Does Visual Token Pruning Improve Calibration? An Empirical Study on Confidence in MLLMs

Kaizhen Tan
Carnegie Mellon University
Pittsburgh, PA, USA
kaizhent@cmu.edu

Abstract

Visual token pruning is a widely used strategy for efficient inference in multimodal large language models (MLLMs), but existing work mainly evaluates it with task accuracy. In this paper, we study how visual token pruning affects model calibration, that is, whether predicted confidence matches actual correctness. Using LLaVA-1.5-7B on POPE and ScienceQA-IMG, we evaluate Expected Calibration Error (ECE), Brier score, and AURC under several pruning strategies, including SCOPE with different saliency weights, saliency-only pruning, FastV, and random pruning, across multiple token budgets. Our results show that pruning does not simply trade reliability for efficiency. On POPE, a pure-coverage setting in SCOPE achieves substantially lower ECE than the full unpruned model while maintaining similar accuracy. An internal α-sweep further shows a consistent trend: reducing the saliency weight improves calibration at all tested token budgets, while accuracy changes only slightly. In contrast, saliency-based pruning leads to worse calibration, and real FastV causes severe performance degradation in our setting. On ScienceQA-IMG, pruning also reduces ECE, with accuracy remaining stable or slightly improving. We additionally study the gap power exponent in coverage-based selection and find that its default setting is not always optimal. Overall, our results suggest that visual token pruning should be evaluated not only by accuracy, but also by confidence quality, especially for multimodal systems that need reliable decisions.

1 Introduction

Multimodal Large Language Models (MLLMs) [10, 11] typically process hundreds of visual tokens for a single image, which makes inference increasingly expensive as the input length grows. Visual token pruning reduces this cost by selecting a smaller subset of tokens before they are passed to the language model. As a result, it has quickly become a standard efficiency technique for MLLMs. Recent methods explore different selection principles, including saliency [3], coverage [7], diversity [1, 16], and transport-based objectives [4].

Existing work on visual token pruning is evaluated almost entirely by task accuracy. This evaluation protocol is useful, but incomplete. In many deployment settings, especially those involving multimodal decision-making, it is also important to know whether model confidence is reliable. A system can keep similar accuracy after pruning, yet become more overconfident or less selective about its errors. From this perspective, pruning should be evaluated not only by how often the model is correct, but also by whether its confidence remains aligned with correctness.

Calibration measures this alignment between confidence and actual correctness. For multimodal systems, poor calibration is a practical problem rather than a purely descriptive one. A model that gives a highly confident answer to a visually grounded question may still be wrong, and this mismatch directly affects downstream decisions such as whether to answer, abstain, or request additional information. This issue is particularly relevant when MLLMs are used as components in larger reasoning systems or multimodal agents.

Prior work on compression and calibration in CNNs reports mixed conclusions. Depending on the pruning strategy and compression ratio, pruning may either hurt or improve calibration [14, 15]. However, visual token pruning in MLLMs differs from weight pruning in an important way. Weight pruning removes parameters while keeping the input structure unchanged, whereas token pruning removes image patches or patch-level representations before language modeling. This change can alter not only task performance but also how much visual evidence the model retains for making confident predictions.

In this paper, we study how visual token pruning affects calibration in MLLMs. We use LLaVA-1.5-7B and evaluate several pruning strategies on POPE and ScienceQA-IMG with calibration-oriented metrics, including ECE, Brier score, and AURC. Our results show that pruning does not simply introduce a trade-off between efficiency and reliability. Under our setting, moderate coverage-based pruning can reduce calibration error while keeping accuracy nearly unchanged, and in some cases it performs better than the full unpruned model on calibration. In contrast, saliency-based pruning gives worse calibration, and real FastV causes severe degradation in our experiments.

Our main contributions are as follows:

1. We provide, to our knowledge, the first empirical study of how visual token pruning affects calibration in MLLMs, moving the evaluation of pruning beyond accuracy alone.

2. On POPE, we find that moderate pruning can improve calibration. In particular, a pure-coverage setting in SCOPE at $K{=}128$ reduces ECE from 0.041 to 0.016 while maintaining similar accuracy.

3. Through an internal α-sweep within SCOPE, we observe a consistent trend across $K \in \{64, 128, 192\}$: reducing the saliency weight improves calibration, while accuracy changes only slightly. Random pruning does not show the same behavior, which suggests that the selection rule matters, not only the compression ratio.

4. We further analyze the gap power exponent p in coverage-based selection and find that its default setting is not always optimal. Based on these results, we argue that future visual token pruning work should report calibration together with accuracy.

2 Related Work

Visual Token Pruning for MLLMs. Visual token pruning has become a common way to reduce the inference cost of MLLMs. Existing methods mainly differ in the selection signal they use. Coverage-based methods include SCOPE [7], FLoC [6], and OTPrune [4]. Saliency-based methods include FastV [3] and SparseVLM [18]. Diversity-based methods include DivPrune [1] and CDPruner [16]. Some recent work also explores adaptive or data-dependent pruning policies, such as AgilePruner [2] and PruneSID [8]. Although these methods differ substantially in design, evaluation is still dominated by task accuracy and efficiency, and calibration is rarely reported.

Calibration in MLLMs. Calibration has received increasing attention in multimodal large language models, especially in relation to overconfidence and hallucination. Chen *et al.* [5] study calibration across multiple MLLMs and show that overconfidence remains a persistent issue even after visual instruction tuning. Zhou *et al.* [19] improve calibration through training-time objectives, introducing calibrated self-rewarding for vision-language models. Zhang *et al.* [17] estimate uncertainty through semantic-preserving perturbations on both visual and textual inputs, with a focus on hallucination detection. These studies show that confidence quality is an important aspect of MLLM reliability. However, they mainly examine training or inference-time uncertainty estimation, rather than the effect of token compression itself. In contrast, our work focuses on a simple but practically important question: how calibration changes when visual evidence is explicitly reduced before language modeling.

Compression and Calibration. The interaction between compression and calibration has been studied more extensively in conventional vision models. Prior work reports mixed findings. Misra *et al.* [14] show that sparsity can worsen calibration in transfer learning settings, whereas Mitra *et al.* [15] find that unstructured pruning may improve calibration while structured pruning can hurt it. BRC [13] further shows that reliability may degrade faster than accuracy after compression, and proposes post-hoc recalibration for compressed models. These results suggest that compression should not be evaluated by accuracy alone. At the same time, they do not directly answer the question studied here. Visual token pruning in MLLMs differs from weight pruning in image classifiers because it removes part of the input evidence rather than model parameters. For this reason, its effect on confidence can be qualitatively different. Our work provides an empirical study of this issue in the setting of MLLM visual token pruning.

3 Methodology

3.1 Model and Pruning Methods

We evaluate LLaVA-1.5-7B [10] with CLIP-ViT-L/14-336, which produces 576 visual tokens for each image. Our main goal is to compare pruning strategies that place different emphasis on coverage and saliency.

We use SCOPE [7] as the main framework because it allows us to vary this trade-off within a unified formulation. In SCOPE, each candidate token is scored by

$$\text{score}(v) = \Delta_{\text{cov}}(v; \mathcal{S}) \cdot a(v)^{\alpha},$$

where $\Delta_{\text{cov}}(v; \mathcal{S})$ denotes the facility-location coverage gain of adding token v to the current selected set $\mathcal{S}$, and $a(v)$ denotes its CLS attention score. The exponent α controls how strongly saliency affects selection. When $\alpha{=}0$, the score reduces to pure coverage. When $\alpha{=}1$, it recovers the default SCOPE setting. Larger α places more weight on saliency. This formulation is based on our reading of the open-source SCOPE implementation.

We compare the following pruning strategies. SCOPE with $\alpha{=}0$ keeps tokens using only the coverage objective. SCOPE with $\alpha{=}0.5$ uses a softer saliency weighting. SCOPE with $\alpha{=}1$ is the default hybrid setting in the original method [7]. These three settings allow us to study the effect of changing the saliency weight while keeping the rest of the pipeline fixed.

We also include three additional baselines. Saliency-only pruning selects the top-K tokens by CLS attention without any coverage term. This provides a saliency-focused reference outside the SCOPE interpolation. FastV (real, 2-pass) [3] ranks tokens using attention from LLM layer 2 and drops low-scoring tokens for the remaining layers. Random pruning selects tokens uniformly at random. For random pruning, we report the mean and standard deviation over three seeds.

The α-sweep serves as a controlled comparison within the same code path, with the same attention source and the

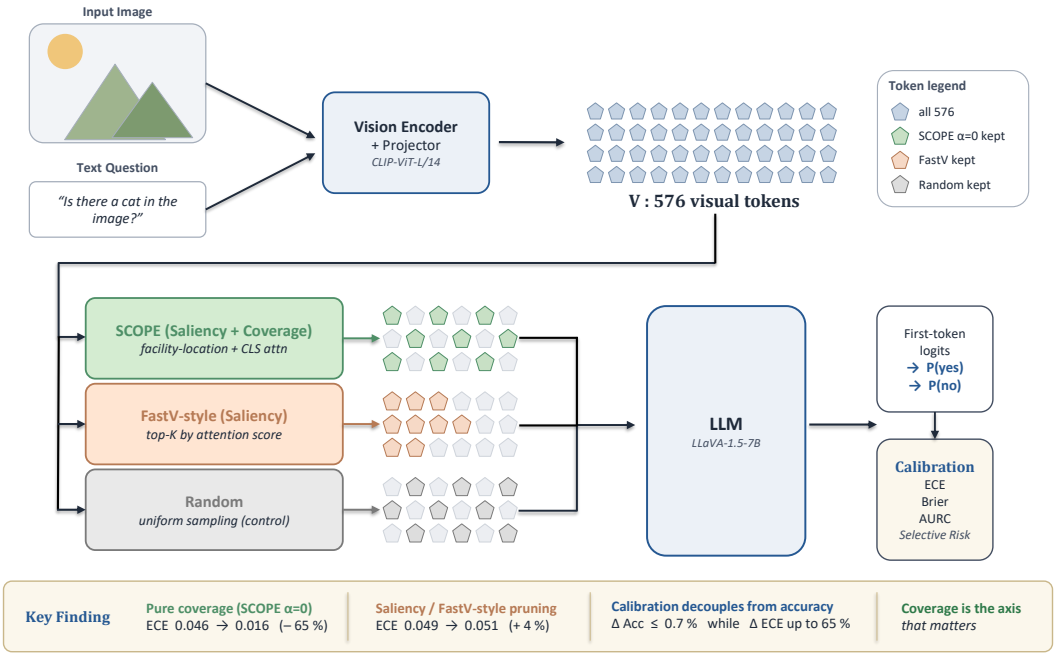

Figure 1. **Study design.** An image-question pair is encoded into V=576 visual tokens. We compare five token selection strategies: SCOPE with α=1 (default hybrid), SCOPE with α=0.5, SCOPE with α=0 (pure coverage), saliency-only pruning based on CLS attention, and random pruning. The selected K tokens are then passed to the LLM. We extract first-token class probabilities and evaluate calibration with ECE, Brier score, and AURC. The α-sweep provides a controlled comparison within the SCOPE framework, while FastV and random pruning serve as external reference points.

same coverage objective. FastV and random pruning provide additional reference points based on different selection rules. We evaluate the α-sweep on POPE at $K \in \{64, 128, 192\}$, and we evaluate the other pruning settings at the same token budgets where applicable.

3.2 Calibration Metrics

Confidence extraction. All experiments use greedy decoding (`do_sample=False`) and the same prompt format across different pruning settings. For each POPE example, we call `model.generate()` with `output_scores=True` and extract the logits at the first generated token position. We then compute the probabilities of "yes" and "no" by summing the softmax probabilities of the corresponding token IDs (including both lowercase and uppercase forms), and normalize them within this answer set:

$$\text{confidence} = \frac{\max(P_{\text{yes}}, P_{\text{no}})}{P_{\text{yes}} + P_{\text{no}}}.$$

For ScienceQA, we extract the probabilities of the answer options A, B, C, and D from the first-token logits, normalize them within the four candidates, and define confidence as $\max_c P(c)$. This first-token confidence is used as a simple measure of the model's initial categorical preference under each pruning condition.

Expected Calibration Error (ECE). Our primary calibration metric is Expected Calibration Error (ECE), computed with 15 equal-width bins:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|.$$

ECE measures the average mismatch between confidence and empirical accuracy across bins.

Additional metrics. To complement ECE, we also report Brier score, negative log-likelihood (NLL), and Area Under the Risk-Coverage curve (AURC). These metrics provide additional views of confidence quality and selective prediction behavior. All 95% confidence intervals are computed from 1,000 bootstrap samples.

3.3 Benchmarks

POPE [9] contains 9,000 yes/no object-existence questions divided into three splits: random, popular, and adversarial. Because each example has a binary answer with a clear correctness label, POPE provides a clean setting for calibration analysis.

ScienceQA-IMG [12] contains 2,017 image-based multiple-choice science questions. Each example has four answer

Table 1. Calibration under default SCOPE (α=1) on POPE (9K samples). Moderate pruning improves calibration, with the best ECE observed at K=128. We report 95% bootstrap confidence intervals for ECE. Results from the α-sweep are presented separately in Sec. 4.2.

K	Acc	ECE↓	Brier↓	NLL↓	AURC↓	T_{opt}
576	86.9	.041	.099	.331	.047	1.30
192	87.3	.027	.093	.307	.040	1.25
128	**86.9**	**.024**	**.095**	**.311**	**.042**	**1.10**
64	85.5	.031	.104	.337	.048	1.20
32	82.7	.045	.123	.391	.064	1.30

Table 2. α-sweep within SCOPE on POPE (9K). The exponent α controls how strongly saliency weights the coverage gain in the SCOPE score $\Delta_{\text{cov}} \cdot a^{\alpha}$, and α=0 corresponds to pure coverage. Across all tested token budgets, lower α is associated with lower ECE, while accuracy changes remain small. The best ECE in this table is achieved at K=128, α=0.

	K=64		K=128		K=192	
α	Acc	ECE	Acc	ECE	Acc	ECE
0.0	**85.2**	**.024**	**87.1**	**.016**	**87.6**	**.018**
0.5	86.0	.026	87.3	.017	87.5	.020
1.0	85.5	.032	86.9	.023	87.3	.027

options ($A/B/C/D$), which allows us to extend the analysis from binary decisions to multi-class answer selection. For this benchmark, confidence is defined as the maximum normalized probability among the four answer options.

4 Results

4.1 Default SCOPE: U-Shaped Curve

Table 1 shows a clear U-shaped trend for calibration under default SCOPE pruning on POPE. As the token budget decreases from the full setting to K=128, calibration improves consistently across ECE, Brier score, NLL, and AURC. When pruning becomes more aggressive, this trend reverses.

Moderate pruning gives the best trade-off in this setting. At K=128, ECE decreases from 0.041 to 0.024 ($p < 0.001$, bootstrap 95% CI: [0.018, 0.030] vs. [0.035, 0.048] for the full model), while accuracy remains unchanged at 86.9%. The corresponding optimal temperature is also closer to 1.0 than in the full-token setting, which suggests that the pruned model requires less post-hoc correction.

However, stronger pruning eventually hurts both confidence quality and task performance. At K=32, ECE increases to 0.045, which is slightly worse than the full model, and accuracy drops to 82.7%. This result suggests that moderate pruning can remove redundant visual information, but overly aggressive pruning starts to discard evidence that is still useful for prediction.

4.2 Alpha Sweep: Saliency vs. Coverage

Table 2 and Fig. 2 show how calibration changes as we vary the saliency weight within the SCOPE formulation. This comparison is controlled in the sense that the code path, attention source, and coverage objective remain the same, and only α is changed.

A consistent trend appears across all three token budgets. As α decreases from 1.0 to 0.0, ECE also decreases at K=64, 128, and 192. Compared with the default setting α=1, pure coverage (α=0) reduces ECE from 0.032 to 0.024 at K=64, from 0.023 to 0.016 at K=128, and from 0.027 to 0.018 at K=192. This pattern suggests that, under our setting, placing less weight on saliency leads to better calibration.

At the same time, the corresponding accuracy changes are small. Across all values of α and K, the difference in accuracy stays within 0.9%. This gap between calibration and accuracy is important: changing the saliency weight has a clear effect on confidence quality even when standard task performance remains almost unchanged.

The best result in this sweep is obtained at K=128 and α=0, where ECE reaches 0.016. This value is lower than the ECE of the full unpruned model (0.041), while accuracy remains comparable. Taken together, these results indicate that moderate coverage-based pruning can improve calibration without requiring a noticeable accuracy trade-off in this setting.

4.3 Comparison with External Baselines

Table 3 compares the best SCOPE setting from the α-sweep with several external baselines that use different token selection rules.

At both K=128 and K=64, pure-coverage SCOPE achieves lower ECE than the default SCOPE setting, saliency-only pruning, and random pruning. The difference is especially clear at K=128, where SCOPE with α=0 reaches 87.1% accuracy and 0.016 ECE, compared with 86.9% and 0.023 for default SCOPE, 84.4% and 0.051 for saliency-only pruning, and 83.6% and 0.046 for random pruning. These comparisons suggest that the improvement is related to how tokens are selected, rather than to compression alone.

FastV behaves very differently from the other methods in our evaluation. At K=128, its accuracy drops to 50.1% and ECE increases to 0.326, indicating severe degradation in both prediction quality and calibration. Since FastV uses attention from early LLM layers as its pruning signal, this result suggests that task-conditioned saliency may be much less stable than coverage-based selection in our setting.

Taken together with the internal α-sweep, these external comparisons support the same overall pattern: methods that

Table 3. Comparison with external baselines on POPE (9K), together with the best SCOPE configuration from the internal α-sweep. In our experiments, pure-coverage SCOPE achieves the lowest ECE among the compared methods at both K=128 and K=64. FastV shows severe degradation under this evaluation setting.

K	Method	Acc ↑	ECE ↓	Brier ↓	Overconf. (%)
576	Full (no prune)	86.9	0.041	0.099	+3.6
128	**SCOPE α=0 (pure coverage)**	**87.1**	**0.016**	**0.094**	+1.3
	SCOPE α=1 (default hybrid)	86.9	0.023	0.095	+2.2
	Saliency-only (CLS top-K)	84.4	0.051	0.113	+5.1
	Random (3 seeds)	$83.6_{\pm 0.2}$	$0.046_{\pm 0.002}$	0.119	+4.5
	FastV, real (LLM layer-2 attention)	50.1	0.326	0.360	+32.6
64	**SCOPE α=0 (pure coverage)**	**85.2**	**0.024**	**0.105**	+2.4
	SCOPE α=1 (default hybrid)	85.5	0.032	0.104	+3.2
	Saliency-only	80.5	0.090	0.147	+8.7
	Random	79.5	0.069	0.144	+7.0

Table 4. Calibration on ScienceQA-IMG (multiple-choice, 2K samples). As the token budget decreases, ECE also decreases, while accuracy remains stable or slightly improves.

K	Acc	ECE↓	Conf	Overconf
576 (full)	63.9	.179	81.8	+17.9
192	63.4	.177	81.2	+17.7
128	64.0	.170	81.0	+17.0
64	64.6	.164	81.0	+16.4
32	**65.1**	**.162**	81.3	+16.2

rely more heavily on saliency tend to show worse calibration than coverage-based selection, while pure-coverage SCOPE gives the most favorable calibration results among the methods we compare here.

4.4 Visualization and Analysis

Figure 3 summarizes the main results on POPE from several complementary views. Panel (a) shows the U-shaped calibration trend under default SCOPE: ECE decreases from the full-token setting to K=128, and then increases again when the token budget becomes smaller. Panel (b) compares SCOPE with random pruning and shows that the two methods behave differently even at the same compression ratio.

Panels (c) and (d) present reliability diagrams for the full model and for SCOPE at K=128. The main difference appears in the high-confidence region (0.9–1.0), where most predictions are concentrated. After pruning to K=128, the confidence-accuracy gap in this region becomes smaller, which is consistent with the lower ECE reported in Table 1.

4.5 Cross-Benchmark: ScienceQA

Table 4 extends the analysis to a multiple-choice benchmark. Compared with POPE, ScienceQA shows a different trend: as the token budget decreases, ECE also decreases, and accuracy remains stable or improves slightly. The best result in this table is obtained at K=32, where accuracy increases from 63.9% to 65.1% and ECE decreases from 0.179 to 0.162.

This result suggests that, for ScienceQA under our setting, pruning does not introduce the same calibration-accuracy trade-off observed under aggressive pruning on POPE. One possible explanation is that the visual information used by this benchmark is more redundant with respect to the model's decision process, although verifying this more carefully would require additional analysis.

It is also worth noting that the absolute ECE values on ScienceQA are consistently higher than those on POPE. This difference is expected because ScienceQA involves four answer options rather than a binary decision, making calibration more challenging overall.

4.6 Per-Split Analysis (POPE)

The calibration improvement at K=128 is consistent across all three POPE splits. The ECE reduction is largest on the adversarial split (0.037→0.018), followed by the popular split (0.040→0.021) and the random split (0.045→0.031). This result shows that the gain is not driven by a single subset of the benchmark.

4.7 Temperature Scaling (Cross-Validated)

We further apply 5-fold cross-validated temperature scaling. Post-hoc scaling reduces ECE for all token budgets, and K=128 remains the best setting after scaling, reaching 0.012 ECE versus 0.017 for the full-token model. This result suggests that the advantage of moderate pruning is preserved after a standard post-hoc calibration step.

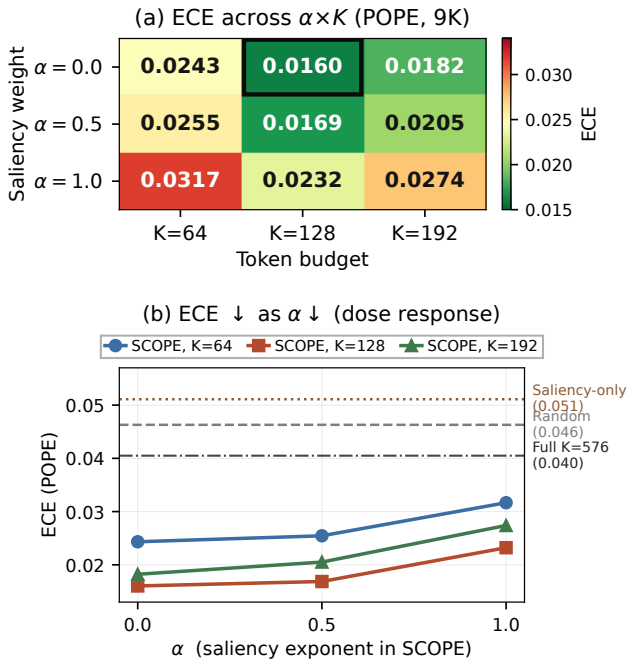

Figure 2. (a) ECE heatmap across α and K on POPE. The lowest ECE at each tested token budget is achieved at α=0, and the best overall setting in this sweep is (α=0, K=128). (b) ECE decreases as the saliency weight is reduced for all three token budgets. For reference, we also show the random baseline, the saliency-only baseline at K=128, and the full unpruned model.

4.8 Selective Prediction

We also evaluate selective prediction by abstaining on low-confidence samples. At 80% coverage, SCOPE K=128 reaches 94.2% accuracy on the covered subset, compared with 92.8% for the full model, and AURC decreases from 0.047 to 0.042. This indicates that the calibration gain at K=128 is also reflected in a simple confidence-based abstention setting.

5 Gap Power Analysis

Beyond the calibration results above, we also examine one design choice inside coverage-based selection: the *gap power exponent* p in the SCOPE objective. This analysis is intended as a complementary study of the coverage term itself. While the main part of the paper focuses on calibration, the gap power results help us better understand how the shape of the coverage objective affects downstream performance.

Existing coverage-based methods typically use p=1 by default. Here we generalize the objective to $p \in \{1.0, 1.2, 1.5, 2.0\}$ and evaluate it across 7 benchmarks and 7 token budgets, for a total of 196 experiments.

Table 5. MME score under different gap power values p across token budgets. Bold indicates the best result in each row.

K	p=1.0	p=1.2	p=1.5	p=2.0	SCOPE
32	**1640**	1630	1629	1638	1648
48	1663	1659	1654	**1668**	1650
64	1698	1695	**1719**	1715	1698
96	1758	**1762**	1761	1748	1763
128	1773	**1783**	1775	1766	1776
192	1797	1801	**1806**	1794	1804
256	**1791**	1780	1782	1786	1769

$$v^* = \arg\max_{v \in \mathcal{V} \setminus \mathcal{S}} \sum_{u \in \mathcal{V}} [\max(0, \mathrm{sim}(u, v) - C(u, \mathcal{S}))]^p \cdot A_v^\alpha \quad (1)$$

The exponent p changes how strongly the selector emphasizes large uncovered regions. When p=1, the gain is linear in the remaining coverage gap. When p>1, larger gaps receive relatively more weight, so the selector becomes more sensitive to under-covered regions. From an algorithmic perspective, this parameter controls how evenly the selected tokens are encouraged to cover the original feature set.

Table 5 and Fig. 4 show that the default setting p=1 is not always the strongest choice. On MME, different token budgets favor different values of p. The largest improvement appears at K=64, where p=1.5 improves the score from 1698 to 1719. At K=192, p=1.5 also gives the best result, though with a smaller margin. In contrast, at some other budgets, such as K=32 or K=256, the gains are limited or inconsistent. These results suggest that the effect of p is real, but not uniform across compression regimes.

To test whether this pattern generalizes beyond MME, Table 6 compares p=1.5 with the default SCOPE setting on seven benchmarks at K=64 and K=192. The improvement is most visible on MME, while the changes on other benchmarks are generally small. For example, at K=64, p=1.5 improves MME by 21 points, but the differences on MMB, SQA, TVQA, POPE, SEED, and GQA remain within a narrow range. A similar pattern appears at K=192, where p=1.5 matches or slightly improves upon the default setting on most benchmarks without causing large regressions.

Overall, these results suggest that the gap power exponent is a non-negligible design parameter in coverage-based pruning. In particular, values around p=1.2 or p=1.5 can be competitive with, and sometimes better than, the standard linear choice. At the same time, this analysis is more modest in scope than our main calibration results: it shows that the default coverage objective is not always optimal, but it does not by itself establish a universal best choice for all models, tasks, or token budgets.

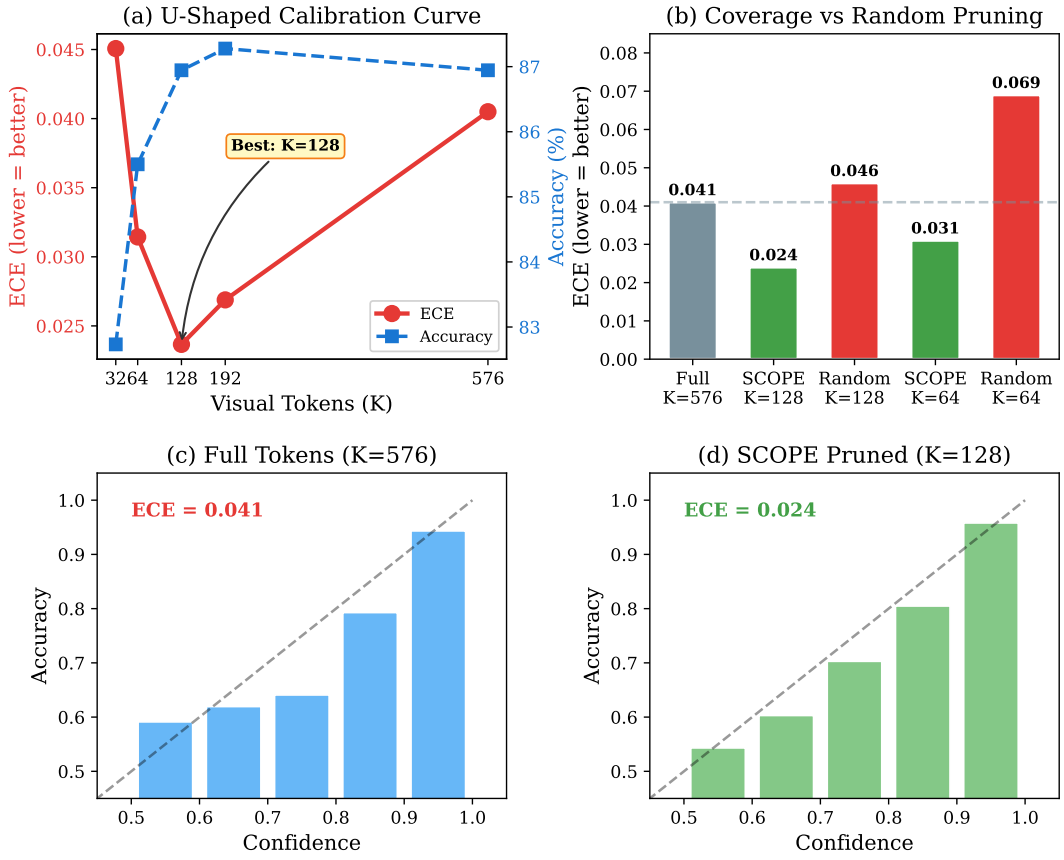

Figure 3. **Main results on POPE.** (a) ECE under default SCOPE across token budgets, with accuracy shown for reference. The best calibration is observed at K=128. (b) Comparison between SCOPE and random pruning at two token budgets. (c,d) Reliability diagrams for the full model (K=576) and SCOPE pruning at K=128. The high-confidence region shows a smaller confidence-accuracy gap after pruning.

Table 6. Comparison between p=1.5 and the default SCOPE setting at K=64 and K=192 across seven benchmarks.

	K=64			K=192		
	SCOPE	p=1.5	Δ	SCOPE	p=1.5	Δ
MME	1698	**1719**	+21	1804	**1806**	+2
MMB	61.7	61.0	−0.7	63.6	63.7	+0.1
SQA	68.8	68.8	0.0	68.8	68.8	0.0
TVQA	56.6	56.7	+0.1	57.7	57.8	+0.1
POPE	83.9	83.8	−0.1	86.4	86.5	+0.1
SEED	56.3	55.9	−0.4	58.7	58.7	0.0
GQA	58.3	57.9	−0.4	60.1	60.0	−0.1

6 Discussion

Why might coverage help while saliency hurts calibration? Our results suggest that coverage-based and saliency-based pruning behave differently with respect to confidence quality. A possible explanation is that coverage-based selection tends to preserve a more representative subset of the visual input, while saliency-based selection focuses more heavily on a small set of highly scored regions. Under this view, coverage may reduce redundancy without removing too much supporting evidence, whereas a stronger saliency bias may make the retained token set less balanced.

This interpretation is consistent with several observations in our experiments. Within SCOPE, reducing the saliency weight improves ECE at all tested token budgets, while the changes in accuracy remain small. Random pruning does not show the same behavior, which suggests that the effect is related to the selection rule rather than to compression alone. The reliability diagrams also show that, after moderate pruning, the confidence-accuracy gap becomes smaller in the highest-confidence range.

At the same time, we do not claim that saliency is always harmful or that coverage is universally sufficient. Our results are limited to the models, benchmarks, and pruning settings studied here. A more cautious conclusion is that, in our experiments, placing too much weight on saliency is associated with worse calibration, while coverage-based selection gives more favorable confidence behavior.

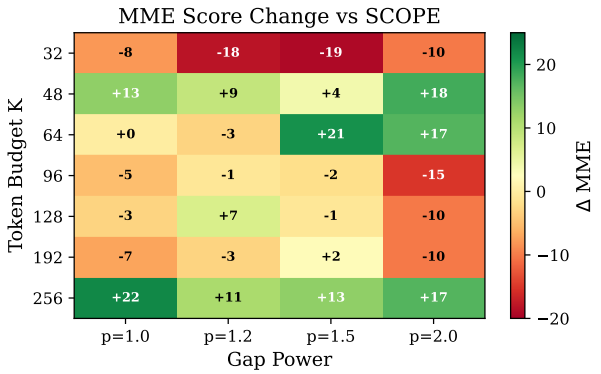

Figure 4. Change in MME score relative to the default SCOPE setting across gap power values and token budgets. Positive values indicate improvement over the baseline configuration. The largest gain is observed at K=64 with p=1.5.

Relation to the gap power exponent. The gap power analysis provides a complementary view of the coverage objective. Our results show that the default linear choice p=1 is not always best, and values around p=1.2 or p=1.5 can improve performance at some token budgets. This supports the broader point that the design of the coverage term still leaves useful room for optimization.

Implications for multimodal reasoning systems. For multimodal systems that rely on confidence to decide whether to answer, abstain, or defer, calibration matters in a practical sense. Our selective prediction results show that the calibration improvement at K=128 is reflected not only in ECE, but also in confidence-based abstention. This makes the result relevant for downstream settings where a model's confidence is used as part of a decision rule.

More broadly, our results suggest that efficiency-oriented modifications to MLLMs should not be evaluated only by task accuracy. Two pruning methods may have similar accuracy, yet differ meaningfully in how reliable their confidence scores are. For multimodal reasoning systems, this difference can matter even when the final accuracy change is small.

Practical recommendations. We recommend that future pruning work report calibration together with accuracy, and that hybrid selectors expose the saliency weight as a tunable parameter. In our setting, moderate coverage-based pruning provides the best trade-off between efficiency and confidence quality.

Limitations and future work. This study has several limitations. First, the main α-sweep is conducted only on POPE, so the calibration trend should be tested more broadly on additional multiple-choice and open-ended benchmarks. Second, we evaluate a single base model, LLaVA-1.5-7B. It remains unclear how well the same pattern transfers to other MLLMs with different visual encoders, projectors, or training data. Third, our confidence definition is based on first-token answer probabilities over a restricted verbalizer set. This choice is simple and consistent across pruning settings, but other confidence definitions may reveal additional details. Finally, our findings concern visual token pruning in isolation. It would be useful to study how calibration changes when token pruning is combined with other efficiency techniques such as quantization or KV-cache pruning.

7 Conclusion

In this paper, we study how visual token pruning affects calibration in multimodal large language models. Using LLaVA-1.5-7B on POPE and ScienceQA-IMG, we show that pruning should be evaluated not only by task accuracy, but also by confidence quality. Under our setting, moderate coverage-based pruning can improve calibration while maintaining similar accuracy, and on POPE the best configuration achieves lower ECE than the full unpruned model.

Our internal α-sweep within SCOPE shows a consistent trend across multiple token budgets: reducing the saliency weight improves calibration, while the corresponding accuracy changes remain small. External comparisons with random pruning, saliency-only pruning, and FastV further support the observation that token selection strategy matters for confidence quality, not only the compression ratio. We also find that the gap power exponent in the coverage objective is not always best at its default value, which suggests additional design space inside coverage-based pruning.

Overall, our results suggest that efficiency and reliability do not necessarily conflict in visual token pruning. More broadly, they indicate that confidence-aware evaluation should become a standard part of pruning research for MLLMs, especially in settings where model confidence may be used to decide whether to answer, abstain, or defer.

References

[1] 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*, 2025.

[2] Changwoo Baek, Jouwon Song, Sohyeon Kim, and Kyeongbo Kong. AgilePruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models. *arXiv preprint arXiv:2603.01236*, 2026.

[3] 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 *Proceedings of the European Conference on Computer Vision*, 2024.

[4] Xiwen Chen, Wenhui Zhu, Gen Li, Xuanzhao Dong, Yujian Xiong, Hao Wang, Peijie Qiu, Qingquan Song, Zhipeng Wang, Shao Tang, Yalin Wang, and Abolfazl Razi. OTPrune: Distribution-aligned visual token pruning via optimal transport. *arXiv preprint arXiv:2602.20205*, 2026.

[5] Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

[6] 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.

[7] Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. SCOPE: Saliency-coverage oriented token pruning for efficient multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2025.

[8] Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in VLMs via synergistic importance-diversity. In *International Conference on Learning Representations*, 2026.

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

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

[12] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

[13] Runyu Ma, Songqing Chen, and Shuochao Yao. Better reliability compression: Model pruning with calibrated uncertainty estimation for mobile deep learning applications. In *2025 IEEE 3rd International Conference on Mobility, Operations, Services and Technologies (MOST)*, 2025.

[14] Diganta Misra, Muawiz Chaudhary, Agam Goyal, Bharat Runwal, and Pin-Yu Chen. Uncovering the hidden cost of model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[15] Pallavi Mitra, Gesina Schwalbe, and Nadja Klein. Investigating calibration and corruption robustness of post-hoc pruned perception CNNs: An image classification benchmark study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[16] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. CDPruner: Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *Advances in Neural Information Processing Systems*, 2025.

[17] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. VL-Uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

[18] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[19] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems*, 2024.