

4.1 DATA PARTITIONS AND EVALUATION HISTORY

We use pinned DocVQA mirrors (Mathew et al., 2021). Of 39,463 official-training questions on 10,194 pages, 10,346 are quarantined for source-document or exact-pixel overlap with validation. A seeded source split yields 26,143 training questions (6,787 page IDs, 6,786 RGB images, 4,274 sources) and 2,974 development questions (750 images, 475 sources), with no active-partition overlap in question, pixels or source. Table 4 gives complete accounting.

Official validation has 5,349 questions, 1,286 page IDs, 1,285 RGB images and 938 sources. All fixed checkpoints use the complete cohort. Known prior repository exposure covers 300 questions from 300 validation sources; backbone exposure is unknown. Source isolation applies to selector training and development.

4.2 TRAINING AND CONTROLS

The primary backbone is frozen Qwen3.5-9B BF16 (Qwen Team, 2026), capped at 1,024 visual tokens with $K = \lfloor N/2 \rfloor$. Each paired seed uses identical initialization, question order and one complete training pass: 3,268 AdamW updates per arm. Appendix A.1 records software versions, all three seed IDs, optimizer settings and schedules.

Table 2 lists the official-validation arms. DivPrune (Alvar et al., 2025) has native support-parity qualification. PG shares our scorer and data but incurs different reward-evaluation training costs (Appendix E). GDN-beta keeps five full-token blocks before deletion. Development adds even-stride, question-only and initial-scorer controls. Stable ties in the zero-initialized final layer select a prefix, not a random support (Appendix B).

Greedy generation uses a 128-token cap. We report question-averaged ANLS (strict normalized edit distance below 0.5) and normalized exact match. The primary state-minus-attention contrast uses 10,000 paired source-bootstrap draws: 938 official sources or 475 development sources, with

Table 2: Complete official validation: DocVQA (D, 5,349 questions, ANLS) and TextVQA (T, 5,000 questions, soft VQA accuracy), scores on a 0–100 scale. Backbones: Qwen3.5-9B and Qwen3.8-27B. R: qualified reference port; P: author-core port; A: paper adaptation; C: our control. All scores are executed here. Learned rows are fixed individual runs; three 9B seeds are summarized separately. Entry pruning retains half the visual rows; GDN-beta retains five full-token blocks. Resize changes the encoder and token budget. Costs and training data are not matched across all methods. Dashes mean unmeasured.

Method	Type	9B D	9B T	27B D
Full input	Reference	94.68	84.52	95.33
DivPrune	R	82.52	82.67	84.25
MMTok	P	72.20	72.96	—
DiffPrune	A	77.96	—	—
Physical-deletion PG	C	93.05	83.66	—
SC-VTP	Ours	**93.67**	**84.18**	**90.43**
Attention-only training	Ablation	84.85	80.80	80.06
GDN-beta	A	75.88	71.01	—
Lower-resolution input	C	92.55	83.35	91.90

95% intervals. Other intervals are descriptive unless separately designated; training-seed variation is reported separately. Appendix A.1 specifies each statistical family.

5 Comparison with prior pruning methods

Table 2 compares prior-method implementations, direct-deletion learning and controls: 48,141 DocVQA generations across nine 9B methods and 40,000 TextVQA generations across eight. Figure 3 adds budget transfer and all three training pairs. Appendix Tables 23 and 21 retain detailed workload counts and separate cross-pipeline published references.

Task quality at half retention. SC-VTP reaches 93.67 DocVQA ANLS and 84.18% TextVQA accuracy at half retention. The MMTok author-core port reaches 72.20/72.96: paired gains are +21.47 [19.83,23.16] and +11.22 [10.05,12.37], with prespecified 97.5% cluster-bootstrap intervals. This tests a backbone port, not the original paper setting (Appendix W).

Full input reaches 94.68 ANLS, so state pruning loses 1.01 points $[-1.44, -0.58]$. Its differences from DivPrune and native lower-resolution input are +11.15 [10.07, 12.26] and +1.12 [0.53, 1.71], respectively; these secondary intervals are unadjusted. Lower resolution changes encoder work and has a nearby, unequal token budget. No speed–quality advantage is inferred.

Comparison with direct-deletion learning. State minus PG is +0.62 [0.15, 1.09] official ANLS, versus -0.32 $[-0.77, 0.13]$ on development. PG uses the same scorer/data and five reward forwards per question; training costs differ. Neither PG nor the GDN-beta adaptation reproduces a complete published system (Appendix E). Further diagnostics remain in Appendices B, C and I.

Comparison with DiffPrune. The paper-based DiffPrune adaptation reaches 77.96 ANLS: SC-VTP gains +15.72 [14.47, 16.98]. It shares the backbone, training data and retention, but has a different objective and 30.48M deployed scorer. This supplementary one-seed comparison does not isolate state correction or reproduce author code; PG remains the closest learned comparator (Appendix N).

Larger-checkpoint replication. Qwen3.8-27B repeats the paired training protocol with a 0.66M scorer and the same complete official cohort. SC-VTP reaches 90.43 ANLS versus attention-only 80.06: +10.38 [9.15,11.62] (95% source interval). DivPrune reaches 84.25, full input 95.33 and resize 91.90. This independently trained pair supports the operator's utility at a larger checkpoint in the same hybrid family; it is not an independent-architecture test. Figure 4d shows all four paired contrasts; Appendix Z retains all 26,745 outputs.

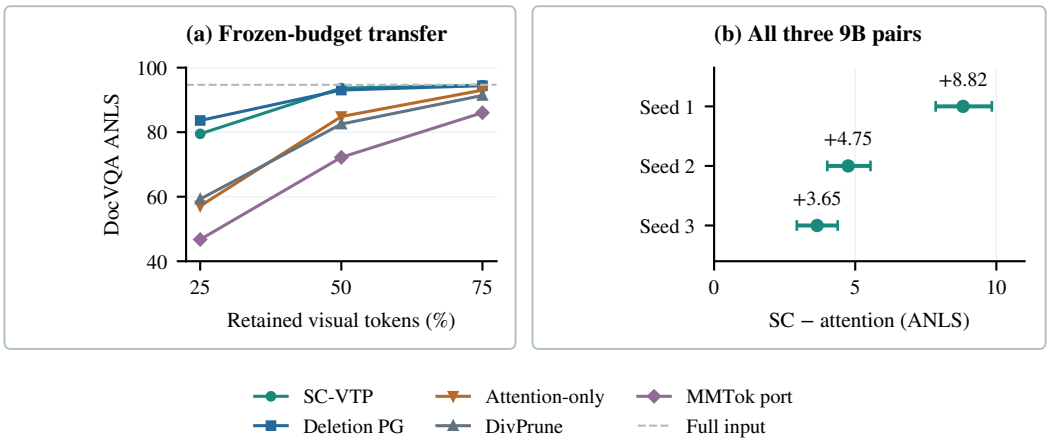

Figure 3: Budget transfer and seed replication on all 5,349 DocVQA validation questions. (a) Fixed 9B scorers at 25/50/75% retention, with every deletion-based comparator and full input. (b) All three predeclared SC-minus-attention gains, with the common supplementary 95% source-bootstrap intervals; these condition on each trained pair. Original protocols and complete controls appear in Appendices Y and G.

Table 3: Native deployment latency on 32 A100 sources, three repeats per method/workload (576 generations). Values average per-source medians in ms. Total includes preparation and model execution; model time includes encoder, selector and decoder. Both pruners retain half the visual rows. Fixed-output workloads and shared-host conditions are detailed in Appendix X.

	One output token (ms)		32 output tokens (ms)	
Method	Total	Model	Total	Model
Full input	469.3	387.8	1882.9	1799.0
SC-VTP	404.9	322.5	1818.5	1735.5
Deletion PG	404.2	322.9	1822.8	1735.4

5.1 Transfer to TextVQA without task-specific training

The fixed DocVQA scorers evaluate all 5,000 TextVQA questions (3,166 images) without tuning: 40,000 complete, uncapped outputs (Table 2). State-minus-attention is +3.38 [2.69,4.07] points; state-minus-PG is +0.51 [0.06,0.96] under its separate unadjusted protocol. State-minus-full is -0.35 [-0.75,0.04], without an equivalence margin. Paired resampling uses image IDs. All 336 previously exposed images (536 questions) remain included. This one-seed transfer concerns text-image QA; Appendix D retains token counts and preprocessing conditions.

5.2 Frozen scorers transfer across retention budgets

The original half-retention scorers also run at 25% and 75%, without retraining, on every official DocVQA question. Figure 3a compares five methods under matched native deletion. SC-VTP reaches 79.49/93.67/94.57 ANLS at 25/50/75%; its gains over attention-only learning at the new budgets are +22.36 [20.62,24.15] and +1.53 [0.95,2.13], using the frozen two-endpoint 97.5% source intervals. Thus the training-operator benefit transfers beyond its training budget.

PG is stronger at 25%: state minus PG is -4.13 [-5.48,-2.81], versus +0.16 [-0.14,0.46] at 75%. Deletion-consistent endpoints therefore do not ensure budget-optimal learned rankings. Appendix Y retains all arms, resize controls and recovery accounting. Published cross-backbone scores remain separate references, not a pooled ranking (Appendix T); LearnPruner and ET-Prune remain further matched-comparison targets.

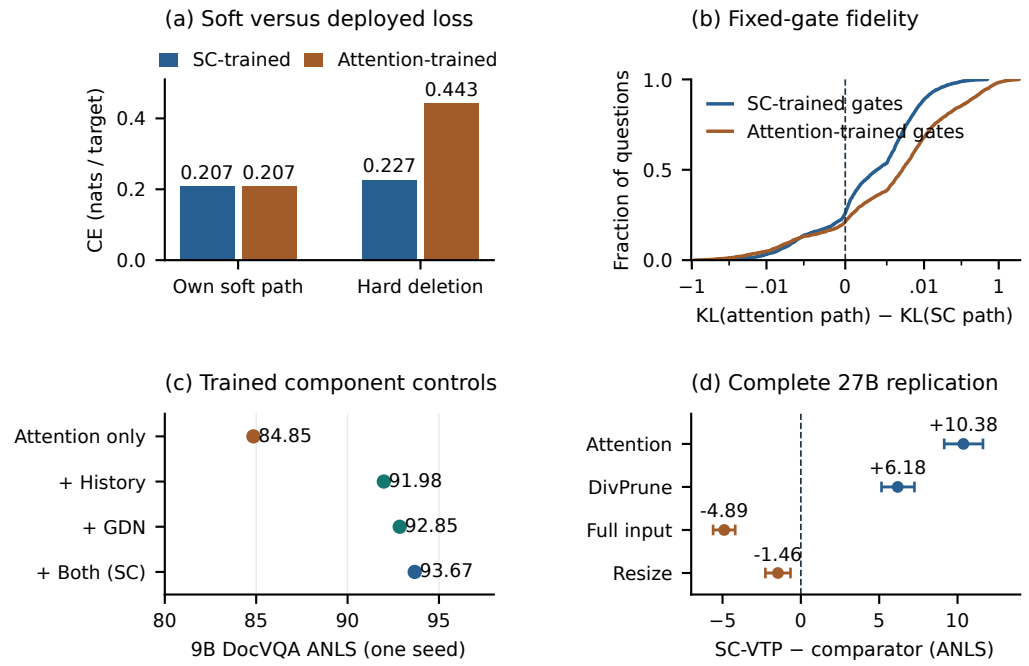

Figure 4: Execution fidelity and learned utility. (a) Original scorers have similar own-soft losses but different hard-deletion losses on all 2,974 development questions. (b) Complete question-wise KL differences hold each scorer's gates fixed: positive values favor the SC operator; the horizontal axis is symmetric log scale. Teacher-forced fidelity is not a mediation test. (c) The complete 5,349-question trained factorial (one seed); all arms gate attention. (d) All four paired SC-minus-comparator contrasts on the complete 27B cohort, with 95% source intervals; attention is primary, others unadjusted. Resize changes encoder work and budget. Figure 13 preserves the budget curves.

5.3 Inference workload at the shared deployment budget

Using the pinned projection shapes, full-input decoder projections at visual rows cost 6.83 TMAC per question; entry deletion at half retention costs 3.41 TMAC (Table 23). The scorer's two linear maps add 0.00052 TMAC. PG and attention-trained selectors share the same native deployment workload. This analytical component count establishes removed projection work, not a method-specific speed advantage. End-to-end cost also includes encoding, text and mixing.

Measured native deployment latency. Table 3 reports a fixed 32-source A100 sample (three repeats); full-input/SC-VTP geometric total-latency ratios are 1.160 [1.145,1.172]× for one generated token and 1.035 [1.030,1.040]× for 32, including image encoding and selection. PG shares native gathering; neither its timing contrasts nor these measurements establish a SC-specific inference advantage or soft-training efficiency. Appendix X gives all times, the four-comparison 98.75% interval family and shared-host conditions.

6 Ablations and execution fidelity

Matched attention-only training. Identical capacity, initialization, data and deployment isolate the training operator. The original primary state-minus-attention gain is +8.82 [7.86, 9.78] official ANLS; development gives +7.43 [6.34, 8.49].

Replication across training seeds. Figure 3b reports all three predeclared pairs, each completing training and both evaluations. Official ANLS means (sample SD) are 93.50 (0.15) for state and 87.76 (2.58) for attention; paired gains average +5.74 (2.73), with positive source intervals for every

seed. Development gains average +4.55. These three-run summaries do not estimate population-level training stability. Appendix G retains every pair and interval; the original primary result is unchanged.

Historical-exposure sensitivity. Excluding 1,713 questions from 300 previously studied sources preserves positive intervals for all three attention contrasts and the original PG contrast; it does not certify a pristine holdout (Appendix L).

Both state corrections contribute on official validation. The completed 2×2 training ablation holds attention gating fixed and varies history and GDN correction (Figure 4c; Appendix Table 12). All four arms use the same initialization, order, data and hard-deletion deployment. The joint arm exceeds history-only by +1.69 [1.04, 2.35] and GDN-only by +0.82 [0.25, 1.39] ANLS, using prespecified 97.5% source-bootstrap intervals. The positive increments support retaining both corrections in this regime, not statistically independent mechanisms. The factorial response is subadditive: the exploratory interaction is -6.32 [-7.34, -5.28]. On development, the joint-minus-GDN interval includes zero (+0.13 [-0.35, 0.62]). Partial arms use one seed; the endpoints were frozen after the original two-arm results were known (Appendix J).

Crossed execution exposes surrogate optimism. The frozen original scorers receive all three execution paths on 2,974 development questions (17,844 complete-answer forwards, $\tau = 0.1$). Own-soft CE is 0.20707/0.20706 nats/target for state/attention, but hard-deletion CE is 0.22656/0.44272 (Figure 4a). The hard-loss gap is +0.21616 [0.18588,0.24730], versus a matched-soft gap of -0.000014 [-0.00349,0.00345] (Equation 14; post-hoc unadjusted 95% source intervals). Similar soft losses conceal a measured deployment discrepancy; they do not establish equivalence.

The prespecified KL alignment advantages are 0.00485 [0.00360, 0.00627] and 0.07778 [0.06480, 0.09137] nats/target (97.5% intervals). The complete distributions appear in Figure 4b. Advantages occur on 73.94%/78.82% of questions and persist across four gate-ambiguity strata per scorer (Appendix R). Neither per-question dominance nor mediation of generation gains is established.

7 Discussion

Matched attention-only comparisons and component ablations connect execution consistency to task gains. Budget transfer distinguishes that operator property from learned ranking quality. Task evidence covers two checkpoints in one hybrid family and short-answer QA; other mixer checks establish operator semantics. Independent architectures and transcription remain open. Training costs differ; timing covers 32 shared-host sources. Three seeds assess the full operator, with partial-arm variability and mediation unresolved. Source-exposure accounting applies throughout.

8 Conclusion

SC-VTP aligns selector training with native deletion through coordinated attention, history and memory control. Three 9B pairs improve DocVQA over attention-only training by 5.74 ANLS on average; the 27B pair gives +10.38. TextVQA and budget transfer extend the comparison beyond training. These results support deletion-consistent training of visual-token selectors for hybrid decoders.

Reproducibility statement

Frozen protocols, model/data revisions, all designated outputs, exposure accounting and paired numerical replays are retained in the supplement. The appendices distinguish saved-score verification from native execution.

AI use statement

Generative AI assisted research design, experiments, analysis and writing. Appendix O details its use and verification scope.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *CVPR*, 2025. https://arxiv.org/abs/2503.02175.

Victor Campos, Brendan Jou, Xavier Giro-i-Nieto, Jordi Torres, and Shih-Fu Chang. Skip RNN: Learning to skip state updates in recurrent neural networks. In *ICLR*, 2018. https://arxiv.org/abs/1708.06834.

Joonmyung Choi, Sanghyeok Lee, Jongha Kim, Sehyung Kim, Dohwan Ko, Jihyung Kil, and Hyunwoo J. Kim. DocPrune: Efficient document question answering via background, question, and comprehension-aware token pruning. In *CVPR*, pp. 3543–3552, 2026. https://arxiv.org/abs/2604.22281v1.

Xuanming Cui, Shlok Kumar Mishra, Wentao Bao, Aashu Singh, Zihao Wang, Xiangjun Fan, Jun Xiao, Ser-Nam Lim, and Jianpeng Cheng. MoEMB: Scaling universal multimodal embeddings with efficient mixture-of-experts models. *arXiv preprint arXiv:2609.08663*, 2026. https://arxiv.org/abs/2609.08663.

Zizhong Ding, Junxian Li, Kai Liu, Shaoqiu Zhang, Xiao Xiao, Linghe Kong, and Yulun Zhang. ET-Prune: Evidence-aware dynamic budgeting for visual token pruning in text-rich MLLMs. *arXiv preprint arXiv:2608.01979*, version 2, 2026. https://arxiv.org/abs/2608.01979v2.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. MMTok: Multimodal coverage maximization for efficient inference of VLMs. In *ICLR*, 2026. https://arxiv.org/abs/2508.18264v2.

Landi He, Mingde Yao, Shawn Young, and Lijian Xu. DiffPrune: Differentiable information throttling for token pruning in vision-language models. *arXiv preprint arXiv:2608.01985*, 2026. https://arxiv.org/abs/2608.01985.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. Nüwa: Mending the Spatial Integrity Torn by VLM Token Pruning. *ICLR*, 2026. https://arxiv.org/abs/2602.02951.

Yuanyuan Jia, Shunpu Tang, and Qianqian Yang. S^2Prune: Spatially structured visual token pruning for multimodal large language models. *arXiv preprint arXiv:2609.01224*, version 1, 2026. https://arxiv.org/abs/2609.01224v1.

Jindong Jiang, Amala Sanjay Deshmukh, Kateryna Chumachenko, Karan Sapra, Zhiding Yu, Guilin Liu, Andrew Tao, Pavlo Molchanov, Jan Kautz, and Wonmin Byeon. Stateful token reduction for long-video hybrid VLMs. *arXiv preprint arXiv:2603.00198*, 2026. https://arxiv.org/abs/2603.00198.

Jiwan Kim, Kibum Kim, Wonjoong Kim, Byung-Kwan Lee, and Chanyoung Park. Why and when visual token pruning fails? A study on relevant visual information shift in MLLMs decoding. *arXiv preprint arXiv:2604.12358*, version 1, 2026. https://arxiv.org/abs/2604.12358v1.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop: Deep Reinforcement Learning Meets Structured Prediction*, 2019. https://wouterkool.github.io/publication/buy-4-samples-free-baseline/.

Xudong Liu, Bicheng Wan, and Yulin Jin. LayoutLite: Token-level implicit layout analysis for efficient document OCR. *arXiv preprint arXiv:2607.22200*, 2026. https://arxiv.org/abs/2607.22200.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, 2021. https://arxiv.org/abs/2007.00398.

MMF Authors. TextVQAAccuracyEvaluator, MMF source repository, pinned commit `b0ffc7fba39f210f833c127362146aec01a2f259`. Publisher evaluation implementation, accessed September 2026. https://github.com/facebookresearch/mmf/blob/b0ffc7fba39f210f833c127362146aec01a2f259/mmf/utils/m4c_evaluators.py.

Seong Yeol Park, Min Jung Kwon, Xianghua Piao, and Yeong Hyeon Gu. DTP: Delta-guided two stage pruning for Mamba-based multimodal large language models. In *ICLR*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/78bee25eb4309dedc6f2066eb0e2a76c-Abstract-Conference.html.

Qwen Team. Qwen3.5-9B model card, 2026. Experiment revision `c202236235762e1c871ad0ccb60c8ee5ba337b9a`. https://huggingface.co/Qwen/Qwen3.5-9B.

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*, volume 34, 2021. https://proceedings.neurips.cc/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019. https://arxiv.org/abs/1904.08920.

Rinyoichi Takezoe, Yaqian Li, Zihao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. LearnPruner: Rethinking attention-based token pruning in vision language models. In *ICLR*, 2026. https://arxiv.org/abs/2604.23950v1.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. https://doi.org/10.1007/BF00992696.

Mengxuan Wu, Zekai Li, Zhiyuan Liang, Moyang Li, Xuanlei Zhao, Samir Khaki, Zheng Zhu, Xiaojiang Peng, Konstantinos N. Plataniotis, Kai Wang, Wangbo Zhao, and Yang You. Dynamic vision Mamba. *arXiv preprint arXiv:2504.04787*, 2025. https://arxiv.org/abs/2504.04787.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *ICLR*, 2025. https://arxiv.org/abs/2412.06464.

Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, and Yanzhi Wang. Rethinking token reduction for state space models. In *EMNLP*, pp. 1686–1697, 2024. https://aclanthology.org/2024.emnlp-main.100/.

Hansen Zhang, Landi He, Mingde Yao, and Lijian Xu. StepPrune: Adaptive sequential visual token selection across multimodal large language models. *arXiv preprint arXiv:2609.13804*, version 1, 2026. https://arxiv.org/abs/2609.13804v1.

A REPRODUCIBILITY AND RECOVERY

The supplementary components implement State-Consistent Visual Token Pruning (SC-VTP); SC in plots and the archived `state` arm denote this method.

Pinned public data. The public dataset mirrors and model are fixed to the following revisions:

Repository	Revision
`pixparse/docvqa-single-page-questions`	`33136ef456fa5a3fe68568d6e31dda4eeff95b9b`
`lmms-lab/DocVQA`	`539088ef8a8ada01ac8e2e6d4e372586748a265e`
`Qwen/Qwen3.5-9B`	`c202236235762e1c871ad0ccb60c8ee5ba337b9a`

Page identity and decoded-image identity are accounted for separately. The official test split and hidden test labels are not used. Source-split seed 2026091709 is fixed before selector learning.

Table 4: Complete DocVQA accounting. Page IDs can differ while decoded RGB pixels are identical. Quarantined questions are excluded from both selector training and internal development.

Partition	Questions	Page IDs	RGB images	Sources
Selector training	26,143	6,787	6,786	4,274
Internal development	2,974	750	750	475
Quarantined official training	10,346	2,657	2,653	523
Official validation	5,349	1,286	1,285	938

The source and data manifests are frozen and hash bound. Training interruptions, original failed statuses and recovery checkpoints are preserved. Recovery restores scorer parameters, optimizer state, CPU/CUDA RNG and absolute learning-rate/temperature progress. Existing BF16 state-gradient non-determinism prevents a claim that recovered training is bitwise identical to an uninterrupted counterfactual. Completed immutable journals cover each arm's full question order and final evaluation cohort. The two arms contain 26,766 combined final development calls, including duplicated initial controls.

Archived tensors, checkpoints and execution logs support reproduction of the reported analyses. Separate CPU checks cover label flow, image/source accounting, historical exposure and alternative aggregation, and recompute saved-output metrics and the primary interval. These numerical replays verify archived outputs; native execution checks are reported separately.

Recorded training cost. A separate audit verifies the complete 26,143-example step journals and 3,268 updates of each original training arm. Summed recorded wall-time segments are 20.97 hours for state training and 18.60 hours for attention training; median recorded per-question segments are 2.88 and 2.55 seconds. These include image reads and preparation but exclude initialization, routine journal/checkpoint writes, development evaluation and discarded recovery work. There is no explicit end-of-step CUDA synchronization. Process-level peak allocated memory in those records is 18.66 and 18.62 GiB, respectively; the counters can include earlier qualification activity. These historical records describe completed work and are not an isolated GPU benchmark or the total computational cost of the research program.

A.1 ADDITIONAL TRAINING AND EVALUATION DETAILS

The primary backbone is Qwen3.5-9B (pinned revision in Appendix A), with native BF16, torch 2.5.1+cu124 and transformers 5.3.0. Image preparation has a maximum of 1,024 visual tokens. We set $K = \lfloor N/2 \rfloor$. State and attention arms use identical scorer initialization, question order and a complete training pass. The original comparison uses seed 2026091709; the supplementary replication adds predeclared seeds 2026091902 and 2026091903 under the same training protocol. AdamW uses learning rate 3e-4, zero weight decay, eight-question gradient accumulation, norm clipping at one, a 5% warmup and cosine decay to 10% of the peak rate. There are 3,268 updates per arm.

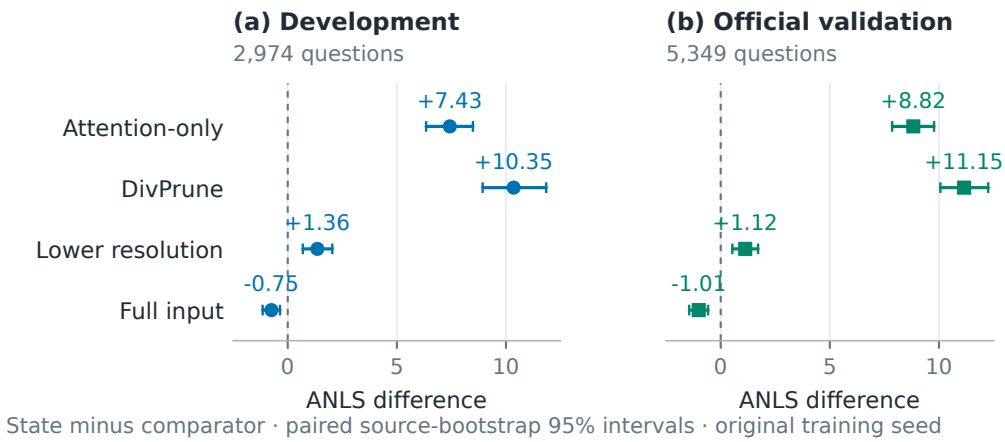

Figure 5: Paired state-minus-comparator ANLS differences on complete internal development (2,974 questions; 475 sources) and official validation (5,349 questions; 938 sources). Intervals resample source documents and condition on one trained selector pair. State-minus-attention is primary in each cohort; other intervals are descriptive and unadjusted. The lower-resolution arm has a nearby token budget and reduces encoder work. Official validation has known prior repository exposure.

All outputs are generated greedily with a 128-token limit. The primary score is question-averaged ANLS with a strict normalized edit-distance threshold below 0.5; we also report normalized exact match. The primary official-validation contrast is final state minus final attention ANLS, with 10,000 paired source-document bootstrap replicates over all 938 sources and a 95% interval. The complete development comparison uses the same contrast over its 475 sources. Secondary intervals are descriptive and unadjusted for multiplicity. Training-seed variation is reported separately. Separate frozen supplementary protocols designate state minus PG on each DocVQA cohort and state minus attention on TextVQA; their intervals are not simultaneous across protocols.

Dense history weights implement the differentiable recurrence, preserving native pre-activation rounding. This qualification implementation uses quadratic sequence storage; it is not an optimized training kernel.

B Complete internal-development diagnostics

The internal cohort contains 2,974 questions on 750 RGB images and 475 source documents, disjoint from selector training and official validation by source and exact decoded pixels. This section preserves the complete historical development panel and diagnostics. Its confidence intervals condition on the same trained pair as the official results.

The initial scorer was evaluated independently in both arms and produced identical supports, image grids and generated token IDs for all development questions. All arms cover every development question. Only the question-only arm has length-capped outputs (five); all of them remain in its score.

State-consistent training improves ANLS over matched attention-only training by **7.43 points**, with a source-document bootstrap 95% interval **[6.34, 8.49]**. It is 0.75 points below full input, with interval [-1.15, -0.35]. Relative to DivPrune the improvement is 10.35 points [8.93, 11.84]. Relative to native lower-resolution input the difference is 1.36 points [0.69, 2.04]. The lower-resolution arm uses a nearby rather than exactly matched visual budget and also reduces encoder work, so it remains a particularly important practical control.

Exact matching is less forgiving: state pruning answers 2,672/2,974 questions exactly, full input 2,744, and attention pruning 2,309. Of the 302 nonexact state answers, 221 receive positive edit similarity and 81 receive zero ANLS. These counts prevent interpretation of 95.85 ANLS as 95.85% error-free OCR.

Table 5: Complete internal DocVQA development cohort: 2,974 questions, 750 images, 475 source documents. All learned methods use one seed. ANLS is reported on a 0–100 scale. Token counts are question-weighted means; they do not measure latency.

Method	Retained tokens	ANLS	Exact match (%)
Full input	985.61	96.60	92.27
State-consistent training, hard deletion	492.77	95.85	89.85
Physical-deletion PG adaptation	492.77	96.16	91.06
Matched attention-only training, hard deletion	492.77	88.42	77.64
DivPrune, hard deletion	492.77	85.50	72.80
Even-stride deletion	492.77	81.20	56.42
Initial scorer / prefix-half support	492.77	77.11	71.25
Native lower-resolution input	485.36	94.49	87.12
Question only	0	3.44	1.61

The supplementary PG comparator uses the same full development cohort and achieves 96.16 ANLS and 91.06% exact match, with no capped outputs. State minus PG is -0.32 $[-0.77, 0.13]$, using 10,000 source-document resamples with seed 2026091709. This interval includes zero and is not an equivalence test. The final checkpoint was fixed by protocol; no checkpoint was selected using this contrast.

A post hoc aggregation sensitivity check gives state-minus-attention differences of +6.60 ANLS when each distinct image is weighted equally and +6.31 when each source document is weighted equally. These descriptive results retain the sign of the frozen question-weighted endpoint. They are not additional independent experiments or new confirmatory tests.

The 12,800-question midpoint was evaluated on the same development cohort. Its result is preserved as training diagnostics; the final checkpoint is the endpoint reported here. We do not treat midpoint and final evaluation as independent replication, nor select the larger checkpoint margin for the headline.

C COMPLETE ORIGINAL QUESTION-TYPE ANALYSIS (POST HOC)

We reuse all 5,349 official-validation questions and the nine original DocVQA question-type annotations (Mathew et al., 2021). These labels describe the information needed to answer a question; they were not constructed from the model's predictions. The groups overlap: 1,067 questions have multiple types, giving 6,493 type memberships. Every original type is retained, including Yes/No with 28 questions. Tables 6 and 7 report all five completed arms. The native lower-resolution control uses a nearby, unequal budget and changes encoder work.

This analysis was specified after the aggregate official-validation result was known and is *post hoc and descriptive*. It does not change the primary endpoint, trained checkpoints or evaluation cohort. For each type, 10,000 paired bootstrap draws resample the source documents represented in that type, carrying all of that source's questions with the label together. The statistic remains question-weighted. Both contrasts use the same draws. Each type's seed is the unsigned big-endian integer from the first eight bytes of SHA256 of `b1-docvqa-type-v1:` followed by its exact label. The 95% percentile intervals are unadjusted across types and comparisons; overlapping types are not independent replications.

Table 6: All nine original question types: complete five-arm ANLS on the 0–100 scale. QA, RGB and source counts describe each overlapping group and must not be summed as independent units. No type or output is excluded.

Type	QA	RGB	Sources	State	Attn.	Full	DivPrune	Low res.
Image/Photo	98	58	46	88.52	85.19	89.83	86.90	84.22
Yes/No	28	21	19	92.86	75.00	100.00	78.57	100.00
figure/diagram	265	86	73	92.52	90.30	94.30	83.67	92.62
form	1021	375	329	96.21	87.34	96.38	87.02	93.82
free_text	765	382	309	90.66	79.82	92.30	82.53	90.68
handwritten	319	126	110	90.43	83.04	91.26	81.01	88.36
layout	1981	884	701	93.60	84.79	93.86	88.65	92.86
others	236	109	101	94.05	82.95	95.51	88.67	91.41
table/list	1780	538	434	93.78	84.57	95.28	73.10	92.66

Table 7: Normalized exact match (%) for the same complete type groups and arms. Normalization strips leading/trailing whitespace and lowercases strings; it does not collapse internal whitespace. Group counts are given in Table 6.

Type	State	Attn.	Full	DivPrune	Low res.
Image/Photo	79.59	73.47	83.67	77.55	73.47
Yes/No	92.86	75.00	100.00	78.57	100.00
figure/diagram	89.43	83.02	90.57	78.87	87.17
form	88.44	74.34	90.30	72.09	83.94
free_text	83.66	68.63	85.49	69.80	81.31
handwritten	75.55	67.08	78.06	64.58	74.29
layout	87.83	74.00	88.89	78.50	86.07
others	89.41	74.58	90.25	78.81	85.17
table/list	88.65	76.29	91.57	61.35	87.36

All nine matched-control point estimates are positive, but the Image/Photo, Yes/No and figure/diagram intervals include zero. State-minus-full point estimates are negative in every type. Only free_text, -1.64 [-3.29, -0.10], and table/list, -1.51 [-2.26, -0.76], have intervals entirely below zero (Figure 6). These marginal intervals do not establish simultaneous coverage or differences between type effects. They condition on one trained pair and retain known official-validation exposure and unknown backbone pretraining. The annotations neither measure isolated transcription quality nor identify a separate causal contribution from convolutional history or recurrence. This diagnostic does not replace the unresolved strong learned-baseline comparisons.

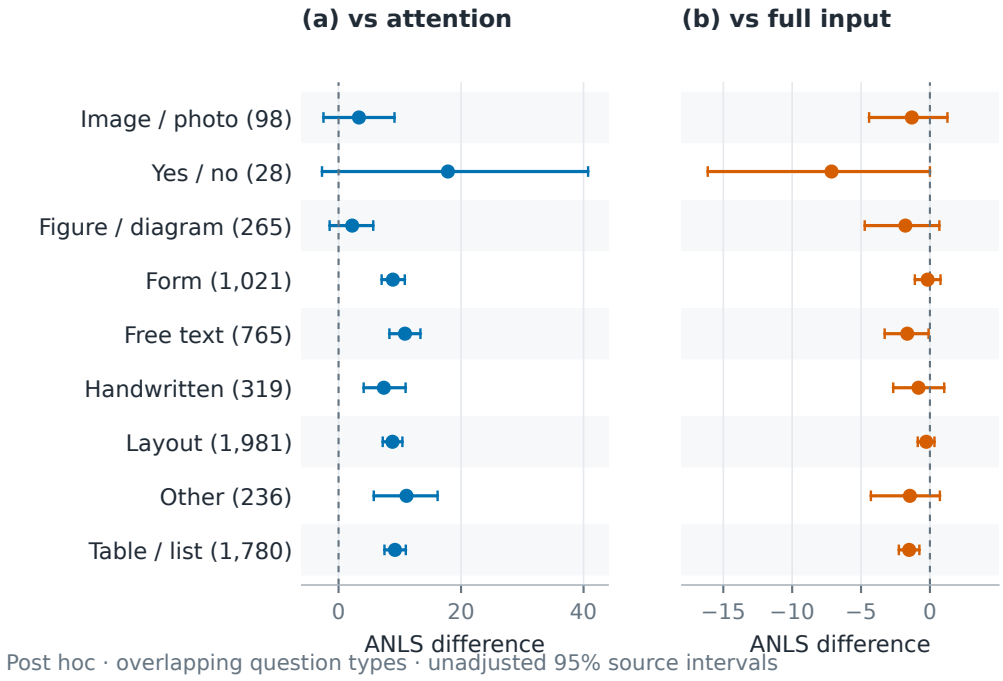

Figure 6: Post hoc state-minus-comparator ANLS differences across all nine original DocVQA types. Parenthesized counts are questions. Each 95% interval uses 10,000 paired resamples of source documents within that type. Types overlap; intervals are unadjusted across types and comparisons and condition on the same single trained pair. They do not give simultaneous or multi-seed confidence, and official validation has known prior exposure.

D COMPLETE CROSS-TASK EVALUATION

DocVQA-to-TextVQA transfer. The transfer protocol uses all 5,000 official TextVQA 0.5.1 validation questions on 3,166 image IDs (Singh et al., 2019). Every image ID has a distinct decoded RGB hash in the prepared cohort. The two learned scorers are the frozen final DocVQA scorers after 26,143 training questions. No TextVQA selector training, checkpoint selection, prompt tuning or retention search is allowed. The original protocol gives every question state, attention, full input, DivPrune and native 512-token-budget input: 25,000 scientific generation calls. Separately fixed GDN-beta and physical-deletion PG references each add 5,000 calls on the identical cohort, with the same preparation and official metric: 35,000 scientific generations in total. Preparation preserves four official MPO images under their original filenames, using the same default first-frame conversion as native image preparation. The later frozen MMTok supplement adds 5,000 questions without changing those original runs, bringing the eight-method total to 40,000 (Appendix W).

Gold-free generation and primary metric. The input contains the image and original question followed by "Answer the question with a single word or phrase." No OCR words, annotations, answer options or gold answers enter preparation or generation. Answers are scored only after greedy decoding, with a 128-token cap and no selective retry. The primary metric is the publisher's pinned MMF evaluator (MMF Authors, 2026). With ten normalized annotations and m matches to a normalized prediction, its per-question accuracy is

$$a(m) = \frac{m \min(1, (m-1)/3) + (10-m)\min(1, m/3)}{10}, \qquad m \in \{0, \ldots, 10\}. \tag{4}$$

The first term is zero at $m = 0$. Thus one, two, three and at least four matches receive 0.3, 0.6, 0.9 and 1.0. This is the average of ten held-out-annotator agreement scores, rather than the simplified $\min(m/3, 1)$ rule.

Uncertainty and exposure. The primary transfer contrast is question-averaged state minus attention soft accuracy. A paired bootstrap resamples the 3,166 official image-ID groups 10,000 times, retaining all questions of each sampled image, with seed 2026091904. The remaining three state-comparator intervals are exploratory and unadjusted. There are 336 images in two identified earlier repository cohorts, covering 536 full-validation questions; 336 exact question IDs were previously selected. All remain in the main endpoint. Descriptive exposure strata do not establish unseen-backbone evaluation. Preparation found no exact decoded-RGB matches to the current DocVQA train, internal-development or official-validation partitions.

Falsifiable transfer criterion. A nonpositive primary estimate or an interval containing zero does not confirm transfer of the mechanism advantage. Such a result must be retained without tuning on TextVQA. The completed primary estimate is +3.38 [2.69, 4.07], meeting this fixed criterion. It establishes same-backbone, cross-task evidence for one seed and does not replace further learned-baseline comparisons, independent-architecture evaluation or end-to-end timing.

Table 8: All 5,000 TextVQA validation questions on 3,166 images, with frozen DocVQA scorers. The official VQA soft metric uses all ten annotations. All 40,000 outputs are included; none is capped or empty. GDN-beta uses full input in the first five blocks; its retention count applies only afterward. Token counts are question-weighted means, not latency.

Method	VQA soft accuracy (%)	Retained tokens
Full input	84.52	741.80
State-consistent	84.18	370.90
Physical-deletion PG	83.66	370.90
Attention-only	80.80	370.90
DivPrune	82.67	370.90
Lower resolution	83.35	488.31
GDN-beta (after block 4)	71.01	370.90
MMTok (author-core port)	72.96	370.90

The secondary state-minus-full difference is -0.35 [-0.75, 0.04], state-minus-DivPrune is +1.51 [0.93, 2.09], and state-minus-native-lower-resolution is +0.83 [0.16, 1.50] percentage points. These 95% intervals are unadjusted. The full-input interval does not demonstrate equivalence or noninferiority. The lower-resolution arm retains 488.31 tokens versus 370.90 for state pruning and changes encoder work. No latency conclusion follows from these accuracy differences.

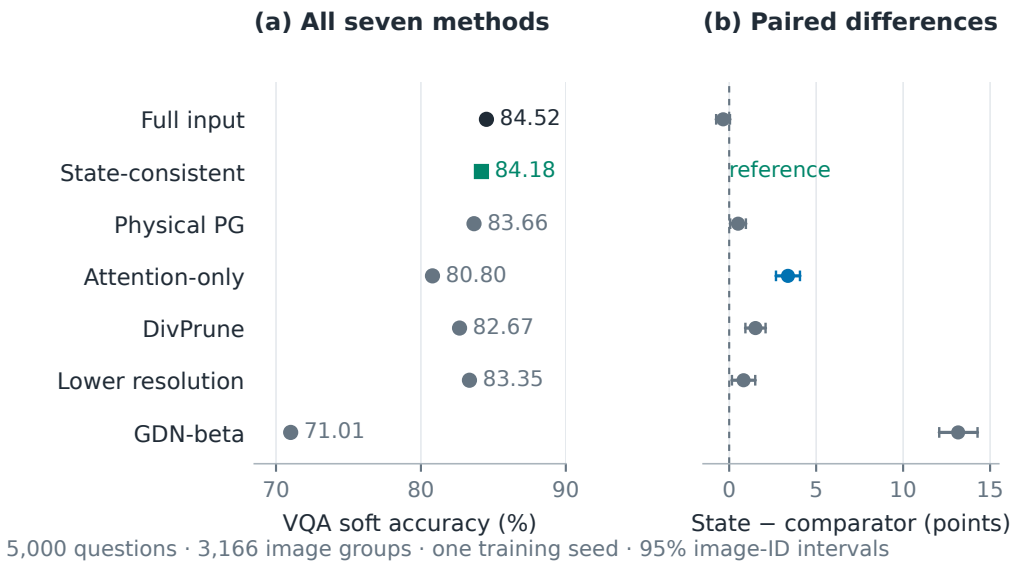

Figure 7: Complete TextVQA transfer. Left: all seven arms on the same 5,000 questions. Right: paired state-minus-comparator differences with their originally fixed image-ID bootstrap 95% intervals; only state minus attention is the transfer primary. PG and GDN are supplementary additions; GDN retains full-input computation in five early blocks. One trained pair, known exposure and unadjusted secondary intervals limit interpretation.

E Supplementary comparator methods and complete results

Direct policy-gradient learning with physical deletion. A supplementary learned baseline uses the same scorer, original initialization, training order, frozen backbone and complete 26,143-question training population. For each question, five independent Gumbel top-K draws sample ordered Plackett–Luce selections at temperature one. Their supports are gathered in original sequence order, and the native decoder computes the mean answer-token cross entropy, including EOS. The scorer is updated by the REINFORCE estimator (Williams, 1992) with the mean cost of the other four draws as a leave-one-out baseline, following the established multiple-sample variance-reduction approach (Kool et al., 2019):

$$\widehat{\nabla_\theta L} = \frac{1}{5}\sum_{r=1}^{5}\left(\ell_r - \frac{1}{4}\sum_{s\neq r}\ell_s\right)\nabla_\theta \log p_\theta(\pi_r).$$

Costs are detached in this score-function calculation. No differentiable masking, variance normalization, clipping of probability ratios or entropy bonus is added. The optimization schedule and final-checkpoint rule remain fixed. The five native decoder evaluations per training question make this a comparison with matched data and capacity, not matched computation. This baseline is a direct physical-deletion policy-gradient adaptation, not an official reproduction of LayoutLite or DiffPrune. The full training pass completed 26,143 questions on 6,786 images, 3,268 optimizer updates and 130,715 frozen-decoder reward forwards. Complete development (2,974 questions) and official validation (5,349 questions) give 96.16 and 93.05 ANLS, respectively. The main text reports both contrasts with state pruning. Historical instrumented microstep durations are 11.11 hours for PG, 20.97 for state and 18.60 for attention; these are not isolated matched-run timing measurements.

Five forwards per question do not by themselves imply greater elapsed cost than differentiating through the frozen decoder. No training-speed or deployment-speed advantage is inferred.

The frozen extension transfers the identical final DocVQA PG checkpoint to all 5,000 TextVQA validation questions without TextVQA training, tuning or checkpoint selection. The protocol was fixed after the six existing arms were known and before any PG transfer generation. All 5,000 scientific generations and two native generation qualification calls are complete. PG achieves 83.66% official VQA soft accuracy, with no empty or capped output. State minus PG is +0.512 percentage points, with a paired image-ID bootstrap 95% interval [0.061, 0.960] from 10,000 draws and fixed seed 2026092001. This prespecified supplementary endpoint is unadjusted and does not replace the original state-minus-attention transfer primary. Both methods retain 370.90 visual tokens on average and share the native hard-deletion inference path; their training computations differ. An independent CPU replay verifies all 35,000 seven-arm records and this fixed contrast, reusing the official answer normalizer. A separate portable score export replays all 25 frozen DocVQA/TextVQA contrasts; neither check reruns the models. The positive official margins and reversed development point ordering do not establish uniform superiority or reproduce LayoutLite.

Native GDN-beta comparison. A paper-based MoEMB adaptation (Cui et al., 2026) reads the native sigmoid write gates at zero-based decoder layer four, averages heads in FP32, and physically retains the top half of visual rows at the entry to layer five. Stable ties follow original position. Blocks zero through four use the full prompt; early attention caches and recurrent states retain that computation. Later blocks receive the compact prompt with retained logical positions. The single-layer reduction follows the paper's specific appendix setting; its less-specified cross-layer aggregation is not silently assumed. This autoregressive question-answering adaptation is not the full embedding/MoE system. It is evaluated on all 5,349 official-validation questions only after real-model execution qualification. Early full-token work and retained state information must be disclosed when comparing its cost and performance with decoder-entry pruning. The original real-model qualification failed a BF16 cached-versus-cache-free hidden-state tolerance and is retained as failed. Before full-cohort evaluation, a separate recovery protocol strengthens BF16 cached-versus-independent-cached comparisons to exact equality and requires an FP32 decoder reference check with fixed BF16 image features and supports. Native unpruned BF16 execution shows the same type of cached-versus-cache-free discrepancy on the engineering inputs. The amended qualification passes; it does not retroactively pass the original BF16 criterion. The inference adapter, ranking and BF16 evaluation are unchanged. The complete evaluation produces 5,349 answers, including 0 capped and 1 empty outputs, all retained in the reported metrics. An independent CPU replay verifies all saved string scores, selected supports, early/late cache lengths and the five originally specified ANLS contrasts. It does not rerun the model or establish correctness on arbitrary inputs.

The mean full and compact prompt lengths are 1023.28 and 529.56. Summing input rows over decoder blocks gives 19414.64 for this GDN path and 16946.07 for half-retention entry pruning, a ratio of 1.146. This is structural accounting only: blocks have different operations, and input-row sums are neither FLOPs, elapsed time nor an information-removal measure. The additional GDN evaluation shared a GPU with training and supplies no timing evidence.

The supplementary TextVQA extension froze the identical GDN method before inspecting DocVQA GDN accuracy, while the original five-arm TextVQA results were known. All 5,000 new scientific generations on 3,166 image IDs are complete, with no capped or empty output. The first fixed question also passes native versus all-keep generated-token equality (two separate qualification calls). GDN-beta reaches 71.01% official VQA soft accuracy; state minus GDN is +13.16 [12.08, 14.28] percentage points, using the fixed 10,000 image-ID bootstrap draws and seed 2026091922. This supplementary comparison is unadjusted and does not replace the original transfer primary. An independent CPU replay verifies all 30,000 six-arm records, integer ten-annotation agreement and all nine existing transfer contrasts; it reuses the official normalizer and does not repeat GPU qualification. There is no new training or hyperparameter selection, and no shared-GPU timing conclusion.

F BINARY ENDPOINT OF THE IMPLEMENTED TRAINING OPERATOR

Why an attention mask alone cannot guarantee this endpoint. An abstract scalar mixer illustrates the remaining state paths. Start from $S_0 = 0$, keep tokens one and three, and discard token two.

Set ungated decay $a = e^g = 1/2$, write strength $\beta = 1/2$, $k = q = 1$, and values $(2, 3, 0)$. These are idealized mixer inputs; joint realizability by the pretrained Qwen checkpoint is not asserted. The retained first token establishes state one. Compact deletion gives output $1/4$ at token three; leaving only the discarded step's decay active gives $1/8$, leaving only its write active gives $1/2$, and leaving both active gives $7/16$. A width-two convolution supplies another example: with projected values $(1, 2, 0)$ and only the previous-token coefficient nonzero, copying history gives preactivation one at token three, whereas advancing through the discarded token gives two. Within this abstract mixer class, these algebraic examples refute a uniform equivalence guarantee for the incomplete operators. They establish neither failure on every input nor a causal explanation of the measured accuracy gap. Identity-versus-update gating itself is established prior work (Campos et al., 2018).

Setting and scope. Consider one unpadded, unpacked sequence in cache-free prefill of the studied hybrid decoder. Fix a binary retention vector $z \in \{0, 1\}^T$, shared by all relevant layers, with $z_1 = 1$ and every text gate equal to one. Write its retained indices as $r_1 < \cdots < r_M$, and put $n(t) = \sum_{j=1}^{t} z_j$, $n(0) = 0$. The native compact execution receives exactly $(h_{r_a}, p_{r_a})_{a=1}^{M}$: embeddings, order and logical positional coordinates are preserved, while physical sequence indices become $1, \ldots, M$.

Assume deterministic exact arithmetic with finite unmasked activations and parameters; attention masks use the usual extended-real $-\infty$ convention. The only cross-token operations are causal attention, the causal convolution, and the Gated DeltaNet recurrence specified below; all remaining maps (including normalization, residuals and output modulation) are positionwise. There is no additional length- or compact-index-dependent positional transform: the same logical coordinates produce the same rotary transforms in both runs. Each layer starts from the same native empty history and zero recurrent state. There are no padding masks, packed-sequence resets or stochastic dropout.

Claim. For these conditions and this operator, at every layer ℓ, $h^{\ell}_{r_a} = \widetilde{h}^{\ell}_{a}$ for $a = 1, \ldots, M$, where a tilde denotes native execution on the compact sequence. Consequently, positionwise final normalization and the language-model head also agree at retained positions.

Proof. The claim holds at the input by construction. Assume it at a layer's input. All positionwise projections and rotary transforms then agree at retained positions. For a retained query r_a, adding $\log z_j$ to each causal key logit, with $\log 0 = -\infty$, gives support

$$\{j \le r_a : z_j = 1\} = \{r_b : b \le a\}.$$

This is exactly the compact causal support. Its logits, values and softmax denominator agree term by term; hence its attention output agrees. The retained first token ensures nonempty support even for queries at discarded positions.

For a recurrent block, let H_t be its $w - 1$ projected history vectors *before* original position t, newest first. The native convolution uses the current projection x_t and H_t, after which the gated history is updated:

$$H_{t+1} = (1 - z_t)H_t + z_t[x_t, H_t^{1:w-2}],$$
$$\bar{S}_t = e^{z_t g_t} S_{t-1}, \qquad S_t = \bar{S}_t + z_t \beta_t k_t (v_t - \bar{S}_t^{\top} k_t)^{\top}.$$

Here S_t is the state *after* position t; equations are per head. Use the inner induction

$$H_t = \widetilde{H}_{n(t-1)+1}, \qquad S_{t-1} = \widetilde{S}_{n(t-1)}.$$

It holds initially. If $z_t = 0$, the history is copied and both decay and write are neutralized: $H_{t+1} = H_t$ and $S_t = S_{t-1}$. Since $n(t) = n(t-1)$, the invariant continues. If $z_t = 1$, then $t = r_a$ with $a = n(t)$. The input projection and prior history equal their compact counterparts, so the convolution, its activation and the resulting queries, keys and values agree. Positionwise decay/write parameters agree as well. The history advances once, and $z_t g_t = g_t$, $z_t \beta_t = \beta_t$ give precisely the native transition. Thus the updated state and its native query readout agree, proving the invariant and output equality at each retained position.

Remaining positionwise operations preserve that equality, completing the layer induction. Discarded-position outputs need not agree with anything; their effects on retained tokens are excluded by the three gated mixing paths. ∎

Dimensions and limits. Per head, g_t, β_t, z_t are scalars, $q_t, k_t \in \mathrm{R}^{d_k}$, $v_t \in \mathrm{R}^{d_v}$ and $S_t \in \mathrm{R}^{d_k \times d_v}$. Thus $S_t^\top k_t \in \mathrm{R}^{d_v}$, the outer-product update has shape $d_k \times d_v$, and $S_t^\top q_t \in \mathrm{R}^{d_v}$; native query scaling is included in q_t. This conditional endpoint identity does not establish BF16 equality, cached-generation compatibility, equality of soft gates to expected random deletion, or improved learning or task accuracy. It is not a general pruning theorem for other token mixers.

G Complete development and official validation across training seeds

The original seed 2026091709 and two predeclared seeds 2026091902 and 2026091903 each complete both training arms: 26,143 questions, 6,786 distinct RGB images, and 3,268 updates per arm. Within a seed, scorer initialization and example order match; both differ across seeds. No development-based checkpoint selection occurs. Every final checkpoint is evaluated on the same 2,974 questions, 750 images and 475 source documents. The six closed journals, final checkpoints, exact orders and scores are checked before aggregation.

Table 9: All planned pairs on the reused development cohort. Each interval uses 10,000 source-document resamples, seed 2026091901. Intervals quantify data sampling conditional on that trained pair; they are not simultaneous or population-over-seeds intervals.

Seed	State ANLS	Attention ANLS	Difference	95% source interval
2026091709	95.85	88.42	+7.43	[6.36, 8.50]
2026091902	95.92	91.74	+4.19	[3.34, 5.03]
2026091903	95.66	93.64	+2.03	[1.19, 2.88]

The mean paired difference is 4.5469 points and sample standard deviation across three differences is 2.7182. The positive signs satisfy the fixed directional replication criterion, but three runs give limited information about training variability. Development reuse and unknown backbone exposure remain. The seven-method official table and the TextVQA comparison still use the original pair. The older separately reported first-seed interval used its original analysis seed; this table recomputes all three with the common supplementary seed.

Complete official-validation replication. The same six fixed final checkpoints are evaluated on all 5,349 official questions, 1,285 distinct RGB images and 938 source documents. The original pair contributes 10,698 archived generations; the two additional pairs contribute 21,396 new generations, separately from four qualification calls. Closed journals, checkpoint bindings, native qualifications and terminal cleanup receipts are verified before aggregation. No checkpoint or seed is selected using these scores. Shared-resource execution is not used for timing claims.

Table 10: All predeclared training pairs on the complete official-validation cohort. Each interval uses 10,000 source-document draws with common supplementary seed 2026091921. The original primary interval remains [7.86, 9.78]; the small difference is Monte Carlo resampling, not a change to scores or examples.

Seed	State ANLS	Attention ANLS	Difference	95% source interval
2026091709	93.67	84.85	+8.82	[7.85, 9.84]
2026091902	93.42	88.67	+4.75	[4.01, 5.54]
2026091903	93.40	89.75	+3.65	[2.93, 4.38]

The mean paired official difference is 5.7408 points and its sample standard deviation is 2.7269. These unadjusted supplementary intervals condition on each trained pair and do not quantify population-over-seeds uncertainty. Independent arithmetic replay reproduces all six arm means and all three intervals from 32,094 saved scores. Prior repository exposure and unknown backbone pre-training exposure remain unchanged.

H COMPLETE FINITE-GATE VERSUS HARD-DELETION ALIGNMENT PROBE

The supplementary protocol was frozen after task-accuracy outcomes were known but before this probe's output distributions were measured. It uses all 2,974 development questions, fixed original final scorers, native BF16 Qwen3.5-9B, an input budget of 1,024 visual tokens, half retention and $\tau = 0.1$. Scores and top-K supports are held fixed within each scorer. Every question receives native hard deletion, attention-only soft gating and coordinated soft gating under each scorer: six complete-answer forwards per question, 17,844 in total. Two exposed engineering fixtures are excluded from this count.

Each branch teacher-forces the first accepted reference plus EOS, with fresh state and no target truncation. Per-position KL uses the full vocabulary, averages over target positions within a question, then over questions. This avoids restricting the distribution to answer candidates. Define D as question-mean $\mathrm{KL}(p_{\mathrm{hard}} \| p_{\mathrm{attention\ soft}}) - \mathrm{KL}(p_{\mathrm{hard}} \| p_{\mathrm{state\ soft}})$. Twenty thousand common source-document bootstrap draws, seed 2026092104, carry all questions of each sampled source; 97.5% intervals address the two fixed-scorer endpoint family. This is a question-weighted estimate, not an average over independent answer tokens.

Table 11: KL from native deletion to each soft path, in nats per target position. Both fixed-scorer contrasts satisfy the prespecified positive-interval criterion.

Fixed scorer	Attention KL	State KL	D	97.5% interval
State-trained	0.03463	0.02978	0.00485	[0.00360, 0.00627]
Attention-trained	0.35332	0.27554	0.07778	[0.06480, 0.09137]

Closed immutable chunks and a separate arithmetic replay reproduce the complete cohort and intervals. The result is conditional on these scorers, temperature, reference prefixes and reused development population. It does not establish the gradient's relation to a discrete selection objective, a causal mediator of generation accuracy, or an efficiency benefit. All forwards use the original dense history implementation; the failed checkpointed-history engineering prototype is not substituted.

I ARCHIVED DOCUMENT SUPPORTS AND COMPLEMENTARY OUTCOMES

Figure 8 displays two original-seed official-validation examples, selected by an explicit post-hoc rule. Among examples with a correct full-input answer, we separately consider state-correct/attention-incorrect and state-incorrect/attention-correct outcomes. We restrict question length to at most 110 characters and every displayed prediction and first reference to at most 55 characters; 743 and 87 questions, respectively, remain eligible. We select the minimum SHA256 of a fixed figure-specific prefix and question ID within each category. These are complementary illustrations, not representative or independent evaluation samples.

Question 57533 has reference "Dream Cream": full input and state training generate it, while attention training generates "Dark Fantasy". Question 2007 has reference "no effect": full input and attention training generate it, while state training generates "low". The archived image hashes, feature grids, retained-index lists and response strings bind each plotted panel. The 2×2 correctness table in the main figure covers all 5,349 questions: 3,865 are correct in both arms, 520 incorrect in both, 831 correct only under state training, and 133 correct only under attention training. This descriptive accounting uses the saved exact-match endpoint and adds no model calls.

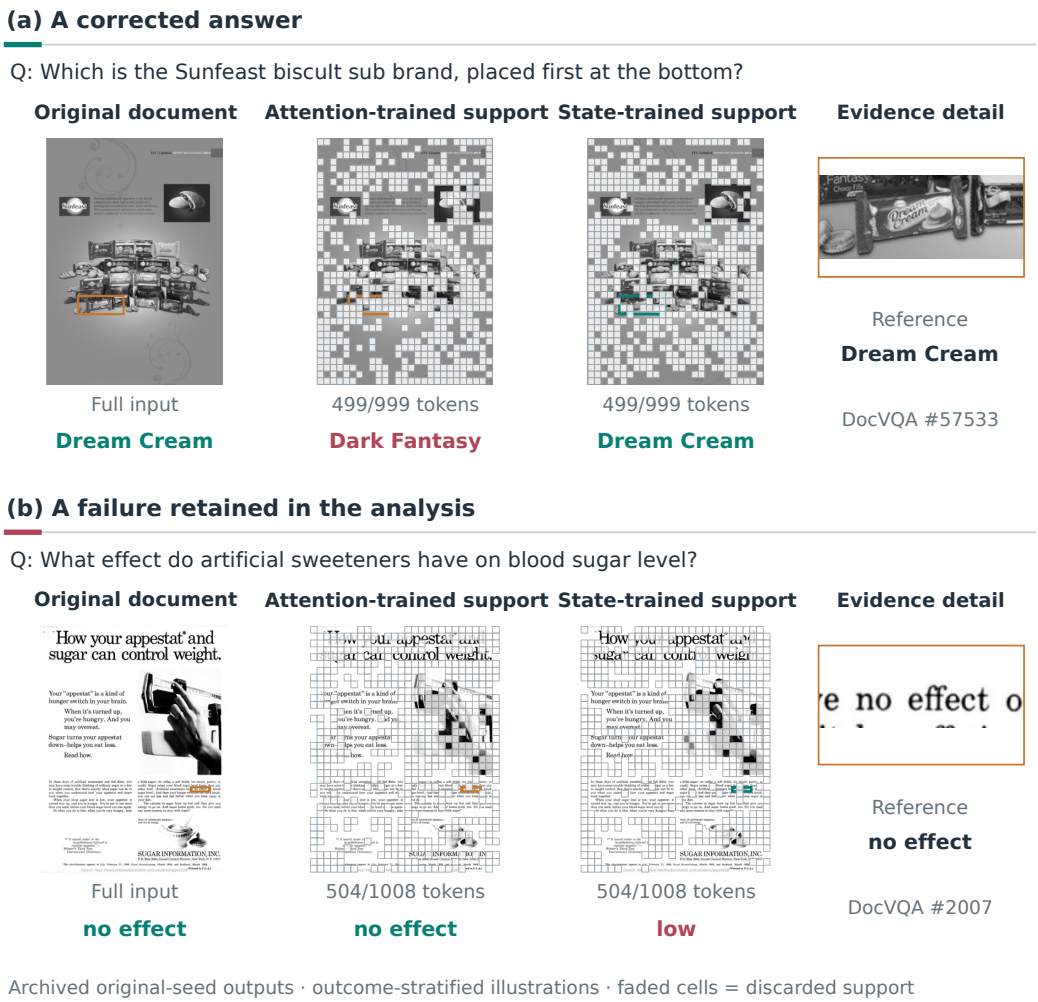

Figure 8: Real documents, archived retained indices and generated answers. Faded cells denote discarded post-merge token positions mapped to the original page grid; the vision encoder saw the full image. Spatial supports are not localized receptive fields or causal explanations. The orange reference boxes and magnified crops are manually chosen display annotations, never selector inputs. Both selectors use identical budgets: 499/999 tokens in (a), 504/1,008 in (b). The success and failure are selected within declared outcome categories; neither changes the complete-cohort statistics.

Table 12: Trained component ablation. Every arm gates attention; columns show the two additional corrections. All arms train on 26,143 questions and deploy half-retention hard deletion. Complete development / official validation contain 2,974 / 5,349 questions. ANLS is on a 0–100 scale; one seed.

Training arm	History	GDN	Dev.	Official
Attention only	No	No	88.42	84.85
History correction	Yes	No	93.75	91.98
GDN correction	No	Yes	95.72	92.85
Joint correction	Yes	Yes	95.85	93.67

J FACTORIAL ABLATION OF STATE CORRECTIONS

The supplementary protocol fills a 2×2 history/GDN design with attention gating present in every cell. Attention-only uses neither correction; history-only gates history advancement; GDN-only gates log decay and write strength with native history; joint uses both. All four 524,545-parameter scorers share initialization, seed 2026091709, the complete 26,143-question order, 3,268 updates, optimizer, schedule, labels, preparation and final-checkpoint rule. All deploy half-retention hard deletion. Different training operators have different computation; equal steps do not imply equal GPU time.

Every arm completes all 2,974 development questions (750 RGB images; 475 sources) and 5,349 official-validation questions (1,285 RGB images; 938 sources), without capped outputs. Support lengths, positions and grids are checked per arm; equal budgets need not yield identical learned supports. Original attention/joint generations are reused without changing their primary analysis. The two new arms fill missing cells, not four independent replications.

Table 13: Complete four-cell means. All arms gate attention and deploy hard deletion. EM uses the original normalized string rule.

	Development		Official validation	
Arm	ANLS	EM (%)	ANLS	EM (%)
Attention only	88.4219	77.6395	84.8480	74.7429
History correction	93.7452	86.8527	91.9842	84.9318
GDN correction	95.7185	90.5514	92.8532	86.8200
Joint correction	95.8484	89.8453	93.6722	87.7921

The two primary contrasts, official joint-minus-history and joint-minus-GDN, use 20,000 paired source-document bootstrap draws, seed 2026091914, with 97.5% percentile intervals each (nominal Bonferroni 95% family coverage). All other intervals are exploratory 95%, without multiplicity adjustment. Resampling retains all questions within each sampled source and question weighting. Coverage is approximate and conditional on one trained quartet.

Both primary intervals exclude zero, supporting each correction's utility conditional on the other in this regime. Negative interactions indicate overlapping rather than superadditive gains. Development joint-minus-GDN includes zero without a prespecified equivalence margin; universal component necessity is not established. The original two-arm three-seed experiment does not estimate partial-arm seed variability.

The protocol postdates the original attention/joint results; multiplicity adjustment does not remove research-history adaptation, known exposure or unknown pretraining. A separate CPU audit checks 1,350 closed chunks, independently recomputes ANLS/EM for 33,292 answer records, and replays all means and intervals. It verifies completion, initialization and order bindings; full trajectory auditing remains the original analysis. The numeric supplement supplies source-grouped scores and replay code.

Table 14: All ten frozen contrasts. Only the two official joint-minus-partial rows form the primary family. The interaction is negative on both cohorts; development joint-minus-GDN includes zero.

Cohort	Contrast	ANLS points	Interval	Level
Dev.	Joint − history	2.103	[1.381, 2.844]	95%
Dev.	Joint − GDN	0.130	[-0.351, 0.617]	95%
Dev.	History − attention	5.323	[4.255, 6.414]	95%
Dev.	GDN − attention	7.297	[6.237, 8.350]	95%
Dev.	Joint − history − GDN + attention	-5.193	[-6.255, -4.146]	95%
Official	Joint − history	1.688	[1.043, 2.351]	97.5%
Official	Joint − GDN	0.819	[0.255, 1.394]	97.5%
Official	History − attention	7.136	[6.158, 8.112]	95%
Official	GDN − attention	8.005	[7.048, 8.954]	95%
Official	Joint − history − GDN + attention	-6.317	[-7.338, -5.282]	95%

K EXECUTION DISCREPANCY: CONDITIONAL THEORY AND FULL-COHORT DIAGNOSIS

K.1 BUDGET DIFFERENTIATION

Implicit differentiation of the budget constraint gives, for upstream gate derivatives u_i and $a_i = z_i(1 - z_i)$,

$$\frac{\partial L}{\partial s_i} = \frac{a_i}{\tau}\left(u_i - \frac{\sum_j a_j u_j}{\sum_j a_j}\right).$$

Thus the derivative accounts for the threshold's dependence on all scores. We anneal the temperature from 1.0 to 0.1 with a fixed cosine schedule. Threshold-adjusted budgeted soft top-K gates are established tools, including in DiffPrune (He et al., 2026); our novelty claim concerns the downstream state/history operator. This gate is a continuous relaxation; it is not an unbiased gradient of hard top-k selection.

K.2 FIXED-STREAM PROPAGATION IDENTITY AND BOUND

We analyze discrepancy propagation through a GDN recurrence conditional on fixed input streams. Native recurrence is due to Yang et al. (2025); copy/update transitions are established (Campos et al., 2018). Fix k_t, v_t, g_t, β_t and initial S_0 in the two executions. Convolution, earlier layers and attention can change these streams in a real model; that feedback is outside this result.

For $E_t = S_t(z) - S_t(m)$, subtract the recurrences to obtain $E_t = A_t(z_t)E_{t-1} + \delta_t$, with δ_t from Equation 13 and $E_0 = 0$. Expanding recursively yields the stated ordered product, whose empty case is identity. Submultiplicativity gives

$$\|E_T\|_F \le \sum_{t=1}^{T}\left(\prod_{j=t+1}^{T}\|A_j(z_j)\|_2\right)\|\delta_t\|_F. \tag{5}$$

When $g_t \le 0$, $\beta_t \ge 0$, $\beta_t\|k_t\|^2 \le 2$ and $u \in [0,1]$, $I - u\beta_t k_t k_t^\top$ has eigenvalues 1 orthogonal to k_t and $1 - u\beta_t\|k_t\|^2$ along it. Thus $\|A_t(u)\|_2 \le 1$. The zero-key case follows directly. Differentiating gives

$$A_t'(u) = g_t e^{ug_t}(I - u\beta_t k_t k_t^\top) - e^{ug_t}\beta_t k_t k_t^\top, \tag{6}$$

$$b_t'(u) = \beta_t k_t v_t^\top. \tag{7}$$

Hence $\|A_t'(u)\|_2 \le |g_t| + \beta_t\|k_t\|^2$. Integrating between m_t and z_t gives

$$\|\delta_t\|_F \le C_t|z_t - m_t|, \tag{8}$$

$$C_t = (|g_t| + \beta_t\|k_t\|^2)\|S_{t-1}(m)\|_F + \beta_t\|k_t\|\|v_t\|, \tag{9}$$

$$\|E_T\|_F \le \sum_t C_t|z_t - m_t|. \tag{10}$$

A fixed-query output difference is bounded by $\|q_T\|\|E_T\|_F$. These coefficients depend on state and projected streams; small mean gate distance alone is insufficient. The bound may be loose, does not cover full-decoder feedback, and does not predict task accuracy. The archived probe does not estimate these constants; it independently measures full-model output fidelity.

Attention-only masking does not force $A_t = I, b_t = 0$ at dropped recurrent steps. A simple exact-arithmetic example starts from zero state: a retained token establishes state 1; a dropped step with decay $1/2$ and no write leaves state $1/2$, whereas deletion preserves 1. A later retained step with effective state multiplier $1/4$ reads $1/8$ instead of $1/4$. Likewise, advancing a convolution buffer at a dropped step can replace the retained predecessor. These are counterexamples to a uniform guarantee, not proof of pointwise necessity or the cause of the observed accuracy gain. Binary consistency also does not mean the soft path equals expected random deletion: already $e^{zg} \ne (1 - z) + ze^g$ for an interior gate and nonzero decay.

K.3 Complete exploratory deployment-loss accounting

Use the original fixed checkpoints, all 2,974 development questions (750 images; 475 sources), and the original six scorer/path branches. Every branch uses the same reference-answer prefixes and EOS target. Per-question target means are averaged over questions, not pooled by answer length. The new analysis verifies every closed chunk, joins question/source/pixel identities, and rechecks target loss and KL means. No model calls, checkpoint selection or changed gates are introduced.

Table 15: Complete crossed loss matrix, in nats per answer target. A soft path's loss is not a generation accuracy.

Frozen scorer	State-soft CE	Attention-soft CE	Hard CE
State-trained	0.207070	0.212018	0.226565
Attention-trained	0.226118	0.207055	0.442724

For $O = L_h - L_s$, the state scorer has mean +0.01950 [0.01400, 0.02517] and attention scorer +0.23567 [0.20478, 0.26736]. Their difference is +0.21617 [0.18633, 0.24734], completing Equation 14 alongside the main-text soft and hard differences. All five new intervals use 10,000 paired source resamples, seed 2026092201, at exploratory unadjusted 95% confidence. The analysis was chosen after aggregate means were known; this adjustment does not address research-history selection. It does not explain the ANLS gain causally.

State gating has lower hard-to-soft KL on 73.94% [71.94, 75.95] of questions for the state scorer and 78.82% [76.95, 80.62] for the attention scorer, using the same exploratory bootstrap. Median KL advantages are 0.000762 and 0.002619; their 10th percentiles are negative (-0.001810, -0.002352). We retain all questions and the full empirical distributions rather than selecting a favorable subgroup. Original primary 97.5% intervals remain unchanged in Appendix H. The numeric supplement contains every question-level cell and all exploratory contrasts.

K.4 Original complementary evidence summary

This preserved figure adds seed replication, transfer and complete-cohort error accounting; it is not an independent dataset or an additional hypothesis family.

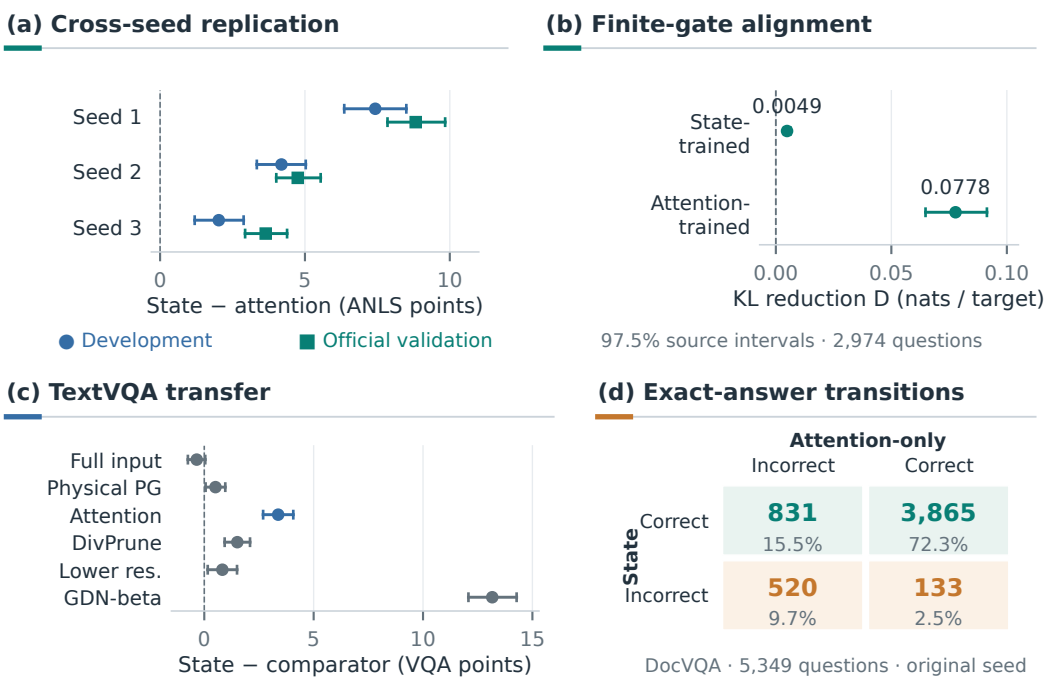

Figure 9: Four views of the completed evidence. (a) Three predeclared seeds on complete development and official validation; 95% source intervals condition on each pair. (b) Fixed-scorer soft–hard KL contrasts; 97.5% intervals, reference-prefix conditioning. (c) State-minus-comparator TextVQA differences; 95% image intervals, one seed, unequal comparator computation. (d) All 5,349 original-seed official questions by exact correctness: 831 become correct and 133 become incorrect relative to attention-only. Panel (d) is a post-hoc cross-tabulation; the task-level results do not establish component-specific causal mediation.

L SOURCE-EXPOSURE SENSITIVITY AND IMPLEMENTATION CHECKS

L.1 KNOWN HISTORICAL-SOURCE EXCLUSION

The earlier frozen cohort contains one question from each of 300 official-validation source documents. We exclude *all* current questions belonging to those sources, not only the 300 previously selected questions. This leaves 3,636 questions, 892 distinct images and 638 sources; the excluded stratum contains 1,713 questions, 393 images and 300 sources. The full main cohort remains unchanged. Selection uses source identities only and was specified before computing these stratum results, after the full-cohort outcomes were known.

Table 16: All seven original-seed ANLS means after the source-identity partition. This is a post-hoc sensitivity of saved outputs, not a new independent test set.

Original-seed method	Remaining sources	Known exposed sources
State-consistent	93.85	93.30
Attention-only	84.59	85.40
Full input	94.74	94.54
DivPrune	82.69	82.16
Lower resolution	92.59	92.47
Physical PG	93.15	92.85
GDN-beta	75.94	75.77

Table 17: Exploratory 95% source-bootstrap intervals. Seed suffixes identify 2026091709, 2026091902 and 2026091903. All strata and contrasts remain reported.

Paired contrast	Remaining sources	Known exposed sources
State–attention, seed 1709	9.26 [8.05, 10.48]	7.90 [6.34, 9.53]
State–attention, seed 1902	4.52 [3.66, 5.41]	5.25 [3.89, 6.72]
State–attention, seed 1903	4.01 [3.09, 4.91]	2.88 [1.72, 4.12]
Original state–PG	0.70 [0.16, 1.27]	0.45 [−0.38, 1.30]

Each stratum uses 20,000 common source-document resamples, seed 2026092203, carrying every question of a sampled source and retaining question weighting. The intervals are unadjusted and condition on the trained checkpoints. They do not remove adaptive research history, unrecorded repository exposure or unknown backbone pretraining exposure. In particular, the remaining sources are not certified pristine. The portable supplement provides all 5,349 per-question records with source identities, exposure flags, seven original-arm scores and all three trained pairs.

L.2 EXECUTABLE OPERATOR AND NATIVE-PATH CHECKS

The accompanying implementation archive supplies the byte-identical frozen operator, an independent sequential convolution reference, the native adapter, training/data/evaluation interfaces and qualification commands. With the pinned PyTorch 2.5.1 and Transformers 5.3.0 environment, run the following from its `core_replay` directory:

```
export CUDA_VISIBLE_DEVICES=''
export HF_HUB_OFFLINE=1
python -B -m b1_repro.core.qualify_cpu --output cpu.json
```

Use a persistent output directory. The qualifier independently checks the dense history operator against a sequential recurrence, FP64 finite-difference gradients for history and budget gates, compact causal convolution at binary gates, native hybrid endpoints at lengths 11 and 70, and a full tiny-hybrid directional derivative through soft gates. It also checks hook restoration and the scorer parameter count.

The additional `random_endpoint_check.py` repeats these checks and tests 18 randomly selected binary masks: lengths 11/70, three retention fractions (0.25/0.5/0.75), and three masks per setting, while preserving the first and final text tokens and original positions. A fresh CPU execution passes, with maximum retained-state absolute error 2.39×10^{-7} at tolerance 2×10^{-5}. The decoder is a randomly initialized 32-wide, four-layer native FP32 hybrid model; these engineering tests are not pretrained-9B, BF16, arbitrary-input or task-accuracy certification. Existing pretrained qualifications retain their original scope. No full GPU training or qualification is claimed by a numeric replay.

The completed scientific experiments train at half retention and evaluate at three retention ratios on one fully evaluated backbone. Temperature sweeps, trained alternative endpoint-consistent interpolations, multi-seed partial-arm ablations and equal-time published-method comparisons remain distinct experiments; the checks above do not substitute for them.

M General deletion contract and alternative soft paths

Assumptions and proof. Consider deterministic, causal prefill with unpadded batch-one sequences, finite activations, no dropout, identical initial states and retained logical positions. Each mixer must satisfy Equation 1, and its state must include all quantities carried between token steps. Pointwise normalization, feed-forward transformations and residual additions commute with gathering. If the retained inputs to a module agree with compact execution, its contract makes its retained outputs agree; induction over modules proves agreement at every retained position. For recurrent modules, discarded steps are identities and retained steps are native updates, so induction over retained steps proves the local contract. A causal attention module has the same retained keys, values, coordinates and causal order after inaccessible keys are removed from its softmax. An initial retained text key avoids an empty softmax support.

This sufficient condition excludes unhandled cross-token normalization, packing-dependent operations, global routing capacity, renumbered logical coordinates and external state that advances on discarded positions. It concerns prefill, not arbitrary generation-cache compaction. A model with such operations requires a separate adapter and qualification. The affine transition in Table 1 is valid only when A, b are computed from retained inputs that agree under the contract; any preceding causal convolution must be controlled too. For nonlinear cells, freezing a visible output alone is insufficient: all hidden and cell states must freeze. These observations explain why gating a model-specific write coefficient without auditing its other state is inadequate.

Continuous paths are not identified by endpoints. For $g \le 0$, convexity and the linear-interpolation error bound for $f(z) = e^{zg}$ give

$$0 \le (1-z) + ze^{g} - e^{zg} \le \tfrac{1}{2} g^2 z(1-z), \qquad 0 \le z \le 1. \tag{11}$$

The upper bound uses $f''(z) = g^2 e^{zg} \le g^2$. It bounds only the scalar decay factor, not a full decoder, learned selector, or task loss. Interpolation of the entire native transition, Equation 2, is a third admissible path. Different paths can have different gradients inside $(0, 1)$. Therefore exact endpoint consistency is a design constraint, not evidence that the exponential path is optimal. Comparing their trained selectors and temperature schedules remains a separate empirical question.

Executable reference qualification. The supplement evaluates 64 random binary masks for each of six float64 reference systems: causal attention, causal convolution, affine state space, nonlinear recurrence, gated delta update, and a composed attention/convolution/delta stack. All 384 gather-versus-skip comparisons pass with maximum absolute error below 9×10^{-16}. Six finite-difference gate-gradient checks pass. A 1,001-point grid verifies the scalar interpolation bound at $g \in \{-0.01, -0.1, -1, -5\}$. A negative control makes the omission concrete: a state of one becomes e^{-1} when an excluded step still decays it, instead of remaining one. These are small synthetic implementation checks, not new pretrained-model experiments or evidence of downstream superiority on the other architecture families.

M.1 Layer-level discrepancy and measured deployment loss

For a single GDN layer, *fix* its projected k_t, v_t, g_t, β_t streams. Write $S_t = A_t(u_t) S_{t-1} + b_t(u_t)$, where

$$A_t(u) = e^{u g_t}(I - u\beta_t k_t k_t^{\top}), \qquad b_t(u) = u\beta_t k_t v_t^{\top}. \tag{12}$$

For soft z and binary m, the state difference obeys the exact identity

$$E_T = \sum_{t=1}^{T} \underbrace{A_T(z_T)\cdots A_{t+1}(z_{t+1})}_{\text{propagation}} \underbrace{\big[(A_t(z_t) - A_t(m_t))S_{t-1}(m) + b_t(z_t) - b_t(m_t)\big]}_{\delta_t\text{: local discrepancy}}. \tag{13}$$

When $g_t \le 0$ and $0 \le \beta_t \|k_t\|^2 \le 2$, these transitions are nonexpansive. Appendix K derives $\|E_T\|_F \le \sum_t C_t |z_t - m_t|$, with explicit state-dependent C_t. Thus gate closeness alone does not determine fidelity. This conditional layer analysis does not bound the full input-dependent decoder.

We measure the real model separately. Let $L_s(\theta)$ be held-out reference-prefix cross entropy under scorer θ's own soft training operator, $L_h(\theta)$ the same loss under hard deletion, and $O(\theta) = L_h(\theta) -$

$L_s(\theta)$. For attention-trained a and state-trained s,

$$L_h(a) - L_h(s) = \underbrace{L_s(a) - L_s(s)}_{\text{surrogate difference}} + \underbrace{O(a) - O(s)}_{\text{deployment discrepancy difference}} . \tag{14}$$

This is loss accounting, not a generalization or causal-mediation bound; O need not be nonnegative on each question.

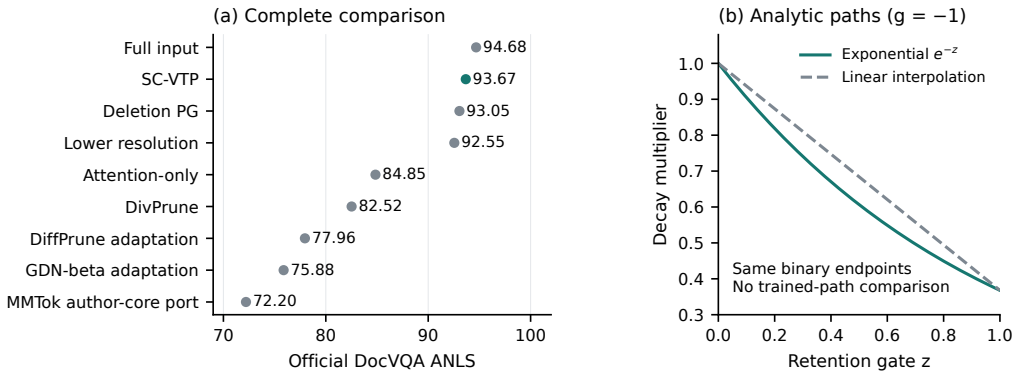

Figure 10: New comparison and design-space clarification. (a) All nine methods on every official DocVQA validation question, including the MMTok author-core port. PG remains close to SC-VTP; published-method adaptations do not have equal training capacity or cost. The plain points are one-seed means without confidence bars; paired uncertainty is reported in the text. (b) Two analytic scalar decay paths share binary endpoints but differ inside the interval. This panel is not a trained-model experiment and does not favor either path empirically.

N Complete-component DiffPrune adaptation

The comparator is a paper-based adaptation of He et al. (2026); no verified author implementation was available to this experiment. It uses two vision-configuration transformer scorer blocks before patch merging. Four patch scores are averaged per merged token. Constrained gates drive $\sqrt{\alpha}x + \sqrt{1-\alpha}\epsilon$ with unit Gaussian noise, followed by one diagonal-attention repair block and the frozen merger/decoder. Deployment removes noise and repair, scores clean features and gathers original merged features in score-selected, original sequence order. This includes all named components and is materially different from treating our attention-only ablation as DiffPrune.

The adaptation completes one pass over the same 26,143 training questions, 6,786 RGB images and 3,268 optimizer updates. Its deployed scorer has 30,480,161 parameters; the repair adds 15,239,504, including 2,656,512 structurally inactive Q/K parameters under diagonal attention. The resulting parameterized total is 45,719,665. Backbone, image preparation, question order, half-retention deployment and greedy answer limits are shared with the comparison; capacity, objective and optimization dynamics are not matched. The protocol records its hyperparameter mapping. A failed earlier native attention-mask qualification is retained historically; the completed version corrects that interface before training, without changing the comparator architecture.

Table 18: Paired source-document bootstrap, 10,000 draws; supplementary unadjusted intervals with one training seed. Both official empty predictions and the single capped generation are retained.

Complete cohort	SC-VTP ANLS	Adaptation ANLS	Difference [95% CI]
Development: 2,974 questions	95.85	80.73	15.12 [13.51, 16.75]
Official: 5,349 questions	93.67	77.96	15.72 [14.47, 16.98]

A closed-record replay verifies all question identities, answers, source documents, image/prompt hashes, scored predictions, stable top-K support, gathered positions and native generation accounting. It exactly reproduces the archived contrasts and intervals and exports an 8,323-row portable paired-score CSV. This numerical replay does not repeat model execution. The large advantage is specific to this complete-component port and protocol; the smaller PG comparison remains essential to interpreting practical gains. We claim neither an author-code reproduction nor a speed/quality advantage at equal training resources.

O FULL REPRODUCIBILITY AND ASSISTANCE DISCLOSURE

Reproducibility. The method section states the retention operator and its endpoint assumptions. The executable supplement contains the frozen operator, native adapter, training/evaluation interfaces and CPU qualification; Appendix L states the command and its scope. The experimental protocol and appendices specify dataset revisions, source-document partitions and exclusions, scorer architecture, optimization, final-checkpoint selection and evaluation metrics. Frozen manifests record source and checkpoint hashes; immutable journals retain all designated examples and their generated outputs. Statistical units, resampling procedures and known prior evaluation exposure are stated with the relevant results. A portable numeric supplement reproduces the complete seven-arm tables for both datasets and 25 frozen ANLS/VQA contrasts from per-question scores; it requires no model or original filesystem. An additional portable CSV and replay reproduce the three official-seed contrasts. The component supplement replays four-arm means and ten contrasts. A separate depth supplement replays the complete crossed losses, exploratory source-bootstrap intervals and conditional mathematical checks. Numeric replays do not repeat model execution.

AI assistance. Generative AI assistance was used in literature discovery and synthesis, hypothesis and method development, experimental design, implementation, formulation and checking of mathematical claims and proof sketches, data preparation, result interpretation, scientific figures, manuscript drafting and translation. The recorded workflow includes source checks against primary literature, separate implementation checks, checks of data identity and metric calculations, and preservation of unsuccessful and interrupted experiments. Separate AI agent contexts assisted some implementation and audit passes; independence in these checks refers to code paths or contexts. The experimental sections distinguish completed model executions, CPU verification and planned work, and identify the scope of each validation.

P OBSERVED FINITE-GATE PROFILES

The complete closed alignment probe records per-question mean Bernoulli gate entropy and mean absolute distance from the fixed hard top-K mask. We verify all 2,974 question records and their chunk hashes. Each number below first averages over that question's visual positions; population summaries then weight questions equally, matching the main probe. Quantiles are across question means, not across individual tokens. These are post-hoc summaries of existing native model executions; there are no new forwards or changed temperatures.

Table 19: Actual gate profiles at the common evaluation temperature $\tau = 0.1$. These summarize two learned score distributions, not matched-score temperature interventions.

Scorer	Gate diagnostic	Mean	10th	Median	90th
State	Entropy (nats)	0.29055	0.23854	0.28302	0.35780
Attention	Entropy (nats)	0.48188	0.44078	0.48199	0.52384
State	Mean $\lvert z - m \rvert$	0.13525	0.10409	0.13056	0.17619
Attention	Mean $\lvert z - m \rvert$	0.25234	0.22463	0.25159	0.28004

State-minus-attention mean differences are -0.19133 nats in entropy and -0.11710 in gate distance. Exploratory unadjusted 95% source-bootstrap intervals are $[-0.19451, -0.18811]$ and $[-0.11913, -0.11501]$, using 10,000 common draws over 475 sources, seed 2026092303. The lower uncertainty and smaller hard-mask distance describe these trained selectors. Different score scales can change these statistics; they do not identify causal mediation or prove the exponential interpolation optimal. The crossed-path probe remains important because it compares operators while holding a scorer and its gates fixed. This connects the observed finite gates to the empirical fidelity question without turning an endpoint proof into a task-performance guarantee.

Q ANALYTICAL DEPLOYMENT PROJECTION ACCOUNTING

This CPU analysis counts a specified subset of native prefill arithmetic, not hardware runtime or total model FLOPs. One multiplication with accumulation is one MAC. With hidden size $d = 4096$, MLP width $u = 12288$, attention query width $q = 4096$, key/value width $v_a = 1024$, GDN key width $k_g = 2048$, value width $v_g = 4096$ and 32 value heads, the per-row coefficients are

$$c_{\mathrm{MLP}} = 3du, \qquad c_{\mathrm{att}} = d(3q + 2v_a), \qquad c_{\mathrm{GDN}} = d(2k_g + 3v_g + 2 \cdot 32).$$

The attention query projection includes its native output gate. The GDN terms include qkv, z, a, b input projections and the output projection. Thus

$$C = 32c_{\mathrm{MLP}} + 8c_{\mathrm{att}} + 24c_{\mathrm{GDN}} = 6{,}918{,}504{,}448 \ \mathrm{MAC/visual\ row}.$$

Multiplication by each arm's recorded question-weighted visual-row count gives Table 23. GDN-beta counts original rows through zero-based block four and retained rows in the remaining blocks, giving 3.950 TMAC. Lower resolution gives 3.355 TMAC but also changes encoder work. The state scorer's two linear projections add $N(4096 \cdot 128 + 128)$ MAC, or 0.518 GMAC per question. Its nonlinearities are not counted.

The count excludes the vision encoder/merger, text-row projections, attention score/value products, recurrent state products, convolution, normalization, activations, language-model head and autoregressive generation. Other selectors' overhead is not added to their decoder-only column; in particular the larger DiffPrune scorer is excluded. The common retained-row count therefore does not establish equal total inference cost across selectors. For an unchanged number T of text rows, decoder projection work alone is $(N + T)C$ versus $(K + T)C$ before scorer cost, so even that broader ratio is not generally one half. No token-squared attention saving is assigned to the entire hybrid decoder. Frozen model configuration, module shapes, source hashes and exact input means accompany the portable calculation.

R FIDELITY ACROSS OBSERVED GATE AMBIGUITY

The binary contract alone does not address finite gates. We therefore supplement the complete crossed-path probe with a post-hoc stratified analysis, using all 2,974 questions from 475 source documents. For each frozen scorer separately, we partition examples by empirical quartiles of their mean $|z - m|$. Threshold ties remain together; the strata need not have equal sizes. Within every question we compare the attention-only and coordinated-state operators with the *same scorer, gates, projected entry features and reference prefix*. Define the advantage as $\mathrm{KL}(p_{\mathrm{hard}} \| p_{\mathrm{attention}}) - \mathrm{KL}(p_{\mathrm{hard}} \| p_{\mathrm{state}})$, in nats per target token.

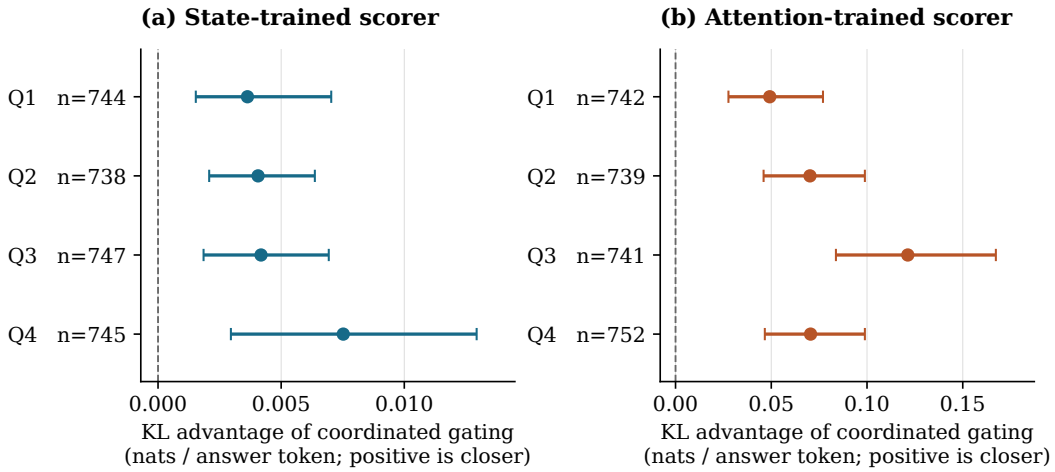

Figure 11: Coordinated gating is closer to hard deletion across the observed ambiguity range. Q1 has the smallest gate-to-hard distance and Q4 the largest, separately within each scorer. Whiskers are 99.375% source-bootstrap intervals (Bonferroni adjustment for eight comparisons). Axes have different scales. This is a post-hoc teacher-conditioned analysis, not eight independent replications or a causal hardness intervention.

Table 20: Complete observed strata; no unfavorable questions are removed.

Scorer	Quartile	Questions	Mean advantage	99.375% interval	Positive (%)
State	1	744	0.00363	[0.00153, 0.00703]	76.6
State	2	738	0.00406	[0.00208, 0.00637]	75.1
State	3	747	0.00418	[0.00185, 0.00693]	73.2
State	4	745	0.00752	[0.00296, 0.01294]	70.9
Attention	1	742	0.04921	[0.02763, 0.07701]	81.4
Attention	2	739	0.07021	[0.04599, 0.09888]	78.9
Attention	3	741	0.12130	[0.08377, 0.16729]	78.5
Attention	4	752	0.07053	[0.04663, 0.09889]	76.5

All eight adjusted lower bounds are positive. This supports a finite-gate fidelity advantage across these observed strata, rather than only in nearly binary examples. It does not show that increasing gate ambiguity causes a larger effect; indeed, the attention-scorer means are not monotone. We resample all 475 source documents jointly with 20,000 draws, seed 2026092308, retaining the observed stratum boundaries. The intervals are conditional on those boundaries, and this exploratory analysis does not correct the wider adaptive research history. No new model forward or accuracy measurement is introduced.

S A LOCAL FINITE-GATE PERTURBATION BOUND

The following standard norm argument makes the limits of mean gate distance explicit. It applies to a single recurrent primitive with *fixed* projected inputs across the two paths, not the complete decoder. Use the state orientation of Section 3 and assume $\|k_t\|_2 \le 1$, $g_t \le 0$, $\beta_t \in [0,1]$ and $z_t \in [0,1]$. Then

$$T_z(S) = e^{zg}(I - z\beta kk^\top)S + z\beta kv^\top.$$

The eigenvalues of $I - z\beta kk^\top$ lie in $[0,1]$, so T_z is e^{zg}-Lipschitz in S under the Frobenius norm. Differentiating at a fixed state gives

$$\|\partial_z T_z(S)\|_F \le (|g| + \beta)\|S\|_F + \beta\|v\|_2.$$

For identically initialized soft and binary trajectories on the same input sequence, put $e_t = \|S_t^z - S_t^m\|_F$ and $c_t = (|g_t| + \beta_t)\|S_{t-1}^m\|_F + \beta_t\|v_t\|_2$. Adding and subtracting $T_{z_t}(S_{t-1}^m)$, then unrolling the recurrence, yields

$$e_t \le e^{z_t g_t} e_{t-1} + |z_t - m_t| c_t \tag{15}$$

$$\le \sum_{i=1}^{t} \exp\left(\sum_{j=i+1}^{t} z_j g_j\right) |z_i - m_i| c_i. \tag{16}$$

For a fixed read query, output error is at most $\|q_t\|_2 e_t$. Thus the same unweighted gate distance can have different effects depending on state/value magnitude and the remaining decay horizon. Binary equality and finite-gate error control are different requirements.

For the fixed-input history update $H'_z = (1-z)H + z(PH + e_0x^\top)$, where P shifts the buffer and $\|P\|_2 \le 1$, the analogous local inequality is

$$\|H_t^z - H_t^m\|_F \le \|H_{t-1}^z - H_{t-1}^m\|_F + |z_t - m_t|\|(P-I)H_{t-1}^m + e_0x_t^\top\|_F.$$

For attention with fixed logits and values, let $a_i = z_i e^{\ell_i}$, $a'_i = m_i e^{\ell_i}$ and positive normalizers Z, Z'. Expanding the normalized difference using either denominator gives

$$\|p_z - p_m\|_1 \le \min\left(2, \frac{2\|a - a'\|_1}{\max(Z, Z')}\right).$$

Multiplication by $\max_i \|v_i\|$ bounds the corresponding output discrepancy. These local bounds do not cover changes in projected inputs across decoder layers, rounding, or task accuracy; substituting the observed mean gate distance into them would not certify the native model. They explain why endpoint correctness is necessary yet insufficient. Bounded CPU checks cover 120 synthetic instances of each primitive, including 5,040 recurrent steps; these are algebra/implementation checks, not OCR experiments or a training-speed result.

Table 21: Selected paper-reported TextVQA results, not reproduced here. Blocks are independent protocols. K/N denotes retained fraction where reported. The ET-Prune paper's "keep" column is measured on OCRBench-v2 (‡), not TextVQA; it is shown with that restriction. Different pruning sites, training data and preprocessing preclude cross-block best-score claims.

Method	Reported budget	TextVQA
Qwen2.5-VL-7B; MMTok Table 3 (Dong et al., 2026)		
Full input	$K/N = 100\%$	77.72
VisionZip	$K/N = 20\%$	70.43
DivPrune	$K/N = 20\%$	65.86
MMTok	$K/N = 20\%$	70.49
Qwen2.5-VL-7B; DiffPrune Table 3 (He et al., 2026)		
Full input	$K/N = 100\%$	84.80
FastV	$K/N = 33.3\%$	77.90
HoloV	$K/N = 33.3\%$	78.90
DiffPrune	$K/N = 33.3\%$	79.00
Qwen3-VL-8B; ET-Prune Table 1 (Ding et al., 2026)		
PyramidDrop	Keep $50.0\%^{\ddagger}$	74.63
SparseVLMs	Keep $50.0\%^{\ddagger}$	77.30
ET-Prune	Keep $50.3\%^{\ddagger}$	78.30

T PUBLISHED COMPARISONS: ATTRIBUTION AND BUDGET DEFINITIONS

Paper-reported values are legitimate comparison evidence under their original protocols. We retain the reporting source and table rather than presenting them as our reproductions. An original paper's baseline entry remains attributed to that reporting paper. No paired confidence interval or measured latency is inferred from a published aggregate. Our local DivPrune reference port, MMTok author-core backbone port, DiffPrune paper-based adaptation, attention-only ablation, and literature scores have distinct provenance. Table 2 contains only completed local results; Table 21 contains only literature results.

Learned selector reference. Takezoe et al. (2026), Table 1, report the following TextVQA values on LLaVA-1.5-7B. The pruned rows use 128 visual tokens averaged across decoder layers, not 128 tokens from decoder entry. Full input has 576. These are reported values, not matched Qwen3.5 results.

Table 22: A selected single-benchmark extract from LearnPruner Table 1. The complete selected budget block includes competitors above and below LearnPruner.

Method	TextVQA	Method	TextVQA
Full input	58.2	FastV	50.6
SparseVLM	54.9	DivPrune	56.1
DART	56.3	VisPruner	57.0
VisionZip	56.8	VisionZip (fine-tuned)	57.0
TwigVLM	57.8	LearnPruner	57.3

OCR-specific reference. LayoutLite uses FireRed-OCR on 1,651 OmniDocBench images: its Table 1 reports Omni scores 92.753 with full input and 91.432 at 50% compression (Liu et al., 2026). This is structured OCR, with text/formula/table evaluation; it is not DocVQA ANLS. ET-Prune Table 1 reports soft TextVQA scores on a 0–1 scale; Table 21 multiplies them by 100 and retains the original OCRBench-v2 budget qualification. DocPrune evaluates a retrieval-augmented document pipeline (Choi et al., 2026); its document setting is not automatically single-page DocVQA validation.

Table 23: Complete official DocVQA validation: 5,349 questions, 1,285 distinct RGB images and 938 source documents. Learned methods use one seed and fixed final checkpoints; PG and the DiffPrune adaptation have different training computation; the latter also uses a larger scorer. ANLS is on a 0–100 scale. Counts are question-weighted. † counts only decoder linear projections at visual rows during prefill (TMAC: 10^{12} multiply-accumulates); it excludes the encoder, text rows, mixing products and selector overhead. GDN-beta uses five full-token blocks. This is not total compute or latency (Appendix Q).

Method	ANLS	EM (%)	Retained tokens	Proj. TMAC†
Full input	94.68	89.81	987.28	6.831
State-consistent	93.67	87.79	493.57	3.415
Attention-only	84.85	74.74	493.57	3.415
DivPrune	82.52	71.00	493.57	3.415
Lower resolution	92.55	85.62	484.86	3.355
Physical-deletion PG	93.05	87.34	493.57	3.415
GDN-beta (after block 4)	75.88	64.03	493.57	3.950
DiffPrune (paper-based adaptation)	77.96	68.05	493.57	3.415
MMTok (author-core port)	72.20	65.68	493.57	3.415

Comparability requirements. A direct ranking requires agreement on model revision, input resolution, task/split, metric, prompting and decoding, retained-token convention, and intervention site or comparable measured work. Reported scores can join a compatible table without full reproduction. Different settings remain separate blocks. Relative score retention can describe a within-protocol trade-off, but does not remove differences in task difficulty. Missing entries are unmeasured, not zero.

U Complete same-backbone task and workload table

This detailed table supplements the two-task main comparison, including the completed MMTok port. It includes all original arms and retains the narrower meaning of the analytical projection count.

V Complete mechanism overview of the measured results

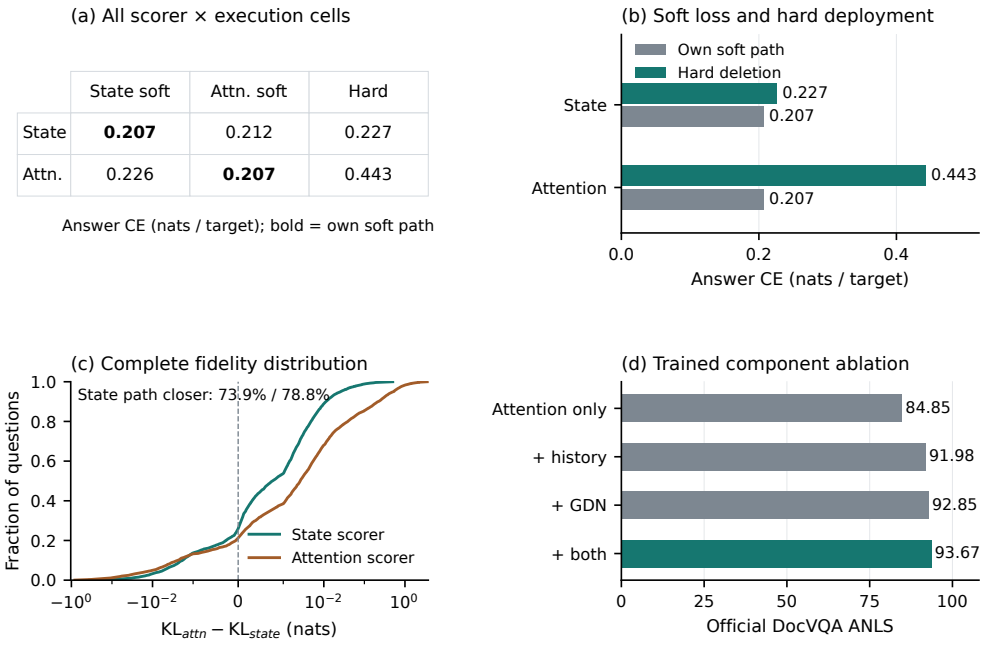

Figure 12: Approximation fidelity and selector utility, shown with direct numerical comparisons. (a) All six scorer–execution cells on 2,974 development questions; bold entries are each scorer's own training path. (b) Matched soft losses round to 0.207, but hard-deletion losses separate. (c) The complete empirical distribution of hard-to-soft KL differences; positive values favor coordinated state gating. The symmetric-log axis preserves negative differences and outliers. (d) All four trained component arms on 5,349 official questions; every arm gates attention. Paired component intervals are reported in the text. Panels (a)–(c) condition on reference prefixes and are not generation accuracies.

W MMTok author-core port: implementation and complete results

Comparison question. MMTok (Dong et al., 2026) selects tokens by joint question–vision and vision–vision coverage. This training-free, question-conditioned method complements the image-only learned SC-VTP selector. We execute the unmodified author selection core at commit `8e2686836ed5`, adapting its Qwen2.5 feature interface to our pinned Qwen3.5-9B backbone. This is a same-backbone author-core port, not a reproduction of the original paper's Qwen2.5 scores. The frozen defaults are $\alpha = 0.5$, text–vision temperature 0.01, and vision–vision temperature 0.2, with neither keyword cleaning nor padding exclusion. We do not tune them against either validation set.

Feature and deployment contract. The author's Qwen pathway embeds the complete user question/instruction, prefixed by `Question:`, through the frozen language embedding layer. It uses merged visual vectors for text–vision coverage and the mean of each four pre-merger patch vectors for vision–vision coverage. We preserve this construction; Qwen3.5 does not require the Qwen2.5 window-order reversal. Captured merger outputs equal the decoder-entry visual rows. The unchanged core returns ordered unique indices and the corresponding original feature vectors. Physical gathering preserves all text tokens and original logical mRoPE coordinates, followed by the same greedy 128-token decoder. Eight engineering conditions on the first two fixed questions per task verify equality of generated token IDs with native generation; these questions remain in the complete cohorts. Float32/BF16 CPU checks additionally verify JIT/Python selector parity. These checks qualify integration, not accuracy or generalization.

Frozen complete-cohort analysis. The supplement contains 21,047 designated generations: all 5,349 DocVQA questions at 25%, 50%, and 75% retention, plus all 5,000 TextVQA questions at 50%. The same public versions, preprocessing, input cap, scoring and known prior exposure apply. Every saved chunk is hash-verified. Half-retention comparisons additionally check image and pixel hashes, prompt hashes, visual counts and grids against the original SC-VTP outputs. Two prespecified primary endpoints compare SC-VTP with MMTok at 50%: 20,000 paired bootstrap draws over 938 DocVQA sources and 3,166 TextVQA images, with 97.5% intervals (seed 2026092401). These intervals account for the two-endpoint family but condition on the fixed trained SC-VTP selector.

Table 24: All completed MMTok conditions. Score is DocVQA ANLS or official TextVQA soft accuracy, on a 0–100 scale. Token counts are question-weighted means. Length cap counts outputs reaching the 128-token limit; all outputs are included. The complete paired comparator panel is now reported in Appendix Y; the additional state–MMTok contrasts remain descriptive.

Dataset	Retained fraction	Score	Retained tokens	Length cap
DocVQA	25%	46.74	246.69	3
DocVQA	50%	72.20	493.57	0
DocVQA	75%	86.04	740.40	0
TextVQA	50%	72.96	370.90	0

At 50% retention, SC-VTP exceeds the port by 21.47 ANLS points [19.83,23.16] on DocVQA and 11.22 accuracy points [10.05,12.37] on TextVQA. This supports comparative task quality in the fixed target pipeline. It neither attributes the gap solely to state correction nor establishes equal training cost: SC-VTP learns from DocVQA, whereas MMTok is training-free and uses question features at inference. The port also depends on the new backbone's embedding geometry. Shared-GPU elapsed times do not support latency claims. DivPrune and direct-deletion PG remain stronger local accuracy comparators than this particular MMTok port.

X QUALIFIED NATIVE DEPLOYMENT TIMING

Measured native deployment latency. A fixed 32-source cohort uses one question/page per source and three repeats after warm-up. Native generation emits exactly one or 32 tokens; one-token completion is a first-token proxy, not streaming TTFT. Total time includes preprocessing, transfer, encoding, selection, decoding and detokenization from resident image bytes, without cached features. Table 3 uses within-source median times on A100-SXM4-80GB with native unfused GDN. The timing GPU is exclusive; GPU1 and the shared host remain active.

Table 3 in the main text reports all three methods and both output lengths.

Paired full-input/SC-VTP geometric latency ratios are 1.160 [1.145, 1.172]× and 1.035 [1.030, 1.040]× for one/32 tokens. Corresponding PG/SC-VTP ratios are 0.999 [0.990, 1.006]× and 1.002 [0.997, 1.009]×. Intervals are 98.75% paired-source bootstrap intervals for the four total-latency comparisons. Both methods deploy the same scorer architecture and native gathering. Intervals containing one do not establish equivalence. Appendix X gives qualification, workloads, repeat handling and memory measurements.

Cohort and estimand. The existing timing sample (seed 2026091904) contains 32 source documents, one image/question per source, selected without timing or correctness. Accuracy results predate this timing protocol. This source-balanced sample does not estimate question-weighted latency on the full validation cohort. Repeats are technical measurements, not independent sources.

Native pipeline and qualification. The pinned Qwen3.5-9B BF16 model uses PyTorch 2.5.1+cu124 and Transformers 5.3.0. Fast GDN is unavailable: results use the native unfused fallback. Input cap 1,024, batch one, greedy decisions and final SC-VTP/PG scorers remain fixed; all trained scorers remain resident in every arm. Native cached generation uses ordered physical gathering and retained logical positions. Every call recomputes the full encoding.

Two fixed sources check three arms in greedy, one-token and 32-token modes: all 18 native/manual output checks pass (36 generations); six greedy outputs also match archived complete-validation token IDs. Every measured source matches its archived first generated token, prompt digest and token count. Repeated outputs within each cell are identical.

An empty EOS-ID list and disabled forced EOS ensure exactly one/32 outputs, even after an ordinary answer would end. These are standardized workloads, not new answer-quality scores. Qualification failures preceding measurement—a GPU UUID representation mismatch and model-default EOS restoration when set to `None`—and their versioned corrections remain archived.

Timing and resource conditions. Per source, six method/workload blocks are randomly ordered (seed $2026092404 + i$). Each has one excluded warm-up and three measured trials. CPU wall clocks synchronize CUDA at boundaries. Total time includes resident-image decoding/preparation, transfer, encoder, selection, native generation and detokenization; it excludes file reads, model loading, audit probes and result writes. Model time starts after synchronized transfer. Each table entry averages the 32 within-source medians.

The GPU is checked empty before loading; its CUDA UUID matches NVML. A single non-MPS, non-MIG compute context is verified after loading, after qualification and before/after each source. Exclusive scheduling and boundary checks do not certify every instant. GPU1 jobs continue, host CPU/storage are shared and clocks are not locked. Clocks, power, temperature and memory are recorded outside timed calls. The worker runs no training or compilation; results do not establish production throughput or training-cost ordering.

Paired uncertainty and workload interpretation. We geometrically average the 32 baseline/SC-VTP ratios of median times. Paired-source bootstrap uses 20,000 draws (seed 2026092405), with 98.75% intervals for the four total-time contrasts and descriptive 95% model-time intervals. All sources, conditions and repeats remain included; neither a ratio of table means nor 576 nominal trials replaces the paired-source estimate.

Mean total-time savings are 64.4 ms for both output lengths. Subtracting one-token from 32-token mean time and dividing by 31 gives 45.60 ms per additional output token for both full input and

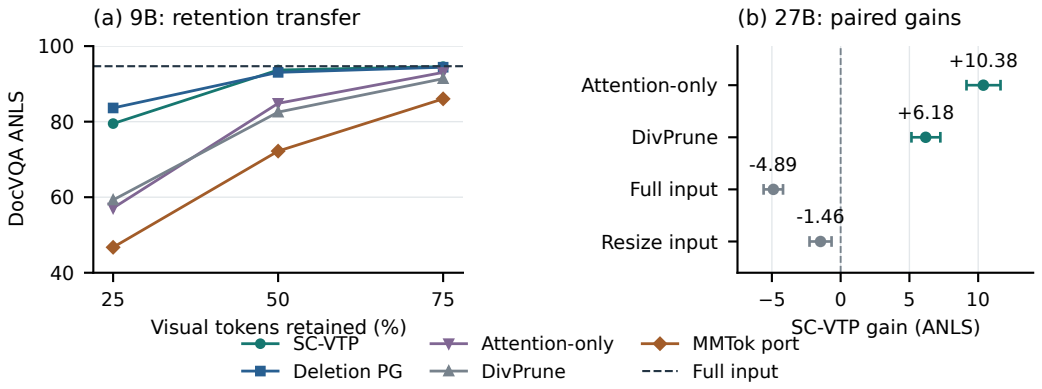

Figure 13: Complete 5,349-question DocVQA evaluations. (a) Qwen3.5-9B: fixed half-retention scorers at three budgets; PG is stronger at 25%. (b) Qwen3.8-27B: all four SC-VTP-minus-comparator contrasts, 95% source-bootstrap intervals. Attention is primary; the other contrasts are secondary and unadjusted. Resize has a nearby, unequal token budget and different encoder work. Both checkpoints belong to the same hybrid family.

SC-VTP. This post-hoc workload comparison is consistent with input-processing savings, but does not directly profile individual decoding steps.

Table 25: Means of per-source median allocator peaks. All scorers remain resident; these are PyTorch allocations, not total process or device memory. Incremental peak subtracts live allocations at each trial start.

Method	Output tokens	Peak allocated (GiB)	Incremental peak (GiB)
Full input	1	17.896	0.344
Full input	32	17.896	0.344
SC-VTP	1	17.788	0.237
SC-VTP	32	17.788	0.237
Deletion PG	1	17.788	0.237
Deletion PG	32	17.788	0.237

Y Complete frozen-scorer budget transfer

Question and fixed design. Does the coordinated training operator retain its advantage when a scorer trained at 50% must select a different number of tokens? Before the additional-budget calls, the protocol fixed the original state, attention and physical-deletion PG checkpoints, the full 5,349-question official DocVQA validation cohort and the 25%/75% endpoints. All 1,285 RGB images and 938 source documents remain included, with the prior exposure described in Appendix L. This is transfer of fixed scorers, not budget-specific retraining or a new holdout. Each pruner uses the same per-question image, prompt, grid, native decoder and retained-row count. Original 50% results and full input come from complete archived runs; MMTok comes from its separately qualified author-core port.

All conditions. Table 26 reports every designated arm. DivPrune and MMTok are training-free ports. Resize caps change the vision encoder input and have unequal actual budgets, so they are separate controls. Full-resolution mean visual length is 987.28; pruned means are 246.69, 493.57 and 740.40 rows. All generated outputs, including caps, enter the scores. Shared-resource execution supplies no multi-budget timing inference.

Table 26: Complete official DocVQA budget panel. Every row contains the same 5,349 questions. EM uses the frozen normalization; capped means reaching the 128-token generation limit. Learned scorers are the original 50%-trained checkpoints. Resize controls have distinct encoder work and are not matched-budget pruning.

Method	Budget	ANLS	EM (%)	Mean rows	Capped
SC-VTP	25%	79.49	67.92	246.69	0
Attention-only	25%	57.13	41.20	246.69	1
Deletion PG	25%	83.62	75.10	246.69	1
DivPrune	25%	59.27	43.77	246.69	0
MMTok port	25%	46.74	39.86	246.69	3
SC-VTP	50%	93.67	87.79	493.57	0
Attention-only	50%	84.85	74.74	493.57	0
Deletion PG	50%	93.05	87.34	493.57	0
DivPrune	50%	82.52	71.00	493.57	0
MMTok port	50%	72.20	65.68	493.57	0
SC-VTP	75%	94.57	89.61	740.40	0
Attention-only	75%	93.03	86.61	740.40	0
Deletion PG	75%	94.41	89.42	740.40	0
DivPrune	75%	91.38	84.11	740.40	0
MMTok port	75%	86.04	80.35	740.40	0
Full input	100%	94.68	89.81	987.28	0
Resize	cap 256	85.41	72.69	244.93	0
Resize	cap 512	92.55	85.62	484.86	1
Resize	cap 768	93.91	88.37	737.80	0

Table 27: Paired source-document contrasts. The first four rows use 97.5% intervals in two separately frozen two-endpoint families: primary attention comparisons and supplementary PG comparisons. The last two rows are supplementary, descriptive 95% intervals, not new primary tests.

Contrast	Retention	ANLS difference	Interval
SC-VTP − attention	25%	+22.36	[20.62, 24.15]
SC-VTP − attention	75%	+1.53	[0.95, 2.13]
SC-VTP − PG	25%	−4.13	[−5.48, −2.81]
SC-VTP − PG	75%	+0.16	[−0.14, 0.46]
SC-VTP − MMTok	25%	+32.75	[31.04, 34.48]
SC-VTP − MMTok	75%	+8.53	[7.47, 9.64]

Uncertainty and criteria. We resample all 938 source groups, retaining every question within each draw; question-weighted ANLS is the estimand. Both original families use 20,000 replicates with seed 2026092306. The attention-baseline transfer criterion requires both adjusted lower bounds above zero and is met. The corresponding PG-dominance criterion is not met: PG is stronger at 25%, while the 75% interval crosses zero. The supplementary MMTok joins use 20,000 draws and seed 2026092403; their plan was fixed after MMTok completion while the comparator panel was incomplete. They do not redefine the original primary endpoints. A narrow full-input gap at 75% is not an equivalence test.

Retention-budget effects. A module's binary deletion contract applies to any retained support satisfying its assumptions. This algebraic property is distinct from whether a scorer learns the best ranking for a changed budget. The same fixed state scorer remains stronger than its matched attention-only counterpart across all tested budgets, yet loses to direct-deletion learning under stronger compression. These observations localize the supported contribution to the coordinated differentiable training path; they do not imply all-budget dominance or identify why PG's ranking transfers better at 25%.

Recovery and numerical replay. The ten-arm additional-budget execution completes 53,490 calls. After an output-pipe interruption, the recovery reuses 3,001 verified closed question records and completes the remaining 2,348 (23,480 new calls). The original interrupted status is retained. This is one completed panel, not two independent experiments. Joining these outputs with the closed half-retention references and MMTok produces 19 complete arms (101,631 score cells). Checks cover identities, pixels, prompts, grids and retained counts for every matched pruning arm. The portable supplement exports source counts and arm-wise score sums for all 938 sources; its replay reconstructs all 19 means, capped counts and six additional-budget intervals without a model or the original filesystem. This numerical replay does not repeat model inference or replace the native implementation qualifications.

Z COMPLETE LARGER-CHECKPOINT REPLICATION

Question and scope. This separately frozen replication tests whether coordinating state transitions improves a learned selector over attention-only training in a larger hybrid checkpoint. It does not test a different mixer family. The model is `Qwen/Qwen3.8-27B`, revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`, whose pinned native implementation is `Qwen3_5`. It has 64 decoder layers and width 5120. The image-only scorer has 655,617 parameters (5120–128–1). The backbone remains frozen in BF16.

Paired training and complete evaluation. State and attention use seed 2026091709, identical initial scorer weights, question order and optimizer schedule. Each completes the same 26,143-question training set and 3,268 updates; all 2,974 internal-development questions are evaluated. Final checkpoint hashes and the five-arm official-evaluation binding are fixed before official inference. The two source-disjoint execution shards cover 2,733 and 2,616 questions (469 source documents each). Together they contain 26,745 generations: all five arms on all 5,349 official questions, 1,285 decoded RGB images and 938 sources. The official validation cohort has the same prior repository exposure described in Section 4.

Every learned or DivPrune arm retains $\lfloor N/2 \rfloor$ original visual rows in order, with retained logical positions and a native cached decoder. Full input has a 1,024-visual-token cap; the lower-resolution control uses 512 and changes encoder work. Generation is greedy with a 128-token limit. One state output reaches this limit and remains included; all other arms have zero length-capped outputs.

Table 28: Every completed 27B condition, each containing all 5,349 questions. Counts are question-weighted. Resize is not matched-budget pruning. PG, MMTok and TextVQA are not evaluated on this checkpoint.

Method	ANLS	Exact match (%)	Mean visual rows	Capped
Full input	95.33	89.38	987.28	0
SC-VTP	90.43	81.70	493.57	1
Attention-only training	80.06	66.26	493.57	0
DivPrune	84.25	71.19	493.57	0
Lower-resolution input	91.90	83.36	484.86	0

Table 29: All paired 27B contrasts. We resample 938 source groups with all questions retained per draw, 10,000 times, seed 2026091901. Secondary intervals are unadjusted. The estimand is question-weighted ANLS conditional on this trained pair.

SC-VTP minus	ANLS difference	95% source interval	Role
Attention-only	+10.38	[9.15, 11.62]	Primary
DivPrune	+6.18	[5.15, 7.25]	Secondary
Full input	−4.89	[−5.60, −4.19]	Secondary
Lower-resolution input	−1.46	[−2.26, −0.66]	Secondary

Interpretation and evidence. The positive primary interval extends the attention-only comparison to a larger checkpoint, with DivPrune as a second practical comparator. Full input and lower resolution retain higher accuracy in this cohort. The two checkpoints belong to one hybrid family, so their comparison does not isolate parameter scaling. This single pair measures a checkpoint-specific effect; the three-seed analysis remains specific to 9B. Independent-architecture and further learned-method comparisons are separate extensions.

The supplement binds all closed shard journals, model revisions, final checkpoints, input hashes and exact supports. A compact 26,745-row export preserves every prediction, reference answer, source ID, decoded-image hash, termination status and token count. Portable code independently recomputes ANLS/exact match from strings and all four source-bootstrap intervals. This verifies

saved evidence, not a new native execution. Historical failed scheduler jobs remain preserved and are not reclassified by this completed replication.

AA INTEGRATED EXPERIMENTAL EVIDENCE

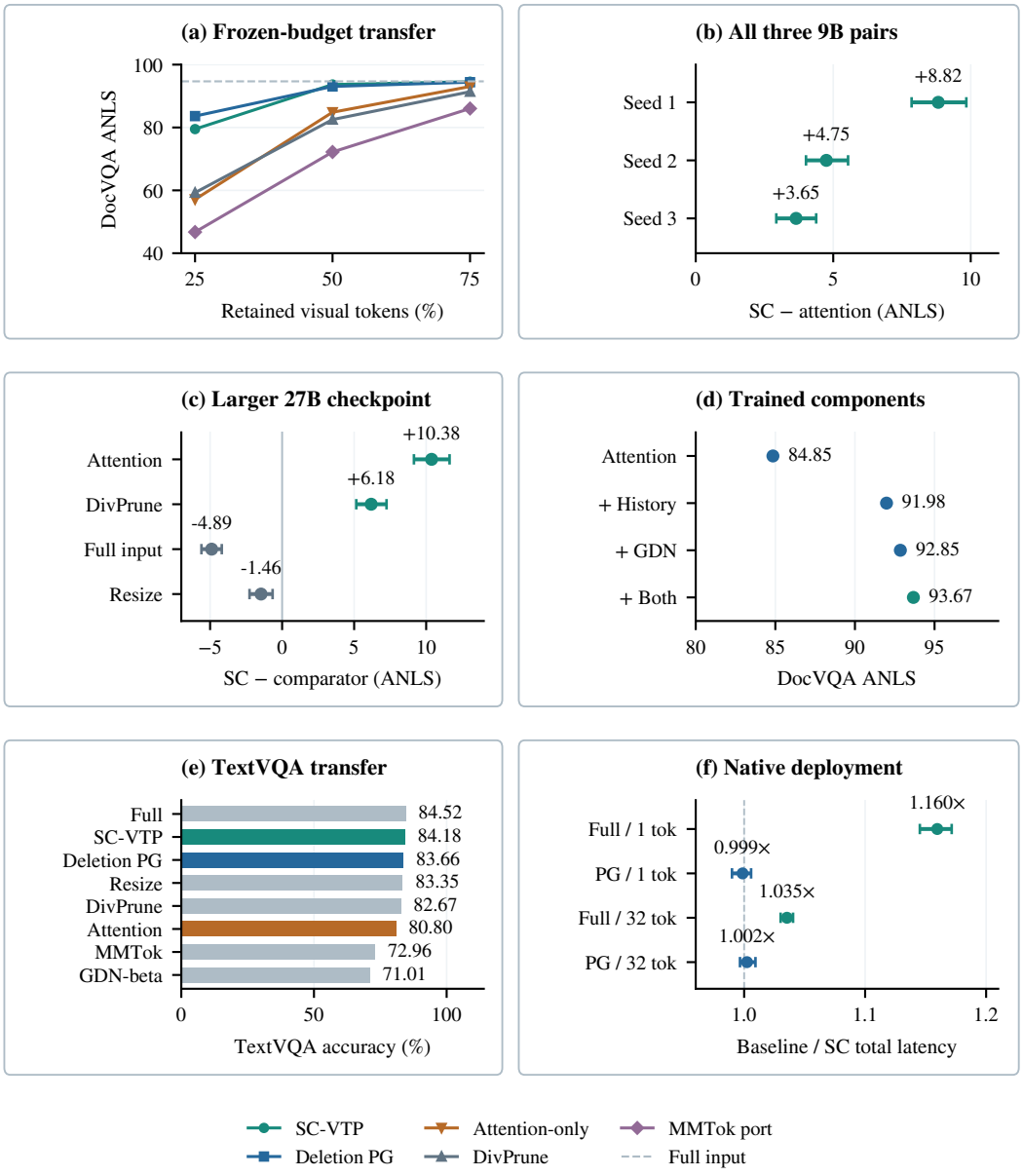

Figure 14: Complete evidence overview from archived evaluations. (a) All five deletion-based budget curves and full input; the scorer is fixed across budgets. (b) Every predeclared 9B training pair with the common supplementary 95% source intervals. (c) All four 27B paired comparisons, 95% source intervals; attention is primary and the others descriptive. (d) Every trained history/memory component arm. (e) All eight complete TextVQA methods; preprocessing and token budgets follow Table 2. (f) Every prespecified total-latency contrast, geometric baseline/SC ratios with 98.75% intervals, 32 sources and three repeats. These panels retain the comparison-specific conditions and do not pool their uncertainty.