

When Does Visual Token Pruning Improve Calibration? The Role of Evidence Coverage in MLLMs

Kaizhen Tan¹, Yang Feng², Heqing Du², Hanzhe Hong¹, Siru Tao¹, Xin Xu¹

¹Carnegie Mellon University, Pittsburgh, PA, USA

²Columbia University, New York, NY, USA

Abstract

Visual token pruning is widely used to reduce the inference cost of multimodal large language models (MLLMs), but it is usually evaluated only by accuracy. We study how pruning affects *calibration*, defined as the agreement between confidence and correctness, and show that the selection rule matters more than the token budget alone. On POPE with LLaVA-1.5, coverage-based pruning from 576 to 128 tokens reduces expected calibration error from 0.041 to 0.016 without a statistically significant accuracy loss. In contrast, attention-based selection preserves confidence while accuracy deteriorates, becoming less calibrated than random pruning at aggressive budgets. Across pruning conditions, kept-set coverage is strongly associated with accuracy (Spearman $\rho=+0.89$) but not with mean confidence ($\rho=-0.03$), producing a strong inverse relation with overconfidence ($\rho=-0.92$); controlled kept-set interventions support this evidence-coverage account. Two boundaries limit it: query-conditioned FastV is more overconfident than its coverage predicts, and on language-prior-dominated ScienceQA coverage ceases to order calibration at all. The selector ordering otherwise generalizes to GQA and LLaVA-NeXT, and coverage beats random on Qwen2-VL, although the calibration gain over the unpruned model is task- and model-dependent. We also identify an evaluation pitfall in FastV: zeroing rather than removing pruned tokens can reduce accuracy to chance. Visual token pruning therefore changes confidence quality as well as efficiency, and calibration should be evaluated alongside accuracy when comparing pruning methods.

Introduction

Multimodal Large Language Models (MLLMs) (Liu et al. 2023, 2024) spend most of their inference cost on hundreds of visual tokens per image. Visual token pruning reduces this cost by selecting a subset of tokens before or inside the language model, and a rapidly growing line of work explores selection signals based on saliency (Chen et al. 2024; Zhang et al. 2025b), coverage (Deng et al. 2025; Cho et al. 2026), diversity (Alvar et al. 2025; Zhang et al. 2025a, 2026a), optimal transport (Chen et al. 2026a), and adaptive policies (Ye et al. 2025; Zhang et al. 2026b; Baek et al. 2026). A recent survey catalogues more than a hundred such methods (Shao et al. 2026). Their evaluation protocol, however, is nearly uniform: task accuracy at a given token budget.

Accuracy alone is an incomplete measure of what pruning does to a model. In deployment, MLLM confidence is used to

decide whether to answer or abstain (Geifman and El-Yaniv 2017), defer to a human, or trigger a retrieval step; calibration, the agreement between confidence and correctness, is the property those decisions rely on. MLLMs are known to be overconfident (Chen et al. 2025), and a growing body of work measures and improves multimodal confidence quality (Zhou et al. 2024; Zhang, Zhang, and Zheng 2024; Bhattacharya, Kulkarni, and Ganapathy 2025; Lau et al. 2026; Du et al. 2025). Yet none of this work asks what happens to confidence when the *visual evidence itself* is compressed, and to our knowledge no pruning paper reports calibration. The question matters because token pruning is unlike weight pruning: it removes part of the model’s input evidence, not its parameters, so its effect on confidence can be qualitatively different from the mixed findings reported for compressed CNNs (Misra et al. 2024; Mitra, Schwalbe, and Klein 2024; Ma, Chen, and Yao 2025).

This paper studies the pruning–calibration interaction systematically. SCOPE (Deng et al. 2025) serves as a controlled testbed: its selection score $\Delta_{\text{cov}}(v) \cdot a(v)^\alpha$ interpolates between pure coverage ($\alpha=0$) and saliency-weighted coverage ($\alpha=1$) inside one code path, so the saliency exponent α sweeps the coverage–saliency axis with everything else held fixed. We compare it against saliency-only, random, FastV (Chen et al. 2024), and VisionZip (Yang et al. 2025) selection at matched token budgets K , on POPE (Li et al. 2023), ScienceQA-IMG (Lu et al. 2022), and GQA (Hudson and Manning 2019), with LLaVA-1.5-7B, LLaVA-NeXT-Vicuna-7B, and Qwen2-VL-7B (Wang et al. 2024).

Our contributions are:

- **A calibration-centric evaluation protocol for token pruning**, covering six confidence-quality metrics, multiple confidence definitions, binning-sensitivity checks, and index-paired bootstrap tests, so that selection rules can be compared on confidence quality as rigorously as on accuracy.
- **An empirical finding**. On POPE, moderate coverage-based pruning improves calibration at statistically unchanged accuracy (ECE 0.041→0.016 at $K=128$); less saliency weight helps at every budget, and saliency-only selection is significantly worse calibrated than random at aggressive budgets. The cross-selector ordering holds on GQA, on LLaVA-NeXT, and on Qwen2-VL, where coverage beats random at both ratios.

- **An evidence-coverage account.** Pruning trims confidence mostly on errors, which amounts to an implicit temperature scaling plus gains that survive post-hoc scaling. Measuring what each selector keeps shows why: accuracy tracks kept-set coverage ($\rho=+0.89$) while confidence is independent of it ($\rho=-0.03$), so overconfidence is the gap that opens as coverage falls ($\rho=-0.92$; robust to independent feature spaces; monotone when we manipulate the kept set within a selector). The account also predicts its own boundary: where a task is answered from language priors, coverage stops ordering calibration, as ScienceQA shows.
- **An evaluation-pitfall diagnosis.** Implementing FastV by zeroing pruned hidden states, instead of removing them, collapses POPE accuracy to chance because zeroed keys still absorb attention mass; our corrected implementation restores sensible behavior and reveals the honest failure mode: flat, unshakeable confidence over collapsing accuracy.

We will release the unified evaluation runner, the corrected FastV implementation with its no-drop equivalence check, and the offline kept-set coverage tooling.

Related Work

Visual token pruning for MLLMs. Training-free pruning methods differ mainly in the selection signal. Saliency-based methods rank tokens by attention: FastV uses early-LLM-layer attention (Chen et al. 2024), SparseVLM uses text-conditioned attention (Zhang et al. 2025b), and VisionZip uses vision-encoder attention with contextual merging (Yang et al. 2025). PyramidDrop schedules pruning across depth (Xing et al. 2025). Coverage- and diversity-based methods instead optimize how well the kept set represents all tokens: SCOPE greedily maximizes facility-location coverage weighted by saliency (Deng et al. 2025), FLoC applies facility location to video (Cho et al. 2026), DivPrune and CDPruner maximize diversity (Alvar et al. 2025; Zhang et al. 2025a), and OTPrune aligns kept and original distributions via optimal transport (Chen et al. 2026a). Recent 2025–2026 work moves toward adaptivity and bias-correction, with instance-adaptive budgets (Ye et al. 2025; Zhang et al. 2026b), empirical attention–diversity trade-offs (Baek et al. 2026), importance–diversity hybrids and debiased importance (Fang et al. 2026; Zhang et al. 2026a), and globally optimized or early-stage schedules (Li et al. 2025a; Wang et al. 2026; Chen et al. 2026b). All of these are evaluated by accuracy and efficiency; none reports confidence quality. Our study is orthogonal: we ask what any such selector does to calibration, and find that the coverage–saliency axis they differ along also organizes it.

Calibration and uncertainty in MLLMs. Chen et al. (2025) document persistent overconfidence across MLLMs; Zhou et al. (2024) improve confidence via calibrated self-rewarding training; VL-Uncertainty perturbs both modalities to detect hallucination (Zhang, Zhang, and Zheng 2024). More recent work estimates MLLM uncertainty by functionally-equivalent input sampling (Bhattacharya, Kulkarni, and Ganapathy 2025) or the semantic volume of

sampled responses (Lau et al. 2026), trains models to verbalize confidence (Li et al. 2025b), and studies calibration empirically in medical VQA (Du et al. 2025); Liu et al. (2025) survey LLM uncertainty quantification. These works estimate or improve confidence while the visual evidence the model answers from stays intact; we instead study how an efficiency intervention that removes part of that *input evidence* changes calibration, and find the token selector is itself a calibration knob that operates before any post-hoc method.

Compression and calibration. For CNNs, weight pruning can hurt or help calibration depending on structure and ratio (Misra et al. 2024; Mitra, Schwalbe, and Klein 2024), and reliability can degrade faster than accuracy (Ma, Chen, and Yao 2025). Token pruning differs in kind: parameters are untouched and the *evidence* is reduced. That difference is what our account turns on, since the confidence a pruned MLLM keeps depends on which evidence survives, a variable weight pruning does not have.

Study Design

Models and pruning methods. Our main model is LLaVA-1.5-7B (Liu et al. 2023) with CLIP-ViT-L/14-336 (Radford et al. 2021) (576 visual tokens); generalization runs use LLaVA-NeXT-Vicuna-7B (Liu et al. 2024) (AnyRes, ~ 2880 tokens) and Qwen2-VL-7B (Wang et al. 2024), an architecturally distinct family (dynamic-resolution ViT without a CLS token, 2×2 token merger, M-RoPE, Qwen2 LM), where we port input-level selection: chosen merged vision tokens keep their original 3D positions, and when no token is dropped the pipeline reproduces the unmodified model’s standard `generate()` outputs exactly, which validates the port. SCOPE (Deng et al. 2025) scores candidates by $\Delta_{\text{cov}}(v; \mathcal{S}) \cdot a(v)^\alpha$, where Δ_{cov} is the facility-location coverage gain over penultimate-layer CLIP features and $a(v)$ is CLS attention (summed over heads, same layer). The exponent α interpolates from pure coverage ($\alpha=0$) to the default hybrid ($\alpha=1$) *inside one code path*, holding the attention source, feature space, and pipeline fixed. External baselines: *saliency-only* (top- K by the same CLS attention), *random* (seeds $\{0, 1, 2\}$), *FastV* (Chen et al. 2024) (corrected implementation, below), and *VisionZip* (Yang et al. 2025).

Benchmarks. POPE (Li et al. 2023) asks 9,000 yes/no object-existence questions over COCO images, evenly split into random, popular, and adversarial subsets; it is almost entirely visually grounded, which makes it a clean calibration testbed. ScienceQA-IMG (Lu et al. 2022) contains 2,017 multiple-choice science questions with images; as we show later, most of its performance is driven by language priors. GQA (Hudson and Manning 2019) adds a strongly visual task with an *open* answer vocabulary (we evaluate 6,000 testdev-balanced questions); confidence there is the maximum softmax probability of the first generated token and correctness is exact match on the normalized answer. The triple spans the visual-dependence and answer-space axes.

A FastV implementation pitfall. FastV keeps all tokens through the first $L=2$ LLM layers, ranks image tokens by the attention they receive at layer $L-1$, and *removes* low-ranked

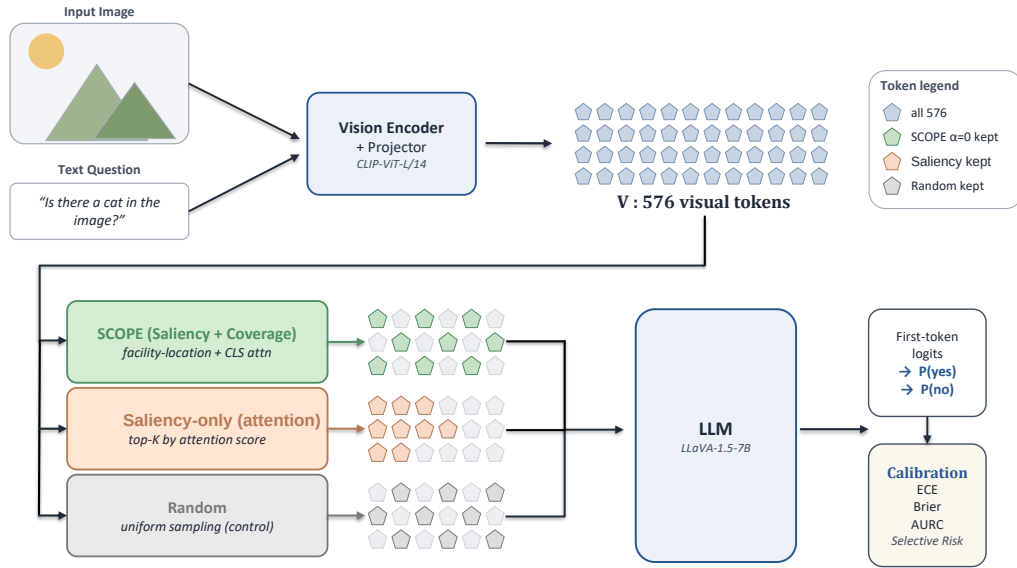


Figure 1: **Study design.** An image is encoded into $V=576$ visual tokens. Selectors differing in coverage/saliency emphasis (SCOPE $\alpha \in \{0, 0.5, 1\}$, saliency-only, random; FastV and VisionZip are evaluated as external baselines) pass the kept K tokens to the LLM; we extract first-token answer probabilities and evaluate calibration.

tokens for all later layers. A tempting shortcut, zeroing the hidden states of pruned positions, is not equivalent: a zeroed position still contributes a key, its attention logit is 0 rather than $-\infty$, and after softmax hundreds of zeroed positions absorb a large share of every attention distribution. In our audit this artifact collapsed POPE accuracy to 50.1% (chance) at $K=128$; an earlier version of this study reported such numbers as “FastV”, and the shortcut is easy to write without noticing, which is precisely why we document it. Our corrected implementation physically removes tokens from the sequence and per-layer KV cache, keeps original rotary positions, and reproduces vanilla outputs exactly when no token is dropped (argmax agreement on 50/50 audited samples, max logit deviation 0.016 at fp16). All FastV numbers below use the corrected version, evaluated both at matched budgets ($K \in \{64, 128\}$) and at its native operating point (50% retention, $K=288$).

Confidence extraction. All runs use greedy decoding and identical prompts across pruning settings. For POPE we read the first generated token’s logits and compute $P_{\text{yes}}, P_{\text{no}}$ by summing softmax probabilities of case variants; confidence is $\max(P_{\text{yes}}, P_{\text{no}})/(P_{\text{yes}} + P_{\text{no}})$. For ScienceQA-IMG we normalize over the option letters. A restricted verbalizer could in principle leave probability mass unaccounted for. It does not: in every condition that logged them, the unnormalized masses $P_{\text{yes}} + P_{\text{no}}$ average above 0.999 and never fall below 0.9, so normalizing over the answer set is close to a no-op, and conclusions are unchanged under the *unnormalized* confidence $\max(P_{\text{yes}}, P_{\text{no}})$ wherever it was logged. Verbalized self-confidence is analyzed in the robustness subsection.

Metrics. Primary: Expected Calibration Error (ECE) with 15 equal-width bins; we also report *overconfidence*, mean

Table 1: **Moderate pruning is a free calibration win.** Default SCOPE ($\alpha=1$) on POPE (9K), LLaVA-1.5-7B; deltas vs. the full model with paired-bootstrap 95% CIs. At $K \in \{128, 192\}$ calibration improves while accuracy is statistically unchanged; below that, accuracy starts to pay. Lower AURC is better.

K	Acc	$\Delta\text{Acc CI}$	ECE	$\Delta\text{ECE CI}$	AURC
576	86.9	—	.041	—	.047
192	87.3	$[-0.1, 0.7]$	.027	$[-.019, -.009]$	.040
128	86.9	$[-0.4, 0.4]$	.024	$[-.022, -.012]$	.042
64	85.5	$[-2.0, -0.9]$	.031	$[-.014, -.004]$	.048
32	82.7	$[-4.9, -3.6]$	.045	$[-.002, +.011]$	.064

confidence minus accuracy. Robustness: 10/15/20 bins, equal-width and equal-mass (Nixon et al. 2019). Proper scores: Brier and negative log-likelihood (NLL). Ranking: the area under the risk-retention curve (AURC) (Geifman and El-Yaniv 2017) and the AUROC of confidence for predicting correctness. Post-hoc: 5-fold cross-validated temperature scaling (Guo et al. 2017). All confidence intervals are 1,000-sample percentile bootstraps; comparisons against the full model are *index-paired* bootstrap tests on the same 9,000 POPE questions.

Results

Moderate Pruning Improves Calibration

Table 1 shows the U-shaped calibration curve under default SCOPE. From the full model down to $K=128$, ECE and AURC improve steadily (Brier and NLL follow the same pattern); the ECE reduction at $K \in \{64, 128, 192\}$ is significant under index-paired bootstrap ($p < 0.01$), while the accuracy

Table 2: **Less saliency weight, better calibration.** α -sweep within SCOPE on POPE (9K). ECE improves monotonically as α falls at every budget, while accuracy moves ≤ 0.9 points (ECE CIs $\approx \pm .005$). The $\alpha=1$ rows independently replicate Table 1 and agree within .001, which bounds run-to-run noise.

α	$K=64$		$K=128$		$K=192$	
	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	.021
1.0	85.5	.032	86.9	.023	87.3	.027

change at $K \in \{128, 192\}$ is not. Only at $K=32$ does the trend reverse: accuracy drops 4.2 points and the ECE difference vs. full becomes insignificant (CI $[-0.002, +0.011]$). Under coverage-based selection, moderate pruning is thus a strict calibration improvement, not a trade; as the baselines below show, this is a property of the selector, not of pruning per se.

The Selection Rule Matters: α -Sweep

Holding everything fixed except α , calibration improves monotonically as saliency weight decreases (Table 2; heatmap in the supplementary material): at $K=128$, ECE falls from .023 ($\alpha=1$) to .016 ($\alpha=0$), and analogously at $K=64$ and $K=192$. Accuracy changes stay within 0.9 points; at $K=192$, $\alpha=0$, accuracy is in fact significantly *above* the full model ($+0.69$, CI $[0.24, 1.11]$). Confidence quality therefore responds to the selection rule even where accuracy barely moves; accuracy-only evaluation cannot see this axis.

External Baselines

Table 3 places the sweep endpoints among external selectors. Coverage-based SCOPE dominates on calibration at both budgets; index-paired pairwise tests at $K=128$ confirm that $\alpha=0$ beats every alternative (Δ ECE vs. $\alpha=1$: $-.007$, $p=.04$; vs. VisionZip $-.028$, saliency-only $-.035$, random $-.032$, FastV $-.131$, all $p<.01$). Saliency-only selection is the interesting failure: it keeps the most “important” tokens by the model’s own attention, yet is significantly worse calibrated than random at $K=64$ ($+.018$, $p<.01$) and no better at $K=128$ ($+.003$, n.s.), while keeping confidence essentially at full-model levels (.895 vs. full .906 mean confidence).

VisionZip sits exactly where the coverage account predicts: its attention-selected dominant tokens are augmented with *merged contextual* tokens that restore part of the lost coverage, and its calibration lands between saliency-only and coverage-based selection (ECE .044; significantly better than saliency-only and significantly worse than pure coverage, both $p<.01$).

Corrected FastV carries the pattern to its extreme. At the 50% retention it was designed for it behaves reasonably, but squeeze it to the budgets we compare everything else at and its accuracy collapses while its confidence does not move at all, staying *above* the full model’s and opening a $+21.4$ -point overconfidence gap at $K=64$. Its variants shift the

Table 3: **The selector, not the budget, sets confidence quality.** External baselines on POPE (9K) at matched budgets. Random is the mean over seeds $\{0, 1, 2\}$ (across-seed sd ≤ 0.13 accuracy, $\leq .0012$ ECE). FastV uses our corrected implementation, at matched budgets and at its native 50%-retention point ($K=288$); the zeroing-artifact row is kept for the audit.

K	Method	Acc $\uparrow$	ECE $\downarrow$	Brier $\downarrow$
576	Full (no prune)	86.9	.041	.099
288	FastV ($L=2$, native 50%)	84.5	.068	.121
128	SCOPE $\alpha=0$ (coverage)	87.1	.016	.094
	SCOPE $\alpha=1$ (default)	86.9	.023	.095
	VisionZip	84.9	.044	.111
	Saliency-only (CLS top- K)	84.4	.051	.113
	Random (3 seeds)	83.5	.048	.119
	FastV ($L=2$)	77.8	.147	.180
	FastV, zeroing artifact	50.1	.326	.360
64	SCOPE $\alpha=0$ (coverage)	85.2	.024	.105
	SCOPE $\alpha=1$ (default)	85.5	.032	.104
	Saliency-only	80.5	.090	.147
	Random (3 seeds)	79.3	.071	.145
	FastV ($L=2$)	71.0	.215	.242

magnitude without touching the signature: pruning on a later layer’s attention helps monotonically ($L=2/3/5$), and the official last-token ranking clearly beats a mean-over-queries one (full table in the supplementary material), yet every version preserves the tokens that drive confidence rather than the context that keeps the model honest. The mechanism section quantifies exactly this.

Robustness of the Measurement

The ordering is stable across bin counts and binning schemes (full table in the supplementary material); equal-mass binning, robust to the shape of the confidence distribution, agrees with equal-width. It survives per-split analysis too, with the $K=128$ improvement appearing on all three POPE splits and largest on the adversarial one (.037 $\rightarrow$.011).

Confidence definitions matter more in principle, less in practice. In the binary POPE setting, margin and entropy are monotone transforms of max-probability confidence, so every ranking metric is identical by construction; the substantive alternatives are the unnormalized readout, which changes nothing (AUROC 0.8018 vs. 0.8024 for the full model), and *verbalized* self-confidence, which we elicited with a second prompt (“respond with an integer between 0 and 100”). Verbalized confidence turns out to be nearly degenerate at this scale: 94% of answers say “90” whatever the question, so its ECE merely tracks the accuracy gap and cannot register selection-level effects at all. That is consistent with the reported weaknesses of verbalized confidence in open models (Li et al. 2025b), and it leaves first-token probabilities as the channel with something to say.

Table 4: **Pruning takes confidence mostly from errors.** Mechanism evidence on POPE (9K). Fitted T (5-fold CV) moves toward 1; after temperature scaling the default hybrid matches the full model while pure coverage stays ahead; AUROC (confidence predicting correctness; higher is better) improves in every pruned setting.

Condition	conf @err	conf @corr	log- odds	T	AUROC	ECE after T
Full $K=576$	.782	.924	3.23	1.34	.802	.013
$\alpha=1$, $K=128$	.754	.912	2.88	1.15	.818	.013
$\alpha=0$, $K=128$	.741	.905	2.79	1.10	.815	.005
$\alpha=0$, $K=192$	.749	.911	2.92	1.13	.814	.008

Why Does Coverage Help? An Evidence-Coverage Account

Observation 1: asymmetric confidence shrinkage. Pruning does not cost the model confidence evenly. At $K=128$ it gives up three to four points of confidence on the answers it gets wrong but only one to two on the ones it gets right, and its mean absolute log-odds shrink from 3.2 to 2.9 (Table 4). Calibration improves precisely because the confidence the model loses is concentrated where that confidence was unwarranted.

Observation 2: implicit temperature scaling, plus effects temperature cannot reproduce. A natural suspicion is that pruning merely does temperature scaling in disguise. It partly does: through the moderate range the fitted temperature falls toward 1 (Table 4), so pruning performs much of the correction post-hoc scaling would apply, though it turns back up once pruning grows aggressive enough to cost accuracy. Scaling both models settles the rest, and separates the selectors. The default hybrid’s advantage dissolves, its raw-ECE gain being implicit rescaling and nothing more; pure coverage survives optimal scaling of both, yielding confidence no temperature on the full model reproduces. Two further effects lie outside temperature’s reach. Scaling is monotone and cannot reorder anything, yet every pruned setting ranks its own errors better (AUROC in Table 4; AUC in Table 1), and abstaining on the least confident fifth of POPE leaves the pruned model more accurate on what remains (93.0–93.4% vs. 92.7%). The gap widens under distribution shift, where scaling is weakest: a temperature fitted on the adversarial split still leaves the full model at .021 ECE on the adversarial split, while the pruned model sits at .011 unscaled. Coverage pruning therefore does more than imitate temperature scaling, and unlike scaling it needs no held-out split to fit.

Observation 3: why saliency fails. Saliency-only selection holds its confidence at nearly full-model level while shedding 2.5 points of accuracy, so its overconfidence *widens* rather than shrinks. Corrected FastV shows the same signature amplified: squeezed from 288 tokens down to 64, it loses 13.5 points of accuracy without conceding any confidence at all, ending +21.4 points overconfident. The reason is almost tautological once stated. High-attention tokens are the tokens that drive the answer logit, so keeping them and drop-

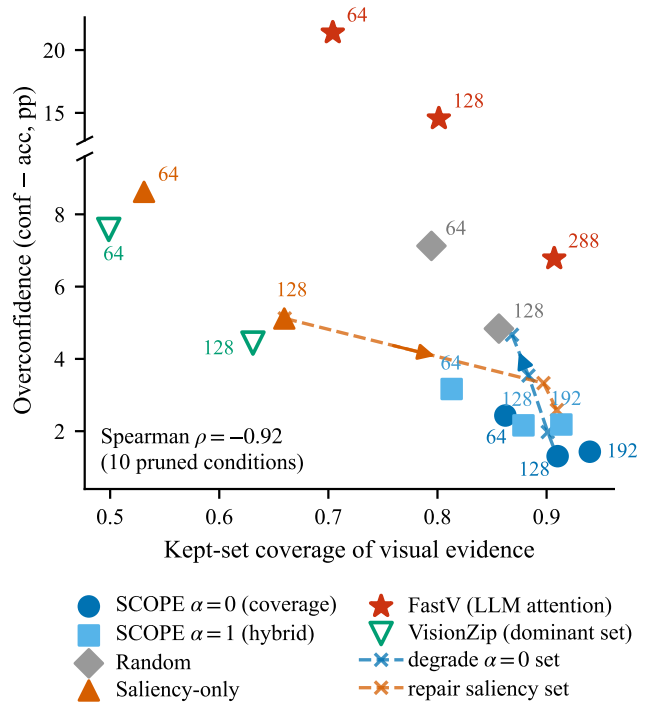


Figure 2: **Overconfidence tracks kept-set coverage.** Mean coverage of each selector’s kept set (recomputed for all 9,000 POPE questions) against overconfidence; point labels give the budget K , colored by selector. Encoder-level selectors follow one relation; query-conditioned FastV sits far above it. Arrowed dashed trajectories are the $K=128$ kept-set interventions, each starting at its own selector’s point; $\times$ marks the progressively swapped sets. The y -axis is broken to widen the dense band below 10 pp.

ping everything else preserves the model’s conviction while removing the context that would corroborate or contradict it. That fits what is known about high-attention positions, which often act as attention sinks or register-like aggregators rather than carriers of local visual evidence (Xiao et al. 2024; Darcet et al. 2024): selecting by attention optimizes for what the model already believes, not for what would check it.

An evidence-coverage account, tested. What property of the kept set tracks these effects? We recompute every selector’s choice offline for all 9,000 POPE questions, from the same features and attention the selectors themselves use, and measure three candidates: how well the kept set *covers* the full token set ($\frac{1}{N} \sum_u \max_{s \in S} \cos(u, s)$), how redundant it is internally, and how much of the CLS attention it captures. Coverage is the one that lines the selectors up (Fig. 2), and splitting overconfidence into its two terms says why.

Across the ten pruned conditions, accuracy rises with coverage (Spearman $\rho=+0.89$) while mean confidence shows no rank association with it at all ($\rho=-0.03$). That asymmetry is the finding. A selector that covers the image poorly costs the model correctness, but the model’s confidence does not register the loss, because the focal high-attention evidence

that anchors it survives almost any selection rule (Observation 1). Overconfidence is the gap those two terms leave, so it falls as coverage rises ($\rho = -0.92$), and coverage-based pruning improves calibration by maximizing the very quantity the gap tracks. Two rival readings do not survive. Internal redundancy is not the driver: random selection keeps the most self-similar set yet is not the most overconfident, and per-sample redundancy barely tracks confidence within a condition ($\rho \approx +0.03$ to $+0.06$). Nor does the relation depend on measuring coverage in the space SCOPE optimizes. Recomputing it from the features of an independent encoder, DINOv2 (Oquab et al. 2024), gives $\rho = -0.72$, and a model-free spatial version, asking only how far each patch sits from the nearest kept one, gives $\rho = +0.69$; both are over the twelve encoder-level conditions, which add VisionZip’s dominant sets to the ten above. All rank correlations here are descriptive summaries over design conditions, not draws from a population.

Manipulating the kept set. Because that correlation is measured across selector families, coverage could be standing in for family membership. So we intervene *within* a fixed family at $K=128$ (dashed trajectories in Fig. 2). Starting from the $\alpha=0$ selection, we swap a growing random fraction $p \in \{.25, .5, .75\}$ of kept tokens for random ones: coverage falls and overconfidence climbs monotonically ($+1.3$ to $+4.7$), converging on the random baseline. Starting from saliency-only we do the reverse, replacing its lowest-attention picks with facility-location ones: coverage rises and overconfidence falls just as steadily ($+5.1$ to $+2.2$). Six manipulated levels and their two anchors trace the same relation the observational conditions do, though a small family residual remains: at matched coverage, repaired-saliency sets stay about a point more overconfident than native ones. Family membership alone therefore cannot explain the pattern. Nor does this isolate coverage, since swapping tokens moves a set’s semantic content, attention mass, and spatial spread along with its coverage; we therefore read coverage as a strong and manipulable predictor of overconfidence rather than its sole cause.

Where the account ends: query-conditioned selection. FastV is the informative exception (Fig. 2). Its three points are monotone in coverage but sit far above the encoder-level relation: at matched coverage it is $+16.2$, $+10.5$, and $+4.2$ points more overconfident than the fit predicts at $K=64/128/288$, an excess that shrinks as retention approaches the 50% point FastV was designed for. Coverage is therefore not the whole story. What the rest of it is, these data cannot say: FastV is at once the only selector here that is query-conditioned, the only one ranking inside the LLM rather than the encoder, and the only one pruning mid-forward, and the three are perfectly confounded. The attention-sink reading makes query conditioning the natural suspect, since selecting on what the model already attends to should preferentially keep the evidence confirming its forming answer; separating it out would need a query-independent selector at the same layer.

Task and Model Dependence

ScienceQA-IMG: where the selection rule stops mattering. ScienceQA-IMG receives the same grid we ran on POPE (full table in the supplementary material), and the contrast is the point. On POPE the coverage–saliency axis separates methods by up to $3\times$ in ECE and FastV collapses at aggressive budgets. On ScienceQA the coverage axis simply stops predicting anything. Accuracy lands within ± 0.8 points of the full model for every method at every budget, even FastV at $K=64$; the monotone α -trend breaks down, with $\alpha=0.5$ falling behind $\alpha=1$; and saliency-only is, if anything, marginally the best calibrated. Calibration is not perfectly flat across selectors here (the ECE spread at $K=64$ is about .025, comparable to our headline POPE effect), but it no longer lines up with coverage, and overconfidence sits stubbornly at $+16$ to $+18$ points whatever we do.

A no-image (blind) baseline explains the contrast by quantifying each benchmark’s visual dependence. Blind accuracy is 50.4% on POPE (chance: the task is entirely visual) but 58.9% on ScienceQA against 63.9% with the image; relative to this benchmark’s 35.5% multi-choice chance rate, the blind model recovers 82% of the above-chance performance. Where answers come mostly from language priors, *which* visual tokens survive stops mattering, and the account says why: with accuracy no longer riding on the visual evidence, coverage has nothing to predict. What remains is the other half of the story. Every encoder-level selector still trims a little unwarranted confidence, cutting ECE below the unpruned model at both budgets while accuracy holds flat, so the confidence-trimming effect survives even where the ordering does not. Only FastV, which prunes on the LLM’s own attention, ends up worse calibrated than not pruning at all. The floor is high in any case: overconfidence sits at $+16$ to $+18$ points whatever we do, because the miscalibration that dominates here lives in the language prior, beyond a token selector’s reach.

A third benchmark: open-vocabulary GQA. GQA is as visually dependent as POPE (blind accuracy 37.5% against 62.7% with the image) but answers in an open vocabulary rather than a binary (full table in the supplementary material). The ordering survives the move intact: at $K=128$, coverage leads at ECE .066, then random .071, saliency-only .082, and FastV .103 with its confidence again pinned near the full model’s while accuracy falls away.

What GQA adds is a boundary. Relaxing the budget improves pruned calibration monotonically toward the full model ($.066 \rightarrow .061 \rightarrow .060$ for pure coverage at $K=128/192/288$), and coverage stays ahead of random throughout, yet even at half the tokens nothing quite overtakes the unpruned .058. On a task this fine-grained the accuracy cost of dropping evidence outweighs the confidence the pruning trims, so here the ordering generalizes and the free lunch does not. The same budget axis explains a detail that would otherwise look odd: $\alpha=0$ and $\alpha=1$ sit level at $K=128$, but the coverage advantage re-emerges as soon as compression softens ($.060$ vs. $.062$ at $K=288$).

A second backbone: LLaVA-NeXT-Vicuna-7B. We repeat the core POPE grid on LLaVA-NeXT, whose AnyRes

encoding produces ~ 2880 visual tokens split across patches (full table in the supplementary material). What transfers is the ordering: at both budgets, coverage-based SCOPE beats random, which beats attention-based FastV, on accuracy and ECE at once. FastV’s signature transfers with it, and on this architecture it is starker still, confidence pinned near 0.92 while accuracy slides from 87.6 to 67.9, a +23.0-point gap. Coverage buys real efficiency here too: $\alpha=1$ at $K=288$ matches the full model’s calibration with $10\times$ fewer tokens.

Those two budgets (10% and 4.4% keep ratios) sit beyond the $\sim 22\%$ sweet spot observed on LLaVA-1.5, and there pruning no longer improves on full-model ECE and the internal α -trend reverses (with only ~ 25 tokens per AnyRes patch, attention weighting concentrates the budget on informative patches, whereas unweighted facility location also spends it on near-empty padding regions). At the *matched* moderate ratio, however, the LLaVA-1.5 picture returns in full: at $K=640$ (22%), $\alpha=0$ reaches ECE .027 against the full model’s .043 at equal accuracy (87.5 vs. 87.6), $\alpha=1$ attains 88.1 accuracy with ECE .035, and the α ordering un-reverses. The calibration free lunch and the α -trend thus transfer across backbones when budgets are matched by keep ratio; whether the regime boundary is universally ratio-aligned across architectures remains to be established on further families.

A third, architecturally distinct family: Qwen2-VL. Qwen2-VL shares almost nothing with the LLaVA recipe: a dynamic-resolution vision tower with no CLS token, a 2×2 merger, M-RoPE positions, a different language model. It is also the most overconfident of our three backbones to begin with, which makes it a useful test of whether pruning can take that surplus away (full table in the supplementary material). It can. Coverage selection at the 22% ratio cuts ECE from .064 to .040 and beats random at the same ratio; at 10% the two separate further (.020 vs. .033) as pruning keeps trimming the surplus. The account holds inside Qwen’s own feature space as well, where the coverage-selected sets cover more at each ratio and are the better calibrated for it.

Against the unpruned model the trade is real rather than free: index-paired, the 22% setting costs 1.1 accuracy points (CI $[-1.6, -0.6]$) to halve ECE ($\Delta\text{ECE} = .023$, CI $[-.029, -.018]$), and the 10% setting costs 2.4. Against random at a matched budget, though, coverage simply wins on both axes at both ratios. (Images are capped at ~ 768 vision tokens for memory; the $\sim 2\%$ that exceed it are skipped in the pruned runs, so every comparison here is index-paired over the 8,820 questions shared by all conditions, the full model included.)

Discussion and Practical Guidance

Efficiency of the compared selectors. Per-sample latency and peak memory are logged by our runner (RTX 4090D, fp16, batch 1). On prefill-dominated POPE, latency per question drops from 131ms (full) to 72–94ms at $K \in \{64, 128\}$ for random/VisionZip/saliency/FastV ($1.4\text{--}1.8\times$). SCOPE’s greedy selection adds ~ 40 ms (only $1.1\times$ here) and its $O(N^2)$ similarity matrix grows $25\times$ on NeXT; FastV is slightly slower at matched K because its first two layers process all 576 tokens. The coverage free lunch is therefore in con-

fidence quality, not raw speed; savings for all selectors grow with generation length and batch.

Guidance. **For pruning research:** report calibration next to accuracy; two selectors with equal accuracy can differ by $3\times$ in ECE (Table 3), and hybrid selectors should expose their saliency weight as a free calibration knob. **For deployment:** where confidence gates downstream decisions, moderate coverage-based pruning buys cheaper inference and better-ranked confidence at little or no accuracy cost, composing with (not subsumed by) post-hoc temperature scaling. Validate the operating ratio on the target task, however: GQA is the cautionary case, where coverage still beat the alternatives at every budget but never overtook the unpruned model’s calibration at any ratio we tried. **For evaluation:** pruned baselines are easy to get silently wrong (the FastV zeroing artifact yields chance accuracy while looking plausible in code); we ship a no-drop-equivalence sanity check.

Limitations

Our strongest evidence is on POPE with LLaVA-1.5-7B; the LLaVA-NeXT and Qwen2-VL grids omit some selectors (attention mapping does not port directly to AnyRes; Qwen omits saliency-only and FastV), and RLHF-aligned, proprietary, or larger ($>7B$) models remain untested. The coverage interventions manipulate token identity at a fixed budget and selector; interventions on the image itself (occlusion, redundancy injection) would probe the account at a second level. First-token verbalizer confidence, audited for mass leakage and normalization sensitivity, does not cover long-form generation; extending to open-ended answers (sequence likelihood, semantic-equivalence sampling (Lau et al. 2026)) is future work.

Conclusion

Visual token pruning changes not just how much MLLMs compute but how much they *believe* their answers. Which tokens a method keeps, not only how many, decides whether a pruned model stays honest: on POPE, coverage-based selection improves calibration at unchanged accuracy, acting as an implicit temperature scaling plus gains no post-hoc temperature reproduces, while attention-based selection holds confidence steady as accuracy falls away. Measuring the kept sets accounts for most of that pattern without needing to invoke anything exotic. Accuracy follows how completely the kept set covers the image; confidence does not follow it at all; overconfidence is the gap between them. Coverage is not the whole story, and we mark where it runs out: query-conditioned selection is overconfident beyond what its coverage predicts, and on ScienceQA, answered mostly from language priors, coverage orders nothing at all. The account travels to a second and a third backbone and to a second visual benchmark; its practical payoff travels less well, costing accuracy on Qwen2-VL and never overtaking the unpruned model on GQA. Confidence quality should join accuracy and FLOPs as a standard axis for evaluating token compression in multimodal models.

References

- Alvar, S. R.; Singh, G.; Akbari, M.; and Zhang, Y. 2025. DivPrune: Diversity-Based Visual Token Pruning for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Baek, C.; Song, J.; Kim, S.; and Kong, K. 2026. AgilePruner: An Empirical Study of Attention and Diversity for Adaptive Visual Token Pruning in Large Vision-Language Models. In *International Conference on Learning Representations*.
- Bhattacharya, D.; Kulkarni, A.; and Ganapathy, S. 2025. FESTA: Functionally Equivalent Sampling for Trust Assessment of Multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. 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*.
- Chen, X.; Zhu, W.; Li, G.; Dong, X.; Xiong, Y.; Wang, H.; Qiu, P.; Song, Q.; Wang, Z.; Tang, S.; Wang, Y.; and Razi, A. 2026a. OTPrune: Distribution-Aligned Visual Token Pruning via Optimal Transport. *arXiv preprint arXiv:2602.20205*.
- Chen, Y.; Shan, B.; Ye, X.; and Chen, C. 2026b. EvoPrune: Early-Stage Visual Token Pruning for Efficient MLLMs. *arXiv preprint arXiv:2603.03681*.
- Chen, Z.; Hu, W.; He, G.; Deng, Z.; Zhang, Z.; and Hong, R. 2025. Unveiling Uncertainty: A Deep Dive into Calibration and Performance of Multimodal Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*.
- Cho, J.; Lee, J.; Hayat, M.; Hwang, K.; Porikli, F.; and Choi, S. 2026. FLoC: Facility Location-Based Efficient Visual Token Compression for Long Video Understanding. In *International Conference on Learning Representations*.
- Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. In *International Conference on Learning Representations*.
- Deng, J.; Li, W.; Zhou, J. T.; and He, Y. 2025. SCOPE: Saliency-Coverage Oriented Token Pruning for Efficient Multimodal LLMs. In *Advances in Neural Information Processing Systems*.
- Du, Y.; Wang, Y.; Kong, M.; Liang, T.; Long, Q.; Chen, B.; and Zhu, Q. 2025. Confidence Calibration for Multimodal LLMs: An Empirical Study through Medical VQA. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*.
- Fang, Z.; Lyu, P.; Zhang, C.; Lu, G.; Yu, J.; and Pei, W. 2026. Prune Redundancy, Preserve Essence: Vision Token Compression in VLMs via Synergistic Importance-Diversity. In *International Conference on Learning Representations*.
- Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*.
- Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Lau, G. K. R.; Dao, H.; Lin, N. K. H.; and Low, B. K. H. 2026. Uncertainty Quantification for Multimodal Large Language Models with Incoherence-adjusted Semantic Volume. *arXiv preprint arXiv:2602.24195*.
- Li, K.; Chen, X.; Gao, C.; Li, Y.; and Chen, X. 2025a. Balanced Token Pruning: Accelerating Vision Language Models Beyond Local Optimization. In *Advances in Neural Information Processing Systems*.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Li, Y.; Xiong, M.; Wu, J.; and Hooi, B. 2025b. ConfTuner: Training Large Language Models to Express Their Confidence Verbally. In *Advances in Neural Information Processing Systems*.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>. Technical report.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.
- Liu, X.; Chen, T.; Da, L.; Chen, C.; Lin, Z.; and Wei, H. 2025. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems*.
- Ma, R.; Chen, S.; and Yao, S. 2025. 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)*.
- Misra, D.; Chaudhary, M.; Goyal, A.; Runwal, B.; and Chen, P.-Y. 2024. Uncovering the Hidden Cost of Model Compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- Mitra, P.; Schwalbe, G.; and Klein, N. 2024. 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*.
- Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*.
- Shao, K.; Tao, K.; Zhang, K.; Feng, S.; Cai, M.; Shang, Y.; You, H.; Qin, C.; Sui, Y.; and Wang, H. 2026. A Survey of Token Compression for Efficient Multimodal Large Language Models. *Transactions on Machine Learning Research*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, T.; Guo, Y.; Huang, R.; Lu, Z.; Zhang, Q.; Li, C.; Zhang, Y.; Cao, J.; Shen, Z.; Du, Y.; Gan, G.; Wang, W.; Cong, L. W.; and Zhang, S. 2026. TOPS: First-Principles Visual Token Pruning via Constructing Token Optimal Preservation Sets for Efficient MLLM Inference. *arXiv preprint arXiv:2606.27161*.
- Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2024. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*.
- Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; and Lin, D. 2025. Pyramid-Drop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Ye, X.; Gan, Y.; Ge, Y.; Zhang, X.-P.; and Tang, Y. 2025. ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zhang, E.; Yu, F.; Wu, A.; Wen, Z.; Yan, K.; Ding, S.; Qi, B.; and Zhang, L. 2026a. D²Pruner: Debiased Importance and Structural Diversity for MLLM Token Pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zhang, H.; Lyu, M.; He, C.; Ao, Y.; and Lin, Y. 2026b. TrimTokenator: Towards Adaptive Visual Token Pruning for Large Multimodal Models. In *Findings of the Association for Computational Linguistics: ACL 2026*.
- Zhang, Q.; Liu, M.; Li, L.; Lu, M.; Zhang, Y.; Pan, J.; She, Q.; and Zhang, S. 2025a. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. In *Advances in Neural Information Processing Systems*.
- Zhang, R.; Zhang, H.; and Zheng, Z. 2024. VL-Uncertainty: Detecting Hallucination in Large Vision-Language Model via Uncertainty Estimation. *arXiv preprint arXiv:2411.11919*.
- Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D. A.; Okuno, T.; Nakata, Y.; Keutzer, K.; and Zhang, S. 2025b. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *Proceedings of the 42nd International Conference on Machine Learning*.
- Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024. Calibrated Self-Rewarding Vision Language Models. In *Advances in Neural Information Processing Systems*.