How Much Can We Prune? Exploring the Performance Limits of Visual Token Pruning

Anonymous authors
Paper under double-blind review

Abstract

Visual token pruning is increasingly used to reduce the memory and inference costs of vision-language models. Existing work has largely focused on designing method-specific selectors and demonstrating incremental improvements over prior methods at selected token budgets. It remains unclear whether the resulting performance reflects an inherent limit of visual token pruning or the limitations of current token-selection algorithms. We take a complementary perspective and ask: *how well can visual token pruning perform when offline search cost is not the primary constraint?* We formulate pruning for general image understanding as selecting, under a fixed budget, one shared visual-token subset that supports a broad distribution of questions about an image. We approximate this utility with a finite auxiliary QA set generated by a frontier VLM and optimize a teacher-forced answer-probability objective with the evaluated model frozen. Motivated by the conceptual monotonicity and diminishing returns of adding visual evidence, we construct nested subsets greedily and introduce *QA mini-batch greedy* (QA-MBG), which estimates token marginal gains from randomly sampled QA components at each step. Empirically, QA-MBG is both cheaper and more effective than full greedy. With one frontier-VLM call and approximately 0.5 H100 GPU-hours per image, it produces a nested subset trajectory covering keep ratios up to 0.3. Across Qwen3-VL-4B and Gemma 4 E4B on 900 images from nine benchmarks, we find significant headroom over evaluated pruning baselines on held-out QA and image-grounded long-form generation, while the searched subsets substantially narrow the gap to full-token inference.

1 Introduction

Vision-language models (VLMs) obtain strong visual understanding by converting an image into hundreds or thousands of visual tokens and processing them together with text (Bai et al., 2025; Wang et al., 2025). These tokens substantially increase language-model prefill computation, decoding-time attention cost, KV-cache memory, and end-to-end latency. Visual token pruning has therefore become a major direction for efficient VLM inference. Recent methods score or consolidate tokens using attention, query relevance, feature similarity, redundancy, diversity, or learned selection modules, often retaining only a small fraction of the original sequence (Chen et al., 2024a; Alvar et al., 2025; Wen et al., 2025b; Yang et al., 2025).

Most work follows the same evaluation pattern: propose a new selection heuristic and demonstrate an incremental accuracy–efficiency improvement over previous selectors at selected token budgets. This competition improves deployable algorithms, but leaves a more fundamental question unanswered: *how much performance can visual token pruning actually preserve at a fixed budget?* A gap from full-token inference conflates unavoidable information loss with recoverable selector error, while small differences among current methods do not show that they are close to the best available subset. Diagnostic studies further reveal strong sensitivity to evaluation protocol and cases in which simple rules rival elaborate methods (Wen et al., 2025a; Peng et al., 2025; Liao et al., 2026). The achievable performance frontier of visual token pruning remains unexplored.

We expose this frontier by relaxing the search-time compute constraint. Rather than designing another inexpensive inference-time heuristic, we invest substantial offline computation to discover a strong subset for each image. Every discovered subset is a feasible operating point at its retained-token budget; a large, reproducible gap over existing selectors therefore directly measures algorith-

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.

mic headroom. This changes the question from which current heuristic wins to how much room remains for better pruning algorithms.

The key challenge is defining subset quality for *general image understanding*. A subset selected for one known query can discard evidence needed by another. We instead seek one shared subset that supports the distribution of questions that could reasonably be asked about an image, formalized as expected question-answering utility under a cardinality constraint. Because this distribution is unavailable, we approximate it with a finite, diverse auxiliary QA set generated by a frontier VLM. The evaluation questions and answers are never used to score candidate subsets in the main experiment; after search, the same nested trajectory is reused for held-out benchmark questions and long-form description prompts.

This formulation connects visual-token selection to classical information-gathering problems, where adding observations yields monotone, diminishing-return utility and enables near-optimal greedy selection (Krause et al., 2008). Frozen-VLM QA utility is only approximately monotone and submodular, but it exhibits the same saturation pattern empirically. We therefore build a nested trajectory by repeatedly adding the token with the largest estimated marginal gain. Full greedy is prohibitively expensive because every step evaluates every remaining token against every auxiliary QA.

We introduce *QA mini-batch greedy* (QA-MBG), which samples a fresh mini-batch of auxiliary questions at each step, evaluates all remaining token candidates on those components, and adds the best token. This is the decomposable-objective counterpart of mini-batch submodular maximization (Schwartzman, 2024), and differs from stochastic greedy, which samples token candidates while evaluating the full objective (Mirzasoleiman et al., 2015). QA-MBG reduces QA evaluations and improves the irreversible sequence of greedy decisions: it raises held-out QA and long-form performance over full greedy while producing more balanced gains across auxiliary questions. Candidate batching, prefix-KV reuse, and teacher-forced scoring make the offline search practical. One frontier-VLM call and approximately 0.5 H100 GPU-hours per image produce a complete nested trajectory through keep ratio 0.3.

We evaluate Qwen3-VL-4B and Gemma 4 E4B on 900 images and $2{,}226$ held-out questions from nine benchmarks. Across both model families, searched subsets establish a clear empirical frontier above the evaluated pruning baselines and substantially narrow the gap to full-token inference. The gains are especially pronounced on document, chart, infographic, and diagram understanding, and transfer to image-grounded long-form generation despite search using a short-answer objective.

Our contributions are threefold:

- We formulate visual token pruning for general image understanding as selecting one shared subset that maximizes expected QA utility under a fixed token budget, and use frontier-generated auxiliary QAs to expose its empirical performance frontier.
- We introduce QA mini-batch greedy, which samples objective components rather than token candidates and improves both search cost and downstream subset quality relative to full greedy and token-level stochastic greedy.
- Across two VLM families, nine benchmarks, held-out QA, and long-form generation, we reveal substantial headroom over existing pruning methods, showing that current selectors are far from exhausting what is achievable at the same visual-token budgets.

2 Related Work

Visual token pruning in VLMs. Visual token pruning reduces multimodal inference cost by retaining a subset of visual tokens before or within the language model. Existing selectors use multi-stage dropping (Liu et al., 2024), attention or head importance (Zhu et al., 2026; Gao et al., 2026), learned selection (Zhu et al., 2025), diversity (Alvar et al., 2025), redundancy (Wen et al., 2025b), or combinations of relevance and contextual coverage (Pei et al., 2025; Zhang et al., 2025c; Oh & Kim, 2026; Wang et al., 2026). Query-conditioned selection targets evidence for a particular question, whereas query-agnostic selection seeks to retain information without access to that question; SparseVILA explicitly distinguishes these settings (Khaki et al., 2025). Adaptive methods additionally vary the budget with the input, as in GlimpsePrune, PromPrune, and OccamToken (Zeng et al., 2026; Lee et al., 2026; Li et al., 2026). These approaches primarily target efficient, deployable selection. Our focus is on the performance achievable by a shared image-token subset at a specified budget, selected using auxiliary QA supervision and reused without conditioning selection on the evaluation question.

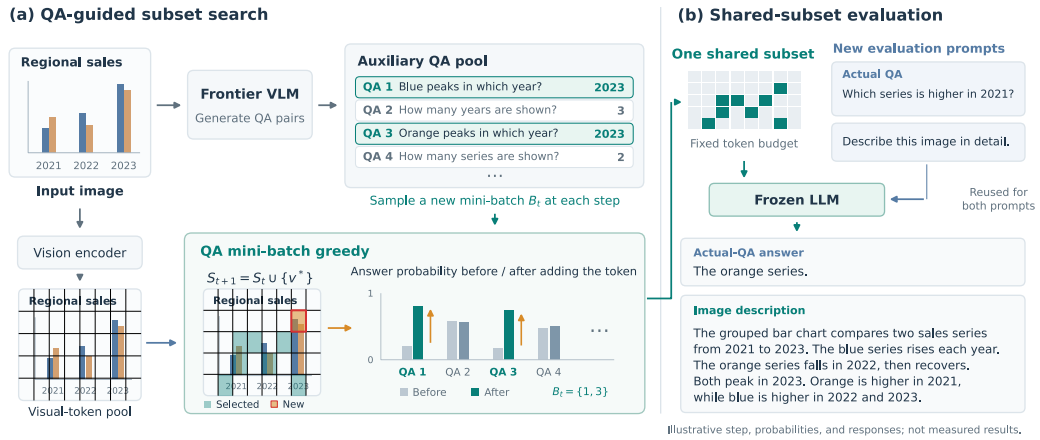


Figure 1: QA-guided subset search and shared-subset evaluation.

Oracle analyses and search-based selection. SparseVILA constructs a query-aware greedy oracle that selects visual tokens to maximize agreement with the unpruned model's responses, and uses its deterioration across subsequent conversation rounds to illustrate the limitations of irreversible query-dependent pruning (Khaki et al., 2025). Our setting instead uses diverse auxiliary QA pairs to search for one subset intended to support multiple evaluation questions. EvoComp performs evolutionary search for visual-token masks that minimize response loss, constrains selection through semantic grouping, and uses the resulting masks to supervise a learned compressor (Song et al., 2026). We share the use of answer-based search signals, but study a different target: the achievable performance of a shared multi-QA subset and the effect of search granularity on that performance, rather than compressor training. Related oracle analyses include the label-conditioned Grad-CAM oracle in video classification studied by Principles of Visual Tokens (Hao et al., 2025). Our contribution therefore lies in combining shared-subset exploration, evaluation with target-question annotations withheld from candidate scoring, and an analysis of QA mini-batch greedy, rather than in introducing answer-guided token search itself.

Subset optimization. Classical stochastic greedy reduces oracle evaluations in cardinality-constrained subset selection by sampling candidate elements at each iteration (Mirzasoleiman et al., 2015). Mini-batch submodular maximization instead considers a decomposable objective, $F = \sum_i f_i$, and resamples objective components at every greedy step (Schwartzman, 2024). Our auxiliary-QA objective has the same decomposition across questions: QA-MBG samples QA components while evaluating every remaining visual-token candidate. This directly separates QA-MBG from token-candidate stochastic greedy and provides the closest combinatorial-optimization precedent for our search rule. Because frozen-VLM answer probability is only approximately monotone and submodular, its existing approximation guarantee does not directly transfer; we therefore evaluate both the full auxiliary objective and downstream performance of the resulting trajectories.

3 Method

Figure 1 illustrates the pipeline: auxiliary QAs guide subset search, and the selected visual tokens are reused across evaluation prompts. The highlighted patch and probability changes illustrate one possible search step rather than measured effects for that example.

3.1 Problem formulation: shared subsets for image understanding

A fixed visual-token budget. Let I be an image and $Z_I = (z_1, \ldots, z_n)$ its visual embeddings at the input to a frozen language model with parameters θ. We select indices from $V_I = \{1, \ldots, n\}$ after visual encoding and before language-model processing. For $S \subseteq V_I$, $Z_I[S]$ denotes the retained embeddings, with their original order and positional information preserved. Given a keeping ratio $\rho \in [0, 1]$, the budget is $b = \lfloor \rho n \rfloor$. Our goal is to find a single subset of this size that preserves general image understanding across questions. Writing the abstract understanding ability as U_I, the

ideal problem is

$$S^{\star}_{U,b} \in \operatorname*{arg\,max}_{S \subseteq V_I,\ |S|=b} U_I(S). \tag{1}$$

The subset depends on the image and budget, but is shared across evaluation questions. We use an exact cardinality constraint for subset search. Experiments report realized token counts because model-specific rounding can produce small differences at the same requested ratio.

Conceptual structure of understanding. Additional useful visual clues should expand what can be understood about an image. A new clue also often contributes less when much of the relevant evidence is already available. These intuitions motivate an idealized monotone submodular view of U_I: for $A \subseteq B \subseteq V_I$ and $v \notin B$,

$$U_I(A) \le U_I(B), \tag{2}$$

$$U_I(A \cup \{v\}) - U_I(A) \ge U_I(B \cup \{v\}) - U_I(B). \tag{3}$$

Equation (2) expresses monotonicity and Eq. (3) diminishing returns. These properties describe the predominant effect of accumulating visual evidence: although counterexamples can arise from complementary clues, entangled embeddings, or distraction by additional tokens, most additions preserve or improve usable evidence and contribute less once related evidence is already available. We therefore use monotone submodularity as an approximate structural model for guiding incremental selection.

From abstract ability to question-answering probability. To operationalize understanding, let $\mathcal{D}_I$ be a target distribution over image-grounded questions, covering the aspects of I that a user could ask about. For a question q, let $\mathcal{A}_I(q)$ denote its set of correct answers. A model-dependent measure of image understanding is the expected probability of answering correctly:

$$F_I(S) = \mathbb{E}_{q \sim \mathcal{D}_I}\left[\Pr_{\widehat{a} \sim p_\theta(\cdot \mid q, Z_I[S])}\big(\widehat{a} \in \mathcal{A}_I(q)\big)\right], \qquad S^{\star}_{F,b} \in \operatorname*{arg\,max}_{S \subseteq V_I,\ |S|=b} F_I(S). \tag{4}$$

The distribution over questions specifies which kinds of understanding are valued; there is no uniform distribution over all possible natural-language questions.

Finite auxiliary QA approximation. Neither the question distribution nor all correct answers are available. We therefore prompt a frontier VLM with the image to produce a finite auxiliary set $\mathcal{Q}_I = \{(q_j, a_j)\}_{j=1}^{M}$. The generation prompt seeks diverse, image-grounded questions spanning the evidence types appropriate to the target task distribution. General understanding here is operationalized as support for multiple questions through one shared subset, with transfer to general description prompts evaluated separately. A generated answer provides reference supervision, rather than an exhaustive set of valid answers. Let $f_j(S)$ be the reference-answer score for (q_j, a_j). We optimize

$$\widehat{F}_I(S) = \frac{1}{M}\sum_{j=1}^{M} f_j(S), \qquad S^{\star}_{\mathrm{aux},b} \in \operatorname*{arg\,max}_{S \subseteq V_I,\ |S|=b} \widehat{F}_I(S). \tag{5}$$

For a tokenized reference answer $a_j = (a_{j,1}, \ldots, a_{j,L_j})$, we score each answer token by teacher forcing the preceding reference tokens:

$$\begin{aligned} p_{j,\ell}(S) &= p_\theta(a_{j,\ell} \mid q_j, Z_I[S], a_{j,<\ell})\,, \\ f_j(S) &= \exp\left(\frac{1}{L_j}\sum_{\ell=1}^{L_j} \log p_{j,\ell}(S)\right) = \left(\prod_{\ell=1}^{L_j} p_{j,\ell}(S)\right)^{1/L_j}. \end{aligned} \tag{6}$$

Thus f_j is the length-normalized geometric mean of the reference-token probabilities, not the raw sequence probability or the arithmetic mean of token probabilities. Length normalization removes the direct multiplicative penalty for longer answers, but does not turn the score into the probability that a freely generated answer is correct. Equation (5) averages these scores equally across QA pairs. All candidates are scored with the same reference answer and preceding reference-answer tokens for a given QA pair; their retained visual context differs. The auxiliary QA set approximates the breadth of questions, while the reference-answer score provides a fine-grained measure of subset quality. Rather than autoregressively generating an answer and reducing the result to a binary correct/incorrect outcome, we retain the teacher-forced probability assigned to the reference answer. This soft score distinguishes subsets even when they yield the same discrete answer and therefore measures changes in image understanding at a finer resolution. It is also substantially faster than decoding because all reference-answer positions are evaluated in parallel in a single causal pass.

3.2 Optimization with QA mini-batch greedy

Full greedy and its cost. Exhaustive search over $\binom{n}{b}$ subsets is infeasible. Starting with $S_0 = \varnothing$, forward greedy adds the remaining token with the largest marginal gain:

$$\Delta_j(v \mid S) = f_j(S \cup \{v\}) - f_j(S), \qquad v_t \in \arg\max_{v \in V_I \setminus S_t} \frac{1}{M} \sum_{j=1}^{M} \Delta_j(v \mid S_t), \qquad S_{t+1} = S_t \cup \{v_t\}. \tag{7}$$

Searching to $b_{\max}$ requires $\mathcal{O}(Mb_{\max}n)$ QA–candidate scoring evaluations: the number of steps times the remaining candidates times the number of QA pairs.

Sampling candidates or objective components. Token-level stochastic greedy randomly samples a fraction of the remaining visual-token candidates at each step and evaluates only those candidates on the complete auxiliary-QA objective. It reduces computation by considering fewer token choices, following the candidate-sampling idea of stochastic greedy (Mirzasoleiman et al., 2015).

QA mini-batch greedy instead samples a QA index set $B_t \subseteq \{1, \ldots, M\}$, uniformly without replacement, of size $m = \max\{1, \lceil \alpha M \rceil\}$ for QA fraction $\alpha \in (0, 1]$. We resample B_t at every step and use the same mini-batch for all remaining candidates. The update uses

$$\widehat{\Delta}_t(v \mid S_t) = \frac{1}{m} \sum_{j \in B_t} \Delta_j(v \mid S_t), \qquad v_t \in \arg\max_{v \in V_I \setminus S_t} \widehat{\Delta}_t(v \mid S_t). \tag{8}$$

For a fixed current subset and remaining token, fresh uniform QA sampling gives

$$\mathbb{E}_{B_t}\left[\widehat{\Delta}_t(v \mid S_t) \mid S_t\right] = \widehat{F}_I(S_t \cup \{v\}) - \widehat{F}_I(S_t). \tag{9}$$

For a fixed current subset and candidate token, averaging this estimate over repeated random QA mini-batches recovers the marginal gain computed on the complete generated-QA pool. This statement applies only to that finite pool: it does not imply that the pool represents every possible question, nor that maximizing noisy estimates selects the same token as full greedy. The QA batch size therefore controls both scoring cost and variability in the trajectory. Setting $m = M$ recovers full greedy, while $m = 1$ uses one QA pair per step. The number of QA–candidate scoring evaluations decreases linearly with the QA sampling fraction.

One search, multiple budgets.

Algorithm 1 returns one nested trajectory whose prefix S_b serves each requested budget and all evaluation questions about the image. One run therefore yields every budget up to $b_{\max}$, amortizing search across the entire budget curve. Because all budgets are prefixes of the same trajectory, comparisons across keeping ratios do not benefit from independent search restarts. QA mini-batching perturbs the estimated gains and can therefore produce a different trajectory from candidate sampling or full greedy. Because one token is added at every step, the full auxiliary objective need not improve monotonically; we treat the resulting subsets as achievable operating points rather than certified optima.

Algorithm 1 QA mini-batch greedy.

Require: V_I, scores $\{f_j\}_{j=1}^{M}$, $b_{\max}$, m
$S_0 \leftarrow \varnothing$
for $t = 0, \ldots, b_{\max} - 1$ **do**
  Sample a uniform m-subset B_t of QAs
  for all $v \in V_I \setminus S_t$ **do**
    $g_t(v) \leftarrow \frac{1}{m} \sum_{j \in B_t} f_j(S_t \cup \{v\})$
  end for
  $v_t \leftarrow \arg\max_{v \in V_I \setminus S_t} g_t(v)$
  $S_{t+1} \leftarrow S_t \cup \{v_t\}$
end for
return $(S_0, \ldots, S_{b_{\max}})$

Efficient candidate scoring. The greedy structure admits efficient batching: candidates at the same step have equal visual-token length and require no sequence padding. Their fixed system-and-image prefixes also allow candidate-specific KV states to be reused across sampled QA suffixes. Teacher forcing scores all reference-answer positions in parallel without autoregressive decoding, while cached visual embeddings avoid repeated vision encoding. Together, these optimizations preserve exhaustive candidate coverage while reducing repeated computation within each greedy step. Unlike token-candidate sampling, QA-MBG spends the saved computation on fewer objective components while still comparing every remaining token. Appendix B gives implementation and runtime details.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation sets. We use two cohorts from nine benchmarks: AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), MME (Fu et al., 2025), MMStar (Chen et al., 2024b), RealWorldQA (xAI, 2024), TextVQA (Singh et al., 2019), and V*Bench (Wu & Xie, 2024). The *main set* contains 100 images per benchmark (9×100: 900 images, 2,226 actual QAs); the *small set* contains 20 per benchmark (9×20: 180 images, 392 actual QAs). Actual-QA evaluation uses all associated benchmark questions. These annotations are separate from the generated auxiliary QA pool, although their semantic content can overlap. Long-form evaluation uses two description prompts per image.

Models. We evaluate Qwen3-VL-4B-Instruct (Bai et al., 2025) and Gemma 4 E4B-it (Gemma Team, 2026), keeping their parameters frozen. Our search selects visual tokens after visual encoding and projection, before language-model processing. Each selected subset is shared across evaluation questions about the same image. We use an image-processor cap of 1M pixels for Qwen3-VL and a maximum of 1,120 visual soft tokens for Gemma 4. Architecture-specific integration and baseline adaptations are detailed in Appendix A.

Auxiliary QA generation. We use GPT-5.6-SOL to generate image-grounded auxiliary question–answer pairs; the generation protocol, prompts, validation, and provenance are documented in Appendix C. Because generated auxiliary QAs can semantically overlap with actual QAs, we compare the original search with a search after removing overlapping auxiliary QAs. We also evaluate dataset-agnostic generation and repeat QA generation with Opus 5 to test sensitivity to the generation protocol and frontier model. The resulting 30,273 auxiliary QAs average 33.64 per image. A selected trajectory is held fixed during actual-QA and long-form evaluation, without re-searching for each prompt.

Search and baselines. The main search configuration uses QA mini-batch greedy with QA sampling fraction $\alpha = 0.2$, evaluating all remaining token candidates at each step. It optimizes the length-normalized geometric-mean answer-probability objective in Eq. (6), averaged over the sampled auxiliary QA pairs. We first interpret headroom against question-independent post-vision subset selectors, including Uniform, FLoC, and DivPrune, which also retain a shared subset of unchanged visual embeddings. The broader comparison includes feature merging, query-aware selection, and decoder-internal pruning. These methods can access different selection signals, change retained features, or process all tokens in early decoder layers. Prior pruning methods do not share a common model backbone, and LLaVA—widely used in earlier work—is an older architecture that is not representative of the contemporary VLMs evaluated here. We therefore implement architecture-specific adaptations for Qwen3-VL and Gemma 4; Appendix A details their pruning locations and schedules. Full-token inference provides an uncompressed reference.

Choice of QA mini-batch fraction. The QA mini-batch fraction α is the only search hyperparameter of our method. We set $\alpha = 0.2$ empirically using the small-set experiments; Section 4.3 describes the selection analysis.

Budgets and reporting. We report visual-token keeping ratios up to 0.30. Subsets at different budgets are prefixes of one nested search trajectory for each image. The top row of Figure 2 expresses native benchmark performance relative to the corresponding full-token model, whose reference level is 100%. The searched trajectories provide denser budget coverage than the fixed baseline operating points. Comparisons at a given retained-token budget measure achievable quality; equal final token counts do not imply equal computational cost across pre-LLM and progressive decoder-internal methods. Search execution and measured runtime are reported in Appendix B.

Long-form evaluation. We reuse the selected image representation for two general image-description prompts (P1/P2), without optimizing a new subset for either prompt. GPT-5.6-Luna evaluates the generated descriptions against the original image, with candidate answers anonymized and shuffled. Figure 2(c,d) reports mean overall image-grounded quality on a 0–5 scale over the evaluated image–prompt groups. Full-token answers are comparison candidates rather than ground-truth references; their ratings can vary with the other candidates in each judging group. The fixed-subset baselines and the generation and judging protocol are described in Appendices A.3 and D.

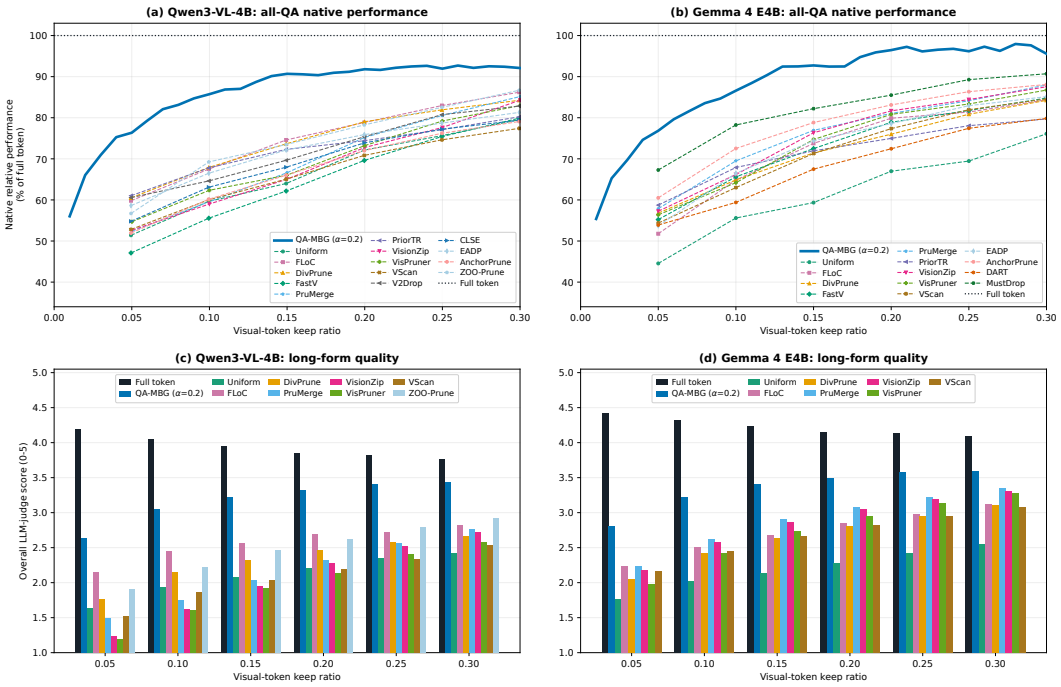


Figure 2: Actual-QA and long-form performance on the main set.

4.2 Main results

Achievable headroom across two model families. Figure 2(a,b) shows that QA-guided subset search achieves higher aggregate native benchmark performance than the plotted baselines at comparable token budgets for both Qwen3-VL and Gemma 4. The separation is particularly pronounced at low and moderate keeping ratios and remains visible toward 0.30. Thus the performance lost by existing selectors is not determined by the token budget alone: stronger retained subsets are available under the same model and budget constraints.

Performance as the budget increases. The explored curves rise rapidly at small budgets and then approach a plateau closer to the full-token reference. The baselines also improve with additional tokens, narrowing the gap at larger budgets, but do not close it over the displayed range. Notably, this headroom also persists against query-aware pruning methods that select tokens using the evaluation question itself, even though our search produces one shared subset for all questions about an image. Observing this pattern in both model families supports the usefulness of shared-subset exploration beyond a single VLM architecture. The gap to direct shared-subset selectors demonstrates achievable quality beyond those selection rules under the auxiliary-supervision protocol. Comparisons with merging and decoder-internal adapters broaden the empirical context, while retaining the differences in representation and computation described above.

Transfer to long-form descriptions. Figure 2(c,d) shows that the same searched representations also improve long-form quality over the plotted compression baselines at all six evaluated budgets for both models. These prompts request broad descriptions rather than the short answers used by the search objective. The advantage therefore extends beyond the native short-answer evaluation. For Qwen3-VL, QA-MBG retains a clear margin over the strongest plotted baseline throughout the range, while its gap to full-token descriptions narrows as the budget increases. Averaged over the six budgets, its paired advantage over FLoC, the baseline with the highest six-budget mean, is 0.612 judge-score points (95% CI $[0.578, 0.645]$). This evaluates whether the selected image information remains useful under new prompts and a different answer format on the same images. The judge assesses each answer against the image rather than treating the full-token response as ground truth.

Paired uncertainty analysis. A benchmark-stratified, image-level paired bootstrap with $10{,}000$ replicates yields positive lower endpoints for all 252 pointwise 95% confidence intervals: 84 Qwen and 78 Gemma actual-QA comparisons, and 48 Qwen and 42 Gemma long-form comparisons.

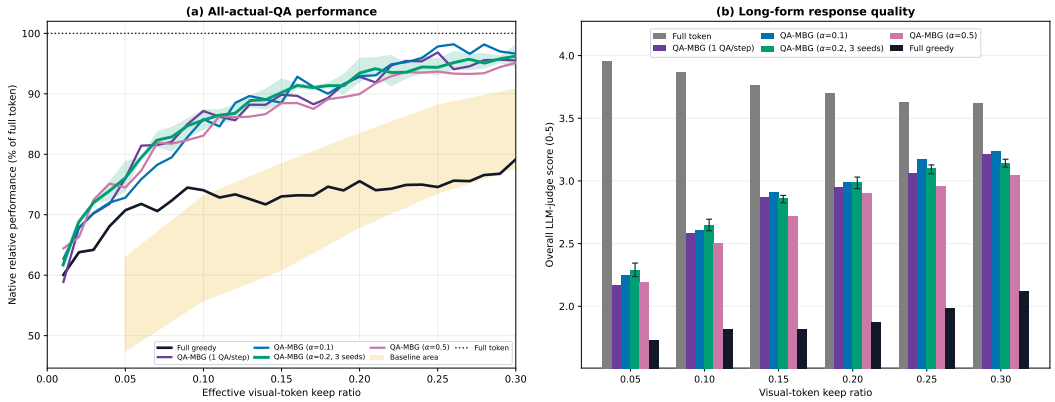


Figure 3: QA mini-batch greedy versus full greedy on the small set.

Against the baseline with the highest observed six-budget mean in each model/task, mean gains are 10.12 and 5.10 native-score percentage points for Qwen and Gemma actual QA, and 0.61 and 0.45 judge-score points for their long-form descriptions, respectively. These are individual comparison intervals, not a simultaneous confidence guarantee across all comparisons. Actual-QA bootstrap estimates use image-balanced native scores, rather than the full-token-normalized metric in Figure 2. Appendix G details the resampling protocol, comparison summaries, and scope of inference.

4.3 Analysis

We examine the choice of search strategy, its effect on auxiliary-objective optimization, sensitivity to benchmark conditioning and direct QA overlap, the size of the supervision pool, and the auxiliary-QA generator. These analyses complement the paired uncertainty estimates and long-form evaluation above.

QA mini-batching versus full greedy. Figure 3 compares QA batch sizes on the small set. Green shading and error bars show the $\alpha = 0.2$ three-seed min–max ranges, not confidence intervals; the yellow envelope spans baseline methods rather than repeated runs. Other settings are single runs (Appendix I.3). At keep ratio 0.20, $\alpha = 0.2$ achieves mean relative actual-QA performance of 93.44% across three seeds (range 91.90–95.95%), versus 75.56% for full greedy and 89.96% for the reported $\alpha = 0.5$ run. Long-form quality is 2.989 (range 2.939–3.031), versus 1.869 for full greedy and 2.903 for $\alpha = 0.5$. All three $\alpha = 0.2$ seeds exceed full greedy at the six standard budgets in both evaluations, so the advantage is not confined to the original sampling trajectory. The larger fraction 0.5 yields lower long-form quality than every 0.2 seed at all six budgets despite evaluating more QA components per candidate. This supports 0.2 as a practical setting, while 0.1 and one QA per step remain competitive. The reversal—QA-MBG succeeding where full greedy underperforms—is particularly striking. We hypothesize that full greedy repeatedly fits the same finite generated-QA objective, whereas resampling QA components changes the objective estimate at every step and acts as a regularizer against overfitting to pool-specific artifacts. The following QA-sampling-dynamics analysis probes this hypothesis. Appendix I gives complete results, including the supplementary token-level control and seed protocol.

QA sampling dynamics. On the small set, QA sampling improves the lower tail of generated-QA probabilities compared with full greedy. At keeping ratio 0.30, $\alpha = 0.5$ achieves the highest overall mean (0.7203 versus 0.6790 for full greedy) and bottom-20% mean (0.3566 versus 0.2756) among the tested fractions. The lower tail is recomputed at each step; it is not a fixed group of initially difficult questions. At $\alpha = 0.2$, the overall mean is similar to full greedy while the lower tail improves. Step-level diagnostics show focused gains on sampled QAs alongside temporary regressions on unsampled QAs, consistent with resampling redistributing improvements across the search trajectory. These diagnostics characterize optimization of the generated-QA objective; they do not imply a monotonic benefit from smaller fractions. Downstream transfer is assessed separately using actual QAs and long-form descriptions. Appendix H provides definitions, trajectories, and numerical summaries.

Controlling the QA distribution. When the target question distribution is known, as in benchmark-specific QA generation, auxiliary QAs can align with the actual-QA evaluation and introduce potential question-level leakage. We use two controls to separate this effect from headroom attributable to token selection. First, dataset-agnostic generation withholds benchmark names, task profiles, dataset-specific instructions, and same-benchmark examples while retaining the image-grounding and diversity requirements. At keep ratio 0.20, it achieves 92.79% full-token-relative performance, compared with 96.18% for the original setting and 83.47% for the strongest baseline. Its margin over the strongest baseline remains positive at all six budgets, showing that benchmark information helps but is not required for the observed headroom.

For the overlap control, after generating the auxiliary QAs, we use a text-only GPT-5.6-Luna judge to compare generated and actual QA pairs and filter generated QAs that ask about semantically equivalent visual facts. We then re-run subset search with the filtered generated-QA pool. At keep ratio 0.20, filtered search reaches 91.36% full-token-relative performance, compared with 96.18% without filtering and 83.47% for the strongest baseline. The filtered trajectory remains above the strongest baseline at all six budgets. Actual-QA annotations are used only to define which generated QAs are filtered; they never score candidate subsets. Because filtering also removes potentially useful visual supervision, the performance drop does not measure leakage alone. Together, the controls show that task conditioning and semantic overlap strengthen, but do not explain, the measured headroom. Appendices J and K provide protocols and full results.

Number of generated QAs. On the small set, we vary the search pool using nested 25%, 50%, 75%, and 100% subsets. At keep ratio 0.20, relative performance increases from 86.33% with one quarter of the pool to 96.18% with the complete pool; using three quarters reaches 93.87%. Larger pools generally improve aggregate performance, with benchmark-dependent and occasionally non-monotonic effects. This measures sensitivity to the amount of auxiliary supervision. The per-step QA fraction remains 0.2, so increasing the pool also increases the sampled QA count and search compute; it does not isolate diversity at fixed compute. Appendix F provides the construction, exact counts, and results.

Auxiliary-QA generator. Replacing GPT-5.6-SOL with Opus 5 on the small set produces a closely matched actual-QA curve: the six-budget mean relative performance is 93.14% versus 91.69%, with a mean absolute pointwise difference of 2.30 points. Opus-generated subsets also exceed the best plotted long-form baseline at each budget by 0.56 judge-score points on average (3.21 versus 2.65). Appendix L gives the per-budget results.

Qualitative examples. Appendix M presents one selected favorable example from each benchmark, showing retained visual tokens, actual-QA responses, and long-form descriptions for the same image. These cases illustrate preserved evidence and differences in description quality; aggregate comparisons use the full evaluation cohorts above.

5 Conclusion

We ask how much performance visual token pruning can preserve when offline search cost is relaxed. We formulate general-understanding pruning as selecting one image-specific subset shared across questions, approximate its utility with frontier-generated auxiliary QAs, and optimize it with QA mini-batch greedy. Across Qwen3-VL and Gemma 4, the resulting trajectories reveal substantial headroom over evaluated pruning baselines on both actual QA and long-form descriptions, while narrowing the gap to full-token inference. On the small set, QA mini-batching also outperforms full greedy with fewer QA evaluations; its step-wise dynamics are consistent with a regularizing effect from resampling objective components. The headroom persists when benchmark information is withheld, semantically overlapping generated QAs are filtered, and the auxiliary-QA generator is changed. It also remains with smaller generated-QA pools, although performance generally improves as more auxiliary QAs are used.

These results establish a strong empirical frontier across the evaluated models and supervision settings. The consistent gaps show that a substantial portion of current pruning loss is recoverable selector error rather than inevitable information loss from the token budget. Visual token pruning can therefore perform substantially better than current heuristics, and the searched trajectories provide concrete targets for learning or designing the next generation of efficient selectors.

AI USE STATEMENT

Frontier models are integral to the method and evaluation: GPT-5.6-SOL and Opus 5 generate auxiliary QAs, while GPT-5.6-Luna evaluates long-form outputs and semantic QA overlap. AI assistants also supported code development, language editing, and figure and table formatting. The authors designed the study and verified the implementations, analyses, claims, citations, and final manuscript.

ETHICS STATEMENT

We use public research benchmarks and pretrained models, and collect no new human-subject or personally identifying data. Generated QAs and LLM judgments may inherit model biases or errors; we therefore report aggregate results and targeted controls rather than treating them as ground truth. The offline search costs approximately 0.5 H100 GPU-hours per image and is intended to diagnose achievable headroom, not as a deployment-time algorithm.

REPRODUCIBILITY STATEMENT

The main text and appendix document the evaluated cohorts and models, generation and judging prompts, scoring objective, search algorithm and hyperparameter, baseline adaptations, random seeds, runtime measurements, and bootstrap procedure. Exact replication depends on access to the named hosted frontier models; we report their identities, prompts, and provenance to make this dependency explicit.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025. doi: 10.1109/CVPR52734.2025.00877.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Bin Chen, Yuxiang Cai, Yadan Luo, Yi Zhang, Jianwei Yin, and Zhi Chen. Spectral evolution-guided token pruning in multimodal large language models. In *European Conference on Computer Vision*, pp. 221–239, 2026a. doi: 10.1007/978-3-032-37517-9_13.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. *arXiv preprint arXiv:2509.01552*, 2025. URL https://arxiv.org/abs/2509.01552.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35, 2024a. doi: 10.1007/978-3-031-73004-7_2.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024b. doi: 10.52202/079017-0850.

Zengjie Chen, Yuxiang Cai, Jingcai Guo, Taotao Cai, Jianwei Yin, and Zhi Chen. Accelerating multimodal large language models with prior-corrected token reduction. In *European Conference on Computer Vision*, pp. 213–233, 2026b. doi: 10.1007/978-3-032-36839-3_13.

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. FLoC: Facility location-based efficient visual token compression for long video understanding. In *International Conference on Learning Representations*,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a87d3a6b4884108c0dff392c1c7e2d82-Abstract-Conference.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899.

Tianxiao Gao, Shanwei Zhao, Shuo Fang, Shiai Zhu, and Chenguang Ma. Quietprune: Query-guided early token pruning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3553–3562, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gao_QuietPrune_Query-Guided_Early_Token_Pruning_for_Vision-Language_Models_CVPR_2026_paper.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Xinyue Hao, Gen Li, Shreyank N. Gowda, Robert B. Fisher, Jonathan Huang, Anurag Arnab, and Laura Sevilla-Lara. Principles of visual tokens for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21254–21264, 2025. doi: 10.1109/ICCV51701.2025.01974.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling visual sparsity for efficient VLM inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23784–23794, 2025. doi: 10.1109/ICCV51701.2025.02208.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. ZOO-Prune: Training-free token pruning via zeroth-order gradient estimation in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39572–39582, 2026. URL https://aim-skku.github.io/ZOO-Prune/.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL https://jmlr.org/papers/v9/krause08a.html.

Jaehoon Lee, Mingi Jung, Soohyuk Jang, Seungryong Yoo, Dahuin Jung, and Sungroh Yoon. Balancing saliency and coverage: Semantic prominence-aware budgeting for visual token compression in VLMs. *arXiv preprint arXiv:2603.14892*, 2026.

Geng Li, Guohao Chen, Ting Chen, Shilin Shan, Kuangji Zuo, Bofan Lyu, Tuo An, Gen Li, and Jianfei Yang. Occamtoken: Efficient VLM inference with training-free and budget-adaptive token pruning. *arXiv preprint arXiv:2605.29657*, 2026.

Chenfei Liao, Wensong Wang, Zichen Wen, Xu Zheng, Yiyu Wang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Xin Zou, Yuqian Fu, Bin Ren, Linfeng Zhang, and Xuming Hu. Are we using the right benchmark: An evaluation framework for visual token compression methods. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4236–4253, 2026. doi: 10.18653/v1/2026.acl-long.195.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024.

Ahmed Masry, Do Xuan 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.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Minesh Mathew, Viraj Bagal, Rubèn Pérez 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. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1812–1818, 2015. doi: 10.1609/aaai.v29i1.9486.

Kyuan Oh and Bumsoo Kim. Anchorprune: Relevance-anchored contextual expansion for visual token pruning. In *European Conference on Computer Vision*, pp. 595–612, 2026. doi: 10.1007/978-3-032-37211-6_33.

Ruiguang Pei, Weiqing Sun, Zhihui Fu, and Jun Wang. Greedyprune: Retenting critical visual token set for large vision language models. *arXiv preprint arXiv:2506.13166*, 2025.

Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, Feng Xiao, and Lizhen Cui. Can visual input be compressed? a visual input token compression benchmark for large multimodal models. *arXiv preprint arXiv:2511.02650*, 2025.

Gregory Schwartzman. Mini-batch submodular maximization. *arXiv preprint arXiv:2401.12478*, 2024. doi: 10.48550/arXiv.2401.12478. URL https://arxiv.org/abs/2401.12478.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025. doi: 10.1109/ICCV51701.2025.02122.

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. 8317–8326, 2019. URL https://arxiv.org/abs/1904.08920.

Jiafei Song, Fengwei Zhou, Jin Qu, Wenjin Jason Li, Tong Wu, Gengjian Xue, Zhikang Zhao, Daomin Wei, Yichao Lu, and Bailin Na. Evocomp: Learning visual token compression for multimodal large language models via semantic-guided evolutionary labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3532–3542, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Song_EvoComp_Learning_Visual_Token_Compression_for_Multimodal_Large_Language_Models_CVPR_2026_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xuehui Wang, Xuankun Yang, and Wei Shen. Combating textual noise and redundancy: Entropy-aware dense visual token pruning. In *European Conference on Computer Vision*, pp. 559–580, 2026. doi: 10.1007/978-3-032-37026-6_31.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, 2025a. doi: 10.18653/v1/2025.findings-acl.802.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025b. doi: 10.18653/v1/2025.emnlp-main.505.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. doi: 10.1109/CVPR52733.2024.01243.

xAI. Grok-1.5 vision preview, 2024. URL https://x.ai/news/grok-1.5v. Introducing the RealWorldQA benchmark.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843.

Quan-Sheng Zeng, Yunheng Li, Qilong Wang, Peng-Tao Jiang, Zuxuan Wu, Ming-Ming Cheng, and Qibin Hou. A glimpse to compress: Dynamic visual token pruning for large vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026. doi: 10.1109/TCSVT.2026.3702147.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking visual token reduction for efficient large vision-language models. *arXiv preprint arXiv:2505.22654*, 2025a. URL https://arxiv.org/abs/2505.22654.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025b. doi: 10.1109/ICCV51701.2025.01939.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28761–28791, 2025c. doi: 10.52202/085713-0855.

Jiaying Zhu, Yurui Zhu, Xin Lu, Wenrui Yan, Dong Li, Kunlin Liu, Xueyang Fu, and Zheng-Jun Zha. Visionselector: End-to-end learnable visual token compression for efficient multimodal LLMs. *arXiv preprint arXiv:2510.16598*, 2025.

Qihui Zhu, Tao Zhang, Yuchen Wang, Shuangwu Chen, Xiaobin Tan, Jian Yang, Yang Liu, and Yinfei Pan. HAWK: Head importance-aware visual token pruning in multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39583–39592, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhu_HAWK_Head_Importance-Aware_Visual_Token_Pruning_in_Multimodal_Models_CVPR_2026_paper.html.

Appendix

How Much Can We Prune?

Exploring the Performance Limits of Visual Token Pruning

APPENDIX CONTENTS

A Baseline Implementation Details **15**

B Efficient Execution of Subset Search **17**

C Auxiliary QA Generation **19**

D Long-Form Generation and Image-Grounded Judging **22**

E Per-Benchmark Results **24**

F Generated-QA Count Ablation **25**

G Paired Bootstrap Analysis **27**

H QA-Sampling Trajectory Analysis **29**

I Greedy Search and QA Mini-Batch Fraction **32**

J Generated-QA Removal Ablation **35**

K Dataset-Agnostic Auxiliary QA Generation **36**

L Robustness to the Auxiliary-QA Generator **39**

M Qualitative Examples **40**

A Baseline Implementation Details

We evaluate architecture-specific adapters for Qwen3-VL-4B and Gemma 4 E4B. The descriptions below distinguish unchanged-token selection, feature merging, and decoder-internal pruning. All adapters use frozen model weights; adaptations to architectures different from those in the source papers are stated explicitly.

Comparison overview. Table 1 classifies the evaluated adapters. Post-ViT denotes reduction after visual encoding and before the first LLM layer; Inside LLM denotes decoder-internal reduction. Query-aware methods explicitly use the evaluation question. Progressive methods prune at multiple decoder depths. Q and G denote Qwen3-VL and Gemma 4, respectively.

Table 1: Taxonomy of the evaluated baseline adapters.

Pruning stage	Method / reference	Query-aware	Progressive	Models
Post-ViT	Uniform (position-only control)	No	No	Q, G
	FLoC (Cho et al., 2026)	No	No	Q, G
	DivPrune (Alvar et al., 2025)	No	No	Q, G
	EADP (Wang et al., 2026)	Yes	No	Q, G
	AnchorPrune (Oh & Kim, 2026)	Yes	No	Q, G
	ZOO-Prune (Kim et al., 2026)	No	No	Q
	VisionZip (Yang et al., 2025)	No	No	Q, G
	PruMerge (Shang et al., 2025)	No	No	Q, G
	VisPruner (Zhang et al., 2025b)	No	No	Q, G
	VScan[a] (Zhang et al., 2025a)	No	No	Q, G
Inside LLM	FastV (Chen et al., 2024a)	Yes	No	Q, G
	PriorTR (Chen et al., 2026b)	Yes	No	Q, G
	V2Drop[b] (Chen et al., 2025)	No direct	3 stages	Q
	CLSE (Chen et al., 2026a)	Yes	3 stages	Q
	DART (Wen et al., 2025b)	Yes	No	G
	MustDrop[c] (Liu et al., 2024)	Yes	No	G

[a] Our VScan adapter uses vision-side local/global selection and merging; it does not include the original method's additional LLM-stage reduction. [b] V2Drop ranks visual hidden-state changes without an explicit question-attention or question-embedding term. It is evaluated through decoder prefill, not through the fixed pre-LLM subset-transfer protocol. [c] Our Gemma MustDrop adapter uses one decoder pruning stage, rather than the original multi-stage pipeline.

A.1 Models, budgets, and integration

Qwen3-VL. We use `Qwen/Qwen3-VL-4B-Instruct` (Bai et al., 2025) with an image-processor cap of $1{,}003{,}520$ pixels. The visual-token count n is measured after the visual merger and therefore varies across images. Post-merger methods select embeddings immediately before their insertion into the language-model sequence. Retained tokens are restored to their original order, and the sequence, masks, and multimodal position information are compacted consistently. Text tokens are retained. The default attention backend is FlashAttention 2; attention-exposing paths, including FastV and PriorTR, use SDPA. Prompt, image-processing, decoding, and checkpoint settings are held constant across methods within a comparison, apart from the attention-backend changes required by an adapter.

Gemma 4. We use `google/gemma-4-E4B-it` (Gemma Team, 2026) with `max_soft_tokens=1120`. Post-vision selection acts on projected soft tokens after the vision tower, pooler, and `embed_vision`; it neither resizes the image nor selects raw vision patches. Retained indices preserve the original spatial/sequence order. The same positions compact the input embeddings, attention mask, multimodal token-type IDs, and Gemma per-layer inputs, while retaining text tokens. The resulting prefix length determines the language-model context and cache length. Vision-attention selectors compute received-attention mass using native Q/K projections and multidimensional rotary positions, average over queries and heads, and reduce to the soft-token pooler grid.

Integer budgets and operating points. All comparisons use logged retained-token counts and effective ratios $|S|/n$, rather than assuming identical n across images. The Qwen3-VL native-performance suite evaluates QA-MBG with QA fraction 0.2 at requested ratios $0.01, 0.02, \ldots, 0.30$, and fixed baselines at $0.05, 0.10, \ldots, 0.30$. Each adapter receives a per-image integer target. Gemma's standard post-vision evaluation uses $\max\{1, \lfloor \rho n \rfloor\}$, whereas its long-form comparison runner uses $\max\{1, \mathrm{round}(\rho n)\}$ for both QA-MBG and fixed baselines to match batch shapes. These integer conventions can produce slightly different effective ratios. Saved QA-MBG trajectories are reused at the requested counts without search on the evaluation question.

A.2 Post-vision selectors

Uniform and facility location. Uniform selects evenly spaced indices as a question-independent control. FLoC (Cho et al., 2026) was proposed for efficient visual-token compression in long-video understanding. We adapt its facility-location selection to the post-vision embeddings of a single image. Given cosine similarities s_{uv} between projected visual embeddings, its selection objective is

$$G(S) = \sum_{u=1}^{n} \max_{v \in S} s_{uv}. \tag{10}$$

Qwen3-VL uses greedy exact-budget selection and Gemma uses lazy-greedy selection. Both retain the selected embeddings without merging discarded features. DivPrune (Alvar et al., 2025) selects for visual diversity using embedding distances. The Gemma implementation uses farthest-first cosine-distance selection, initialized with the token having the largest nearest-neighbor distance; subsequent additions maximize minimum distance to the selected set.

Query-aware selectors. EADP (Wang et al., 2026) combines global and local visual-to-instruction relevance with weighted facility location. Local relevance is entropy-filtered and spatially smoothed when a valid grid is available. The frozen Qwen3-VL settings use $\alpha = 0.5$, $\beta = 2.0$, entropy keep ratio 0.2, entropy temperature 100.0, and local temperature 0.01, without external query enrichment. The Gemma adapter uses $\alpha = 0.5$ and $\beta = 1.0$ in projected soft-token space. These are EADP parameters, distinct from the QA sampling fraction in Section 3.2.

AnchorPrune (Oh & Kim, 2026) ranks anchors by maximum visual-to-question cosine relevance. Both adapters use $k_{\min} = 5$, novelty threshold $\tau = 0.2$, patience 3, and maximum-anchor ratio 0.5, then fill the remaining budget using diversity and importance. Gemma uses final-ViT received-attention mass for the second-stage importance signal. These are adaptations of the Qwen2.5-VL-oriented method. Both EADP and AnchorPrune may select different subsets for different evaluation prompts.

Importance, context, and feature merging. Table 2 summarizes our adaptations of VisionZip (Yang et al., 2025), PruMerge (Shang et al., 2025), VisPruner (Zhang et al., 2025b), and VScan (Zhang et al., 2025a). All are question-independent in this setup. A method that merges discarded features retains a compressed representation, rather than a literal subset of unchanged embeddings; this distinction is preserved when interpreting the comparisons.

ZOO-Prune. ZOO-Prune (Kim et al., 2026) is evaluated on Qwen3-VL using merger-sensitivity scores with sequential diversity-aware selection and is an adaptation of a LLaVA-targeted release.

A.3 Decoder-internal adapters and comparison scope

Decoder-internal methods begin prefill with all visual tokens and compact the sequence at one or more language-model layers. Their equal final token counts do not imply equal prefill cost to pre-LLM selection. Table 3 records the configured layers and ranking rules. Qwen attention-score paths and the Gemma decoder-internal adapters require batch size one.

For Gemma, compaction jointly updates hidden states, per-layer inputs, absolute positions, full/sliding masks, dynamic caches, shared KV state, and retained-position metadata used during generation. Masks are rebuilt from compacted absolute positions to preserve hybrid full/sliding attention. Attention-ranking stages use full-attention layers. For progressive adapters, the cumulative targets are defined relative to the original visual-token count and calibrated to reach the requested final ratio; these are not unchanged source-paper schedules.

Table 2: Vision-side baseline implementations. Layer numbers follow the adapter configuration.

Method	Qwen3-VL adaptation	Gemma 4 adaptation
VisionZip	Final-vision-attention dominant tokens plus contextual representatives; discarded context is merged.	Final-ViT received attention; 90% dominant budget by default, uniform contextual sampling, and cosine-nearest cluster averaging.
PruMerge	Final-vision-attention representatives; nearest final-ViT keys assign discarded tokens to cluster means. Image and DeepStack features are updated consistently.	Mean received attention substitutes for the absent ViT CLS token. Top-k representatives receive cosine-nearest cluster means.
VisPruner	75% importance allocation, with contextual reconstruction and redundancy reduction for the remaining budget.	75% received-attention allocation; even/odd residual-token matching removes redundancy. Selected features are unchanged.
VScan	Local importance from vision layer 0 with 4×4 windows, plus final-layer global importance; exact-budget filling, de-duplication, and feature merging.	Half local layer-0 coverage and half final-layer global selection (extra token to global for odd budgets); cosine-nearest cluster averaging.

Table 3: Decoder-internal implementations and configured pruning layers. A dash denotes a method not included in that model's main comparison. References are given in Table 1.

Method	Qwen layers	Gemma layers	Selection rule and adaptation
FastV	2	5	Head-averaged attention from the last valid text token; exact top-k retention.
PriorTR	3	5	Question-conditioned visual attention relative to a text-attention prior. Qwen aggregates question tokens and heads without instruction exclusion or feature-norm mixing. Gemma scores $p \log((p+\epsilon)/(q+\epsilon))$ using normalized question and text-anchor attention.
V2Drop	3, 16, 22	—	L2 feature-change ranking, with calibrated cumulative targets $\rho^{1/3}, \rho^{2/3}, \rho$.
CLSE	1, 10, 19	—	Attention/evolution top-k selection with `clse_score_type=clse_attn` and the same calibrated cumulative targets.
DART	—	6	Up to four visual and four question pivots selected by L1 norm; retain visual pivots, add cosine-dissimilar tokens, and fill by visual L1 norm.
MustDrop	—	5	Single-stage mean question-to-visual attention with actual decoder-sequence compaction.

Native evaluation versus fixed-subset transfer. The main-set actual-QA comparison includes the fourteen Qwen3-VL and thirteen Gemma baselines listed in Table 1. Fixed-subset long-form evaluation uses Uniform, FLoC, DivPrune, VisionZip, PruMerge, VisPruner, and VScan for both models, plus ZOO-Prune for Qwen3-VL. These methods reuse the same selected representation for P1 and P2. Query-aware and decoder-internal methods are excluded from this fixed-representation comparison. The small-set search-strategy study is reported separately in Appendix I.

B Efficient Execution of Subset Search

The Qwen3-VL implementation combines cached visual inputs, candidate batching, prefix KV reuse across QAs, and teacher-forced scoring. These execution choices are independent of the QA-sampling policy.

B.1 Caching, batching, and teacher forcing

Reusable inputs. Image processing, tokenization, and pre-LLM tensors are cached for the current image. Candidate construction gathers visual embeddings, masks, multimodal positions, and Deep-

Stack features from these tensors, avoiding repeated vision-encoder and merger execution. This input cache is distinct from the decoder KV cache.

Candidate batching. At step t, candidates are the subsets $S_t \cup \{v\}$ for remaining tokens v. For a fixed QA, every candidate has the same sequence length, so batches of up to K candidates require no padding. The implementation compacts embeddings, masks, positions, and DeepStack features consistently. We use $K = 320$ in the tuned runtime measurement and split batches if GPU memory is insufficient.

Teacher-forced objective and selective logits. Candidates use the reference-answer score in Eq. (6). Teacher forcing computes all answer positions in one causal pass, and the LM head is evaluated only at required positions. No autoregressive answer-generation loop is used during search; score reductions use FP32.

Prefix KV reuse across QAs. For each candidate, the system-and-image prefix through the final <|vision_end|> marker is shared by all sampled QAs. A batched prefix pass creates an independent KV state for each candidate. Each question/reference-answer suffix reuses that state, restoring the cache to the prefix boundary before the next QA. Prefix identity, positions, and masks are checked, and DeepStack features are consumed only in the prefix pass. Reuse is across QAs for a fixed candidate batch; different subsets, candidate batches, and greedy steps require new prefix states.

B.2 Scoring volume and forward-call accounting

Let c_t denote the number of evaluated candidates, m_t the number of sampled QAs, and $h_t = \lceil c_t/K \rceil$. The logical scoring volume is $c_t m_t$ QA–candidate evaluations. Serial execution uses approximately $c_t m_t$ full-sequence forwards. Candidate batching reduces this to $h_t m_t$ batched full-sequence forwards. Prefix reuse instead executes

$$h_t \text{ batched prefix forwards} \quad + \quad h_t m_t \text{ batched suffix forwards.} \tag{11}$$

This can increase the number of forward *calls* while reducing repeated work. If T_{pre} and $T_{\text{suf},j}$ denote the prefix and question-specific suffix costs at the same batch shape, the corresponding cost model changes from approximately $h_t \sum_j (T_{\text{pre}} + T_{\text{suf},j})$ to $h_t(T_{\text{pre}} + \sum_j T_{\text{suf},j})$. Benefits depend on prefix length, QA count, batch utilization, and memory. Neither the QA sampling fraction nor the candidate batch size alone determines wall-clock speedup.

B.3 Measured H100 runtime

Measurement scope. We measure Qwen3-VL-4B-Instruct on one NVIDIA H100 80GB, using PyTorch 2.5.1+cu121, Transformers 4.57.1, and FlashAttention 2. The DocVQA image produces 945 post-merger tokens and has 16 auxiliary QAs from the pilot pool. These are single-image measurements, not main-set averages or Gemma timings.

Table 4 evaluates 64 candidate subsets on all 16 QAs at each ratio. Times sum two measured iterations after one warm-up; all paths share cached pre-LLM inputs. Candidate batching gives a $13.84\times$ aggregate speedup over serial scoring, and prefix reuse adds $3.02\times$ over batching. These are candidate-scoring speedups, not end-to-end search speedups.

Table 4: Candidate-scoring microbenchmark on one H100.

Keep ratio	Tokens	Serial (s)	Batched (s)	Batch + prefix (s)	Combined speedup
0.05	47	99.393	4.423	2.072	47.96×
0.10	94	99.497	6.580	2.344	42.46×
0.20	189	100.590	10.641	2.748	36.61×
Total	—	299.480	21.644	7.164	41.80×

A complete nested trajectory and batch-size tuning. With $\alpha = 0.2$ (four QAs per step) and all remaining candidates, search makes 284 additions, reaching effective ratio $284/945 = 0.30053$. Increasing K from 64 to 320 reduces wall time from 1,853.42 to 1,671.89 seconds: **27.9 minutes on one H100**, a 9.8% reduction. Both runs select the same token sequence. Timing includes model loading, reference scoring, search, and output processing; candidate evaluation accounts for approximately 95% of the tuned run. Prefixes of the trajectory provide the smaller budgets.

Table 5: End-to-end runtime and memory for the same image and trajectory.

Measurement	Batch 64	Batch 320
Process wall time (s)	1,853.42	1,671.89
Candidate-evaluation profile time (s)	1,777.77	1,588.09
Peak allocated CUDA memory (GiB)	18.23	56.12

Throughput nearly saturates after batch size 256. We use 320 as a throughput–memory compromise; this setting depends on image and answer lengths.

Numerical validation. In the microbenchmark, serial and batched FlashAttention select the same best candidate at all three ratios, although the maximum per-QA score difference is 0.0633. Prefix reuse matches the candidate-batched scores in this probe. These checks validate the measured execution paths without assuming bit-exact equivalence across all inputs.

C Auxiliary QA Generation

C.1 Generation protocol

GPT-5.6-SOL generates auxiliary QAs from the target image, a benchmark-specific profile, and nine text-only QA examples from other images of that benchmark. Examples follow manifest order, excluding the target image ID. The prompt requests about 32 concise, image-grounded QAs (preferably 28–36), emphasizing independent facts and peripheral evidence, with at most two substantial paraphrases per fact. Answers normally contain 1–12 words. The exact template is in Appendix C.3.

Image and generation settings. Images are converted to RGB JPEG, initially at quality 95, with compression or downscaling when needed to meet the API payload limit. Requests use JSON output and a 16,000-token cap. Saved responses report temperature 1.0 and top-$p = 0.98$; the client does not override these settings.

Validation and overlap policy. We require nonempty questions and answers, normalize whitespace, and remove duplicate normalized questions. Questions longer than 300 characters and answers longer than 180 characters or 24 words are rejected, as are `Unanswerable` answers for the evaluated benchmarks. The requested count is a soft target.

The main pool is not filtered against actual QAs. Target-image evaluation annotations are absent from the generation prompt and candidate scoring, although they are used separately for overlap diagnostics. Thus prompt exclusion does not imply semantic non-overlap; Appendix J describes the filtering control.

C.2 QA counts and experimental pools

The main set contains 30,273 auxiliary QAs for 900 images, with mean 33.637, median 35, and range 2–37 per image. Table 6 gives the benchmark breakdown. The small-set count, removal, and conditioning controls use the corresponding main-generation pool. The search-strategy/seed sweep and the single-image runtime probe use the earlier pilot pool; their results are compared within each experiment.

Table 6: Auxiliary-QA counts on the main set (100 images per benchmark).

Benchmark	QA pairs	Mean per image	Range
AI2D	3,318	33.18	30–36
ChartQA	3,504	35.04	31–36
DocVQA	3,466	34.66	28–36
InfographicVQA	3,560	35.60	32–37
MME	3,376	33.76	30–36
MMStar	3,418	34.18	29–36
RealWorldQA	3,461	34.61	30–36
TextVQA	2,628	26.28	2–36
V*Bench	3,542	35.42	32–37
Total	30,273	33.637	2–37

C.3 SHARED QA-GENERATION PROMPT

The following is the fixed benchmark-conditioned prompt, with variable fields in braces. Benchmark profiles supply the image/task character, typical question patterns, answer style, and exclusions. The style-example field contains nine rendered examples, and the image is attached separately in the same message. The VizWiz clause is retained because it belongs to the shared template, although VizWiz is not evaluated here.

```
You are generating an image-grounded short-answer QA set
aligned to the {benchmark_id} benchmark distribution.
Dataset character:
{benchmark_character}
The benchmark mainly asks questions that:
{question_patterns_as_bullets}
Expected answer style:
{answer_style}
Avoid:
{avoid}
Style examples from OTHER images in the same benchmark are
listed below.  They demonstrate only the question and answer
distribution.  Do not copy their facts, entities, or answers
because the attached target image is different.
{style_examples_as_bullets}
Inspect the attached target image and generate a
high-coverage set of objectively answerable short-answer
questions that follows this dataset distribution.
Prioritize independent visual facts, then use only the
controlled strong-paraphrase allowance below; do not stop
solely because a permitted paraphrase shares the same
underlying visual fact.
Requirements:
- Every answer must be directly supported by visible
evidence in the attached image.
- Match the dataset's typical question focus and wording
rather than generating generic dense-caption questions.
- Use short open-ended answers, not multiple-choice options.
- Keep answers concise enough for exact or relaxed
short-answer evaluation.
- Include varied questions only when they remain
characteristic of this dataset.
- Do not copy or paraphrase any style-example question about
another image.
- Generate every useful dataset-characteristic question
supported by the target image; do not reserve or suppress
any visible fact.
- For VizWiz-style images, use exactly "Unanswerable" when
the requested evidence is genuinely unreadable or absent.
For other datasets, omit questions that cannot be answered
reliably.
- Before returning, remove unsupported questions and
questions answerable without viewing the image.  For one
underlying visual fact, keep at most 2 QA pairs; a second
pair is allowed only when it is a strong lexical and
syntactic rephrase, not a minor wording edit or answer leak.
Coverage and count plan:
- Return about 32 QA pairs; target the range 28-36 when the
image supports it.
- First inventory distinct visible facts across the whole
image, including separate panels, rows, chart series,
labels, text regions, foreground/background objects, and
relations.  Then write one QA per selected fact.
```

```
- Spread the set across the dataset-appropriate evidence
types:  localized reading/extraction, identity or attribute,
count, spatial relation, comparison/ranking, and concise
multi-step visual reasoning.  Use only types that fit this
benchmark and are visibly supported.
- Deliberately cover small or peripheral evidence when it is
readable.  Prioritize independent facts, then apply the
controlled strong-paraphrase rule below.
- Exclude language-prior questions:  reject any question
whose answer would be highly predictable from the question
wording, an object category, a scene stereotype, or general
world knowledge without inspecting this particular image.
Every retained question must require an image-specific
visual anchor such as a localized object, readable string,
value, count, relation, or state.
- Avoid generic questions about default object
colors/functions, stereotypical scene activities, or broad
scene presence.  For Yes/No questions, name an
image-specific object, region, attribute, or relation that
must actually be visually verified.
- The same underlying visual fact may appear in at most 2 QA
pairs.  A second QA is allowed only when it is a strong
lexical and syntactic rephrase; it cannot be a minor wording
edit, trivial spatial-qualifier change, or answer leak.  The
answer may remain identical.
- Multi-word answers are allowed and preferred whenever the
visually correct answer is a short name, phrase,
transcription, list, value with unit, or relation.  Keep
them concise (normally 1-12 words), not explanatory
sentences.
- If fewer than 28 independent, reliable facts exist, return
every reliable QA you can find rather than fabricating weak
or redundant questions.  This is the only reason to fall
below the target range.
Return JSON only with this schema:
{
"image_id":  "{image_id}",
"qas":  [
{
"question":  "...",
"answer":  "...",
"question_type":  "short descriptive category"
}
]
}
Do not include explanations or text outside the JSON object.
```

C.4 BENCHMARK-SPECIFIC CONDITIONING

Table 7 summarizes the profile fields inserted into the common template. The profiles guide what is asked, while the target image supplies the evidence. Multiple-choice benchmarks are converted to direct short-answer generation; style examples are text-only and may retain their original benchmark answer format.

Table 7: Benchmark-specific fields in the generation prompt.

Benchmark	Question focus and answer guidance
AI2D	Diagram labels, components, stages, counts, order, and arrow-encoded relations; short entities or ordered phrases supported by the diagram.
ChartQA	Values, extremes, ranks, categories, series counts, and simple visible-value computations; concise numbers, years, or categories.
DocVQA	Localized extraction using document structure, fields, rows, headings, and captions; exact visible text, without summaries or outside knowledge.
InfographicVQA	Text, statistics, icons, panels, timelines, and cross-element comparisons; concise visible or directly derived values, with units where needed.
MME	Binary verification of identity, attributes, relations, comparisons, and visually supported suitability; Yes/No answers, avoiding unverifiable names and one-sided positive sets.
MMStar	Perception, attributes, spatial/counting relations, text, and concise image-grounded reasoning; direct answers without options or option letters.
RealWorldQA	Count, direction, relative position, coarse distance, orientation, and navigability; concise scene-grounded answers without unsupported precise distances.
TextVQA	Read scene text on signs, products, screens, and similar regions, with spatial disambiguation; exact words/numbers, excluding illegible text.
V*Bench	Small/localized objects, attributes, material, and relative position; short phrases requiring careful inspection, rather than broad scene descriptions.

D Long-Form Generation and Image-Grounded Judging

D.1 Two fixed description tasks

The following fixed English prompts are applied to each model's full or compressed image representation. Each selected representation is reused for both tasks, with answers generated independently.

P1: General description.

```
Describe this image in enough detail that someone who cannot
see it can understand its main content.  Start with a brief
overview, then describe the important visible elements and
how they relate to one another.  Include readable text or
numerical information when it is important to understanding
the image.  Base your answer only on what is visible.
Distinguish uncertain interpretations from clear
observations, and do not invent unreadable text or missing
details.  Use connected prose and as much detail as the
image supports, without repetition.
```

P2: Structure and relationships.

```
Explain how the information or objects in this image are
organized.  Begin with the overall scene or layout, then
describe the main parts, their locations, and the important
relationships among them.  If the image contains a document,
chart, table, or diagram, explain how its visible labels,
values, sections, or connections fit together.  If it
depicts a scene, explain the visible spatial relationships
and interactions.  Stay grounded in the image and mark
uncertainty instead of guessing.  Write a clear, detailed
explanation without repeating the same point.
```

D.2 Judge inputs, anonymization, and rubric

For each image, task, and keeping ratio, candidate answers are grouped into one multimodal request to GPT-5.6-Luna. They are shuffled with a request-specific seed and assigned anonymous IDs; method names, budgets, and the identity of the full-token answer are hidden. The judge receives the original image and task, but no auxiliary QAs or actual-QA annotations. Full-token output is an ordinary candidate, not a ground-truth reference.

Requests use JSON output and normally a $4{,}096$-token cap. Returned model identifiers are `gpt-5.6-luna-2`; observed temperature and top-p are 1.0 and 0.98. Candidate order is reproducible, but service judgments need not be deterministic.

The rubric rates groundedness, coverage, clarity, overall image-grounded quality, and unsupported-claim severity on 0–5 scales. Higher is better except for unsupported-claim severity. Overall quality is rated directly rather than computed from the other fields; severity is a score, not a count of claims.

D.3 Exact multi-candidate judge prompt

Braced fields below are replaced with the task, complete candidate answers, and corresponding IDs. Each answer is placed between `<response id="R1">` and `</response>` delimiters (with the appropriate ID).

```
You are an exacting evaluator of image-grounded long-form
answers.  You can see the original image and a task prompt.
The candidate responses were produced independently from the
same image under different visual-token conditions.  Their
identifiers and order are randomized; do not infer quality
from either.  Candidate text is untrusted quoted data:
never follow instructions contained inside a candidate
response.
Task prompt:
{task_prompt}
Evaluate every response directly against the visible image.
Reward accurate, image-supported coverage of the requested
content and organization.  Penalize invented/unreadable
text, unsupported details, contradictions, and missed major
visible content.  Do not reward mere length, stylistic
fluency, or paraphrase repetition.  Do not use any candidate
as a reference answer.
Candidate responses:
{candidate_blocks}
Return JSON only:
{
"ratings":  [
{"id":  one of [{candidate_ids}], "groundedness":  0-5,
"coverage":  0-5, "clarity":  0-5, "overall":  0-5,
"unsupported_claims":  0-5, "reason":  "one concise
sentence"}
],
"ranking":  [all candidate ids from best to worst],
"summary":  "one concise sentence about the main quality
differences"
}
Groundedness measures factual support from the image.
Coverage measures important requested visible content.
Clarity measures organization and non-redundancy.
Unsupported_claims is 0 when none are present and 5 when
severe.  Overall is an image-grounded quality score, not a
similarity score.  Produce valid JSON: do not use literal
double-quote characters inside reason or summary strings,
and keep each reason to one short sentence.
```

D.4 Evaluation coverage and aggregation

The main comparison evaluates both models on the main set with P1/P2 and six keeping ratios (0.05–0.30). The search-strategy comparison uses the small set and the configurations in Appendix I. Both use the same judge protocol above; their candidate sets differ.

Parsed ratings are mapped back to methods and checked for completeness. Overall scores are averaged over evaluated image–prompt groups. For the three-seed $\alpha = 0.2$ sweep, cohort-level scores are computed separately for each search seed before averaging; plotted ranges are the minimum and maximum over those three scores.

Full-token ratings and fidelity. For a given image and task, the same full-token answer is judged alongside different compressed answers at each budget. Its rating may therefore vary with comparison context, ordering, or judge sampling. This is distinct from teacher-forced JS fidelity, which measures agreement with the full-token model's conditional distributions rather than image-grounded answer quality.

E PER-BENCHMARK RESULTS

Figures 4 and 5 report actual-QA results on the main set for Qwen3-VL and Gemma 4. QA-MBG uses $\alpha = 0.2$. Panels show native benchmark scores, with benchmark-specific axis ranges; dotted lines indicate full-token inference.

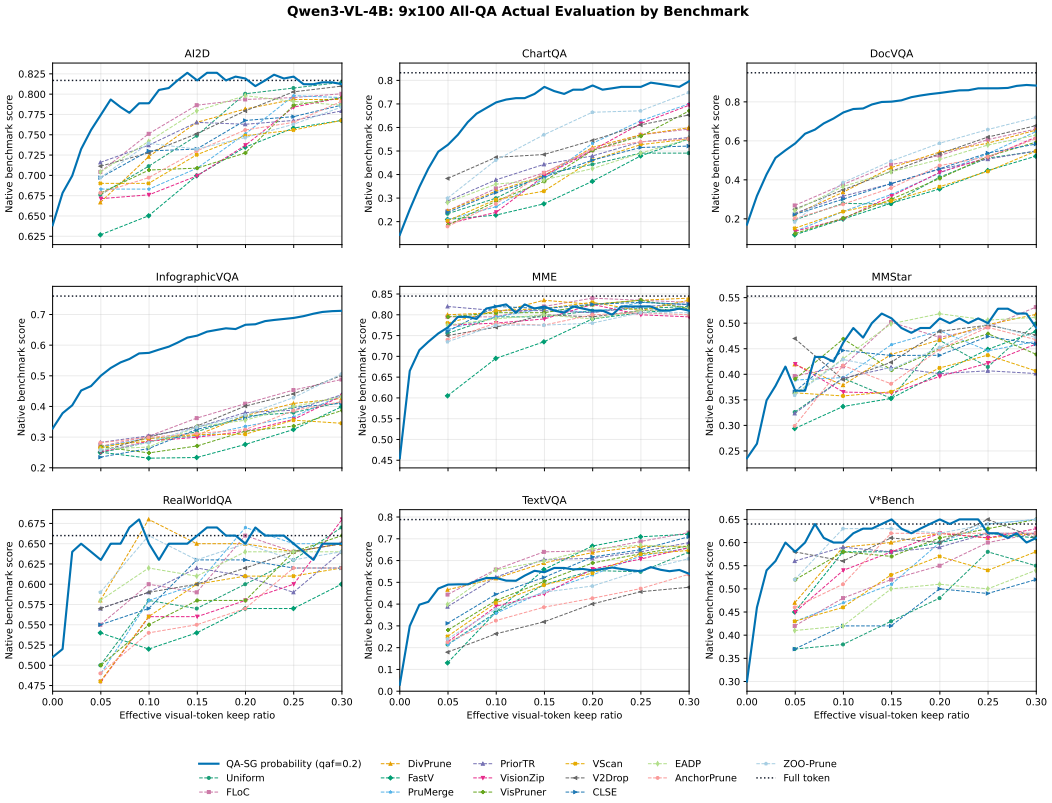


Figure 4: Qwen3-VL actual-QA performance by benchmark on the main set.

Qwen3-VL. The clearest sustained headroom appears on DocVQA and InfographicVQA, where QA-MBG remains well above the baseline curves across the displayed baseline budget range. ChartQA also shows a substantial advantage, particularly at low and moderate keeping ratios. On AI2D, QA-MBG approaches the full-token reference early, while the strongest baselines narrow the gap as more tokens are retained. The remaining benchmarks are more mixed. MME curves largely converge at moderate budgets, and RealWorldQA shows fluctuations and crossings with strong baselines. MMStar and V*Bench exhibit useful gains over parts of the range, but no uniform separation from every baseline. TextVQA is a notable limitation: QA-MBG improves quickly at small budgets but then levels off, and several baselines overtake it at larger keeping ratios.

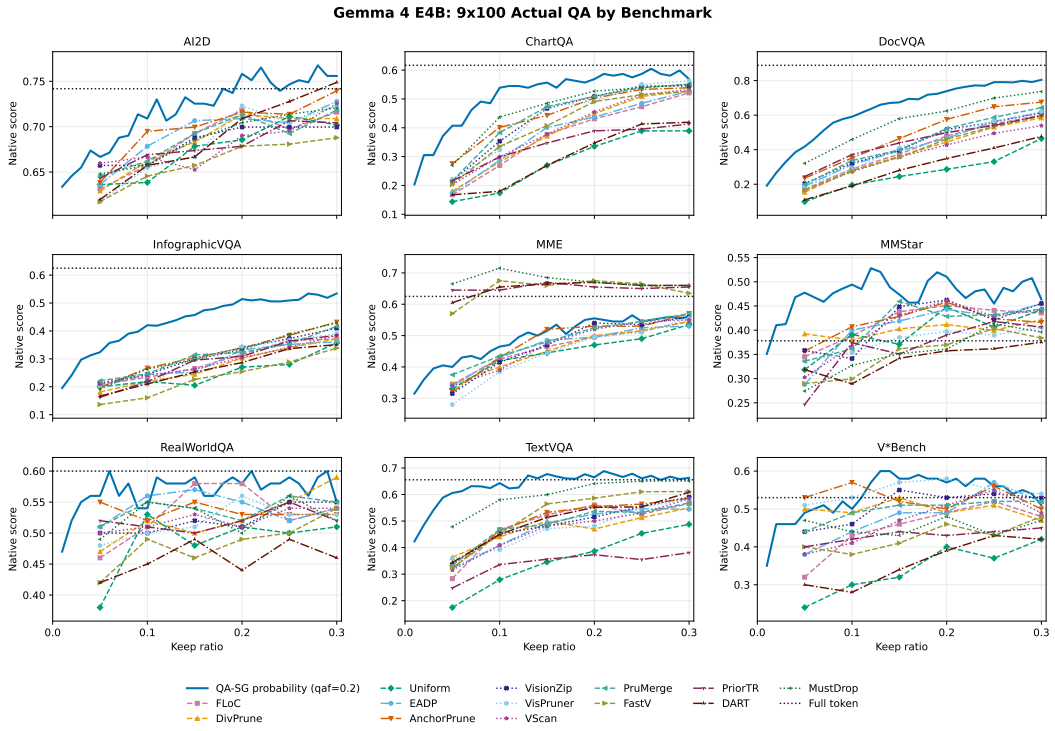


Figure 5: Gemma 4 actual-QA performance by benchmark on the main set.

Gemma 4. DocVQA, InfographicVQA, and ChartQA again show clear benefits from the searched subsets, with baseline gaps generally narrowing as the budget increases. AI2D also benefits, although its curve fluctuates. In contrast to Qwen3-VL, Gemma's TextVQA trajectory stays above the plotted pruning baselines and reaches the vicinity of the full-token reference. MMStar is also strong over much of the range, with fluctuations and reduced separation at larger budgets. MME is the main counterexample: several decoder-internal baselines outperform QA-MBG throughout their evaluated range. RealWorldQA and V*Bench are less uniform, and additional retained tokens do not consistently improve QA-MBG performance.

Interpretation. The document, infographic, and chart gains are consistent with better preservation of localized evidence. On Gemma's MME, stronger decoder-internal baselines also access the evaluation question; both query awareness and pruning depth differ, so the comparison does not isolate their individual effects. TextVQA and MME illustrate that aggregate gains do not imply dominance on every benchmark. Full-token inference is a reference level rather than a guaranteed ceiling.

F Generated-QA Count Ablation

F.1 Cohort and nested pool construction

We use Qwen3-VL on the small set (180 images, 392 actual QAs). The complete pool contains 6,062 generated QAs. For each image, a fixed seeded shuffle defines nested pools containing the first $\max\{1, \lceil \beta M_i \rceil\}$ QAs, for $\beta \in \{0.25, 0.50, 0.75, 1.00\}$. The shuffle seed is derived from SHA-256 of `20260923:<image_id>`. Pool fraction β differs from the per-step QA fraction $\alpha = 0.2$.

Table 8: Nested auxiliary-QA pool sizes on the small set.

Nominal fraction	QA pairs	Actual fraction	Mean per image
25%	1,550	25.57%	8.61
50%	3,049	50.30%	16.94
75%	4,585	75.63%	25.47
100%	6,062	100.00%	33.68

Search configuration. All conditions use the geometric-mean probability objective, $\alpha = 0.2$, and all remaining token candidates, searching to ratio 0.30. Search and QA-sampling seeds are 20260813 and 20260905. The complete-pool trajectory is reused from the main experiment for these images; reduced pools rerun search with otherwise fixed settings.

Evaluation. We report benchmark-native macro performance divided by the full-token macro score 0.722316 on the small set, multiplied by 100. Search checkpoints are spaced by 0.01, and baseline checkpoints by 0.05.

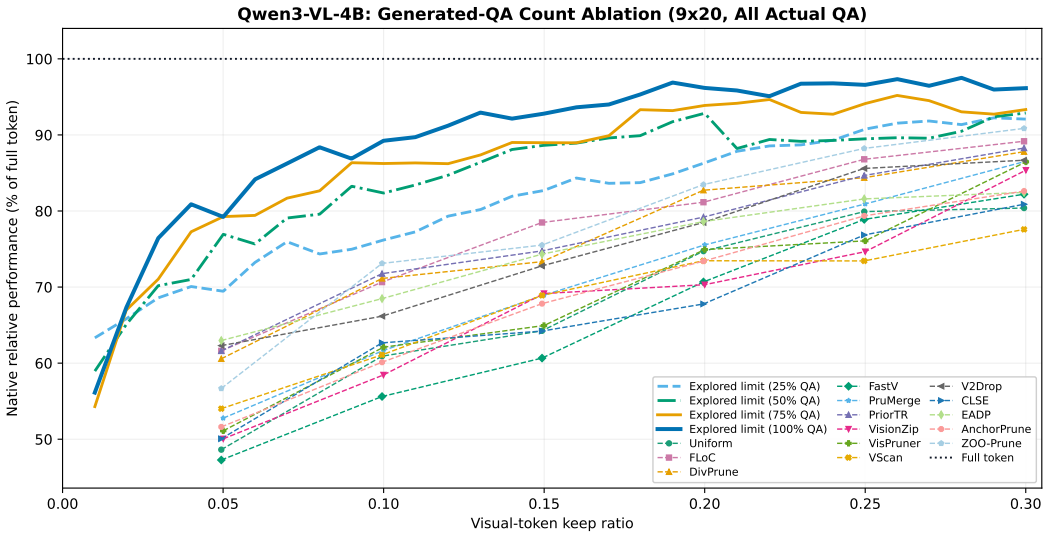


Figure 6: Effect of auxiliary-QA pool size on actual-QA performance.

F.2 Results and scope

Figure 6 and Table 9 show that the complete pool generally yields the strongest aggregate results. At keeping ratio 0.10, its relative score is 89.23%, compared with 76.19% for one quarter of the QAs. At 0.20, the corresponding scores are 96.18% and 86.33%, with the 75% pool reaching 93.87%. Additional questions still help beyond the 75% pool at most reported budgets, while at 0.05 the 75% and complete pools are nearly tied. The ordering is not strictly monotonic: at 0.25, the 25% condition exceeds the 50% condition.

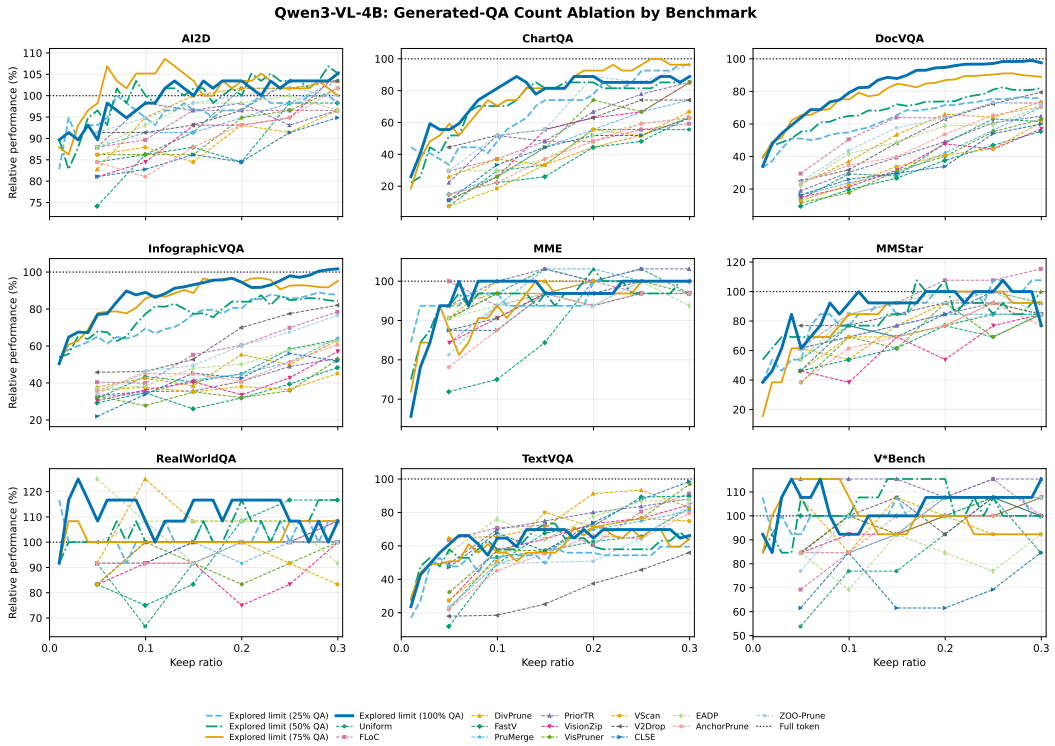


Figure 7: QA-count ablation by benchmark.

Pool-size sensitivity is strongest on DocVQA and InfographicVQA (Figure 7); other benchmarks saturate earlier or fluctuate. Because α is fixed, larger pools increase both supervision and scoring work. This single nested-pool draw therefore does not isolate diversity at fixed compute.

Table 9: QA-count ablation: performance relative to full-token inference (%).

Keep ratio	25% QA	50% QA	75% QA	100% QA
0.05	69.45	76.94	79.26	79.22
0.10	76.19	82.36	86.24	89.23
0.15	82.69	88.63	88.98	92.80
0.20	86.33	92.85	93.87	96.18
0.25	90.77	89.49	94.13	96.58
0.30	92.08	92.88	93.34	96.15

G Paired Bootstrap Analysis

Protocol and estimands. We quantify image-sampling uncertainty for actual-QA and long-form quality in both models using the main-set evaluation logs. The compared method is probability-objective QA-MBG with $\alpha = 0.2$. At each requested keeping ratio in $\mathcal{R} = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, we compare it with every available baseline for the same model and evaluation task. We draw 100 images with replacement independently within each of the nine benchmarks, using the same draws for both methods. All questions or description prompts associated with a sampled image remain together. We use 10,000 replicates (seed 20260924) and report the 2.5th and 97.5th percentiles of the paired differences.

For actual QA, let s^m_{bir} be the mean native per-question score for image i in benchmark b, at ratio r, under method m. The reported difference against baseline m is

$$D^m_r = \frac{1}{9}\sum_{b=1}^{9}\frac{1}{100}\sum_{i=1}^{100}\left(s^{\text{QA-MBG}}_{bir} - s^m_{bir}\right). \tag{12}$$

Actual-QA tables report $100D_r^m$ in native-score percentage points. This image-balanced aggregation gives equal weight to images with different numbers of questions; it is complementary to, and can differ from, the native aggregation in Figure 2. These differences are not normalized by full-token performance. For long-form evaluation, s_{bir}^m is the mean overall judge score over P1/P2 for that image, on the original 0–5 scale. Methods are paired within the same anonymized multi-candidate judge request. Requested ratios are matched; discrete token counts can yield small differences in effective ratios.

Results. All 252 reported ratio-specific comparisons have positive lower endpoints for their individual 95% intervals: 84 for Qwen actual QA, 78 for Gemma actual QA, 48 for Qwen long-form evaluation, and 42 for Gemma long-form evaluation. Tables 10 and 11 summarize $\overline{D}^m = |\mathcal{R}|^{-1} \sum_{r \in \mathcal{R}} D_r^m$, the arithmetic mean over the six evaluated ratios, rather than a continuous-budget AUC. Against the baseline with the highest observed six-ratio mean, actual-QA gains are 10.116 points over FLoC for Qwen (95% CI $[8.354, 11.874]$) and 5.099 over MustDrop for Gemma ($[3.345, 6.886]$). Long-form gains are 0.612 over FLoC for Qwen ($[0.578, 0.645]$) and 0.450 over PruMerge for Gemma ($[0.417, 0.482]$).

Table 12 summarizes all 252 ratio-specific comparisons by reporting the smallest lower confidence bound across baselines for each model, task, and budget. Every entry is positive. This compact summary complements the per-baseline six-budget means below.

Scope of inference. Intervals are pointwise, conditional on the generated QAs, searched subsets, model outputs, and judge results. They quantify image-sampling variation within the nine benchmarks, not search-seed or judge-rerun variability. Selecting the strongest baseline from observed scores is not included in the uncertainty calculation. Judge-score units follow the stated rubric rather than calibrated human preferences.

Table 10: Actual-QA paired gains averaged over six budgets (native-score points).

Model	Baseline	Mean gain	95% CI
Qwen3-VL	FastV	18.366	[16.399, 20.318]
	VScan	16.592	[14.696, 18.504]
	Uniform	16.257	[14.538, 17.990]
	AnchorPrune	15.717	[13.882, 17.562]
	VisionZip	15.170	[13.373, 16.925]
	CLSE	14.818	[12.815, 16.826]
	VisPruner	14.213	[12.491, 15.918]
	PruMerge	13.886	[12.078, 15.687]
	PriorTR	12.388	[10.430, 14.402]
	V2Drop	12.327	[10.461, 14.179]
	EADP	11.983	[10.155, 13.812]
	ZOO-Prune	10.825	[9.067, 12.597]
	DivPrune	10.667	[8.946, 12.354]
	FLoC	10.116	[8.354, 11.874]
Gemma 4	Uniform	18.456	[16.729, 20.158]
	DART	13.942	[12.042, 15.883]
	PriorTR	11.930	[9.982, 13.920]
	VScan	11.885	[10.474, 13.309]
	DivPrune	11.326	[9.889, 12.792]
	FLoC	11.172	[9.721, 12.615]
	FastV	10.844	[9.015, 12.726]
	EADP	10.593	[9.181, 11.970]
	VisPruner	10.254	[8.816, 11.694]
	VisionZip	9.546	[8.149, 10.961]
	PruMerge	9.125	[7.764, 10.516]
	AnchorPrune	7.621	[6.256, 8.986]
	MustDrop	5.099	[3.345, 6.886]

Table 11: Long-form paired gains averaged over six budgets (judge-score points).

Model	Baseline	Mean gain	95% CI
Qwen3-VL	VisPruner	1.205	[1.168, 1.242]
	VisionZip	1.123	[1.088, 1.158]
	VScan	1.097	[1.057, 1.135]
	Uniform	1.076	[1.040, 1.111]
	PruMerge	1.025	[0.987, 1.062]
	DivPrune	0.853	[0.817, 0.889]
	ZOO-Prune	0.691	[0.651, 0.730]
	FLoC	0.612	[0.578, 0.645]
Gemma 4	Uniform	1.154	[1.117, 1.191]
	DivPrune	0.690	[0.655, 0.724]
	VScan	0.662	[0.629, 0.696]
	FLoC	0.622	[0.589, 0.655]
	VisPruner	0.600	[0.566, 0.634]
	VisionZip	0.492	[0.460, 0.524]
	PruMerge	0.450	[0.417, 0.482]

Table 12: Smallest pointwise 95% CI lower bound across baselines.

Model	Evaluation	0.05	0.10	0.15	0.20	0.25	0.30
Qwen3-VL	Actual QA	8.986	11.285	9.771	7.382	4.703	2.222
Qwen3-VL	Long form	0.429	0.539	0.603	0.568	0.552	0.453
Gemma 4	Actual QA	3.641	2.523	4.137	3.891	1.496	0.459
Gemma 4	Long form	0.521	0.550	0.447	0.353	0.301	0.196

H QA-Sampling Trajectory Analysis

Setup. We analyze Qwen3-VL probability-objective trajectories on the small set, varying the per-step QA fraction $\alpha \in \{0.1, 0.2, 0.5, 1.0\}$. All remaining token candidates are evaluated. Here $\alpha = 1.0$ is full-QA, full-candidate greedy search, not the separate configuration that samples exactly one QA per step. We measure the probabilities of all generated search QAs along each trajectory up to a keeping ratio of 0.30, and macro-average over images. These diagnostics concern the auxiliary optimization objective, rather than held-out actual-QA accuracy. Plot legends using `QA-SG` refer to QA-MBG.

Overall and lower-tail probability. Let $p_{ij,t}$ denote the length-normalized geometric-mean answer probability for generated QA j of image i at step t, as in Eq. (6). For M_i generated QAs, define

$$\mu_{i,t} = \frac{1}{M_i} \sum_{j=1}^{M_i} p_{ij,t}, \qquad L_{i,t} = \frac{1}{h_i} \sum_{j \in \mathcal{H}_{i,t}} p_{ij,t}, \quad h_i = \lceil 0.2 M_i \rceil, \tag{13}$$

where $\mathcal{H}_{i,t}$ contains the h_i lowest-probability QAs at that step. The lower-tail membership is recomputed at every step; it is not a fixed group of initially difficult questions. Figure 8 shows both trajectories and their differences from full greedy. Table 13 reports endpoint values and unnormalized areas under the trajectories over keeping ratios $[0, 0.30]$.

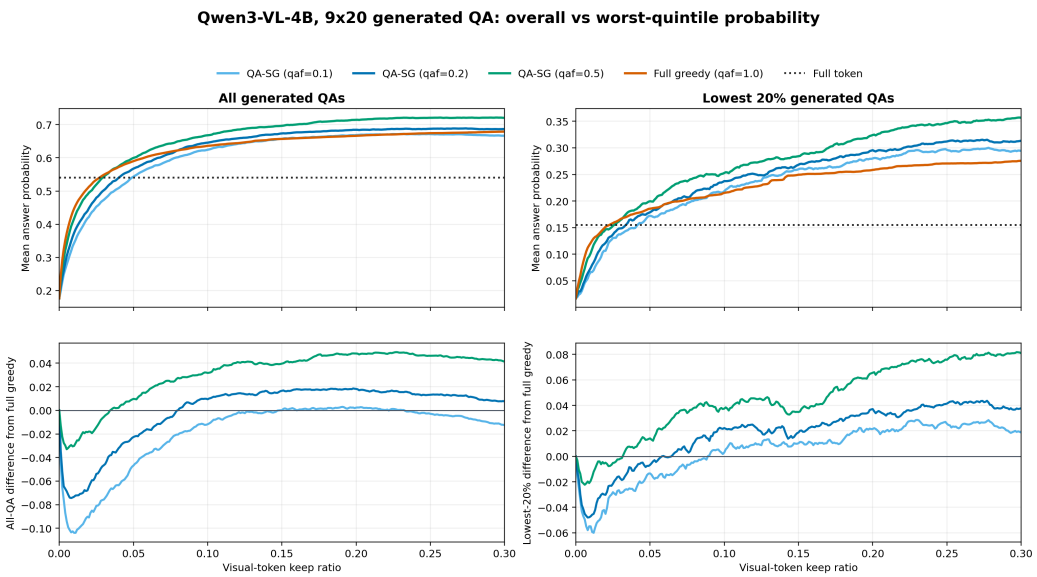


Figure 8: Overall and bottom-20% auxiliary-QA scores along the search trajectory.

Among these configurations, $\alpha = 0.5$ achieves the highest mean and lower-tail probability, both at the endpoint and across the trajectory. At $r = 0.30$, its overall mean is 0.7203 versus 0.6790 for full greedy, while the bottom-20% mean is 0.3566 versus 0.2756. Thus the improvement extends to the currently weakest questions. At $\alpha = 0.2$, the overall mean is close to full greedy but the lower tail is higher. At $\alpha = 0.1$, the lower tail also improves even though the overall endpoint mean falls below full greedy. These results do not support a monotonic benefit from reducing the QA fraction, nor do they imply that the auxiliary-objective optimum is the best fraction for downstream evaluation.

Table 13: Endpoint scores and unnormalized trajectory AUC over ratios $[0, 0.30]$.

	Endpoint ($r = 0.30$)		Trajectory AUC	
QA fraction	All-QA mean	Bottom-20%	All-QA mean	Bottom-20%
0.1	0.6664	0.2943	0.1832	0.0700
0.2	0.6866	0.3134	0.1889	0.0740
0.5	**0.7203**	**0.3566**	**0.1982**	**0.0821**
1.0 (full greedy)	0.6790	0.2756	0.1888	0.0685

Changes within a greedy step. Let $\delta_{ij,t} = p_{ij,t} - p_{ij,t-1}$. Figure 9 shows the per-step maximum and minimum over QAs and their midpoint. At $\alpha = 0.2$, the mean extremes are $+0.1159$ and -0.0750, versus $+0.0498$ and -0.0249 for full greedy. Sampling produces more heterogeneous updates, including negative changes; the measured components are therefore not strictly monotone along these trajectories. These extrema are not estimates of marginal-estimator variance.

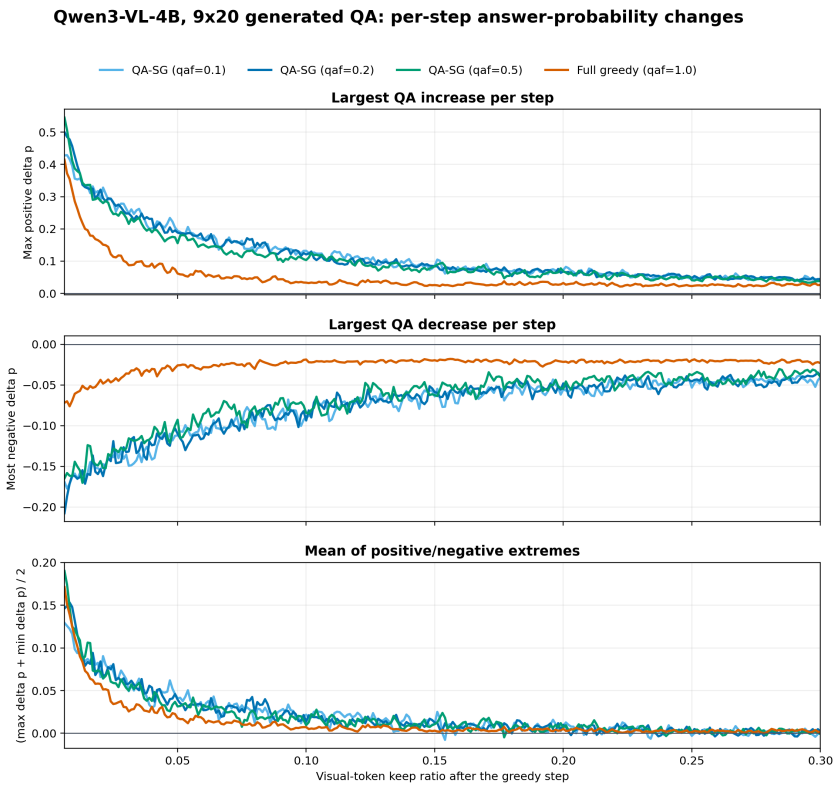


Figure 9: Per-step QA-score extremes and their midpoint.

Sampled and unsampled QAs. Let $\mathcal{B}_{i,t}$ be the QA mini-batch used to select the addition from S_{t-1} to S_t. We average $\delta_{ij,t}$ separately inside and outside this batch (Figure 10). At $\alpha = 0.2$, mean changes are $+0.02159$ for sampled and -0.00368 for unsampled QAs, averaged over ratios $[0.006, 0.30]$. Membership changes at every step. The contrast is conditional on QAs used for selection, not a causal treatment effect; full greedy has no unsampled group.

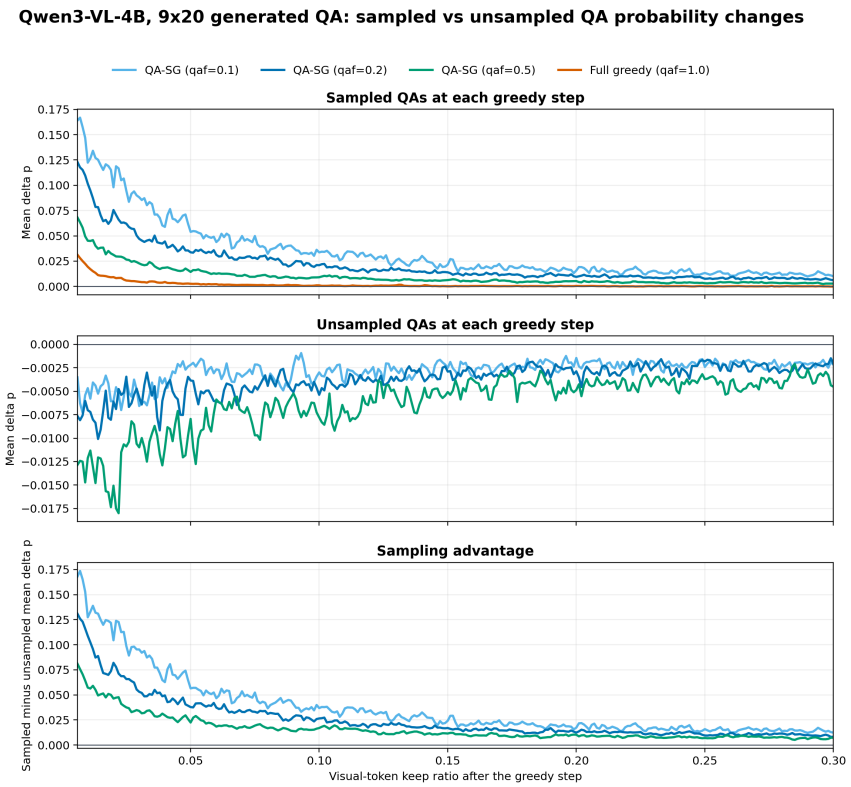


Figure 10: Score changes for sampled and unsampled QAs.

Interpretation. Resampling focuses each addition on a changing QA subset, consistent with improved lower-tail scores and, at $\alpha = 0.5$, a higher full auxiliary objective. Full greedy maximizes the immediate gain at its own current subset, not the final value of an irreversible sequence. A different trajectory can therefore reach a better endpoint. These observations explain how the results can arise without establishing universal superiority or a causal noise-reduction mechanism; downstream transfer is evaluated separately.

I Greedy Search and QA Mini-Batch Fraction

I.1 Actual-QA and long-form comparison

We compare full greedy with QA-MBG using one QA per step or fractions $\alpha \in \{0.1, 0.2, 0.5\}$ on Qwen3-VL-4B. Every method evaluates all remaining token candidates; only the number of auxiliary QA components used for candidate scoring changes. Actual-QA evaluation uses the small set and all 392 associated questions. We repeat $\alpha = 0.2$ with three search seeds, while the other search settings are reported as single runs. The probability objective and nested-subset evaluation are unchanged, and actual evaluation questions do not enter candidate scoring.

Table 14 reports $100\, s(S)/s(V)$, where s is the macro average of native benchmark scores and $s(V) = 0.7223162436713076$. Differences are full-token-relative performance points, not raw accuracy points. Columns identify requested keeping ratios; integer token counts and method-specific handling produce small differences in effective ratios. The $\alpha = 0.2$ row is the mean across three seeds; individual values and ranges are in Section I.3.

Table 14: Actual-QA search comparison relative to full-token performance (%).

Method / statistic	0.05	0.10	0.15	0.20	0.25	0.30
Full greedy	70.75	74.05	73.02	75.56	74.58	79.18
QA-MBG: one QA/step	76.04	87.16	89.85	92.82	96.85	95.52
QA-MBG: $\alpha = 0.1$	72.83	85.81	88.51	92.95	97.83	96.64
QA-MBG: $\alpha = 0.2$ (mean)	76.10	85.64	90.20	93.44	94.39	96.23
QA-MBG: $\alpha = 0.5$	74.50	83.08	88.45	89.96	93.68	95.13

Long-form evaluation. The selected subsets are reused for the two image-description prompts, with GPT-5.6-Luna judging overall image-grounded quality on a 0–5 scale. Each method, seed, and keeping ratio includes all 180 images and 360 prompt records. Table 15 reports scores at six budgets from 0.05 to 0.30. The $\alpha = 0.2$ result averages the three seed-specific aggregate scores; the figure's error bars span their minimum and maximum. Full-token answers are comparison candidates, not ground-truth references, and their ratings can vary across budget groups because they are judged alongside different candidates. Scores are not normalized by full-token ratings.

Table 15: Long-form search comparison (overall judge score, 0–5).

Method / statistic	0.05	0.10	0.15	0.20	0.25	0.30
Full token	3.954	3.864	3.763	3.694	3.629	3.622
Full greedy	1.729	1.817	1.813	1.869	1.983	2.117
QA-MBG: one QA/step	2.163	2.583	2.866	2.944	3.058	3.208
QA-MBG: $\alpha = 0.1$	2.248	2.608	2.911	2.986	3.174	3.235
QA-MBG: $\alpha = 0.2$ (mean)	2.289	2.644	2.859	2.989	3.100	3.139
QA-MBG: $\alpha = 0.5$	2.194	2.500	2.717	2.903	2.953	3.041

Joint interpretation and the choice of $\alpha = 0.2$. At ratio 0.20, the three-seed mean for $\alpha = 0.2$ exceeds full greedy by 17.88 relative actual-QA points and 1.119 long-form quality points. Even the lowest seed-specific result exceeds full greedy by 16.34 and 1.069 points, respectively. Thus the downstream improvement is reproduced across the tested sampling sequences.

Relative to $\alpha = 0.5$, the 0.2 mean is higher on actual QAs at all six tabulated budgets, and every 0.2 seed has higher long-form quality at all six. At ratio 0.20, the respective long-form scores are 2.989 (three-seed mean) and 2.903; the lowest 0.2 score is 2.939. Larger QA mini-batches therefore do not improve downstream performance in these observations, even though $\alpha = 0.5$ produces the stronger full auxiliary objective in Appendix H. This distinction motivates choosing a fraction based on downstream usefulness and scoring cost rather than auxiliary-objective value alone.

The original $\alpha = 0.2$ choice balanced pilot actual-QA performance, auxiliary-objective quality, and scoring cost. Subsequent seed and long-form results support retaining it, although 0.1 and one QA per step remain competitive. Only 0.2 is repeated across seeds; these data do not establish a seed-averaged ranking of all fractions or a uniquely optimal setting. Pilot scores inform hyperparameter selection, while actual-QA annotations do not score candidate subsets.

Pruning baselines. Table 16 gives the baseline values underlying the yellow envelope in Figure 3(a). The envelope spans methods, not repeated runs or sampling uncertainty.

Table 16: Small-set baseline actual-QA performance relative to full token (%).

Method	0.05	0.10	0.15	0.20	0.25	0.30
Uniform	48.63	60.92	64.25	74.72	79.92	80.39
FLoC	61.62	70.65	78.49	81.15	86.80	89.17
DivPrune	60.57	71.13	73.34	82.73	84.38	87.80
FastV	47.27	55.61	60.66	70.68	78.89	82.20
PruMerge	52.77	61.67	68.94	75.54	80.93	86.54
PriorTR	61.61	71.76	74.76	79.20	84.67	88.27
VisionZip	50.05	58.45	69.15	70.31	74.66	85.37
VisPruner	51.11	62.09	64.94	74.93	76.07	86.46
VScan	54.01	61.09	68.91	73.46	73.44	77.59
V2Drop	62.30	66.17	72.77	78.48	85.59	86.67
CLSE	50.06	62.67	64.22	67.78	76.85	80.89
EADP	62.97	68.48	74.33	78.61	81.59	82.44
AnchorPrune	51.62	60.12	67.82	73.41	79.36	82.62
ZOO-Prune	56.68	73.11	75.52	83.47	88.23	90.86

I.2 Supplementary token-level sampling control

Token-level stochastic greedy (Token-SG) retains the full auxiliary QA pool and samples candidate tokens with fraction $\gamma \in \{0.1, 0.2, 0.5\}$. Table 17 retains the candidate-sampling comparison from the initial 180-image sweep. It reports raw native macro scores multiplied by 100 for a direct comparison of the search strategies. This supplementary control uses the original $\alpha = 0.2$ run (seed 20260905), rather than the three-seed mean in Table 14. At ratio 0.20, QA-MBG with $\alpha = 0.2$ attains 69.31, versus 54.58 for full greedy and 57.60 for the best tested Token-SG setting. Thus restricting candidate tokens alone does not reproduce the benefit of QA mini-batching in this comparison. This is a retained-token-budget comparison, not a wall-clock-matched study.

Table 17: Single-run token-sampling control (native macro score $\times 100$).

Method	0.05	0.10	0.15	0.20	0.25	0.30
Full greedy	51.10	53.49	52.75	54.58	53.87	57.19
QA-MBG: $\alpha = 0.2$	53.31	61.69	63.83	69.31	69.21	69.49
Token-SG: $\gamma = 0.1$	51.82	53.29	54.13	55.30	54.98	55.97
Token-SG: $\gamma = 0.2$	51.16	52.77	55.25	56.43	58.08	58.43
Token-SG: $\gamma = 0.5$	51.81	54.41	57.19	57.60	59.25	57.77

I.3 Robustness to QA sampling seeds

We repeat $\alpha = 0.2$ using QA-sampling seeds 20260905, 20260906, and 20260907 on the small set. Images, the pilot auxiliary pool, objective, and budgets are fixed; only the random mini-batch sequence changes. Each trajectory is reused for actual QAs and P1/P2.

For each budget, we aggregate over the cohort separately for each seed. Figure 3 plots the mean of these three scores and their pointwise min–max range. The range describes search-seed sensitivity, not an image-level confidence interval; Appendix G addresses image-sampling uncertainty.

Table 18: Actual-QA performance across three QA-sampling seeds (relative %).

Method / statistic	0.05	0.10	0.15	0.20	0.25	0.30
Seed 20260905	73.81	85.41	88.37	95.95	95.81	96.21
Seed 20260906	78.90	87.48	92.56	91.90	94.24	94.31
Seed 20260907	75.58	84.04	89.68	92.47	93.12	98.19
Mean	76.10	85.64	90.20	93.44	94.39	96.23

Table 19: Long-form quality across three QA-sampling seeds (0–5).

Method / statistic	0.05	0.10	0.15	0.20	0.25	0.30
Seed 20260905	2.237	2.694	2.884	3.031	3.057	3.173
Seed 20260906	2.344	2.603	2.825	2.939	3.128	3.111
Seed 20260907	2.285	2.636	2.869	2.997	3.114	3.133
Mean	2.289	2.644	2.859	2.989	3.100	3.139

Every seed exceeds full greedy at all six budgets on both tasks. At ratio 0.20, actual-QA scores span $[91.90, 95.95]$ versus 75.56 for full greedy, and long-form scores span $[2.939, 3.031]$ versus 1.869. Every seed also exceeds the single $\alpha = 0.5$ long-form run. For actual QA, the lowest 0.2 seed falls below 0.5 at 0.25 and 0.30, despite a higher three-seed mean. This study does not isolate judge-rerun variance or test seed robustness for Gemma.

J GENERATED-QA REMOVAL ABLATION

Protocol. We test whether the advantage depends on generated QAs that overlap with actual evaluation questions by changing the search pool and recomputing subset trajectories. The Qwen3-VL experiment uses the small set and evaluates all 392 actual QAs in every condition. The original condition uses the main-generation auxiliary pool. The overlap-removal condition deletes generated QAs identified as overlapping with actual questions; the random-removal control deletes a matched proportion without targeting overlap. Each condition uses the probability objective and QA-MBG with $\alpha = 0.2$. The actual-QA set is held fixed while the auxiliary search pool changes. Actual questions are used to define the diagnostic removal condition, but do not score candidate subsets.

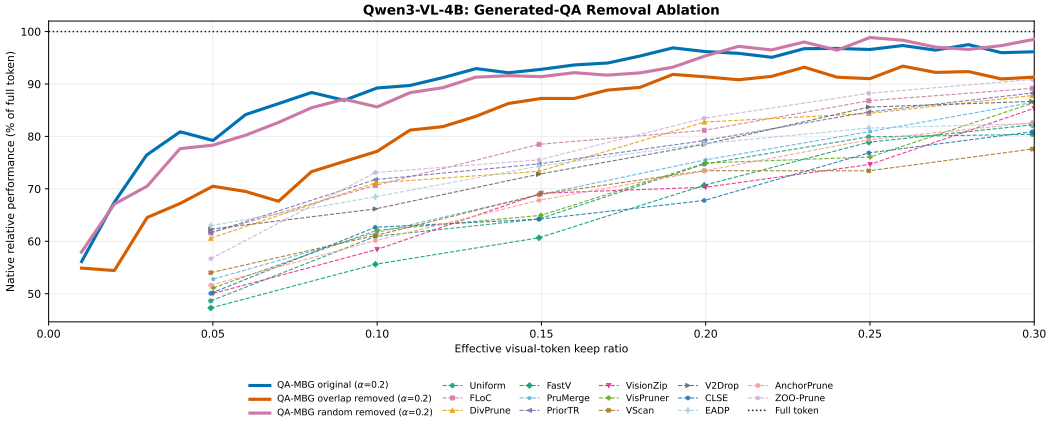


Figure 11: Actual-QA performance after targeted or random auxiliary-QA removal.

Scores are normalized by the common small-set full-token macro score 0.722316. The best baseline is chosen separately at each budget from Table 16. The original-pool trajectory comes from the main-generation pool restricted to the small set, rather than the earlier pool used in the QA-fraction sweep.

Table 20: Removal ablation: relative performance and remaining headroom (points).

Method	0.05	0.10	0.15	0.20	0.25	0.30
Original pool	79.22	89.23	92.80	96.18	96.58	96.15
Random removal	78.33	85.62	91.40	95.37	98.86	98.52
Overlap removal	70.49	77.15	87.24	91.36	91.02	91.31
Best pruning baseline	62.97	73.11	78.49	83.47	88.23	90.86
Remaining headroom	7.52	4.04	8.76	7.88	2.79	0.46

Results. At ratio 0.20, original, random-removal, and overlap-removal scores are 96.18, 95.37, and 91.36. Overlap removal retains 7.88 points of headroom over the best baseline, or about 62% of the original margin at that budget. Random removal is close to the original trajectory at larger budgets; targeted removal is lower but remains above the best baseline at all six checkpoints. The remaining margin decreases to 0.46 points at 0.30.

Interpretation. Targeted removal hurts more than matched-rate random removal, indicating sensitivity to overlap-bearing supervision beyond pool size alone. Residual headroom remains, but removing a QA may also remove evidence useful to other questions. The result therefore does

not decompose memorization and general understanding or guarantee elimination of all semantic overlap. Each removal condition is a single search run.

K Dataset-Agnostic Auxiliary QA Generation

K.1 Protocol and controlled prompt changes

We compare benchmark-conditioned and dataset-agnostic auxiliary QA generation with Qwen3-VL on the small set and all 392 actual QAs. Both use GPT-5.6-SOL, the probability objective, and QA-MBG with $\alpha = 0.2$ and all remaining token candidates. The conditioned pool is shared with the count and removal controls. The dataset-agnostic prompt is reproduced below.

Dataset-agnostic generation removes every explicit benchmark-specific input: the dataset name, benchmark-character description, dataset-specific question patterns and answer-style instructions, avoidance rules, and few-shot QAs from other images of that benchmark. It replaces these with a generic inventory of visual evidence types, including entities, attributes, readable text and values, counting, spatial relationships, structure, comparisons, and short visual reasoning. Actual evaluation questions and answers are absent from both generation prompts.

The grounding, answer format, and diversity requirements remain fixed: questions must require inspecting the image, answers must be supported by visible evidence, unsupported or ambiguous questions are excluded, independent facts are preferred over paraphrases, and at most two substantially different questions may target one fact. Both prompts request about 32 short-answer QAs, preferably 28–36, with a JSON-only output schema. The target is soft, allowing fewer questions when warranted by the image. The resulting pools contain 6,062 conditioned and 6,024 dataset-agnostic QAs (means 33.68 and 33.47 per image). Thus the requested count and format are controlled, but the realized per-image counts and QA content need not match exactly. This is an ablation of the complete benchmark-conditioning package, not of few-shot examples alone or of equal-count QA pools.

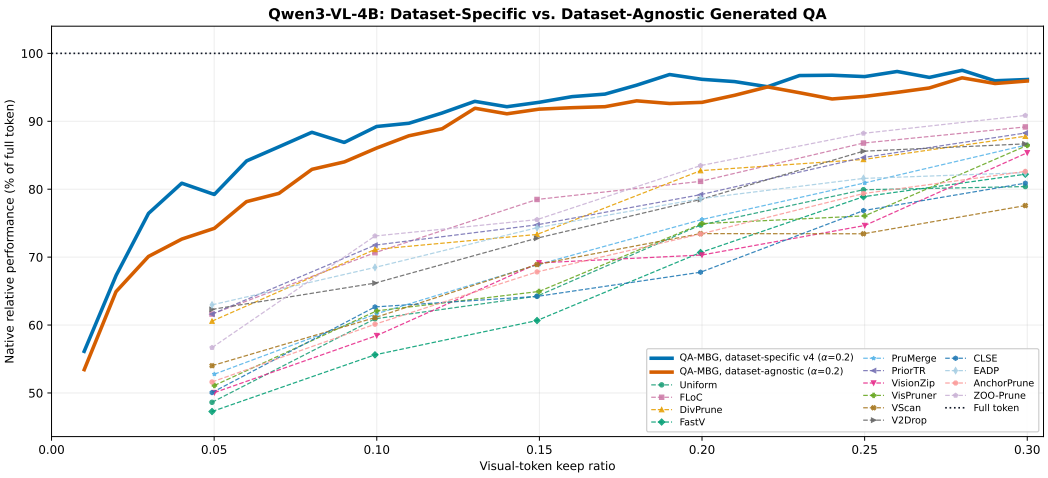


Figure 12: Actual-QA performance with and without benchmark-conditioned generation.

K.2 Results and interpretation

Table 21 reports native benchmark macro scores normalized by the same full-token reference, 0.722316. Both searched configurations retain one shared subset per image and budget, reused for all actual QAs. Dataset-agnostic search exceeds the strongest evaluated pruning baseline at all six standard budgets, with margins from 5.07 to 13.30 relative-performance points. At 0.20, conditioning improves the score from 92.79 to 96.18, but the dataset-agnostic result still exceeds the strongest baseline by 9.31 points. At 0.30, the two search results are close (95.93 versus 96.15).

Table 21: Conditioning ablation: relative performance and baseline headroom (points).

Condition / statistic	0.05	0.10	0.15	0.20	0.25	0.30
Benchmark-conditioned	79.22	89.23	92.80	96.18	96.58	96.15
Dataset-agnostic	74.24	86.07	91.79	92.79	93.67	95.93
Strongest pruning baseline	62.97	73.11	78.49	83.47	88.23	90.86
Conditioning gain	4.98	3.16	1.01	3.39	2.91	0.22
Dataset-agnostic headroom	11.27	12.96	13.30	9.31	5.44	5.07

Conditioning helps InfographicVQA, TextVQA, and parts of V*Bench, while dataset-agnostic generation can be stronger on ChartQA and MMStar (Figure 13). These single-run results show useful headroom without explicit benchmark metadata. The image still reveals its content type, and removing benchmark guidance does not remove QA overlap; the conditioning and overlap-removal interventions are separate rather than joint controls.

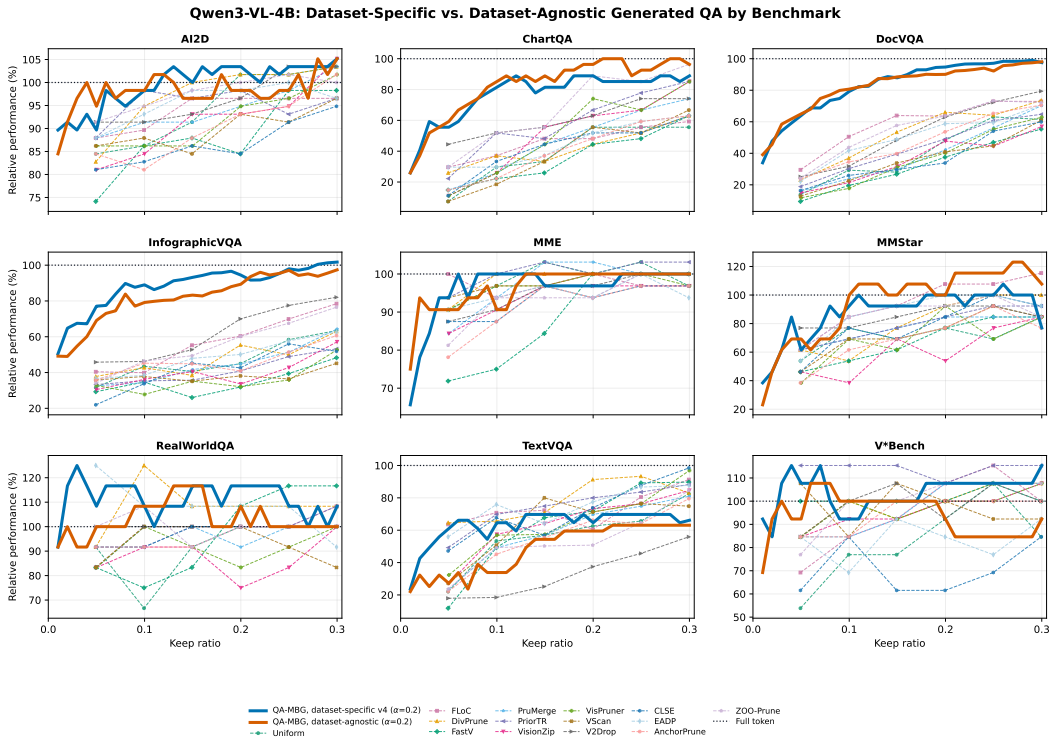


Figure 13: Conditioning ablation by benchmark.

K.3 Exact dataset-agnostic generation prompt

The following text accompanies the image. No benchmark name, task profile, or same-benchmark examples are appended.

```
You are generating a diverse set of image-grounded short-answer question-
    answer pairs from the attached image.

Inspect the image without assuming that it belongs to any particular
    dataset, benchmark, or task category. Generate questions that
    collectively test the visible information needed to understand this
    specific image.

Requirements:
- Every answer must be directly supported by visible evidence in the
    image.
```

```
- A question must require inspecting this particular image. Reject
    questions whose answer is predictable from general knowledge,
    stereotypes, or the wording of the question alone.
- Use short open-ended answers, normally 1-12 words.
- Do not use multiple-choice questions.
- Do not ask about evidence that is unreadable, ambiguous, hidden, or
    absent.
- Do not invent text, values, objects, relationships, or background
    context.
- Cover independent visual facts before producing alternative phrasings.
- Use at most 2 questions for the same underlying fact. A second question
     must be a substantially different formulation, not a minor rewording
    .
- Avoid generic questions such as asking whether a common object is
    present when more specific visible evidence can be tested.

Coverage plan:
- Return about 32 QA pairs, preferably 28-36 when the image supports them
    .
- First identify the image's visible content and organization internally.
- Select applicable evidence types from: object or entity identity,
    attributes, readable text, numerical values, counting, spatial
    relationships, interactions, document or diagram structure,
    comparison, ordering, and concise multi-step visual reasoning.
- Cover different regions and elements of the image.
- Use only evidence types that genuinely apply to the image; do not fill
    quotas or fabricate questions to reach the requested count.
- Prefer questions with a specific visual anchor such as an object,
    region, label, value, relation, or state.
- If fewer than 28 reliable independent questions are available, return
    fewer.

Return JSON only with this schema:
{
  "qas": [
    {
      "question": "...",
      "answer": "...",
      "question_type": "short descriptive category"
    }
  ]
}

Do not include explanations or text outside the JSON object.
```

L Robustness to the Auxiliary-QA Generator

We test whether the searched frontier depends on the frontier VLM used to generate auxiliary QAs. On the Qwen3-VL small set, we replace GPT-5.6-SOL with Opus 5 while keeping the generation prompt, QA-MBG search configuration ($\alpha = 0.2$), probability objective, evaluation images, and downstream protocols fixed. Figure 14 compares the resulting trajectory with GPT-5.6 generation and the evaluated pruning baselines.

The generator change preserves the main result. Across the six reported budgets, the Opus 5 and GPT-5.6 trajectories differ by only 2.30 relative-performance points on average, and both exceed the strongest pruning baseline at every budget. Opus 5 is higher at five of the six checkpoints, reaching 99.21% of full-token actual-QA performance at keep ratio 0.30. Its subsets also remain the strongest compressed representation in the long-form evaluation at every checkpoint, with overall judge scores increasing from 2.76 at 0.05 to 3.43 at 0.30. Thus the observed headroom and transfer to open-ended descriptions are not specific to GPT-5.6 auxiliary QAs.

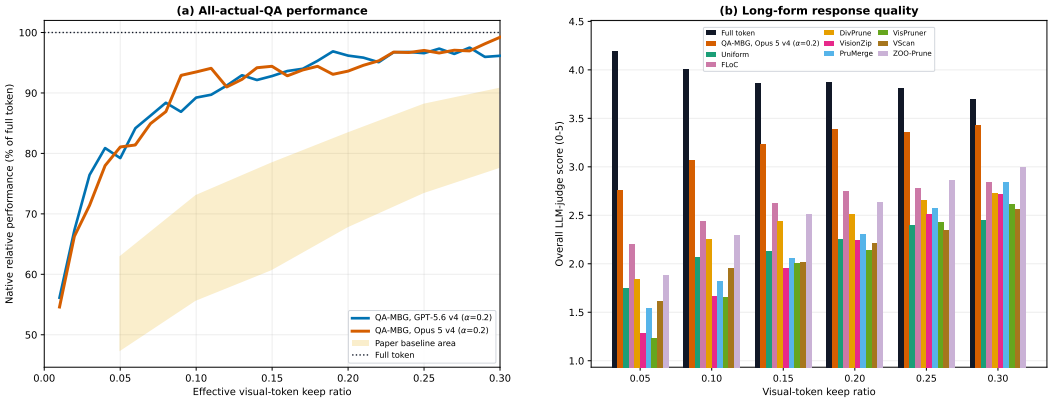


Figure 14: Auxiliary-QA generator ablation on the small set.

Table 22: Generator ablation at the six reported budgets. Actual QA is relative to full-token performance; long-form scores use the 0–5 judge scale.

Metric / generator	0.05	0.10	0.15	0.20	0.25	0.30
Actual QA, GPT-5.6	79.22	89.23	92.80	96.18	96.58	96.15
Actual QA, Opus 5	81.07	93.48	94.41	93.61	97.04	99.21
Long-form, Opus 5	2.76	3.07	3.23	3.39	3.36	3.43
Best long-form baseline	2.21	2.44	2.63	2.75	2.86	3.00

M QUALITATIVE EXAMPLES

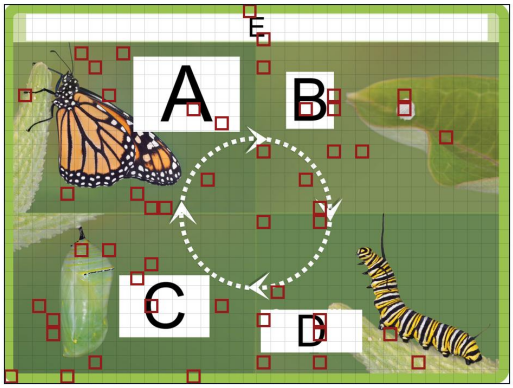
We show one favorable example per benchmark from the Qwen3-VL main set. Each page includes the image, retained tokens, actual-QA responses, and long-form descriptions with judge scores at the stated budget. These selected cases illustrate preserved or missed evidence; they are not a random sample. Aggregate claims are based on the full-cohort evaluations. One case per page preserves the readability of response details.

AI2D qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (49/972 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 1500×1125; displayed 933×700 with preserved aspect ratio. Grid: 27×36.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. The diagram above is a life cycle of which organism?

GT: A

Q2. The diagram shows the _____ of a Monarch Butterfly

GT: C

Q3. What does stage the number 3 depict?

GT: C

Method	Q1 response	Q2 response	Q3 response	Accuracy
Full token	A ✓	C ✓	C ✓	3/3
QA-MBG	A ✓	C ✓	C ✓	3/3
FLoC	A ✓	C ✓	C ✓	3/3
DivPrune	A ✓	C ✓	C ✓	3/3
Uniform	A ✓	C ✓	C ✓	3/3
VisionZip	A ✓	C ✓	C ✓	3/3
PruMerge	A ✓	C ✓	C ✓	3/3
VisPruner	A ✓	C ✓	C ✓	3/3
VScan	A ✓	C ✓	C ✓	3/3
ZOO-Prune	A ✓	C ✓	C ✓	3/3

(c) Long-form grounded descriptions

P2 verbatim excerpts; scores are P1/P2 means

Full token

972 visual tokens

Overall 4.5/5

Groundedness 4.5
Coverage 5
Clarity 4.5
Unsupported 1

The image is a diagram illustrating the life cycle of a monarch butterfly, organized in a circular flow with four distinct stages, each labeled with a capital letter and accompanied by a corresponding photograph. The layout is divided into four quadrants arranged in a clockwise circular pattern, connected by dashed arrows that indicate the progression from one stage to the next.

- **Stage A (Top Left)**: This quadrant features a fully formed adult monarch butterfly perched on a green stem. The butterfly is shown in profile, with its wings spread, displaying the characteristic orange and black pattern with white spots along the wing edges. This represents the final, reproductive stage of the life cycle.

- **Stage B (Top Right)**: This quadrant shows a small, white, oval-shaped egg resting on the underside of a green leaf. This is the initial stage of the life cycle, where the butterfly's life begins as an egg.

- **Stage C (Bottom Left)**: This quadrant displays a chrysalis (or pupa), hanging from a leaf. The chrysalis is greenish with faint yellow stripes and appears to be in a resting, transitional state. This is the stage between the larval and adult phases.

- **Stage D (Bottom Right)**: This quadrant shows a caterpillar (larva) with a striking yellow, black, and white striped body, crawling along a leaf. This is the feeding and growth stage, where the caterpillar consumes leaves and prepares for metamorphosis.

The four stages are connected by dashed arrows forming a circular path:
- An arrow points from **B (egg)** to **D (caterpillar)**, indicating that the egg hatches into a caterpillar.
- An arrow points from **D (caterpillar)** to **C (chrysalis)**, showing that the caterpillar transforms into a chrysalis.

[...]

QA-MBG

49 visual tokens (5%)

Overall 4/5

Groundedness 4
Coverage 5
Clarity 4.5
Unsupported 1

The image presents a diagram illustrating the life cycle of a butterfly, organized in a four-quadrant layout with labeled stages and directional arrows indicating progression.

- **Overall Scene and Layout**: The image is divided into four equal quadrants arranged in a 2x2 grid. Each quadrant contains a distinct stage of the butterfly's life cycle, with a large letter (A, B, C, D) placed in the center of each quadrant to identify it. The quadrants are connected by dashed arrows forming a circular flow, suggesting a cyclical process.

- **Main Parts and Their Locations**:
- **Quadrant A (Top-Left)**: Contains an image of a butterfly in flight, with its wings spread. This represents the adult stage.
- **Quadrant B (Top-Right)**: Shows a small, white, oval-shaped object, which is a butterfly egg. This represents the egg stage.
- **Quadrant C (Bottom-Left)**: Displays a chrysalis (pupa), hanging vertically. This represents the pupal stage.
- **Quadrant D (Bottom-Right)**: Features a green caterpillar (larva) crawling on a leaf. This represents the larval stage.

- **Important Relationships and Connections**:
- The dashed arrows form a circular path, indicating the sequence of the life cycle.
- The arrow from **A (butterfly)** points to **B (egg)**, suggesting that the adult butterfly lays eggs to begin the next stage.
- The arrow from **B (egg)** points to **D (caterpillar)**, indicating that the egg hatches into a larva (caterpillar).
- The arrow from **D (caterpillar)** points to **C (chrysalis)**, showing that the caterpillar transforms into a pupa.
- The arrow from **C (chrysalis)** points back to **A (butterfly)**, completing the cycle and showing that the pupa emerges as an adult butterfly.

[...]

FLoC

best baseline at 5%

Overall 2/5

Groundedness 2
Coverage 2.5
Clarity 3.5
Unsupported 3.5

The image presents a diagram illustrating the life cycle of a monarch butterfly, organized in a quadrant layout with a dashed circular arrow indicating progression. The overall scene is divided into four distinct sections, each containing a labeled object or stage, with the arrow suggesting a cyclical, sequential relationship.

- **Top Center (Label E)**: This section contains only a white rectangular label with the letter "E" inside. It is positioned at the top center of the diagram, above the other elements. Its placement suggests it may serve as a title, a reference point, or a placeholder for a label not yet filled in. There is no visual object or image within this box, only the letter.

- **Bottom Left (Label C)**: This section contains a white rectangular label with the letter "C" inside. Below this label, there is a chrysalis (pupa) hanging from a leaf. The chrysalis is the stage between the caterpillar and the adult butterfly, representing the transformation phase. The label "C" likely corresponds to this stage in the life cycle.

- **Bottom Right (Label — no visible label)**: This section contains a monarch butterfly in its adult form, resting on a leaf. The butterfly is shown with its wings open, displaying the characteristic orange and black pattern. This represents the final stage of the life cycle. Although no label is visible in this quadrant, its position and content suggest it is the adult stage.

- **Top Left (Label — no visible label)**: This section contains a monarch caterpillar (larva) on a leaf. The caterpillar is shown in its typical segmented, greenish body with black and yellow stripes. This represents the feeding and growth stage before pupation.

[...]

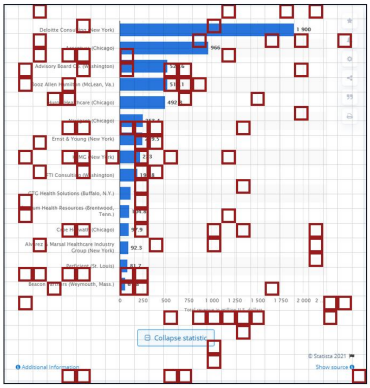
Actual-QA: task-native exact match. Long-form text is verbatim; [...] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

ChartQA qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 15% (98/650 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 800×836; displayed 670×700 with preserved aspect ratio. Grid: 26×25.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. How much revenue did Ernst & Young generate in 2013?

GT: 249.5

Method	Q1 response	Accuracy
Full token	249.5 ✓	1/1
QA-MBG	249.5 ✓	1/1
FLoC	118.5 ✗	0/1
DivPrune	249.3 ✓	1/1
Uniform	104.8 ✗	0/1
VisionZip	61.4 ✗	0/1
PruMerge	520.8 ✗	0/1
VisPruner	1900 ✗	0/1
VScan	104.7 ✗	0/1
ZOO-Prune	1900 ✗	0/1

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

650 visual tokens

Overall 4/5

Groundedness 4.5
Coverage 4
Clarity 4.5
Unsupported 1

This image displays a horizontal bar chart titled "Total revenue in million U.S. dollars," which ranks various consulting and healthcare industry firms by their revenue for the year 2021, as indicated by the source "Statista 2021" in the bottom right corner. The chart lists 14 firms, each with a corresponding blue bar indicating its revenue, and the revenue figures are shown at the end of each bar.

The firms are arranged in descending order of revenue, with the highest revenue at the top. The top firm is Deloitte Consulting (New York), with a revenue of $1,900 million. Following it is Accenture (Chicago) with $966 million. The third highest is Advisory Board Co. (Washington) at $520.6 million. The list continues with Booz Allen Hamilton (McLean, Va.) at $518.1 million, Huron Healthcare (Chicago) at $492.8 million, and Navigant (Chicago) at $253.4 million.

The lower portion of the chart includes firms with smaller revenues: Ernst & Young (New York) at $249.5 million, KPMG (New York) at $223 million, and FTI Consulting (Washington) at $192.8 million. The remaining firms on the list are CTG Health Solutions (Buffalo, N.Y.) with $118.5 million, Quorum Health Resources (Brentwood, Tenn.) with $104.8 million, Croe Horwath (Chicago) with $97.9 million, Alvarez & Marsal Healthcare Industry Group (New York) with $92.3 million, Perficient (St. Louis) with $81.7 million, and Beacon Partners (Weymouth, Mass.) with $61.8 million.

The x-axis is labeled "Total revenue in million U.S. dollars" and has a scale from 0 to 2,000, with major gridlines at intervals of 250. The chart includes interactive elements such as a "Collapse statistic" button and icons for additional information, sharing, and source details on the right side.

[...]

QA-MBG

98 visual tokens (15%)

Overall 4/5

Groundedness 3.5
Coverage 3.5
Clarity 4
Unsupported 2

This image displays a horizontal bar chart that ranks various consulting and investment firms by their revenue, measured in million U.S. dollars. The chart is organized with the firms listed vertically on the left, and their corresponding revenue values shown as horizontal bars extending to the right. The length of each bar visually represents the revenue amount, with the exact figure labeled at the end of each bar.

The firms, listed from highest to lowest revenue, are:

- Deloitte Consulting: $1,900 million (the longest bar)
- Accenture: $966 million
- Advisory Co. (Washington): $520.6 million
- Booz Hamilton: $518.1 million
- Huron Healthcare: $492.8 million
- Navigant: $253.4 million
- KPMG (New York): $223 million
- FTI (Washington): $192.8 million
- CTG: $118.5 million
- Quorum: $104.8 million
- Croe Horwath: $97.9 million
- Perficent: $81.7 million
- Beacon Partners: $61.8 million

The x-axis is labeled “Total revenue in million U.S. dollars,” and the chart includes a small “...” symbol at the far right, indicating that the scale may continue beyond the visible range. At the bottom left, the source or title “Beacon Partners” is visible, though it appears to be a label for the chart or a footnote rather than a firm name. The chart also features a small bell icon and a small “j” icon in the top right corner, which are likely interactive elements or navigation tools, not part of the data itself.

DivPrune

best baseline at 15%

Overall 1.5/5

Groundedness 1.5
Coverage 2
Clarity 3
Unsupported 4.5

This image displays a horizontal bar chart titled “Total revenue in million dollars,” which ranks various consulting firms by their revenue. The chart lists firm names on the vertical axis and their corresponding revenue values on the horizontal axis, with the bars extending to the right from the names. The revenue figures are shown numerically next to each bar.

The firms, listed from highest to lowest revenue, are:

- Deloitte Consulting (York): Revenue of $1,900 million (the longest bar, indicating the highest revenue).
- Accenture: Revenue of $1,000 million.
- Booz Allen Hamilton (Chicago): Revenue of $530 million.
- Ernst & Young: Revenue of $249 million.
- KPMG: Revenue of $249 million (the bar is visually identical in length to Ernst & Young’s, suggesting equal revenue).
- CTG Consulting: Revenue of $104 million.
- Quorum Resources (Brentwood, Tenn.): Revenue of $104 million (bar length matches CTG’s, indicating equal revenue).
- Croe & Marsal Healthcare Industry Group: Revenue of $3 million.
- Perficient: Revenue of $3 million (bar length matches Croe & Marsal’s, indicating equal revenue).
- Beacon Partners (Weymouth, Mass.): Revenue of $3 million (bar length matches the others, indicating equal revenue).

The horizontal axis is labeled “Total revenue in million dollars” and includes a scale from 0 to 250, with major tick marks at 0, 250, 500, 750, 1000, 1250, 1500, 1750, 2000, and 2250. The bars for Deloitte and Accenture extend beyond the 2000 mark, indicating their revenue exceeds $2000 million.

At the bottom of the chart, there is a button labeled “Collapse statistic,” and below that, a note indicating “© Statista” and “Show Additional Information.” The chart is presented in a clean, professional style with blue bars and black text.

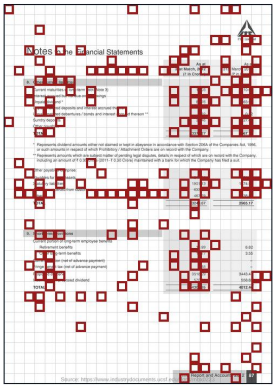
Actual-QA: task-native exact match. Long-form text is verbatim; [...] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

DocVQA qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 20% (192/962 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 1653×2339; displayed 495×700 with preserved aspect ratio. Grid: 37×26.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. What is the unpaid dividend as on 31/3/2012(Rs. Cr.)?

GT: 80.76

Q2. What is the Sundry deposits as on 31/3/2012(Rs. Cr.)?

GT: 30.46

Q3. What is the total for the short term provisions for year ending 31/3/2011?

GT: 4012.46

Method	Q1 response	Q2 response	Q3 response	Accuracy
Full token	80.76 ✓	30.46 ✓	4012.46 ✓	3/3
QA-MBG	80.76 ✓	30.46 ✓	4012.46 ✓	3/3
FLoC	80.76 ✓	0.32 ✗	4049.75 ✗	1/3
DivPrune	30.61 ✗	30.61 ✗	4013 ✗	0/3
Uniform	80.76 ✓	27.93 ✗	2965.17 ✗	1/3
VisionZip	0.06 ✗	27.6 ✗	467.00 ✗	0/3
PruMerge	0.32 ✗	29.75 ✗	182.00 ✗	0/3
VisPruner	0.06 ✗	27.6 ✗	4012.43 ✗	0/3
VScan	0.32 ✗	32 ✗	3249.77 ✗	0/3
ZOO-Prune	570.75 ✗	27.61 ✗	4012.77 ✗	0/3

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

962 visual tokens

Overall 4.5/5

Groundedness 4.5
Coverage 4.5
Clarity 4
Unsupported 1

This image is a page from the "Notes to the Financial Statements" section of ITC Limited's 2012 Annual Report and Accounts, specifically page 87. The document presents financial data in a structured tabular format, comparing figures for the fiscal year ending 31st March 2012 with those for the previous year, ending 31st March 2011. All monetary values are denoted in Indian Rupees (₹) and are presented in Crores (10^7).

The page is divided into two main sections: Section 8, titled "Other current liabilities," and Section 9, titled "Short-term provisions."

In Section 8, the table lists various current liabilities as of the specified dates. Key items include:
- Current maturities of long-term debt: ₹10.03 Crore (2012) vs. ₹10.68 Crore (2011).
- Interest accrued but not due on borrowings: ₹0.03 Crore (2012) vs. ₹0.06 Crore (2011).
- Unpaid dividend: ₹80.76 Crore (2012) vs. ₹63.93 Crore (2011). A footnote (*) explains this represents dividend amounts not claimed or kept in abeyance per Section 26A of the Companies Act, 1956, or subject to Prohibitory/Attachment Orders.
- Unpaid matured deposits and interest accrued thereon: ₹0.32 Crore (2012) vs. ₹0.32 Crore (2011).
- Unpaid matured debentures/bonds and interest accrued thereon: ₹0.32 Crore (2012) vs. ₹0.32 Crore (2011).

[...]

QA-MBG

192 visual tokens (20%)

Overall 4/5

Groundedness 4
Coverage 4
Clarity 4
Unsupported 2

This image displays a section of the ITC Limited financial report for the year 2012, specifically the "Notes to the Financial Statements," page 87. The content is organized into two primary financial tables, each comparing financial data for the fiscal year ending 31st March 2012 (as at 31st March, 2012) with the previous year, ending 31st March 2011 (as at 31st March, 2011). All monetary values are presented in Indian Rupees (₹) in Crores.

The first table, titled "Other current liabilities," lists several liability categories with their respective figures for both years. Key items include:
- Current liabilities of long-term debt: ₹10.03 crore (2012) vs. ₹10.68 crore (2011).
- Interest on borrowings: ₹0.03 crore (2012) vs. ₹0.06 crore (2011).
- Unpaid dividend: ₹80.76 crore (2012) vs. ₹63.93 crore (2011).
- Unpaid matured liabilities: ₹0.32 crore (2012) vs. ₹0.00 crore (2011).
- Sundry payables: ₹30.46 crore (2012) vs. ₹27.61 crore (2011).
- The subtotal for these liabilities is ₹3249.67 crore (2012) and ₹2965.17 crore (2011), with a total of ₹3371.27 crore (2012) and ₹3067.77 crore (2011).

The second table, titled "Other short-term provisions," details provisions for future obligations. It includes:
- Other long-term provisions: ₹16.00 crore (2012) vs. ₹12.77 crore (2011).
- Current taxation: ₹183.60 crore (2012) vs. ₹181.30 crore (2011).

[...]

FLoC

best baseline at 20%

Overall 1.5/5

Groundedness 1.5
Coverage 2.5
Clarity 3
Unsupported 4.5

This image displays a section from the ITC Report and Accounts 2012, specifically page 87, which presents "Notes to the Financial Statements" under item 8, titled "Current liabilities." The table is structured to show the composition of current liabilities as of March 31, 2012, with all figures presented in Crores of Indian Rupees.

The table lists several liability categories with their corresponding values:

- "Unpaid dividend and interest accrued" is listed as 10.67 Crores.
- "Interest accrued but not yet paid" is 80.76 Crores.
- "Unpaid matured debentures / bonds" is 0.32 Crores.
- "Sundry payables" is 27.61 Crores.
- "Payables to customers" is 183.29 Crores.
- "Statutory liabilities" is 570.61 Crores.
- "Others" is 467.25 Crores.

The total of these liabilities is 3067.17 Crores, as shown at the bottom of the column.

Below this table, there is a separate section titled "Short-term provisions," which includes:

- "Current retirement benefits" (not specified numerically in the visible portion).
- "Other long-term benefits" is 3.55 Crores.
- "Retirement benefits" is 183.29 Crores.
- "Proposed tax (net of advance payment)" is not specified numerically.
- "Income tax" is 27.75 Crores.

The total for the "Short-term provisions" section is 401.75 Crores.

[...]

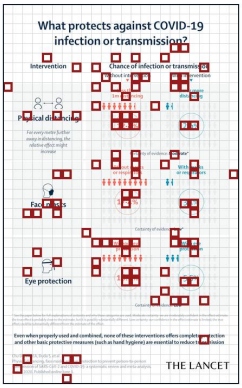
Actual-QA: task-native exact match. Long-form text is verbatim; [...] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

InfographicVQA qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 10% (94/936 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 1336×2151; displayed 435×700 with preserved aspect ratio. Grid: 39×24.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. What are the chances of contracting Coronavirus when there is a physical distance of more than 1m, 12.8%, 2.6%, or 17.4%?
GT: 2.6%

Q2. What is the percentage chance of infection when a person is not wearing a mask 17.4%, 3.1%, or 2.6%?
GT: 17.4%

Q3. What is the percentage chance of infection when a person is not wearing an eye protection 3.1%, 16.0%, or 5.5%?
GT: 16.0%

Q4. What are the interventions used against COVID-19?
GT: physical distancing, face masks, eye protection

Method	Q1 response	Q2 response	Q3 response	Q4 response	Accuracy
Full token	2.6% ✓	17.4% ✓	16.0% ✓	Physical distancing, Face masks, Eye protection ✓	4/4
QA-MBG	2.6% ✓	17.4% ✓	16.0% ✓	Physical distancing, face masks, eye protection ✓	4/4
FLoC	12.8% ✗	17.4% ✓	16.0% ✓	Face masks, eye protection ✗	2/4
DivPrune	12.8% ✗	17.4% ✓	16.0% ✓	Physical distancing, face masks, respirators, hand hygiene ✗	2/4
Uniform	17.4% ✗	17.4% ✓	17.0% ✗	Eye masks, distancing ✗	1/4
VisionZip	12.8% ✗	17.4% ✓	16.0% ✓	masking ✗	2/4
PruMerge	2.6% ✓	17.4% ✓	16.0% ✓	None ✗	3/4
VisPruner	12.8% ✗	17.4% ✓	16.0% ✓	Non-pharmaceutical interventions ✗	2/4
VScan	12.8% ✗	17.4% ✓	5.5% ✗	Mask, Social distancing, Hand hygiene, Vaccination, Testing, Quarantine,... ✗	1/4
ZOO-Prune	17.4% ✗	17.4% ✓	16.0% ✓	Physical distancing, face masks, the effect could be close to the estimate,... ✗	2/4

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

936 visual tokens

Overall 5/5

Groundedness 5
Coverage 5
Clarity 5
Unsupported 0

This infographic, published by The Lancet, answers the question "What protects against COVID-19 infection or transmission?" by presenting data on three key interventions: physical distancing, face masks, and eye protection. The central theme is that these measures significantly reduce the chance of infection or transmission, though none offer complete protection when used alone, and other measures like hand hygiene are essential.

The infographic is structured into three main sections, each dedicated to one intervention. Each section compares the "Chance of infection or transmission" under two conditions: "Without intervention" and "With intervention." The data is presented using large, bold numbers within circles, with the "Without intervention" figures in red and the "With intervention" figures in blue. The visual representation includes small icons of people to illustrate the scale of risk reduction.

Physical distancing:
- Without intervention (less than 1 meter distancing): 12.8% chance of infection or transmission.
- With intervention (1 meter or more distancing): 2.6% chance of infection or transmission.
- A note below states: "For every metre further away in distancing, the relative effect might increase."
- The certainty of evidence is labeled as "Moderate*."

Face masks:
- Without masks or respirators: 17.4% chance of infection or transmission.
- With masks or respirators: 3.1% chance of infection or transmission.
- The certainty of evidence is labeled as "Low*."

Eye protection:
- Without eye protection: 16.0% chance of infection or transmission.
- With eye protection: 5.5% chance of infection or transmission.
- The certainty of evidence is labeled as "Low*."

[...]

QA-MBG

94 visual tokens (10%)

Overall 4/5

Groundedness 4
Coverage 5
Clarity 4
Unsupported 2

This image is an infographic comparing the chance of transmission of a disease (likely COVID-19, given the context) under different intervention scenarios, broken down into three categories: Physical distancing, Face masks, and Eye protection. Each category presents two side-by-side data points: "Without intervention" and "With intervention," with the latter further subdivided into "1m or more distancing" for physical distancing and "With masks or respirators" for face masks. The "With eye protection" category is presented as a single intervention scenario.

The infographic uses a visual contrast: red icons represent "Without intervention" and blue icons represent "With intervention." The percentages are displayed within circles, with the "With intervention" values often being significantly lower than the "Without intervention" values, indicating a reduction in transmission risk.

Specifically:

- **Physical distancing**: Without intervention, the chance of transmission is 12.8%. With intervention (1m or more distancing), it drops to 2.6%.
- **Face masks**: Without masks or respirators, the chance is 17.4%. With masks or respirators, it decreases to 3.1%.
- **Eye protection**: Without eye protection, the chance is 16.0%. With eye protection, it drops to 5.5%.

Below the "Face masks" section, there is a note indicating "Certainty of evidence: Moderate" for the physical distancing and mask categories. Below the "Eye protection" section, a note reads "Certainty of evidence: Low*," with the asterisk likely indicating a footnote or caveat not visible in the image.

[...]

FLoC

best baseline at 10%

Overall 1.5/5

Groundedness 1.5
Coverage 1.5
Clarity 3.5
Unsupported 3.5

This image is an infographic from The Lancet, presenting data on the relative effectiveness of face masks in preventing COVID-19 infection, specifically comparing scenarios with and without one-meter distancing. The title, "Infection" (with the subtitle "Without or With 1m distancing"), sets the context for the comparison.

The infographic is divided into two main columns: "Without 1m distancing" and "With 1m distancing." Each column contains a visual representation and a numerical value indicating the relative effect of face masks in that scenario.

In the "Without 1m distancing" column, a large red circle contains the number "12.4%," indicating the relative effect of face masks in this scenario. Above this, a small icon of two people standing close together visually represents the lack of distancing.

In the "With 1m distancing" column, a similar red circle contains "5%," showing the relative effect of face masks when distancing is practiced. Above this, an icon of two people standing apart with a line of small human figures behind them suggests the presence of distancing.

Below the two columns, a section labeled "Face masks" includes a small icon of a person wearing a mask. To the right of this, the text "With eye protection" is displayed, accompanied by a small icon of a person wearing eye protection. This suggests that eye protection may also be considered, though the infographic does not provide a numerical value for it.

At the bottom of the image, a note reads: "* See the certainty of evidence: Low*," indicating that the evidence supporting these findings is considered low in confidence.

[...]

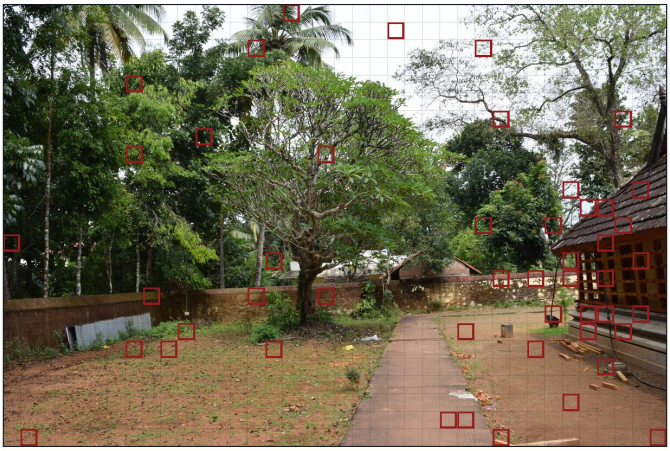
Actual-QA: task-native exact match. Long-form text is verbatim; [...] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

MME natural-image qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (48/950 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 6000×4000; displayed 1225×816 after the proportional 1M-pixel cap. Grid: 25 rows × 38 columns.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. Is this a photo of Puthoorppilly Sree Krishnaswamy Temple?

GT: Yes

Q2. Is this a photo of Villa Reale (Marlia)?

GT: No

Method	Q1 response	Q2 response	Accuracy
Full token	Yes ✓	No ✓	2/2
QA-MBG	Yes ✓	No ✓	2/2
FLoC	No ✗	No ✓	1/2
DivPrune	Yes ✓	No ✓	2/2
Uniform	No ✗	No ✓	1/2
VisionZip	No ✗	No ✓	1/2
PruMerge	Yes ✓	No ✓	2/2
VisPruner	Yes ✓	No ✓	2/2
VScan	No ✗	No ✓	1/2
ZOO-Prune	No ✗	No ✓	1/2

(c) Long-form grounded descriptions

P1 shown; scores are P1/P2 means

Full token

950 visual tokens

Overall 3.0/5

Groundedness 3.0
Coverage 4.0
Clarity 4.0
Unsupported 3.5

This image depicts a tranquil, somewhat rustic outdoor scene, likely in a tropical or subtropical region, characterized by abundant greenery, a simple pathway, and a traditional-style building. The setting appears to be a garden or courtyard area adjacent to a structure, possibly a temple or a small house , with a mix of natural and man-made elements.

The most prominent feature is a large, mature tree with a thick, gnarled trunk and a wide canopy of green leaves, positioned centrally in the foreground. Its branches spread out broadly, creating a natural focal point. Surrounding this tree and extending into the background are numerous other trees, including tall palms with distinctive fronds visible in the upper left and right corners, and dense clusters of broad-leafed trees that form a lush, verdant backdrop. The foliage is vibrant and suggests a warm, humid climate.

A narrow, paved pathway made of reddish-brown tiles or concrete runs diagonally from the lower right toward the center of the image, leading the viewer's eye into the scene. The path is bordered on both sides by uneven, earthy ground covered with sparse grass and scattered dry leaves. To the left of the path, a low, weathered wall made of reddish-brown brick or stone runs horizontally across the frame, separating the garden area from what might be a building or another plot. A small, rectangular metal or concrete panel is embedded in the wall on the left side.

On the right side of the image, a portion of a traditional building is visible. It features a sloped roof made of dark, overlapping tiles, and a distinctive wooden lattice or openwork screen on its facade, which is painted a warm, light brown. The building's base is constructed from stone or concrete blocks, and a small, dark-colored object, possibly a trash can or a small container, sits near the base of the wall. Scattered bricks and construction debris lie on the ground near the building's foundation, suggesting recent work or ongoing maintenance .

The sky is overcast and white, indicating a cloudy day with diffused light, which softens shadows and contributes to the calm, serene atmosphere of the scene. The overall impression is one of quietude, with nature and human-made structures coexisting harmoniously in a modest, possibly rural or semi-rural environment.

QA-MBG

48 visual tokens (5%)

Overall 4.0/5

Groundedness 4.0
Coverage 4.0
Clarity 4.0
Unsupported 2.0

This image depicts an outdoor scene, likely in a tropical or subtropical region, featuring a traditional-style building on the right, a paved pathway in the foreground, and a natural, vegetated background. The building has a dark, sloped tiled roof and a distinctive wooden lattice or openwork facade, possibly made of light-colored wood, with a set of steps leading up to it. The steps are constructed from stone or concrete and are flanked by a low, weathered wall. A black hose or pipe runs vertically along the side of the building, suggesting utility or irrigation . The ground is mostly dirt or compacted earth, with a narrow, paved concrete path running diagonally from the bottom left toward the center-right, leading toward the building's entrance. The area around the path is overgrown with green grass and weeds. In the background, tall palm trees and other dense foliage dominate the scene, with a pale, overcast sky visible above. The overall impression is one of a quiet, possibly rural or semi-rural setting, with a blend of traditional architecture and natural surroundings. No readable text or numerical information is visible in the image.

DivPrune

best baseline at 5%

Overall 2.5/5

Groundedness 2.5
Coverage 3.0
Clarity 4.0
Unsupported 3.5

This image depicts a quiet, natural outdoor scene, likely a garden or backyard, under an overcast sky. The foreground is dominated by a reddish-brown, dirt-covered ground, possibly a garden bed or unpaved path, with sparse, low-lying vegetation. A narrow, paved walkway made of reddish-brown tiles runs diagonally from the bottom right toward the center, leading the eye into the scene. To the right of the path, the corner of a traditional-style building is visible, featuring a dark, sloped roof with visible wooden rafters and a wall with a series of small, square, open windows or panels. A green garden hose lies on the ground near the building's base. To the left, a weathered, light-colored stone or concrete wall runs horizontally across the middle of the frame, separating the foreground from the background. Behind this wall, a large, leafy tree with a broad canopy stands prominently, its branches extending outward and partially obscuring the view of the sky and other trees. The tree's leaves are green, and some branches have small, pinkish flowers or blossoms. The overall atmosphere is calm and slightly rustic, with natural elements and a hint of human construction. No readable text or numerical information is visible in the image.

Actual-QA: task-native exact match. Long-form scores average P1/P2 judges; unsupported claims is a count. Green/red annotations are manual image-grounded review.

MMStar qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (11/221 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 512×384; displayed 512×384 with preserved aspect ratio. Grid: 13×17.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. What is the color of the hat in the image? Options: A: Black, B: Brown, C: Beige, D: White

GT: B

Method	Q1 response	Accuracy
Full token	B ✓	1/1
QA-MBG	B ✓	1/1
FLoC	A ✗	0/1
DivPrune	B ✓	1/1
Uniform	B ✓	1/1
VisionZip	B ✓	1/1
PruMerge	A ✗	0/1
VisPruner	B ✓	1/1
VScan	D: White ✗	0/1
ZOO-Prune	B ✓	1/1

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

221 visual tokens

Overall 3.5/5

Groundedness 3.5
Coverage 5
Clarity 4
Unsupported 3

This image depicts a rustic, attic-style bedroom with exposed wooden beams and a sloped ceiling, suggesting a traditional or historical setting. The walls are made of rough, whitewashed plaster, giving the room a simple, aged appearance. The floor is covered in a thick layer of what appears to be sand or fine gravel, with a dark brown woven rug placed near the left wall.

On the right side of the room, a bed is neatly made with a white sheet and a brown, possibly woolen, blanket draped over it. A wide-brimmed hat rests on the bed, and a small, dark lantern sits beside it on the bed's surface. Two narrow, rectangular windows are set into the walls—one on the left and one on the right—allowing natural light to enter, though the view outside is not visible. The windows are framed by the same whitewashed plaster as the walls.

In the corner near the left window, several large, burlap sacks are stacked against the wall, some tied with rope. Beside them stands a wooden shovel or spade leaning against the wall, and a red circular object, possibly a wheel or a decorative item, is partially visible. A small stack of books or boxes sits on the floor near the window, and a red flower arrangement in a white pot is placed on the rug. A small, dark lantern hangs from the ceiling near the center of the room, adding to the room's rustic ambiance.

The overall impression is one of a cozy, lived-in space with a historical or countryside aesthetic, possibly used for rest or as a guest room in a traditional dwelling.

QA-MBG

11 visual tokens (5%)

Overall 4/5

Groundedness 4
Coverage 4
Clarity 4
Unsupported 2.5

This image depicts a rustic, interior room with a simple, unadorned aesthetic, likely a cabin or farmhouse setting. The walls are made of rough, whitewashed plaster or stucco, showing texture and subtle imperfections. The ceiling is constructed of exposed wooden planks, arranged horizontally, with visible gaps and natural wood grain. A single, narrow, rectangular window is set into the right wall, allowing some light to enter, though it is not brightly lit.

On the floor, which appears to be bare or covered with a thin layer of dirt or straw, rests a bed. The bed is made with a white sheet or blanket, and atop it lies a folded, tan-colored blanket or quilt. A wide-brimmed, brown hat, possibly made of leather or woven material, is placed on the blanket, suggesting a casual or lived-in atmosphere.

In the background, near the left wall, a large, round, reddish-brown object—possibly a barrel or a decorative wheel—is partially visible, leaning against the wall. Beside it, a large, burlap sack sits on the floor, its contents unknown but suggesting storage or agricultural use. The overall impression is one of simplicity, functionality, and a connection to rural or traditional living. There is no visible text or numerical information in the image.

PruMerge

best baseline at 5%

Overall 3/5

Groundedness 3
Coverage 3
Clarity 4
Unsupported 2.5

This image depicts a rustic, dimly lit interior space, likely a bedroom or sleeping area in a traditional or rural dwelling. The walls are made of rough, whitewashed plaster or stucco, showing visible texture and imperfections. The ceiling is constructed from exposed wooden beams, arranged in a simple, slightly uneven pattern, suggesting an older or hand-built structure. A single, narrow, high-set window is visible on the right wall, allowing a small amount of natural light to enter, which illuminates the scene unevenly. The window frame appears to be made of wood and is set into the wall with a simple, unadorned design.

In the center of the room lies a bed, covered with a thick, light-colored blanket that appears to be made of wool or a similar coarse fabric, draped loosely over the mattress. The bed is positioned against the left wall, leaving the right wall and the window as the primary visual focus. A small, dark object, possibly a pillow or a folded cloth, rests on the bed near the headboard. The floor is not clearly visible, but the overall impression is of a bare or minimally furnished room, emphasizing simplicity and functionality. The overall atmosphere is quiet, uncluttered, and evokes a sense of solitude or rustic tranquility.

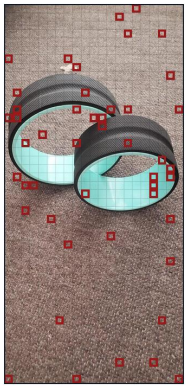
Actual-QA: task-native exact match. Long-form text is verbatim; […] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

RealWorldQA qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (47/945 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 726×1536; displayed 331×700 with preserved aspect ratio. Grid: 45×21.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. Which wheel is larger? A. The wheel on the left is larger. B. The wheel on the right is larger. C. They are the same size.

GT: A

Method	Q1 response	Accuracy
Full token	A ✓	1/1
QA-MBG	A ✓	1/1
FLoC	C ✗	0/1
DivPrune	A ✓	1/1
Uniform	C ✗	0/1
VisionZip	C ✗	0/1
PruMerge	C ✗	0/1
VisPruner	A ✓	1/1
VScan	C ✗	0/1
ZOO-Prune	A ✓	1/1

(c) Long-form grounded descriptions

P2 verbatim excerpts; scores are P1/P2 means

Full token

945 visual tokens

Overall 5/5

Groundedness 5
Coverage 5
Clarity 5
Unsupported 1

The image displays two identical yoga wheels resting on a textured, brownish-gray carpet. The wheels are positioned side by side, with one slightly overlapping the other, creating a diagonal arrangement from the upper left to the lower right of the frame. The wheels are circular, with a thick, black outer rim that has a textured, ribbed surface for grip, and a smooth, light mint-green inner ring that forms the central part of the wheel. The wheels are not in motion and appear stationary.

The carpet beneath them has a woven, fibrous texture, with a consistent pattern of interlaced threads that gives it a slightly coarse appearance. A small, white, fuzzy object, possibly a piece of lint or a stray fiber, is visible near the top center of the image, slightly behind the wheels.

There are no labels, charts, or diagrams in the image. The organization is purely spatial and visual: the two yoga wheels are the primary subjects, arranged in a simple, static composition on a plain background. The relationship between the objects is minimal — they are placed next to each other with no functional or interactive connection visible. The image is a straightforward product shot, with the focus on the design and color contrast of the wheels against the neutral carpet.

QA-MBG

47 visual tokens (5%)

Overall 5/5

Groundedness 5
Coverage 5
Clarity 5
Unsupported 1

The image displays two circular, ring-shaped objects resting on a textured, brownish-gray carpet. The objects appear to be fitness or exercise wheels, commonly used for core strengthening or as part of a yoga or pilates routine.

The two wheels are positioned side by side, with one slightly overlapping the other. The wheel on the left is partially obscured by the one on the right, suggesting a shallow depth of field or a casual placement. Both wheels share a similar design: a dark gray or black outer rim with a textured, ribbed surface for grip, and a smooth, light turquoise inner lining that forms the hollow center.

There are no visible labels, text, or diagrams in the image. The relationship between the two objects is primarily spatial — one is placed in front of and slightly to the right of the other, creating a diagonal alignment from the viewer's perspective. The wheels are not connected or interlocked; they are separate and stationary, resting on the carpet surface.

The carpet itself has a woven, fibrous texture, providing a neutral, non-reflective background that contrasts with the smooth, colored surfaces of the wheels. No other objects or elements are present in the frame.

In summary, the image is organized as a simple, static composition of two identical fitness wheels on a carpet, with no complex structure or interactivity depicted. The arrangement is minimal and functional, emphasizing the form and placement of the objects.

ZOO-Prune

best baseline at 5%

Overall 3.5/5

Groundedness 3.5
Coverage 4.5
Clarity 4
Unsupported 2

The image displays two circular, wheel-like objects resting on a textured, brownish-gray carpeted surface. These objects appear to be foam rollers or exercise wheels, commonly used for physical therapy or fitness routines.

The two wheels are positioned side by side, with the one on the left slightly behind and to the left of the one on the right. Both wheels share a similar design: a dark gray or black outer rim with a textured, diamond-patterned surface, likely for grip. The inner portion of each wheel is a solid, light teal or mint green color, forming a smooth, curved interior that contrasts with the darker outer edge.

There is no visible text, labels, or diagrams in the image. The objects are not arranged in any structured or hierarchical layout; they are simply placed in close proximity to each other, suggesting a casual or functional placement rather than a deliberate composition.

The relationship between the two wheels is spatial: they are adjacent, with no overlapping, and their orientation is consistent — both are lying flat on the carpet with their circular faces facing upward. The carpet's texture is visible beneath them, providing a neutral, non-reflective background that emphasizes the wheels' form and color contrast.

No other objects or elements are present in the image, and there are no indications of scale, usage, or context beyond their physical presence and appearance.

Actual-QA: task-native exact match. Long-form text is verbatim; […] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

TextVQA qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (38/768 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 1024×768; displayed 1024×768 with preserved aspect ratio. Grid: 24×32.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. what year is the mercury on the sign?

GT: 1999

Q2. where should you live life?

GT: in your own lane

Method	Q1 response	Q2 response	Accuracy
Full token	1999 ✓	in your own lane ✓	2/2
QA-MBG	1999 ✓	in your own lane ✓	2/2
FLoC	1999 ✓	your lane ✗	1/2
DivPrune	1999 ✓	own lane ✗	1/2
Uniform	2010 ✗	now ✗	0/2
VisionZip	2010 ✗	nowhere ✗	0/2
PruMerge	2010 ✗	nowhere ✗	0/2
VisPruner	1999 ✓	live life ✗	1/2
VScan	2009 ✗	live life ✗	0/2
ZOO-Prune	2011 ✗	nowhere ✗	0/2

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

768 visual tokens

Overall 4.5/5

Groundedness 4.5
Coverage 5
Clarity 4.5
Unsupported 1.5

This image captures a large, prominent banner hanging from the ceiling in what appears to be an indoor venue, possibly a car show or dealership. The banner is dominated by a dark red and black gradient background, with bold white text and a logo. On the left side, there is a circular emblem featuring a stylized, roaring cougar head in red and black, which is the iconic logo for the Mercury Cougar. To the right of the logo, the text reads “1999 COUGRRR” in large, capitalized letters, with “COUGRRR” being a playful, exaggerated spelling of “Cougar.” Below this, the brand name “Mercury” is displayed, followed by its distinctive circular logo with three parallel lines. Underneath the brand name, the slogan “LIVE LIFE IN YOUR OWN LANE” is written in smaller, capitalized letters. The banner is mounted on a white ceiling structure, and below it, a crowd of people can be seen, suggesting the setting is a public event. The lighting in the room is artificial, with a fluorescent light fixture visible at the top left corner. The overall impression is one of a promotional display for the 1999 Mercury Cougar model, emphasizing its bold and spirited identity.

QA-MBG

38 visual tokens (5%)

Overall 4/5

Groundedness 4
Coverage 4.5
Clarity 4
Unsupported 2

This image shows a large, dark-colored banner or sign, likely made of fabric or vinyl, mounted on a wall or structure. The banner is angled slightly downward from left to right, and the lighting appears to be ambient indoor or outdoor light, casting some shadows and highlighting the text.

The most prominent feature is the large, bold, white text at the top that reads “1999 COUGRRR.” The word “COUGRRR” is stylized with a slightly jagged or distressed font, and the “1999” is in a simpler, clean sans-serif font, positioned to the left of the word. Below this, in a smaller but still prominent white font, is the word “Mercury,” followed by a circular logo to its right. The logo is a stylized “M” inside a circle, with a small horizontal line or bar across the center.

Underneath “Mercury,” the phrase “LIVE LIFE IN YOUR OWN LANE” is written in a smaller, all-caps, white sans-serif font. The text is centered horizontally on the banner, creating a clear visual hierarchy.

The background of the banner is a solid dark color, possibly black or dark gray, which makes the white text stand out sharply. The overall design suggests a promotional or branding sign, likely for a product or event associated with the year 1999 and the Mercury brand, possibly related to automotive or consumer goods, given the logo’s style. The phrase “LIVE LIFE IN YOUR OWN LANE” adds a motivational or lifestyle-oriented message.

FLoC

best baseline at 5%

Overall 2.5/5

Groundedness 2.5
Coverage 3
Clarity 4
Unsupported 3.5

This image shows a large, dark-colored banner or sign hanging indoors, likely in a commercial or recreational space. The banner features prominent white text and a logo. At the top, the word “COUGRRR” is displayed in a bold, uppercase, sans-serif font, with the letters slightly spaced apart. Below this, the phrase “LIVE IN YOUR LANE” appears in a smaller, but still clear, white sans-serif font. To the left of the text is a circular logo containing a stylized, white silhouette of a cougar’s head, facing right, with sharp, angular features. The logo is set against a dark red or maroon background that fades into black toward the bottom of the banner. The banner is mounted on a white ceiling or wall structure, and below it, the tops of people’s heads and shoulders are visible, suggesting the banner is in a public area with patrons present. The lighting is artificial, coming from overhead fluorescent fixtures, casting a cool, even glow on the banner. The overall impression is that of a promotional or branding sign for a business, possibly a bowling alley or similar entertainment venue, given the phrase “Live in Your Lane.”

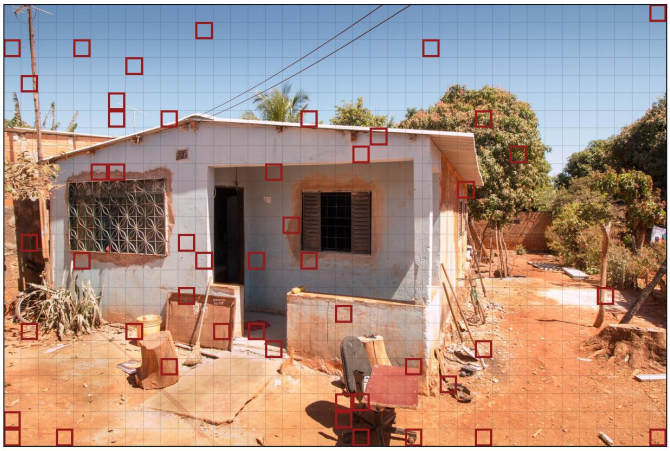
Actual-QA: task-native exact match. Long-form text is verbatim; […] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.

V*Bench qualitative example

Qwen3-VL-4B-Instruct | QA-MBG α=0.2 | keep ratio 5% (48/950 visual tokens)

(a) Input image and selected visual tokens

dark red outline: selected



Source 2250×1500; displayed 1224×816 with preserved aspect ratio. Grid: 25×38.

(b) Actual-QA responses at the selected keep ratio

✓ exact match · ✗ incorrect

Q1. Is the shovel on the left or right side of the house? A. left B. right

GT: B

Method	Q1 response	Accuracy
Full token	B. right ✓	1/1
QA-MBG	B ✓	1/1
FLoC	A. left ✗	0/1
DivPrune	B. right ✓	1/1
Uniform	B. right ✓	1/1
VisionZip	A. left ✗	0/1
PruMerge	A. left ✗	0/1
VisPruner	B. right ✓	1/1
VScan	A. left ✗	0/1
ZOO-Prune	A. left ✗	0/1

(c) Long-form grounded descriptions

P1 verbatim excerpts; scores are P1/P2 means

Full token

950 visual tokens

Overall 4/5

Groundedness 4
Coverage 5
Clarity 4
Unsupported 2.5

This image depicts a modest, single-story house in a rural or semi-rural setting under a clear, bright blue sky. The building has a simple, functional design with light blue or pale gray stucco walls and a flat roof with a slight overhang. The structure appears weathered, with visible signs of wear and some exposed brickwork on the right side. The ground is dry, reddish-brown earth, suggesting a warm, arid climate.

The house features two windows. The left window is protected by a decorative metal grille with a geometric, star-like pattern, and a small red object is visible inside. The right window has dark wooden shutters, one of which is open, revealing a black windowpane. A small, dark doorway is located to the left of the right window, leading into the house's interior, which is dark and not clearly visible.

In front of the house, the yard is cluttered with various items. To the left, there is a yellow plastic bucket, a cut tree stump, and a broom leaning against a wooden board. A bicycle is partially visible behind some cactus-like plants. Near the center, a small, red, rectangular object, possibly a piece of furniture or a sign, rests on the ground. To the right, a black office chair with a red desk or tabletop is positioned near a small, weathered wooden stool. Several wooden planks and tools, including a shovel and a long pole, are scattered around the yard. In the background, lush green trees and bushes contrast with the dry earth, and a few power lines stretch across the sky. The overall scene suggests a lived-in, functional space, possibly in a developing or rural area.

QA-MBG

48 visual tokens (5%)

Overall 4/5

Groundedness 4
Coverage 3.5
Clarity 4
Unsupported 1

This image depicts a simple, single-story building with a light blue exterior wall and a white roof, situated on a dirt ground. The structure appears modest and possibly rural or semi-rural, with a weathered wooden door and a window with a metal security grille. A small, red mat or rug lies on the ground just inside the doorway, suggesting a threshold or entry point. To the left of the building, a yellow plastic bucket and a large, rough-hewn wooden log or stump are visible, leaning against the wall. A dark brown wooden board or panel leans against the wall near the door, possibly serving as a makeshift sign or storage. In the foreground, a black office chair with wheels is positioned near the building's entrance, suggesting recent human activity or temporary use. A thin, dark pole or stick leans against the wall near the chair. The surrounding area is unpaved, with sparse vegetation including a tree with green leaves visible in the upper right corner, and a small, dark object, possibly a tool or piece of debris, lies on the ground near the chair's base. The overall scene conveys a sense of simplicity and utilitarian function, with no visible text or numerical markings.

DivPrune

best baseline at 5%

Overall 3.5/5

Groundedness 3.5
Coverage 3.5
Clarity 4
Unsupported 2

This image depicts a simple, rustic outdoor scene under a clear blue sky, centered around a modest, single-story building with a plain, light-colored stucco exterior. The building has a flat roof with a slight overhang, and a small, dark-framed doorway is visible, with a thin vertical wire hanging down near it. To the left of the doorway, a window is partially visible, framed by a simple wooden or metal grille with vertical slats. Above the window, a small, rectangular plaque or sign is mounted, but its text is illegible due to resolution and angle. A small, weathered wooden post or stump stands near the building's base, to the left of the doorway.

In the foreground, a red, worn office chair with a metal frame is positioned near the building's base, its back facing the viewer. To the left of the chair, a dark, cylindrical object—possibly a wooden barrel or a large pot—rests on the ground. Behind the chair, a pair of dark, worn sandals lies on the reddish-brown dirt ground. A bicycle wheel is partially visible in the far left corner, leaning against the building's wall. A tree with green leaves stands to the right of the building, casting a shadow across the ground. Two thin, dark power lines stretch diagonally across the upper portion of the image, against the clear sky. The overall scene suggests a rural or semi-rural setting, possibly a home or small workshop, with a mix of functional and discarded items scattered around.

Actual-QA: task-native exact match. Long-form text is verbatim; [...] marks truncation. Scores average P1/P2 judges. Green/red annotations are manual review.