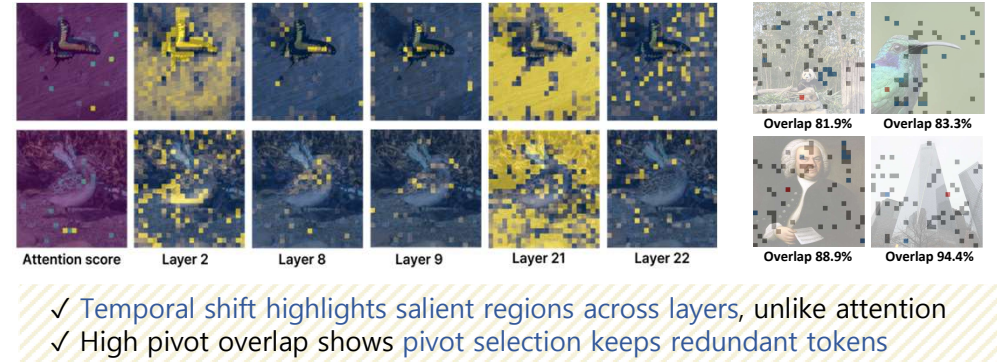


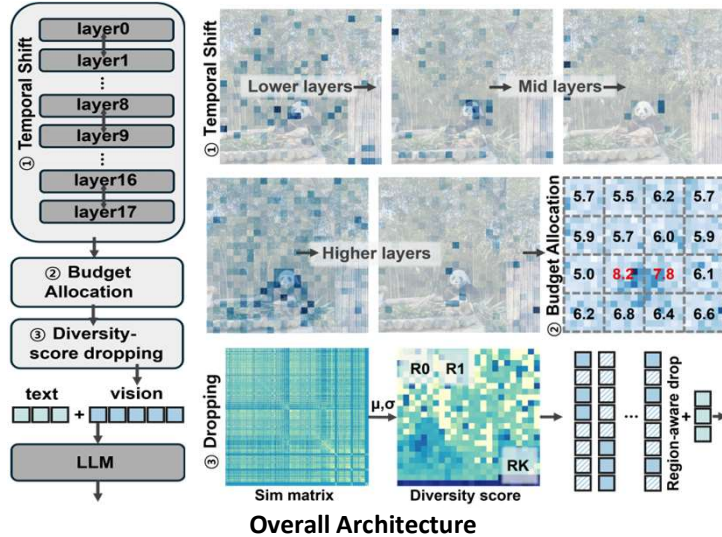
1. Motivation

- Vision tokens are the efficiency bottleneck in VLMs
 - High-res inputs (e.g., LLaVA-Next) yield $\approx 2,880$ vision tokens vs. < 200 text tokens
 - 70–90% of the token is visual $\rightarrow$ attention cost $\mathcal{O}((N+M)^2)$, KV-cache explodes
 - Prior token-dropping has complementary weaknesses
 - Saliency-based (attention scores / hidden norms):
 - $\rightarrow$ Needs dense attention (Lim. 1), Saliency-only (Lim. 2) $\rightarrow$ weak speedup, position bias
 - Diversity-based (pivot selection):
 - $\rightarrow$ Pivots cluster (Lim. 3), Hyperparam-sensitive (Lim. 4) $\rightarrow$ poor local coverage, unstable
- ✓ Can we **preserve both saliency and diversity**?
- ✓ Can we **guarantee local coverage** under tight budgets?

2. Key Idea



3. Method — SPLIT



- Temporal Shift** $\rightarrow$ Solution for Lim.1 & 2
 - Temporal shift: normalized hidden-state shift across model layers
 - Attention-free saliency $\rightarrow$ *efficient and free of position bias*
- Local Budget Allocation** $\rightarrow$ Solution for Lim.3
 - Each region gets its own budget instead of one global budget
 - A uniform floor guarantees every region keeps minimum coverage
 - An importance-weighted term *gives salient regions more budgets*
- Diversity-Score Selection** $\rightarrow$ Solution for Lim.4
 - Rank tokens by *distinctiveness* (σ) and *redundancy* (μ)
 - Keep the most distinctive, non-redundant tokens per region
 - Robust without tuning pivot counts or distance thresholds
- Theoretical Guarantee** $\rightarrow$ Hausdorff coverage
 - A minimum budget makes *every region non-empty*
 - Adaptive budgets give tighter guarantees than uniform splits

$$\frac{\|h_\ell(x) - h_{\ell-1}(x)\|_2}{\|h_\ell(x)\|_2}$$

Eq. Temporal Shift

$$\tilde{B}_k = \frac{B}{K} + B \cdot \frac{I(X_k)}{\sum_j I(X_j)}$$

Eq. Local Budgeting

$$S_{ij} = \left\langle \frac{x_i}{\|x_i\|_2}, \frac{x_j}{\|x_j\|_2} \right\rangle, D(i) = \lambda \sigma_i - \mu_i$$

Eq. Diversity Score

$$d_H(X_k, R_k) = \sup_{x \in X_k} \inf_{r \in R_k} \|x - r\|$$

Eq. Hausdorff distance

4. Experiments

Method	GQA	MMB	MME	POPE	SQA	VQA _{txt}	Avg.
Vanilla	61.2	63.1	1477.7	85.4	68.1	48.7	100%
HiRed	54.0	50.3	1297.5	78.0	66.3	39.2	86.8%
GreedyP	55.0	55.2	1365.1	82.6	64.6	40.3	91.0%
DivPrune	58.4	58.6	1362.1	83.7	67.3	40.8	91.7%
DART	55.1	59.7	1357.6	82.6	67.9	40.7	91.4%
Ours	59.0	60.1	1375.5	84.5	68.6	41.5	92.8%

Comparison on LLaVA-1.5-7B (88.9% token reduction)

Method	Token	Total Time	Prefill Time	FLOPs	POPE	Speedup Total	Speedup Prefill
Vanilla	2880	46:34	28:59	100%	87.7	1.00×	1.00×
GreedyP	320	31:17	12:02	22.3%	85.4	1.49×	2.41×
DivPrune	320	26:34	10:57	19.8%	86.0	1.75×	2.65×
DART	320	27:28	11:19	24.8%	85.6	1.70×	2.56×
Ours	320	25:52	10:31	19.8%	86.2	1.80×	2.76×

Inference costs in LLaVA-Next-7B (POPE)

5. Conclusion

- SPLIT preserves both salient and diverse visual tokens without missing regions uncovered
- Attention-score-free and threshold-free while ensuring *local coverage* under tight budgets
- SOTA *efficiency-accuracy trade-off* under extreme budgets on image & video VLMs
- No training required, *plug-and-play* into existing VLMs