TRACEGAIN: MARGINAL ANSWER UTILITY FOR DOCUMENT VLM REVEAL-OR-STOP CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

At a visited document state, stopping and acquiring another high-resolution region are mutually exclusive actions, yet region saliency and answer confidence score them in incompatible units. TraceGain turns this choice into a measurable control problem by assigning every legal zoom and shift its realized one-step change in official answer score and placing stop at the exact zero of that scale. Exhaustive counterfactual expansion of student-visited states supplies the targets; a centered dense teacher shapes only the offline regression target, and a 1.35M-parameter head batches all legal actions at inference. On DocVQA, InfographicVQA, and ChartQA, raw predictions attain Spearman $\rho = 0.41$–0.47, Utility ECE $= 0.051$–0.062, and one-step stop AUPRC $= 0.78$–0.84. Against DirectPolicy matched for inputs, capacity, teacher privilege, roll-ins, and budget, fixed-cache point estimates show 57% more stop errors and 65% more regret for direct classification. Matched endpoint gains span 0.3–1.1 points with LLaVA and 0.3–1.0 with Qwen2.5-VL; in the DocVQA operating study, TraceGain reaches 81.7 ANLS at 968 reported pre-encoding tokens and 438 ms, compared with 82.4 at 2,016 tokens and 847 ms for Full-Res. A common marginal-utility coordinate thus connects action selection, stopping, calibration, and compute allocation for short-horizon document inference.

1 INTRODUCTION

After a global page view, a document VLM reaches a concrete decision point. It may commit to its current answer, zoom into any still-legal cell, or, after a fine crop, shift the current window; each reveal consumes budget and changes the next answer, history, and action set. Choosing a reveal whose score gain is below that of another legal action incurs one-step regret, while choosing stop when a beneficial reveal exists ends the trajectory prematurely. These errors matter because decisive evidence—a receipt total, legend entry, axis label, or form field—can occupy only a small part of a page, whereas dense encoding pays for the entire page.

Common control signals leave this decision unresolved. Crop and saliency systems rank locations, while confidence and entropy gates evaluate the answer already available. Their scores are not commensurate: a high-saliency region need not change the answer, and an uncertain answer does not identify which acquisition would improve it. The required quantity is action-specific and state-dependent—the answer consequence of each currently legal reveal relative to committing now.

TraceGain (student-trajectory reveal gain) learns this decision coordinate on the states induced by deployment. Every legal one-step zoom and shift is executed counterfactually with the frozen student and labeled by its change in official benchmark score; stop is fixed at zero. Centered dense-teacher shaping reduces target variance offline without changing the realized-utility calibration target. At inference, a compact controller scores the full legal action set in one batch from query, global-image, OCR-density, acquisition-mask, answer-margin, and history features, applies an overlap price, and greedily selects a reveal or the zero-valued stop. Appendix Figure 1 summarizes the resulting student-trajectory loop.

The measurements align the learned target with deployed behavior. On held-out states from DocVQA, InfographicVQA, and ChartQA, TraceGain reaches $\rho = 0.41$–0.47, ECE $= 0.051$–0.062, and one-step stop AUPRC $= 0.78$–0.84. DirectPolicy controls for inputs, capacity, teacher privilege, roll-ins, and serving constraints while comparing utility regression with direct classification; on a fixed state cache, direct classification has 57% higher stop-error and 65% higher regret point estimates. Paired

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.

endpoint gains span four benchmarks and two frozen 7B VLM lineages. The DocVQA operating study further places 81.7 ANLS at 968 reported pre-encoding tokens and 438 ms, close to the 82.4 Full-Res score at 2,016 tokens and 847 ms.

This paper makes three contributions:

1. A state-dependent answer-utility coordinate that assigns realized official-score gain to every currently legal reveal and the exact value zero to stop, making action ranking, calibration, stopping error, and one-step regret directly comparable.
2. TraceGain, a compact batched controller trained by exhaustive expansion of epoch-refreshed student trajectories. Its centered teacher correction is confined to offline target construction; deployment uses the frozen VLM, frozen OCR-density extractor, and controller.
3. A controlled evaluation that compares supervision targets, measures on-policy reveal-or-stop behavior, links value quality to endpoint accuracy and latency, analyzes reveal-value concentration, and replicates the pattern with a second frozen VLM lineage.

2 What Must a Reveal Controller Predict?

High-resolution document encoders preserve text and layout by tiling, adapting aspect ratio, or raising input resolution (Xu et al., 2020; 2021; Kim et al., 2022; Lee et al., 2023; Li et al., 2024; Liu et al., 2024). These designs improve the visual representation but do not by themselves price an additional observation for the current answer. Token pruning and merging act later, after visual encoding, by deciding which representations to retain (Rao et al., 2021; Liang et al., 2022; Bolya et al., 2023).

Adaptive acquisition systems are distinguished by the object their controller predicts. ViCrop localizes one informative crop; CropVLM learns a free-form zoom policy; AdaptVision adapts visual acquisition; and ZoomEye uses tree-structured exploration (Zhang et al., 2025; Carvalho et al., 2026; Lin et al., 2026; Shen et al., 2025). GapSight uses target-model loss gaps at the global state to learn a review gate, expected utility, and a free-form box (Zhu et al., 2026). TraceGain resolves a complementary sequential control object: after every student action, it values the complete currently legal discrete reveal set in official-score units and embeds termination at the zero of that same set. Re-estimating this object on student-visited states makes action calibration, stopping error, and regret well-defined beyond the first crop.

Answer calibration and selective prediction assess the risk of the response already in hand (Guo et al., 2017; Lakshminarayanan et al., 2017; Minderer et al., 2021; Xiong et al., 2024; Wieczorek et al., 2026); value-of-information methods ask what a prospective observation changes (Howard, 1966; Raiffa & Schlaifer, 1961; Foster et al., 2021). TraceGain instantiates value of information at a frozen document VLM's pre-encoding boundary. Its primary learned comparison, DirectPolicy, shares the state, legal action set, capacity, roll-ins, teacher access, optimizer, and serving constraints while replacing utility regression, calibration, and zero-referenced scoring with listwise action classification. This matched-capacity construction tests the paper's decision-coordinate hypothesis without conflating it with the non-isomorphic action geometries of published free-form crop systems.

Appendix Table 12 locates the contribution at the point where a controller must compare mutually exclusive actions. Crop selection ranks locations, answer calibration scores current-answer risk, and token pruning scores encoded content; TraceGain estimates the answer gain of each legal acquisition relative to a zero-utility stop on states induced by the deployed student.

3 Answer-Consequential Reveal-or-Stop Control

3.1 State and actions

Let frozen VLM f_θ answer query q for image I. A mandatory 336×336 global pass defines an $H_g \times W_g = 6 \times 6$ page-coordinate grid with $N = 36$ cells. At step t, the controller observes

$$s_t = (e_q, m_t, \mathcal{R}_t, o_t, h_t),$$

where e_q is the query-conditioned global representation, m_t is the answer margin, $\mathcal{R}_t$ is the acquired-region mask, o_t summarizes cached OCR density, and h_t records crop history and the most recent fine window.

The action-code set contains $\texttt{zoom}(c, s)$ for cell c and scale $s \in \{2\times, 4\times\}$; $\texttt{shift}(c, d)$, which translates the most recent fine window by one cell in direction $d \in \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$ at fixed scale; and $\texttt{stop}$, which commits to the current answer. Cell centers are mapped affinely from the 6×6 grid to page coordinates. The fixed-scale zoom constructor centers its prescribed 672- or 1344-pixel input on that point, and a shift reuses the previous extent at the translated center. The legality mask then removes any reveal whose prescribed extent crosses the page boundary, is fully covered, or would exceed the remaining unique-acquired-token allowance. Thus the base state exposes $2N + 1 = 73$ action codes before legality masking; later states expose the remaining zoom codes, at most four shift codes, and stop. Every reported action denominator counts only the post-mask legal set, and stop is always legal.

3.2 Utility and inference

For a non-stop action, define the population and realized utilities as

$$U(a \mid s) = \mathbb{E}[\ell(s) - \ell(s \oplus a) \mid s, a], \tag{1}$$

$$u^*(s, a) = \ell(s) - \ell(s \oplus a). \tag{2}$$

Here $\ell(s) = 1 - \mathrm{Score}(\hat{y}(s), y)$ after official benchmark normalization: ANLS for DocVQA and InfographicVQA, relaxed accuracy for ChartQA, and VQA consensus accuracy for TextVQA. Decoding is deterministic; the expectation in U is over the document–question and student-visited state population, and $\hat{u}_\phi$ estimates its conditional mean. The instance-level u^* measures realized answer improvement rather than token likelihood. For a reveal, $r(s, a) \in [0, 1]$ is the fraction of its fine-resolution window already represented by the acquisition mask; a fully covered reveal is illegal. We define $u^*(s, \texttt{stop}) = U(\texttt{stop} \mid s) = 0$ and $r(s, \texttt{stop}) = 0$. Every legal action consequently has official-score units, with stop serving as the exact origin of the one-step decision coordinate.

The utility head emits raw $\hat{u}_\phi(s, a)$ for every action code. This is the estimate compared with u^* for Spearman correlation and Utility ECE. Deployment applies a deterministic redundancy price to produce

$$\hat{U}(a \mid s) = \begin{cases} \hat{u}_\phi(s, a) - \eta r(s, a), & a \neq \texttt{stop}, \\ 0, & a = \texttt{stop}. \end{cases}$$

Here η is selected before test evaluation and expresses overlap in the same normalized score range as answer gain. The adjustment can only lower a reveal score, making repeated coverage conservative while preserving stop at the deployed origin. The controller greedily selects the highest legal $\hat{U}$, regenerates the answer after an executed reveal, and terminates when stop wins, the unique-acquired-token allowance is exhausted, or the horizon reaches $T = 4$. This is greedy receding-horizon control: utilities, stop labels, and R_1 evaluate the next decision, while endpoint score evaluates the complete rollout. We keep four quantities distinct: u^* is realized student gain, $\tilde{u}$ is the shaped training target, $\hat{u}$ is the raw calibrated prediction, and $\hat{U}$ is the overlap-adjusted deployment score.

4 TraceGain Controller

4.1 Exhaustive action scoring

TraceGain places a three-layer, 512-hidden-unit MLP above a frozen VLM. For each legal action, the input concatenates the query-conditioned global feature, referenced coarse-cell feature, answer margin, acquisition mask, OCR-density summary, history features, and an action code: 36-way cell, two-way scale, five-way direction including $\texttt{none}$, and a stop bit. The global and cell features are pooled from the frozen backbone representation of the mandatory global view; no VLM high-resolution feature is available before acquisition. For all 336-, 672-, and 1344-pixel inputs, the frozen backbone's native resize-and-padding processor defines the accounting unit: each emitted pre-encoding visual-token position counts once. The answer margin is the difference between the two highest length-normalized sequence scores returned under the fixed answer prompt. A frozen OCR preprocessor runs once on the full page and emits only a normalized 36-cell text-density vector to the controller; recognized strings are discarded. The history block records depth, previous action type, scale or direction, and last-window geometry, while the acquisition mask records covered cells.

A $2\times$ or $4\times$ zoom extracts the corresponding 672- or 1344-pixel input centered on the selected cell; a shift translates the last fine window by one coarse-grid cell at its current scale. We separate two resource quantities. The legality charge is the number of newly covered fine-resolution visual-token positions added by a reveal; the mandatory global pass initializes this union, and cumulative unique positions may not exceed B. Reported processed tokens include the global pass and every reveal token actually passed through the VLM, including re-encoded overlap; latency measures the complete serving path. Exact score ties favor stop, then the lower-charge reveal, then the fixed action-code order. These rules are shared by both learned controllers. The resulting 1.35M-parameter head scores the entire legal set in one batched call.

4.2 Student-visited counterfactual targets

At each visited state, we execute every legal one-step reveal with the frozen student and compute $u^*(s, a)$. The dense teacher is a second frozen copy of the corresponding backbone checkpoint with offline access to the original full-resolution page. It receives a composite raster: cells outside $\mathcal{R}_t$ retain the mandatory global-view resolution, while acquired windows are restored from the full-resolution page. The before call renders $\mathcal{R}_t$; the after call additionally restores the candidate window for $s \oplus a$. Query, ordered history, answer prompt, and deterministic decoder are otherwise identical. This dense composite, rather than the student's sparse multiview encoding, is the teacher privilege. It supplies the centered shaping term

$$\tilde{u}(s, a) = u^*(s, a) + b_{\theta^*}(s, a) - \bar{b}_{\theta^*}(s),$$

where $b_{\theta^*}(s, a) = \ell_{\theta^*}(s) - \ell_{\theta^*}(s \oplus a)$ is the dense teacher's official-score loss reduction under the same action geometry and $\bar{b}_{\theta^*}(s)$ is its mean over the state's legal non-stop actions. Both models are frozen, the centered coefficient is one, and the mean is recomputed for every state. Subtracting the statewise teacher mean removes between-state offsets while preserving the mean target within each state; empirically, pooled target variance falls from .084 to .057. We use $\tilde{u}$ only in the Huber term: the calibration loss, all utility measurements, concentration, one-step stop labels, and regret remain defined by realized student utility u^*. Stop tuples with target zero are inserted at every visited state. The teacher neither selects actions nor participates in inference.

Tuple collection begins with exhaustive base-state counterfactuals. At each epoch boundary, the previous checkpoint is frozen, completes all greedy roll-ins to horizon T, and expands every reached state over its legal actions. Duplicate tuples are keyed by a unique document–question ID, action, and a serialization of the complete observed s_t, including its global/cell representation, OCR density, answer margin, acquisition mask, ordered history, depth, previous action, and last-window geometry; the first occurrence within that refresh is retained. Only after the refreshed cache is complete do controller updates begin. Replacing the cache each epoch aligns training states with greedy deployment, while validation and test caches are fixed after the final checkpoint so held-out tuples never affect fitting or model selection.

4.3 Objective and calibration

For non-stop tuples $\mathcal{T}_r$ and stop tuples $\mathcal{T}_0$, training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{Huber}}(\hat{u}, \tilde{u}) + 0.25\,\mathcal{L}_{\text{softcal}} + 0.10\,\mathcal{L}_0.$$

The Huber transition is 0.05. $\mathcal{L}_0 = |\mathcal{T}_0|^{-1} \sum_{(s,\texttt{stop})} \hat{u}_\phi(s, \texttt{stop})^2$ anchors the raw stop channel even though deployment clamps it exactly. The differentiable calibration term uses 15 equally spaced centers on $[-1, 1]$, triangular assignment with bandwidth $2/15$, and the weighted absolute gap between soft-bin mean $\hat{u}$ and mean u^*; empty bins contribute zero. Evaluation-time Utility ECE clips predictions to the official-score-difference range $[-1, 1]$, compares $\hat{u}$ with u^* in 15 hard equal-width bins, and uses document/question-clustered bootstrap uncertainty. Stop AUPRC and rollout choices use the deployed $\hat{U}$. For score-only comparators, a monotone isotonic map to realized utility is fit on the calibration split and frozen; ECE uses the mapped score, whereas ranking and deployment use each method's original score.

4.4 Fairness-matched learned controller

DirectPolicy is the equally privileged learned comparator used throughout. It receives the identical state tuple, OCR access, action code, legality mask, 1.35M-parameter MLP budget, teacher-corrected

labels, epoch-refreshed student roll-ins, 12-epoch AdamW schedule, and hard acquisition budgets. Its target is the best legal action under $\tilde{u}$, trained by masked listwise cross-entropy; its logits are unconstrained policy scores, with stop represented as an ordinary class. At inference it greedily chooses the highest legal logit. The comparison holds observations, capacity, teacher privilege, roll-ins, optimization schedule, and acquisition limits fixed while jointly changing target parameterization, loss, calibration, and decision score.

5 EXPERIMENTS

5.1 PROTOCOL

We evaluate frozen LLaVA-1.6-Mistral-7B (Liu et al., 2024) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with their released native image processors. Both use the same external 6×6 acquisition interface, 672×672 and 1344×1344 reveal inputs, $T = 4$, fixed answer prompts, deterministic decoding, and official answer post-processing. Controller fitting uses official training questions from DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), ChartQA (Masry et al., 2022), and TextVQA (Singh et al., 2019): 39,463, 23,946, 20,882, and 34,602 questions, respectively. Each official validation set is split by document/image into equal model-selection and calibration halves (2,674/2,675; 1,400/1,401; 528/528; and 2,500/2,500). Endpoint and exhaustive legal-action scoring use untouched official test examples (5,188, 3,288, 2,500, and 5,734). Answer-level test annotations are accessed under benchmark-maintainer authorization solely for final scoring; held-out tuples and endpoint examples never enter fitting, calibration, budget selection, or hyperparameter selection.

Results are means over five paired controller seeds. Endpoint cells marked $\pm$ report seed standard deviation; paired gaps marked $\pm$ report the half-width of a seed-paired 95% interval. Bracketed mechanism intervals use 10,000 document/question-clustered bootstrap resamples. All experiments use 8×A100 80GB GPUs, CUDA 12.4, PyTorch 2.1, FlashAttention-2, and batch-1 latency on one A100. Counterfactual generation takes 18 GPU-hours per dataset–backbone pair and controller fitting takes 6 GPU-hours. Online latency includes OCR, crop extraction, action scoring, re-encoding, and answer decoding.

Full-Res encodes each page with the backbone's native dense-page preprocessing (2,016 emitted visual tokens for the reported LLaVA point), and global-only uses only the mandatory thumbnail. Uniform-Tile scans legal cell centers in top-to-bottom raster order, taking $2\times$ before $4\times$ at each center and then legal shifts in fixed direction-code order. Saliency-Crop ranks a window by mean query-to-image attention over its covered global cells. OCR-First ranks by the same aggregation of cached cell density. Entropy-Refine averages normalized entropy over generated answer tokens to decide whether to continue and uses the Saliency-Crop score to localize. Every pre-encoding baseline receives the same legal windows, unique-acquired-token caps, horizon, deterministic decoder, and model-selection split; thresholded baselines tune thresholds only on that split. Published free-form crop systems use different action geometries and stopping interfaces, so DirectPolicy is the contract-matched learned controller, while Table 12 supplies the method-level contrast. ToMe and DynamicViT operate after encoding and are reported on their native retained-token axis in Appendix Table 13, not in the shared pre-encoding frontier.

Table 1 first places the learned acquisition policy against its dense and light operating references; the following sections then evaluate value prediction, stopping, matched objectives, backbone replication, and serving cost.

Table 1: LLaVA dense-reference context under one serving harness. Tokens are reported pre-encoding visual tokens; latency is batch-1 milliseconds. Table 7 isolates the matched pre-encoding comparison.

Method	Tokens	Latency	Doc	Info	Chart	Text
Full-Res	2016	847	82.4	51.3	73.8	68.1
TraceGain-Light	684	341	80.9	49.1	71.8	66.2
TraceGain	968	438	81.7	50.6	73.1	67.4

Table 2: Held-out non-stop legal reveals on the common TraceGain-visited distribution. Brackets give question-clustered 95% intervals.

Method	Doc	Info	Chart
	Spearman $\rho \uparrow$		
Entropy-Refine	.18	.14	.21
Saliency-Crop	.23	.19	.17
TraceGain w/o calibration	.37	.33	.39
TraceGain	.44 [.41,.47]	.41 [.38,.44]	.47 [.44,.50]
	Utility ECE$\downarrow$		
Entropy-Refine	.142	.168	.131
Saliency-Crop	.119	.137	.148
TraceGain w/o calibration	.091	.103	.087
TraceGain	.058 [.052,.064]	.062 [.056,.068]	.051 [.046,.057]

TraceGain finishes 0.7 score points below Full-Res on each benchmark; its reported operating point uses 48% of Full-Res tokens and 52% of its latency. Subsequent sections report mechanism diagnostics and matched-interface endpoint results separately.

5.2 Can visited-state reveal gains be recovered?

The evaluation spans all held-out visited states and legal reveals reported in Appendix Table 9. TraceGain improves ranking over the strongest proxy by 0.21, 0.22, and 0.26 on DocVQA, InfographicVQA, and ChartQA and approximately halves calibration error. Removing the calibration objective preserves much of the ordering but raises raw-utility ECE by 57–71%. The result establishes that the supervised object is measurable on post-action states, rather than only at the initial page view.

5.3 How do the deployed policies stop?

The one-step oracle-stop event is $\max_{a \neq \texttt{stop}} u^*(s,a) \leq 0$. We rank this event by $-\max_{a \neq \texttt{stop}} \hat{U}(a \mid s)$ for TraceGain and by DirectPolicy's masked stop logit minus its best reveal logit. Premature-stop rate is $\Pr[\hat{a} = \texttt{stop} \mid \max_{a \neq \texttt{stop}} u^*(s,a) > 0]$, whose denominator is Stop$-$ in Appendix Table 9; unnecessary-reveal rate is $\Pr[\hat{a} \neq \texttt{stop} \mid \max_{a \neq \texttt{stop}} u^*(s,a) \leq 0]$, whose denominator is Stop+. StopErr is the union of these errors divided by all evaluated states. The diagnostic regret,

$$R_1(s) = \max_a u^*(s,a) - u^*(s,\hat{a}),$$

measures the decision made at the current state; endpoint score evaluates the full four-step rollout, including shifts enabled by earlier reveals. Table 3 reports each final controller on states generated by its own held-out rollout and therefore measures deployed on-policy behavior. Appendix Table 9 gives TraceGain one-step stop and continuation denominators, while Table 4 fixes the state distribution for the action-value comparison.

Table 3: Held-out on-policy stop point estimates, averaged over five controller seeds. Cells report TraceGain/DirectPolicy on their respective deployed state distributions; Prem. and Unnec. are percentages.

Data	AUPRC$\uparrow$	Prem.$\downarrow$	Unnec.$\downarrow$	Regret$\downarrow$	Depth
Doc	.82/.74	6.8/10.9	9.5/14.6	.017/.025	1.7/2.0
Info	.78/.69	8.1/12.6	11.2/16.8	.021/.030	1.9/2.2
Chart	.84/.76	6.2/9.8	8.7/13.1	.014/.022	1.6/1.9
Text	.61/.58	14.8/16.3	20.9/22.6	.038/.041	2.6/2.7

On their respective deployed state distributions, TraceGain/DirectPolicy premature-stop rates are 6.8/10.9, 8.1/12.6, and 6.2/9.8 on DocVQA, InfographicVQA, and ChartQA; the corresponding

Table 4: Five-seed matched-controller reruns, distinct from the operating study in Table 1. Panel A gives official endpoint score (mean ± seed s.d.); paired gaps are mean ± 95% interval half-width. Panel B evaluates both final controllers on identical held-out TraceGain-visited traces; DirectPolicy ECE uses the frozen isotonic utility map.

Dataset	TraceGain	DirectPolicy	Paired gap
DocVQA	81.6 ± .2	80.8 ± .4	+.8 ± .3
InfographicVQA	50.4 ± .3	49.9 ± .5	+.5 ± .4
ChartQA	72.9 ± .4	71.8 ± .5	+1.1 ± .4
TextVQA	67.2 ± .2	66.9 ± .3	+.3 ± .2

Dataset	ρ (TraceGain/DirectPolicy)	ECE (TraceGain/DirectPolicy)
DocVQA	.44/.39	.058/.071
InfographicVQA	.41/.35	.062/.079
ChartQA	.47/.42	.051/.066
TextVQA	.24/.22	.114/.121

Table 5: Independent DocVQA target-isolation point estimates on a fixed cache and seed list, with architecture, visited states, unique-token cap, and privilege held constant. StopErr is the union of premature stops and unnecessary reveals divided by all cache states; score-only ECE uses calibration-split utility mapping.

Training target	ANLS↑	ρ ↑	ECE↓	StopErr↓	Regret↓
Marginal reveal utility	81.7	.44	.058	8.1	.017
Direct policy class	80.8	.31	.096	12.7	.028
Pairwise rank only	80.9	.42	.121	11.9	.022
Answer-entropy change	79.8	.24	.134	15.1	.035
OCR density/saliency	79.5	.23	.119	14.4	.033

one-step regrets are .017/.025, .021/.030, and .014/.022. These on-policy point estimates summarize complete controller behavior, while the fixed-cache comparison below supplies the common-state objective contrast. On TextVQA, both controllers roll deeper and their rates are closer.

5.4 Matched-capacity utility regression versus classification

The matched reruns in Table 4 share seeds, budgets, side information, and serving harness; Panel B additionally places both controllers on identical TraceGain-visited states. Positive endpoint gaps accompany higher common-trace rank correlation and lower utility-mapped ECE. TraceGain contributes raw utility predictions, whereas DirectPolicy uses its calibration-split isotonic map. Together, the panels connect the supervised coordinate to complete-rollout behavior while keeping fixed-trace and on-policy evidence distinct.

This fixed-cache objective sweep complements the final-controller comparison in Table 4. Rank-only supervision approaches TraceGain's ordering (.42 versus .44) but has ECE .121 versus .058 and StopErr 11.9 versus 8.1. Direct classification performs competitively at the endpoint; its point estimates have 57% more stop errors and 65% more selected-action regret. With states, capacity, and privilege shared, this pattern is consistent with an advantage for the utility-regression objective and its zero-referenced decision rule in the matched setting.

Reveal-value geometry. For analysis, Appendix B defines concentration as the largest positive reveal gain divided by total positive gain. On the LLaVA TraceGain-visited cache, per-state ranking quality is the Spearman correlation across legal non-stop actions between raw $\hat{u}$ and u^*. Across 43,297 positive-utility states, concentration correlates with this quantity at $r_s = .46$ [.42,.50]; mean TraceGain rollout depths for Q1–Q4 are 2.5/2.2/2.3/1.6. This retrospective statistic is never supplied to either controller.

5.5 DOES THE ENDPOINT PATTERN REPLICATE ACROSS VLM LINEAGES?

Table 6: Qwen2.5-VL-7B cross-backbone endpoint replication after controller retraining under the same external interface. Score, ρ, ECE, and regret report TraceGain/DirectPolicy; Gap is the paired score difference (mean $\pm$ 95% interval half-width), while mechanism cells are descriptive point estimates.

Dataset	Score	Gap	ρ	ECE	Regret
DocVQA	88.8/88.1	+.7 $\pm$ .3	.48/.41	.052/.067	.015/.022
InfoVQA	60.8/60.2	+.6 $\pm$ .4	.43/.37	.060/.074	.020/.028
ChartQA	80.6/79.6	+1.0 $\pm$ .4	.50/.44	.049/.062	.013/.019
TextVQA	76.9/76.6	+.3 $\pm$ .3	.27/.25	.108/.115	.035/.038

Table 6 couples endpoint replication with descriptive mechanism diagnostics. On DocVQA, InfographicVQA, and ChartQA, TraceGain retains 0.6–1.0 point paired endpoint gains; the corresponding point estimates show 0.06–0.07 higher rank correlation, 0.013–0.015 lower ECE, and 0.006–0.008 lower regret. TextVQA produces smaller differences. Retraining preserves the endpoint pattern beyond LLaVA most clearly on the other three benchmarks.

5.6 DO VALUE ESTIMATES IMPROVE THE SERVING FRONTIER?

Table 7: DocVQA matched pre-encoding frontier from dedicated five-seed reruns. Each entry is ANLS/latency-ms at one of three displayed anchor caps under a common serving harness.

Method	504 tokens	684 tokens	968 tokens
Uniform-Tile	76.1/286	77.5/315	78.4/398
OCR-First	78.1/304	79.0/329	79.8/411
DirectPolicy	79.1/312	80.2/336	80.8/429
TraceGain	79.8/318	80.9/341	81.6/438

Normalized frontier area is computed per seed over the predeclared grid of all feasible unique-token charges from 504 through 968, $\mathcal{B} = \{B_0, \ldots, B_K\}$, before seed averaging:

$$\mathrm{AUC} = \frac{\sum_{j=0}^{K-1} (B_{j+1} - B_j)\,[S(B_j) + S(B_{j+1})]/2}{100\,(B_K - B_0)}.$$

The three columns display anchor caps from that sweep. TraceGain yields the largest area, .813 versus .805, .792, and .773. Its 6–9 ms controller overhead is visible, while stopping limits unnecessary high-resolution encoding.

Table 8: Independent DocVQA component reruns at the common 968 unique-acquired-token cap. Endpoint entries are five-seed mean $\pm$ s.d.; Tokens are mean processed visual tokens, including the mandatory global pass and re-encoded overlap.

Variant	ANLS	Δ	Tokens	ρ	ECE
TraceGain	81.7 $\pm$.2	—	968	.44	.058
− Utility head (random crops)	77.3 $\pm$.5	−4.4 [−4.9,−3.9]	972	—	—
− Centered teacher correction	80.4 $\pm$.3	−1.3 [−1.7,−0.9]	955	.35	.083
− Calibration loss	81.1 $\pm$.3	−0.6 [−0.9,−0.3]	961	.37	.091
− Redundancy scoring	80.8 $\pm$.4	−0.9 [−1.3,−0.5]	1042	.41	.064
− OCR features	80.2 $\pm$.3	−1.5 [−1.9,−1.1]	974	.39	.072
− Global context features	79.1 $\pm$.4	−2.6 [−3.1,−2.1]	981	.31	.098

The cap counts unique acquired positions, while Tokens counts every processed token; re-encoded overlap explains the 1,042-token row. This independent rerun yields 81.7 rather than the frontier

rerun's 81.6. The random-crop control samples uniformly from the legal reveal set until the common cap or horizon and loses 4.4 points at matched processing. The global-context ablation retains the mandatory global encoding and token charge but masks its pooled controller features. Centered shaping reduces pooled target variance from .084 to .057 and raises ρ from .35 to .44; removing calibration raises ECE from .058 to .091.

The fixed-cache target sweep is the most controlled mechanism comparison: when states, capacity, and teacher privilege are shared, calibrated utility regression has lower zero-crossing error and regret point estimates than classification and rank-only training. Own-rollout stop behavior, matched endpoints, DocVQA serving cost, and Qwen endpoint replication align with that result; TextVQA combines weaker utility ranking with deeper trajectories and smaller gains.

6 Discussion

The convergent mechanism and endpoint results support marginal answer utility as the control coordinate for the tested four-step grid interface. Its evidence assumes exhaustive one-step labels, a dense offline teacher, full-page OCR density, greedy replanning, and per-backbone controller training; counterfactual generation costs 18 GPU-hours per dataset–backbone pair, and overlap can make processed tokens exceed the unique-evidence cap. Longer or complementary reveal sequences induce different state distributions, while TextVQA exhibits weaker utility ranking and smaller controller separation. The legal-gain, zero-stop, and regret definitions remain directly applicable to those regimes.

7 Conclusion

TraceGain places localization and termination on one answer-consequence coordinate: marginal official-score change for each legal reveal and zero for stop. It improves calibrated ranking, matched zero-crossing decisions, and LLaVA rollout endpoints; retrained Qwen controllers reproduce the strongest endpoint pattern on DocVQA, InfographicVQA, and ChartQA. Marginal answer utility is therefore a measurable and effective primitive for short-horizon document acquisition.

AI Use Statement

AI tools assisted with manuscript wording, organization, presentation, formatting, consistency, and bibliography checks during preparation.

Reproducibility Statement

Sections 3–5 define the legal action contract, target construction, controller objectives, data roles, serving harness, metrics, and uncertainty protocol. Appendices A–C record the training algorithm, optimization settings, state-action counts, backbone evidence, and resource-accounting contract.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Miguel Carvalho, Hélder Dias, and Bruno Martins. CropVLM: Learning to zoom for fine-grained vision-language perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11170–11179, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/GRAIL-V/html/Carvalho_CropVLM_Learning_to_Zoom_for_Fine-Grained_Vision-Language_Perception_CVPRW_2026_paper.html.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074. URL https://doi.org/10.1109/TSSC.1966.300074.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, pp. 498–517, 2022. doi: 10.1007/978-3-031-19815-1_29.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. URL https://arxiv.org/abs/1612.01474.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/lee23g.html.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pp. 26763–26773, 2024. doi: 10.1109/CVPR52733.2024.02527.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *ICLR*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_.

Zichuan Lin, Yicheng Liu, Yang Yang, Lvfang Tao, and Deheng Ye. AdaptVision: Efficient vision-language models via adaptive visual acquisition. In *CVPR*, pp. 11923–11932, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lin_AdaptVision_Efficient_Vision-Language_Models_via_Adaptive_Visual_Acquisition_CVPR_2026_paper.html.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://doi.org/10.1109/WACV48630.2021.00225.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022. doi: 10.1109/WACV51458.2022.00264. URL https://doi.org/10.1109/WACV51458.2022.00264.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2106.07998.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2106.02034.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *EMNLP*, pp. 6602–6618, 2025. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8309–8318, 2019. doi: 10.1109/CVPR.2019.00851. URL https://doi.org/10.1109/CVPR.2019.00851.

Tobias Jan Wieczorek, Nathalie Daun, Mohammad Emtiyaz Khan, and Marcus Rohrbach. Variational visual question answering for uncertainty-aware selective prediction. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=jtnMIbJIso.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *ICLR*, 2024. URL https://arxiv.org/abs/2306.13063.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 2579–2591, 2021. doi: 10.18653/v1/2021.acl-long.201. URL https://aclanthology.org/2021.acl-long.201/.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1192–1200, 2020. doi: 10.1145/3394486.3403172.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *ICLR*, 2025. URL https://openreview.net/forum?id=DgaY5mDdmT.

Jinchang Zhu, Rong Fu, Yi Ding, Chenghao Wu, Ying Liu, and Menglin Yang. Learning to look again: Loss-gap supervision for free-form crop routing in vision-language models. *arXiv preprint arXiv:2608.21762*, 2026. URL https://arxiv.org/abs/2608.21762.

A Training and Evaluation Details

Algorithm 1 TraceGain training with student-visited targets

Require: Training data $\mathcal{D}$, frozen student f_θ, dense teacher f_{θ^*}, budget B, horizon T

for each epoch **do**
  Freeze the previous controller checkpoint as the roll-in policy
  Initialize refreshed tuple cache $\mathcal{C} \leftarrow \emptyset$
  for each batch $(q, I) \sim \mathcal{D}$ **do**
    Initialize s_0 from the mandatory 336×336 global view
    for $t = 0, \ldots, T-1$ **do**
      Enumerate legal reveals and append $(s_t, a, u^*(s_t, a), \tilde{u}(s_t, a))$ to $\mathcal{C}$
      Append $(s_t, \texttt{stop}, 0, 0)$ and deduplicate within the refresh
      Score legal reveals; clamp $\hat{U}(\texttt{stop} \mid s_t) = 0$
      Select $a_t = \arg\max_{a \in \mathcal{A}_{\text{legal}}(s_t)} \hat{U}(a \mid s_t)$
      if $a_t = \texttt{stop}$ **then**
        break
      end if
      Execute a_t and update s_{t+1}
    end for
  end for
  Replace the previous cache with completed $\mathcal{C}$
  for each tuple minibatch sampled from $\mathcal{C}$ **do**
    Update ϕ with Huber, soft-calibration, and stop-anchor losses
  end for
end for

Controllers use AdamW with learning rate 3×10^{-4}, weight decay 0.01, batch size 4096 tuples, cosine decay, 500 warmup updates, gradient norm 1.0, and 12 epochs. We select η from $\{0, 0.05, 0.10, 0.20\}$ and budgets from $\{504, 684, 968\}$ acquired tokens on the model-selection half, then freeze them before calibration and test evaluation. DirectPolicy uses the same optimizer, update count, cache refreshes, and seed list.

Table 9: Held-out state-action denominators. Actions exclude stop; Stop+ and Stop− are one-step oracle-stop and continuation states.

Dataset	Examples	States	Legal reveals	Stop+	Stop−
DocVQA	5,188	18,642	821,926	7,436	11,206
InfographicVQA	3,288	13,910	612,744	4,867	9,043
ChartQA	2,500	9,846	441,388	4,128	5,718
TextVQA	5,734	22,507	988,164	5,177	17,330

B Extended Concentration and Backbone Evidence

For positive non-stop realized utilities, define

$$C(s) = \frac{\max_a [u^*(s,a)]_+}{\sum_a [u^*(s,a)]_+}.$$

States with no positive utility form a separate one-step stop stratum; $C(s)$ is never supplied to either controller.

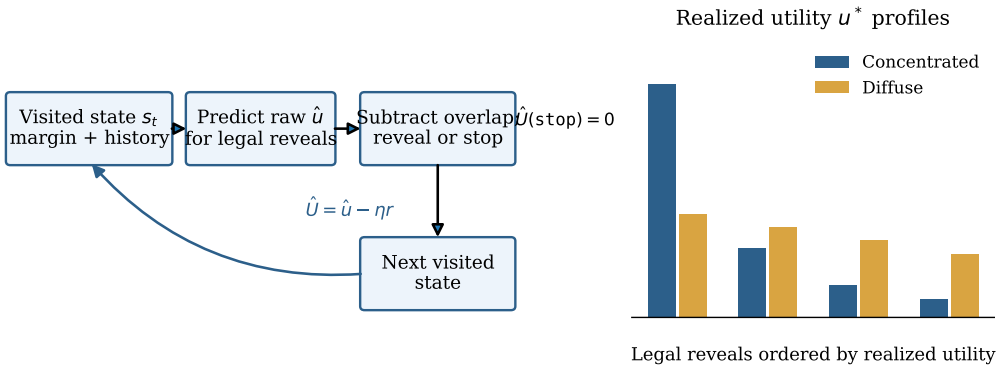


Figure 1: TraceGain's decision coordinate. The controller predicts raw marginal utility $\hat{u}$, applies the overlap price to obtain $\hat{U}$, and compares every legal reveal with zero-valued stop. Executed actions induce the next supervised state; the profiles illustrate the retrospective concentration statistic.

For each dataset independently, TraceGain-visited positive-utility states are sorted by $C(s)$ and divided into four equal-count strata; Table 10 then evaluates both controllers on this shared state cache and pools corresponding strata with equal dataset weight. This construction prevents a dataset's utility scale or state count from defining another dataset's quartile. The separate zero-positive-utility stratum contains 21,608 states, leaving 43,297 positive-utility states.

Table 10: Positive-utility concentration quartiles formed within each dataset before pooling. Each cell reports TraceGain/DirectPolicy on the shared TraceGain-visited state cache.

Stratum	$\rho \uparrow$	ECE↓	StopErr (%)↓	Regret↓
Q1, flatter	.25/.22	.104/.118	15.6/17.1	.038/.042
Q2	.34/.30	.082/.096	12.4/14.7	.029/.034
Q3	.43/.36	.086/.092	10.1/13.0	.021/.029
Q4, concentrated	.52/.44	.052/.071	6.8/9.6	.013/.020

Table 11: Qwen2.5-VL-7B cross-backbone endpoint references and systems quantities. Tokens and latency report TraceGain/DirectPolicy.

Dataset	Global	Full	OCR-First	Tokens	Latency (ms)
DocVQA	82.6	89.5	86.7	912/901	412/404
InfographicVQA	51.6	61.5	57.9	988/976	446/438
ChartQA	72.9	81.3	75.4	836/829	389/382
TextVQA	72.5	77.4	74.8	1046/1031	468/457

The Qwen mechanism evaluation contains 17,908 / 13,284 / 9,311 / 21,846 states and 787,422 / 583,091 / 416,770 / 952,408 legal reveals for Doc/Info/Chart/Text. Mean legal-action counts vary from 43.6 to 44.8, confirming that the interface and legality distribution remain comparable without forcing identical trajectories.

C Evaluation Contract and Additional Operating Points

Table 12: Control objects at adjacent visual-compute boundaries. "Visited" denotes states induced by the deployed student.

Family	Decision object	Evaluation states	Stop	Evaluated quantity
Adaptive crop	Region or resolution	Base/policy states	Optional	Region/task score
Answer calibration	Answer or abstain	Answer states	Abstain	Current-answer risk
Token pruning	Retain or drop token	Encoded states	No	Retention saliency
Value of information	Acquire observation	Model-dependent	Optional	Expected information value
TraceGain	Legal reveal or stop	Student-visited legal states	Yes	Reveal-utility estimate under frozen student

The optimized object is the one-step official-score improvement of a legal pre-encoding reveal. The controller can zoom, shift, or stop within a four-step horizon; page bounds, non-duplication, and the unique-acquired-token cap are enforced by the legality mask. Both backbones remain frozen, OCR is summarized before control, and the dense teacher is restricted to centered offline target shaping. DirectPolicy provides the matched learned comparison because it shares inputs, action space, capacity, roll-ins, teacher privilege, schedule, and legality cap.

Mechanism evaluation uses final-policy states from untouched official test examples and reports action ranking, raw-utility calibration, one-step regret, and stopping behavior. Endpoint evaluation reports official dataset scores, pre-encoding token accounting, and batch-1 end-to-end latency. Table 1 provides the dense-reference context; Table 7 supplies the matched pre-encoding comparison used for the efficiency claim. Acquisition-method tokens count inputs before visual encoding, whereas ToMe and DynamicViT count retained post-encoding tokens; their values in Table 13 are contextual systems measurements rather than a common budget axis. Tables 9 and 11 provide held-out denominators and second-backbone systems measurements.

Table 13: Additional LLaVA operating points under one serving harness. Acquisition methods count pre-encoding tokens; ToMe and DynamicViT count retained post-encoding tokens.

Method	Tokens	Latency	Doc	Info	Chart	Text
Uniform-Tile	504	312	76.1	43.7	65.2	62.8
Saliency-Crop	576	358	77.5	45.9	67.1	63.4
OCR-First	648	381	79.5	47.2	66.8	64.9
Entropy-Refine	612	402	78.9	46.1	67.4	63.7
ToMe	864	425	79.2	47.8	68.3	65.1
DynamicViT	792	398	78.7	46.5	67.9	64.3