NEURAL INFORMATION PROCESSING SYSTEMS

RobustPruner: Decoupled Relevance and Uncertainty for Efficient Visual Token Pruning in MLLMs

Hanwei Zhu*, Junhan Fu*, Xi Zhang, Jiamang Wang, Weisi Lin

NeurIPS 2026 (Poster)

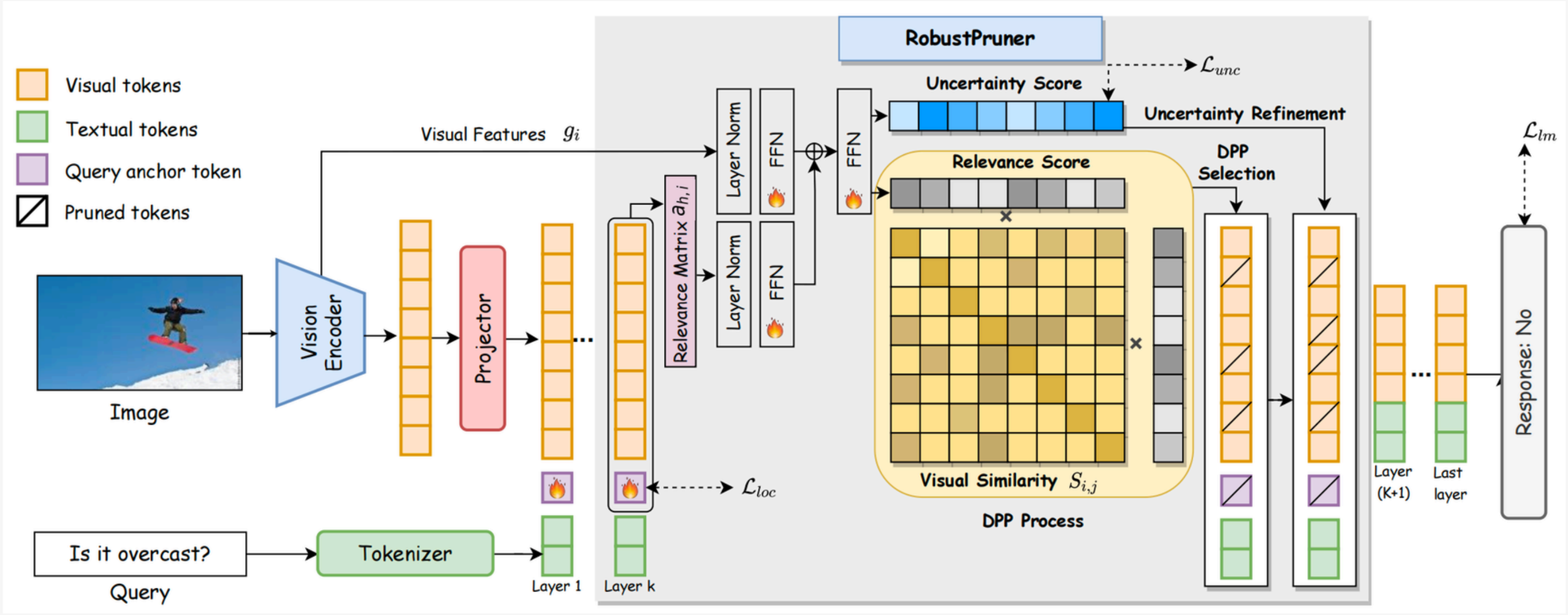
Presenter: Junhan Fu
Email: junhanfu@link.cuhk.edu.cn

*Equal contribution

September 2026

Abstract

We propose RobustPruner, a visual token pruning method for MLLMs. **By combining query relevance, diversity, and uncertainty-aware refinement**, it reduces inference cost while improving robustness to prompt paraphrases and maintaining strong performance across multiple vision-language tasks.



NEURAL INFORMATION PROCESSING SYSTEMS

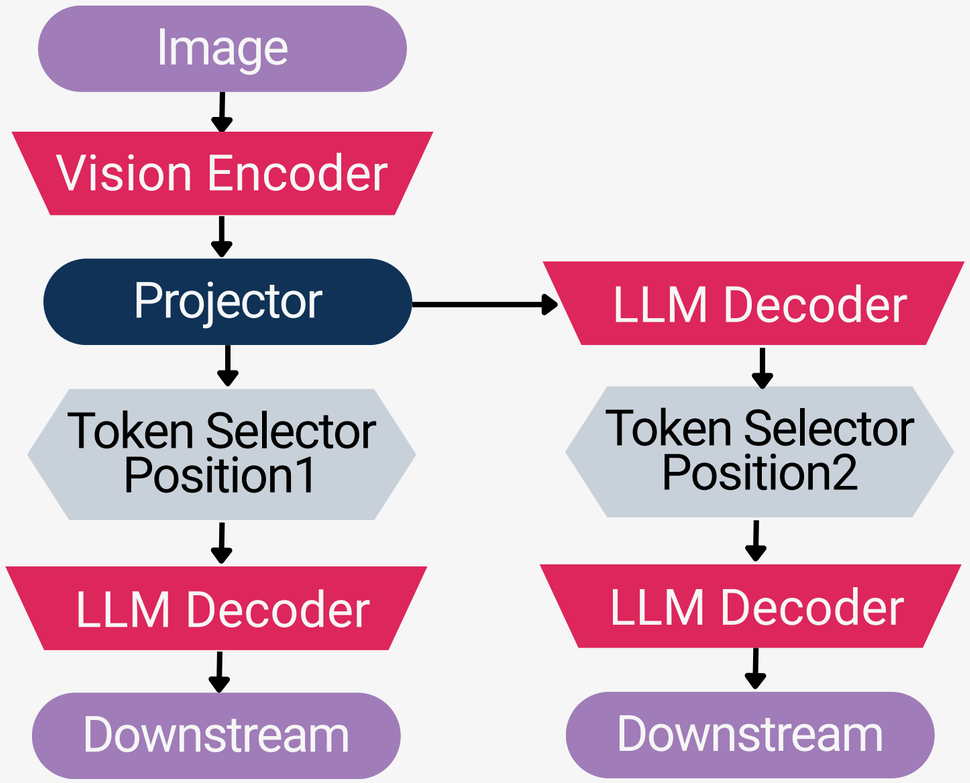
Part 1

Research Motivation

Existing pruners fail under parapharases

In MLLMs, conventional methods use heuristic indicators (attention patterns, similarity measures, and text-guided saliency), inadequate for vision-language reasoning.

Conventional Pruning:
Deterministic Selection



1. **Attention-based pruning:** overemphasize dominant responses
2. **Similarity-based pruning:** fail to adapt to fine-grained instruction semantics
3. **Current methods:** sensitive to prompt perturbations (totally same tasks with parapharasing)

token importance ≠ token dispensibility

Necessary to design an uncertainty-aware visual token pruning

NEURAL INFORMATION PROCESSING SYSTEMS

Part 2 **Empirical Analysis**

Uncertainty Design

The logic of constructing token uncertainty indicator.

The general optimization question in token pruning:

$$Y_\ell^* = \arg \max_{Y \subseteq \{1,\ldots,N\},\, |Y|=m} F_\ell(Y; I, Q)$$

I, Q: Image and Question
N: Original visual token numbers
Y_l : token set remained in the l'th decoding layer
m<<N: targeted token budget

In real cases, we have importance score $\Phi_{l,i}$ **for ranking** and take the **top-m tokens.**

Uncertainty construction: utilizing 4 semantic-preserving textual perturbations (paraphrasing)

$Q^{(1)}$: lexical paraphrase
$Q^{(2)}$: syntactic reformulation
$Q^{(3)}$: semantic-equivalence rewrite
$Q^{(4)}$: instruction-style rewrite

$$P(Q) = \{Q^{(1)}, Q^{(2)}, Q^{(3)}, Q^{(4)}\}$$

$$\phi_{\ell,i}^{(p)} = \phi_{\ell,i}(I, Q^{(p)})$$

$$u_{\ell,i} = \mathrm{Var}_{Q^{(p)} \in P(Q)} \left[\phi_{\ell,i}^{(p)}\right]$$

Low $u_{l,i}$: score is insensitive to paraphrasing, and token relevance is stable;
Hign $u_{l,i}$: score fluctuates with prompt rewriting, pruning evidence is sensitive to paraphrasing.

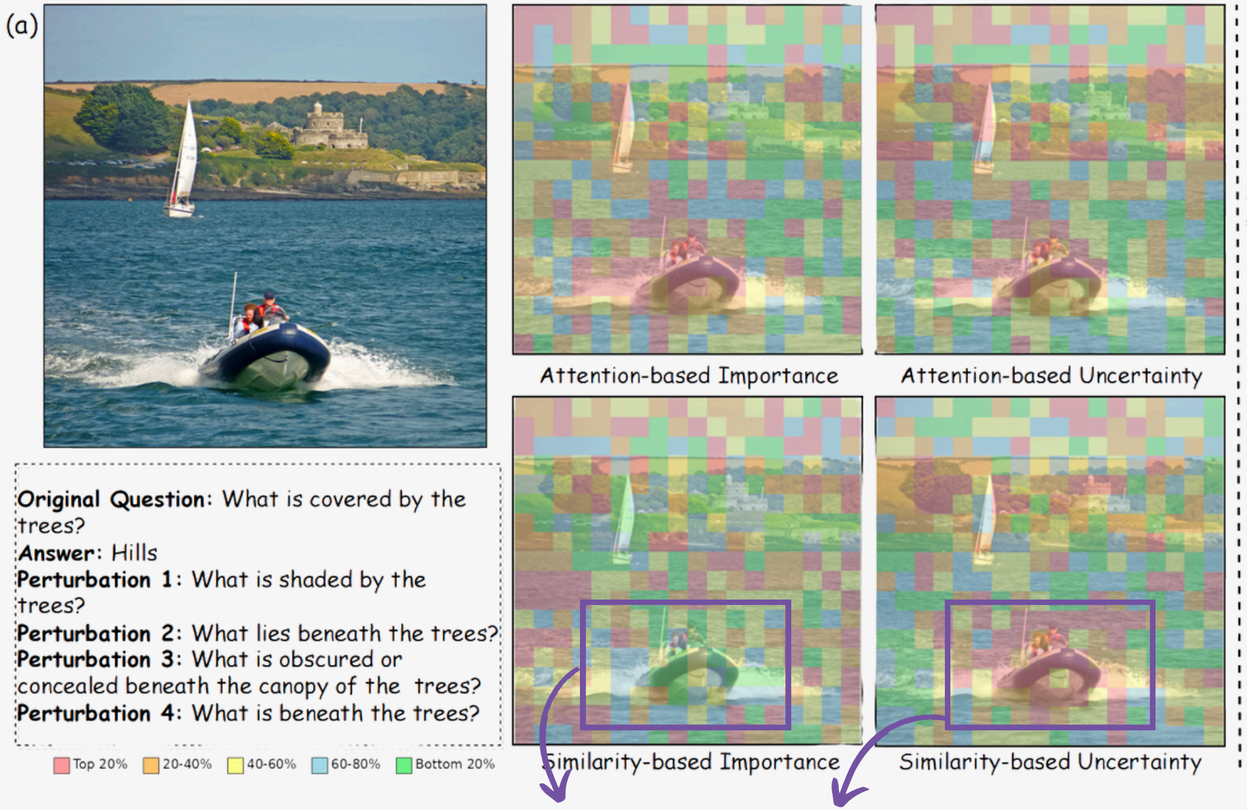
NEURAL INFORMATION PROCESSING SYSTEMS

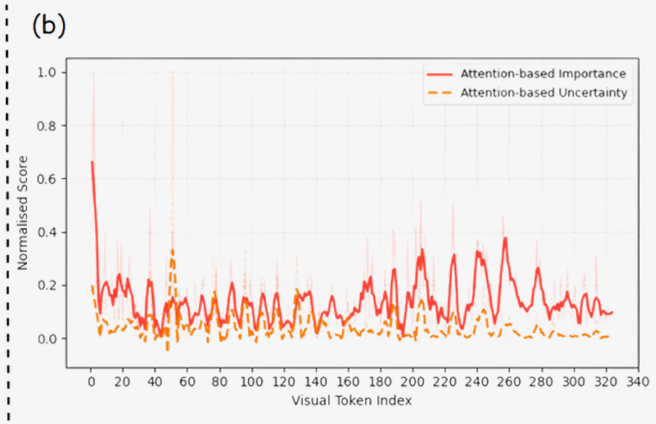
Part 2

Empirical Analysis

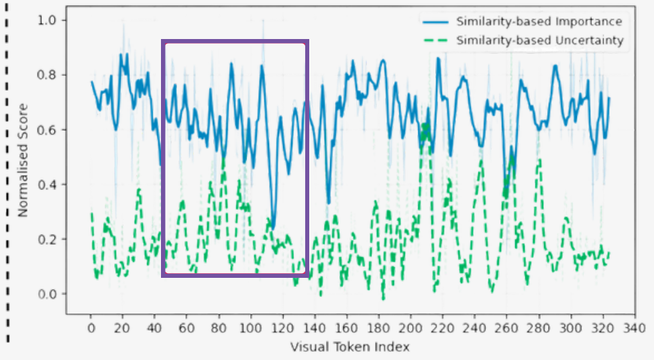
Relevance ≠ Reliability

Examine whether the pruning signals commonly used in MLLMs remain stable under semantics-preserving textual perturbations.





Attention-based uncertainty in MLLMs: aggregated text-to-vision cross-attention



Similarity-based uncertainty in MLLMs: explicit query–visual relevance score

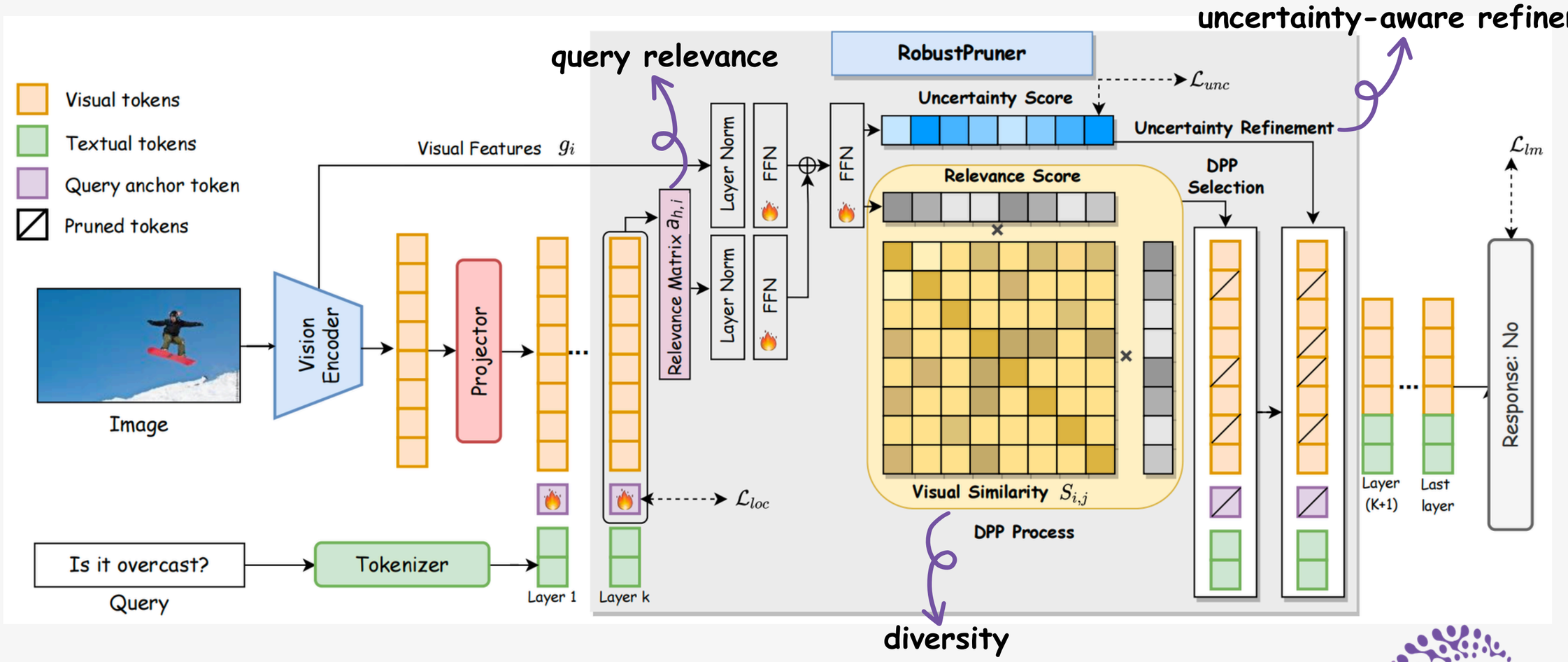
Uncertainty does not collapse to **importance**: many highly relevant tokens remain low-uncertainty, while uncertainty peaks only on a sparse subset of tokens.

NEURAL INFORMATION PROCESSING SYSTEMS

Method

RobustPruner at a glance

decoder layer K / QAT → relevance → DPP candidate selection → uncertainty refinement → decoder layer K+1

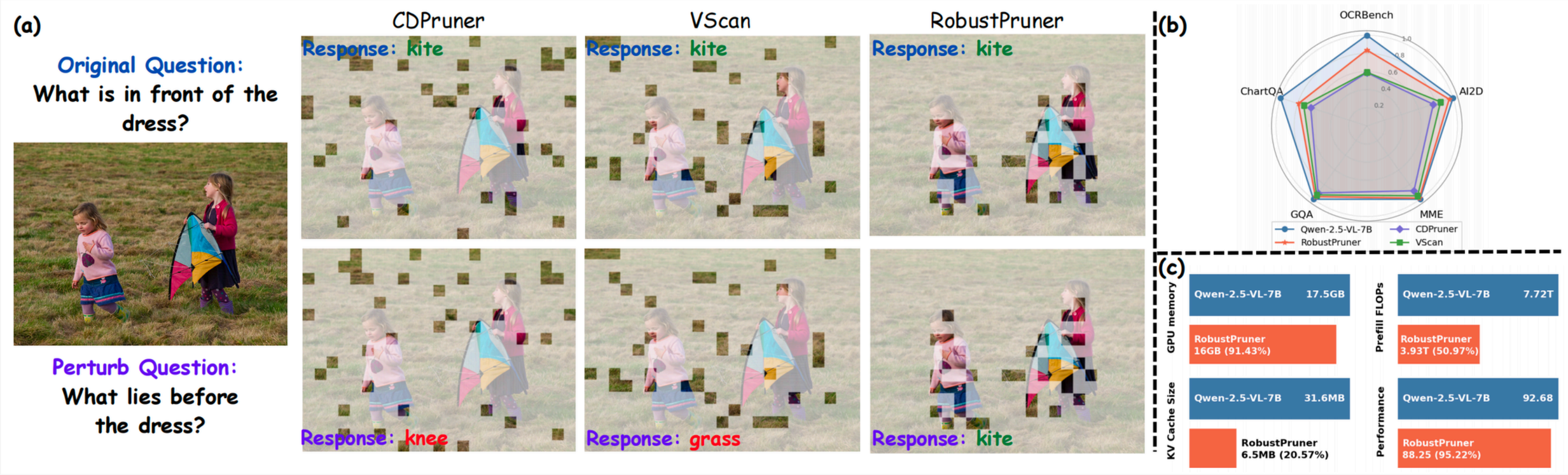


NEURAL INFORMATION PROCESSING SYSTEMS

Part 3 **Method**

Progress of RobustPruner

Comparison with the baseline model and other methods

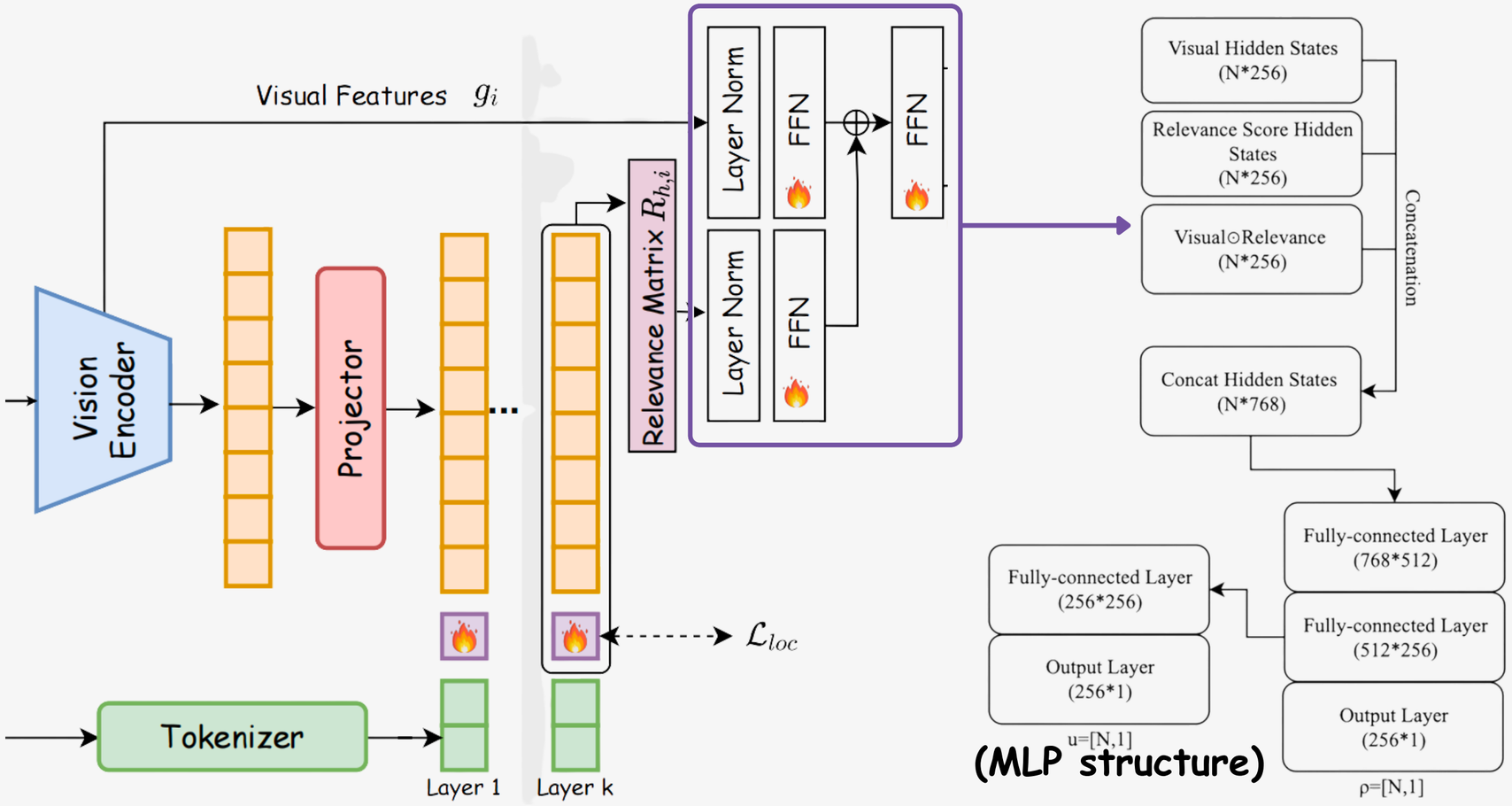


1. Existing MLLM token pruning is uncertainty-blind and perturbation-sensitive, **causing unstable predictions.**
2. RobustPruner is a **query-conditioned, plug-and-play pruner** decoupling relevance, DPP diversity, and uncertainty refinement to retain critical evidence under compression.
3. **3 MLLMs, 10 benchmarks:** better accuracy–efficiency–robustness, lower cost.

NEURAL INFORMATION PROCESSING SYSTEMS

Part 3 **Method**

Component1: Query Anchor Token (QAT) and Query-Conditioned Relevance Prediction



1. Initialize QAT: $q_0 = \frac{1}{M}\sum_{j=1}^{M} t_{0,j}$
2. Use QAT to score the relevance of each visual token:

$$a_{h,i} = \frac{(W_q^h q_K)^\top (W_k^h v_{K,i})}{\sqrt{d_h}}$$

3. Fuse query alignment, visual features, and their interaction:

$$h_i^{\text{rel}} = \text{SiLU}(W_a \hat{a}_i + b_a)$$

$$h_i^{\text{vis}} = \text{SiLU}(W_v \text{LN}(g_i) + b_v)$$

$$x_i = [h_i^{\text{rel}};\ h_i^{\text{vis}};\ h_i^{\text{rel}} \odot h_i^{\text{vis}}]$$

4. Produce a query-conditioned relevance score for DPP:

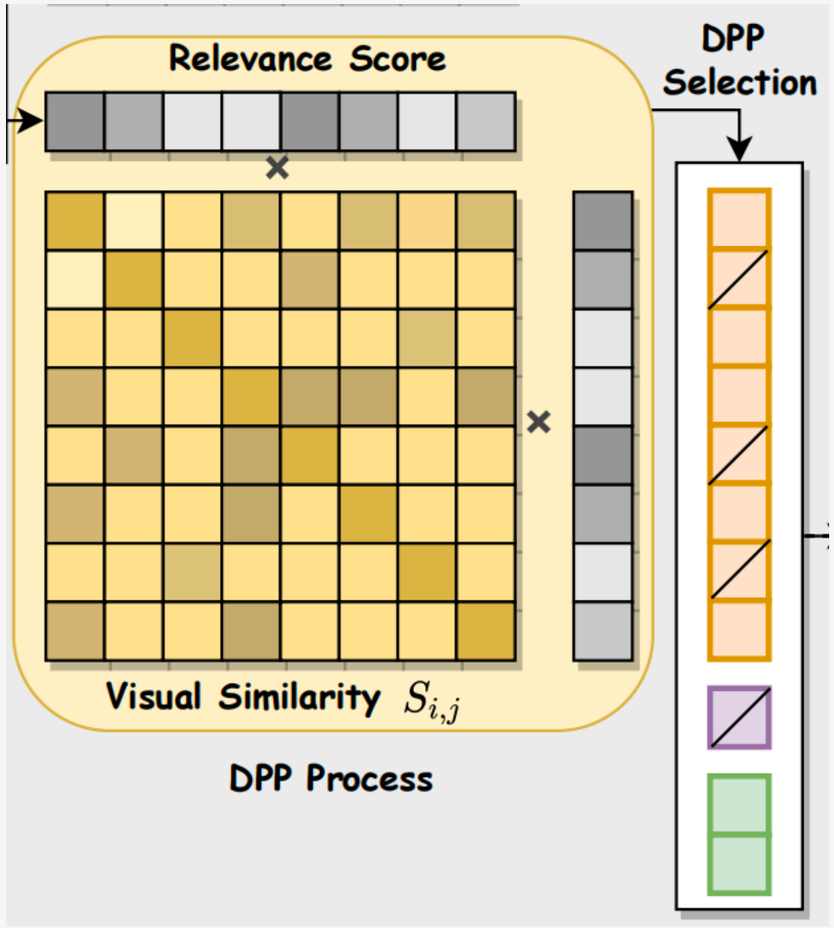
$$c_i = \text{MLP}_c(x_i), \qquad \rho_i = w_r^\top c_i + b_r, \qquad r_i = \sigma(\rho_i)$$

The trainable module is the MLP used to generate the token relevance score and the uncertainty score simultaneously.

NEURAL INFORMATION PROCESSING SYSTEMS

Method

Component2: DPP Candidate Selection for diversity



Target: suppress redundant yet individually relevant tokens

Determinantal Point Process Algorithm:

1. **Initialize** $S \in R^{N*N}$ as a visual-feature similarity matrix where

$$S_{ij} = \frac{v_{K,i}^{\top} v_{K,j}}{\|v_{K,i}\|_2 \, \|v_{K,j}\|_2}$$

2. **Combine** S with the query-conditioned relevance score r_i to form the DPP kernel:

$$L = \mathrm{diag}(r)\, S\, \mathrm{diag}(r)$$

3. DPP can choose the remained token numbers. Given the target budget *m*, first form an n over-complete candidate set of size k = ⌈1.1m⌉:

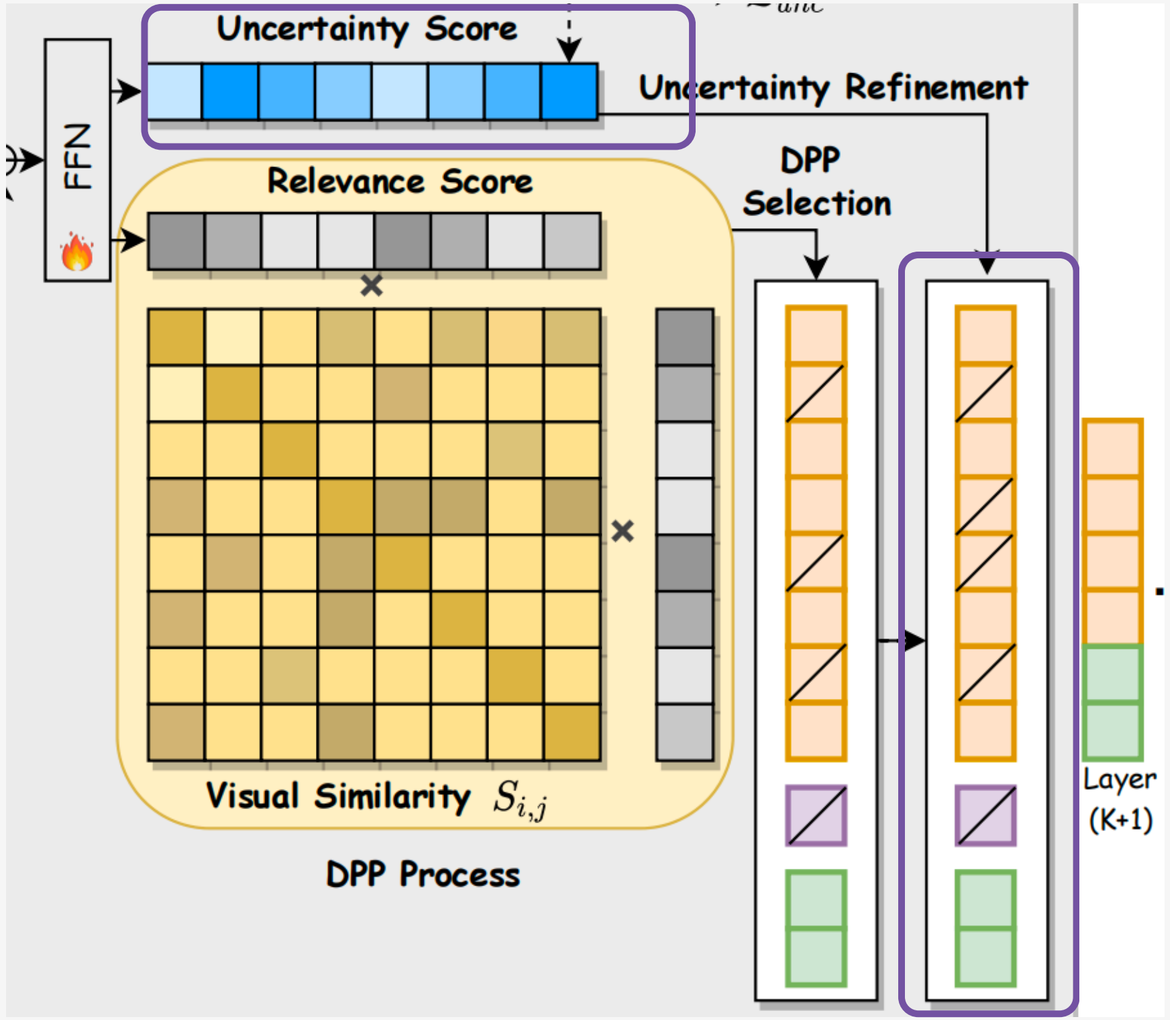
$$Y_{\mathrm{DPP}} = \underset{Y \subseteq \{1,\ldots,N\},\, |Y|=k}{\arg\max} \log\det(L_Y)$$

Greedy MAP Inference to achieve this.

NEURAL INFORMATION PROCESSING SYSTEMS

Method

Component3: Uncertainty-Guided Refinement



Target: **Fine-tuning** of token selection after adding uncertainty score

1. Output u_i from the MLP:

$$u_i = \mathrm{Softplus}(\mathrm{MLP}_p(x_i)), \qquad u_i > 0$$

2. Choose the token set with the least sum of uncertainty scores:

$$Y^\star = \underset{Y \subseteq Y_{\mathrm{DPP}}, |Y|=m}{\arg\min} \sum_{i \in Y} u_i$$

3. The retained token set is passed to the remaining decoder layers together with the textual states.

NEURAL INFORMATION PROCESSING SYSTEMS

Part 3

Method

Training Objects and Evaluation Setups

Training logic and hyperparameters:

1. Grounded GQA supervision - localization loss

$$\hat{b}_i = \sigma(\rho_i), \qquad \mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{Dice}}(\hat{b}, b) + 0.1\mathcal{L}_{\text{BCE}}(\hat{b}, b)$$

2. Language modeling - semantic entropy loss
3. Uncertainty - L_1 loss with the variance

$$\mathcal{L}_{\text{unc}} = \frac{1}{N}\sum_{i=1}^{N} |u_i - \bar{u}_i|$$

Loss combination: $\mathcal{L} = \lambda_{\text{lm}}\mathcal{L}_{\text{lm}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \lambda_{\text{unc}}\mathcal{L}_{\text{unc}}$

Hyperparameter	Value	Hyperparameter	Value
Number of GPUs	4	Training data size	20,000
Optimizer	AdamW	Perturbed question size	6,000
Batch size	1	Data type	bfloat16
Gradient accumulation steps	8	Attention implementation	flash_attention_2
Effective batch size	32	λ_{lm}	1.0
Number of training epochs	1	λ_{loc}	1.0
Learning rate	1×10^{-4}	$\lambda_{\text{uncertainty}}$	0.5
Warmup ratio	0.1		

Evaluation setups: **(3 MLLMs, 10 benchmarks)**

1. **Backbones:** Qwen-2.5-VL-7B, LLaVA-v1.5-7B, LLaVA-OneVision-1.5-8B,
2. **Benchmarks:** GQA, MME, ScienceQA, RWQA, AI2D, ChartQA, DocVQA, TextVQA, MMMU, OCRBench, V*Bench
3. **Token keep ratios:** 10%, 20%, 30%
4. **Baseline Pruning:**
   - **atention-based methods:** FastV
   - **similarity-based methods:** CDPruner, DART, DivPrune
   - **attention&similarity-based methods:** PruMerge+, VisionZip, VScan

NEURAL INFORMATION PROCESSING SYSTEMS

Part 4

Experiments

Performance comparison of different pruning methods on Qwen-2.5-VL-7B.

Method	GQA	MME	SciQA	RWQA	AI2D	ChartQA	DocVQA	TextVQA	MMMU	OCR	V*Bench	Acc.	Rel.
Qwen-2.5-VL-7B	60.46	2323.62	87.46	69.80	93.75	83.08	94.38	82.88	50.67	83.8	70.68	77.7	100.0%
Retain 10% Tokens (↓90%)													
FastV (ECCV24)	47.32	1941.67	81.80	59.21	75.53	44.56	58.71	**70.93**	46.00	43.6	49.21	57.7	75.9%
CDPruner (NeurIPS25)	51.69	1950.13	76.95	53.33	67.07	42.84	30.79	42.55	45.56	36.5	51.83	49.9	68.0%
DART (EMNLP25)	51.72	2011.58	81.80	57.25	66.97	33.48	33.21	53.40	46.11	41.5	54.97	52.0	70.7%
DivPrune (CVPR25)	55.34	2032.39	83.29	60.52	72.05	40.80	54.32	62.96	45.44	47.3	65.97	58.8	78.2%
PruMerge+ (ICCV25)	52.64	1952.94	81.26	62.74	71.05	41.12	41.86	56.61	44.67	41.0	62.83	55.6	74.3%
VisionZip (CVPR25)	54.11	1933.56	83.29	63.40	72.38	43.88	48.44	55.84	45.11	39.5	63.87	57.0	75.8%
VScan (TMLR26)	54.17	2126.76	81.56	60.00	74.16	41.64	54.68	59.02	48.22	38.1	58.64	57.0	76.6%
RobustPruner (Ours)	**57.11**	**2213.69**	**84.33**	**66.27**	**85.43**	**59.00**	**78.41**	69.81	**49.67**	**59.6**	**67.02**	**67.7**	**88.6%**
Retain 20% Tokens (↓80%)													
FastV (ECCV24)	51.44	2131.93	82.70	64.31	82.42	60.56	75.86	**77.92**	49.67	59.7	56.54	66.1	86.2%
CDPruner (NeurIPS25)	55.68	2071.28	80.32	58.56	72.50	53.12	42.51	53.78	47.45	49.6	60.73	57.4	77.0%
DART (EMNLP25)	56.11	2166.96	83.49	63.53	70.44	48.60	50.36	67.48	46.00	54.0	59.69	60.0	80.1%
DivPrune (CVPR25)	58.38	2208.03	84.38	64.05	81.12	58.12	72.82	73.81	48.11	63.5	68.06	67.2	88.1%
PruMerge+ (ICCV25)	56.55	2215.23	**87.46**	64.97	83.06	52.44	60.72	68.65	45.89	56.2	67.54	64.3	84.9%
VisionZip (CVPR25)	57.00	2169.39	83.84	66.66	87.99	61.64	75.42	71.85	47.56	59.3	67.01	67.8	88.5%
VScan (TMLR26)	57.12	2258.84	84.53	65.49	80.38	59.60	77.61	73.91	**50.33**	51.3	61.78	66.2	87.2%
RobustPruner (Ours)	**58.54**	**2291.17**	85.72	**67.97**	**89.51**	**65.72**	**88.01**	76.91	49.89	**71.1**	**70.16**	**72.4**	**94.0%**
Retain 30% Tokens (↓70%)													
FastV (ECCV24)	54.15	2267.16	83.19	66.14	86.95	68.04	84.10	**80.33**	49.00	68.5	62.30	70.3	91.3%
CDPruner (NeurIPS25)	57.47	2161.50	81.41	60.13	76.17	60.64	53.37	60.70	48.45	57.2	67.02	62.3	82.7%
DART (EMNLP25)	57.85	2207.58	83.84	66.41	75.52	57.08	62.98	74.34	47.22	63.4	58.64	64.7	85.4%
DivPrune (CVPR25)	59.14	2245.39	84.88	65.62	87.21	67.08	81.79	77.66	47.89	72.1	69.11	71.2	92.5%
PruMerge+ (ICCV25)	58.21	2276.19	85.57	65.62	83.06	62.60	73.99	73.40	46.67	64.3	66.49	68.0	89.1%
VisionZip (CVPR25)	58.61	2278.95	86.37	66.14	87.99	**72.16**	86.26	77.00	48.89	70.6	65.97	72.0	93.4%
VScan (TMLR26)	58.39	2301.62	86.12	66.93	85.43	71.08	88.22	79.32	49.78	63.0	69.11	71.7	93.4%
RobustPruner (Ours)	**59.51**	**2336.41**	**86.52**	**68.24**	**91.71**	69.52	**90.74**	79.87	**49.89**	**77.7**	**71.20**	**74.5**	**96.5%**

Findings:

- **RobustPruner** gives the best accuracy on Qwen-2.5-VL-7B across almost all token budgets.
- It outperforms **DivPrune**, **VisionZip**, and **VScan**, especially under aggressive compression and on text-rich benchmarks.
- Uncertainty-guided refinement complements relevance and diversity.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

Performance comparison of different pruning methods on LLaVA-OneVision-1.5-8B

Method	GQA	MME	SciQA	RWQA	AI2D	MMMU	OCR	POPE	V*Bench	Acc.	Rel.
LLaVA-OneVision-1.5-8B	62.07	2255.17	98.51	67.97	94.01	56.33	82.7	88.45	76.96	83.0	100.0%
Retain 10% Tokens (↓90%)											
FastV (ECCV24)	48.87	1807.58	81.60	49.41	73.28	49.00	37.2	62.95	37.17	64.5	77.6%
CDPruner	50.99	1645.79	81.46	53.73	72.67	49.78	18.0	73.28	72.25	68.6	82.6%
DivPrune (CVPR25)	56.75	1896.01	87.95	59.87	77.07	51.22	41.9	84.21	72.77	76.3	92.0%
VisionZip (CVPR25)	55.48	–	88.80	**61.17**	71.37	–	–	79.05	–	74.4	89.6%
VScan	52.90	1919.46	86.27	–	73.48	–	24.8	75.77	–	71.6	86.2%
RobustPruner (Ours)	**59.94**	**2175.94**	**95.59**	58.82	**88.70**	**52.44**	**51.7**	**88.64**	**76.96**	**81.4**	**97.9%**
Retain 20% Tokens (↓80%)											
FastV (ECCV24)	51.60	1918.60	88.10	52.94	78.79	51.44	49.2	67.12	43.98	68.9	82.8%
CDPruner	56.06	1820.58	87.90	58.82	79.21	51.22	30.1	82.99	**79.05**	75.6	91.1%
DivPrune (CVPR25)	59.44	2063.25	92.31	63.00	84.52	53.00	**57.0**	87.22	75.96	79.7	96.0%
VisionZip (CVPR25)	58.37	–	91.82	**66.14**	–	–	–	84.83	–	78.3	94.4%
VScan	57.43	2096.05	92.56	–	79.34	–	43.7	84.12	–	78.0	93.9%
RobustPruner (Ours)	**60.80**	**2272.84**	**96.08**	63.01	**89.54**	**53.67**	56.6	**88.38**	76.44	**81.8**	**98.5%**
Retain 30% Tokens (↓70%)											
FastV (ECCV24)	53.05	2017.06	91.12	54.90	83.52	53.67	**59.3**	70.55	48.17	71.6	85.9%
CDPruner	58.81	1937.24	90.23	59.74	83.23	51.89	40.5	86.14	**78.01**	78.4	94.6%
DivPrune (CVPR25)	60.29	2110.49	94.45	64.71	89.09	**54.44**	–	88.10	77.49	80.9	97.5%
VisionZip (CVPR25)	60.12	–	96.22	**67.06**	–	–	–	86.85	–	81.1	97.6%
VScan	59.53	2197.49	95.58	–	84.68	–	57.9	86.64	–	80.6	97.0%
RobustPruner (Ours)	**60.46**	**2268.15**	**96.43**	65.49	**89.64**	52.33	56.8	**88.41**	77.49	**81.8**	**98.4%**

Findings:

- **RobustPruner** consistently achieves the best average accuracy and retention, especially on high-resolution and fine-grained benchmarks.
- This baseline shows its query-aware and diversity-aware selection preserves dense visual evidence, not merely highly salient tokens.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

Performance comparison of different pruning methods on LLaVA-v1.5-7B.

Method	GQA	MME	SciQA	RWQA	AI2D	MMMU	POPE	V*Bench	Acc.	Rel.
LLaVA-v1.5-7B	61.92	1864.84	69.51	55.82	58.81	35.89	85.85	50.26	59.7	100.0%
Retain 10% Tokens (↓90%)										
FastV (ECCV24)	46.88	1272.40	**69.06**	44.71	52.85	34.11	40.15	37.17	46.4	78.6%
CDPruner (NeurIPS25)	58.64	1712.25	68.37	50.85	55.60	35.56	**87.37**	47.12	57.6	95.6%
DivPrune (CVPR25)	57.41	1662.78	68.62	49.54	54.18	35.00	85.45	46.60	56.7	93.9%
VisionZip (CVPR25)	55.61	1672.70	68.77	49.80	53.59	35.00	76.98	36.65	53.8	89.9%
VScan (TMLR26)	56.99	1713.64	68.96	50.85	54.63	**36.22**	81.99	41.88	55.9	93.4%
RobustPruner (Ours)	**59.99**	**1862.69**	67.23	**53.33**	**58.36**	35.67	85.83	**49.74**	**58.6**	**98.3%**
Retain 20% Tokens (↓80%)										
FastV (ECCV24)	49.71	1466.40	68.42	47.71	53.82	35.00	53.14	37.43	49.3	83.5%
CDPruner (NeurIPS25)	59.66	1744.93	68.42	53.07	55.83	35.89	**87.12**	47.12	58.2	96.7%
DivPrune (CVPR25)	59.36	1716.80	68.47	49.67	56.19	35.78	86.65	48.17	57.8	95.9%
VisionZip (CVPR25)	57.61	1742.65	68.82	52.29	55.41	**36.22**	82.04	45.03	56.8	94.9%
VScan (TMLR26)	59.45	1767.78	**68.96**	53.07	55.96	**36.22**	85.66	45.02	57.8	96.3%
RobustPruner (Ours)	**60.43**	**1854.23**	68.77	**54.12**	**58.71**	36.00	85.85	**48.69**	**58.9**	**98.7%**
Retain 30% Tokens (↓70%)										
FastV (ECCV24)	52.72	1571.13	68.47	47.71	54.73	35.22	63.19	38.17	51.5	86.8%
CDPruner (NeurIPS25)	60.30	1762.18	68.96	52.68	55.83	35.67	**86.82**	49.21	58.5	97.4%
DivPrune (CVPR25)	59.85	1744.81	**69.26**	51.50	56.09	36.44	86.62	45.55	57.9	96.3%
VisionZip (CVPR25)	58.90	1778.06	69.06	51.37	56.12	**36.89**	85.29	43.98	57.4	95.9%
VScan (TMLR26)	60.30	1778.79	68.96	52.68	57.19	36.56	86.40	46.07	58.3	97.2%
RobustPruner (Ours)	**61.07**	**1860.76**	68.96	**55.69**	**58.55**	36.11	85.87	**49.22**	**59.4**	**99.4%**

Findings:

- **RobustPruner** preserves full-model performance well even under aggressive pruning and performs strongly on GQA, MME, RWQA, AI2D, and V*Bench.
- **RobustPruner** generalizes to the classic LLaVA backbone and offers a better efficiency–performance trade-off across token budgets.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

Robustness validation under perturbations on Qwen-2.5-VL-7B across five benchmarks.

(To test out-of-distribution robustness, we add a new perturbation5. (Retain 20% tokens)

Method	Original	Perturbation1	Perturbation2	Perturbation3	Perturbation4	Perturbation5	Avg.	Rel.
AI2D								
Qwen2.5-VL-7B	93.75	92.68	93.10	92.20	93.01	93.36	93.02	100.0%
FastV (ECCV24)	83.06	81.83	82.80	81.64	82.35	82.41	82.35	88.5%
CDPruner (NeurIPS25)	72.51	70.95	72.11	71.28	71.50	72.89	71.87	77.3%
DivPrune (CVPR25)	81.86	80.86	81.63	80.67	80.76	81.41	81.20	87.3%
VisionZip (CVPR25)	87.99	81.02	81.76	81.41	81.70	82.19	82.68	88.9%
VScan (TMLR26)	80.38	79.14	80.21	78.40	79.27	79.48	79.48	85.4%
RobustPruner (Ours)	**89.51**	**88.63**	**89.35**	**88.08**	**88.47**	**89.99**	**89.00**	**95.7%**
ChartQA								
Qwen2.5-VL-7B	83.08	77.12	79.68	76.44	68.36	64.48	74.86	100.0%
FastV (ECCV24)	62.52	59.60	61.40	59.80	52.64	49.44	57.57	76.9%
CDPruner (NeurIPS25)	53.12	50.60	51.44	50.80	44.12	45.12	49.20	65.7%
DivPrune (CVPR25)	59.28	56.16	57.40	56.08	49.84	45.96	54.12	72.3%
VisionZip (CVPR25)	61.64	56.88	57.96	57.04	51.04	46.32	55.15	73.7%
VScan (TMLR26)	59.60	56.52	57.76	56.92	50.52	48.24	54.93	73.4%
RobustPruner (Ours)	**65.72**	**61.80**	**63.60**	**61.40**	**52.88**	**52.01**	**59.57**	**79.6%**
GQA								
Qwen2.5-VL-7B	60.46	54.59	53.95	54.91	51.73	59.19	55.80	100.0%
FastV (ECCV24)	51.46	46.54	46.35	46.94	44.40	50.19	47.65	85.4%
CDPruner (NeurIPS25)	55.68	50.13	49.48	50.10	46.49	53.10	50.83	91.1%
DivPrune (CVPR25)	58.46	52.33	**52.13**	53.12	**49.67**	56.16	53.64	96.1%
VisionZip (CVPR25)	56.78	51.05	50.77	51.56	48.35	55.05	52.26	93.6%
VScan (TMLR26)	57.12	51.34	50.94	51.60	48.34	54.48	52.30	93.7%
RobustPruner (Ours)	**58.54**	**52.72**	52.11	**53.14**	49.44	**56.86**	**53.80**	**96.4%**
MME								
Qwen2.5-VL-7B	2323.62	2249.76	2327.96	1879.38	2303.97	2120.76	2200.91	100.0%
FastV (ECCV24)	2155.85	2086.32	2128.24	1722.57	2117.19	1901.81	2018.66	91.7%
CDPruner (NeurIPS25)	2071.28	1988.46	2066.65	1686.62	1993.84	1872.82	1946.61	88.4%
DivPrune (CVPR25)	2194.63	2131.34	2156.13	1747.64	2133.30	1988.10	2058.52	93.5%
VisionZip (CVPR25)	2200.18	2124.87	2178.78	1732.12	2153.43	1999.88	2064.88	93.8%
VScan (TMLR26)	2258.84	2175.52	2206.16	1773.13	2161.16	1900.35	2079.19	94.5%
RobustPruner (Ours)	**2276.22**	**2235.81**	**2263.97**	**1905.92**	**2250.25**	**2201.16**	**2188.89**	**99.5%**
OCRBench								
Qwen2.5-VL-7B	83.80	79.70	82.00	81.50	82.10	84.20	82.22	100.0%
FastV (ECCV24)	59.70	57.50	58.00	58.40	58.70	59.20	58.58	71.3%
CDPruner (NeurIPS25)	49.60	45.90	48.60	48.30	47.40	48.70	48.08	58.5%
DivPrune (CVPR25)	63.50	59.50	62.40	61.30	62.00	62.80	61.92	75.3%
VisionZip (CVPR25)	59.30	56.10	58.80	57.40	59.10	59.50	58.37	71.0%
VScan (TMLR26)	51.30	48.60	48.10	44.50	49.80	48.46	48.46	58.9%
RobustPruner (Ours)	**71.10**	**65.20**	**68.70**	**67.10**	**70.30**	**70.40**	**68.80**	**83.7%**

Findings:

- **RobustPruner** achieves the best average performance at 20% token retention across all five benchmarks under original and perturbed inputs.
- **RobustPruner** outperforms training-free and token-pruning baselines and is less sensitive to perturbations, especially on AI2D, ChartQA, MME, and OCRBench.
- This shows its query and diversity-aware token selection improves accuracy and robustness under cases.

NEURAL INFORMATION PROCESSING SYSTEMS

Part 4

Experiments

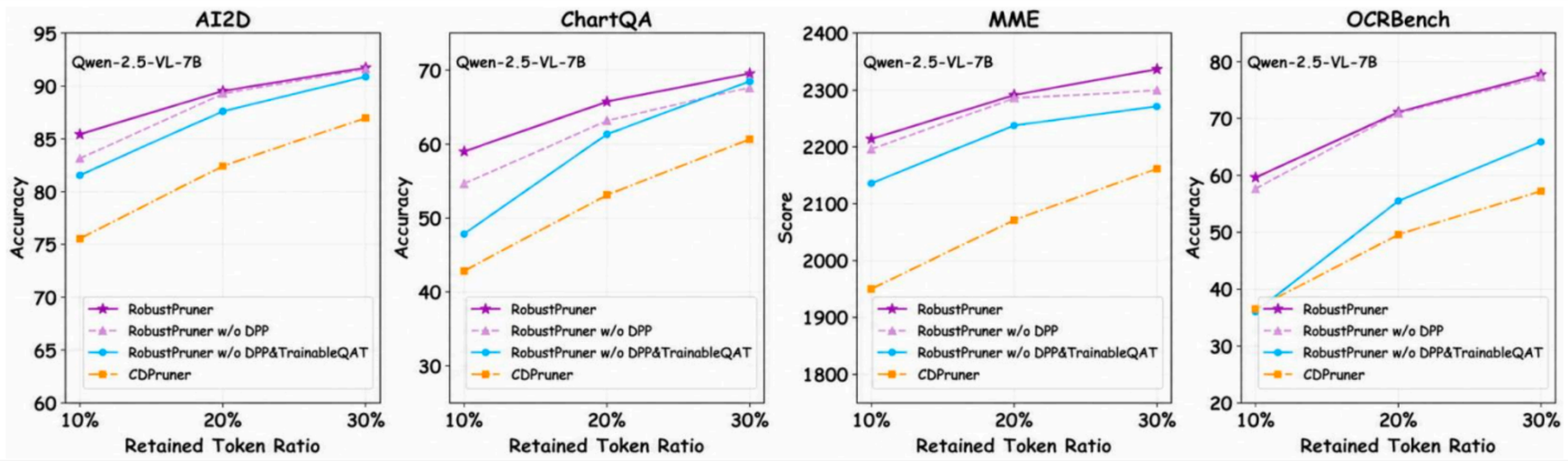
Efficiency: What is the cost?

Space efficiency is the core:

Method	Visual Tokens	Prefill FLOPs	KV Cache	GPU Memory	Avg.Time per Question	Score
LLaVA-OneVision-1.5-8B	523.3	8.53T	83.1MB	19.0GB	265.09ms	94.01
+RobustPruner (Ours)	**52.0 (10.06×)**	**4.41T (1.93×)**	**17.0MB (4.89×)**	**16.5GB (1.15×)**	**218.84ms (1.21×)**	**88.7**
Qwen2.5-VL-7B	510.2	7.72T	31.6MB	17.5GB	154.87ms	93.75
+RobustPruner (Ours)	**50.6 (10.08×)**	**3.93T (2.0×)**	**6.5MB (4.86×)**	**16.0GB (1.09×)**	**145.67ms (1.08×)**	**85.43**

RobustPruner cuts visual tokens by ~10×, reducing prefill FLOPs ~2× and KV cache ~5× at 10% retention.
It delivers clear inference savings while keeping competitive AI2D performance.

Ablation Study: Experiments on the progress.



Comparison:

1. RobustPruner
2. RobustPruner w/o DPP
3. RobustPruner w/o DPP&QAT
4. Training-free CDPruner

NEURAL INFORMATION PROCESSING SYSTEMS

Part 5

Conclusion

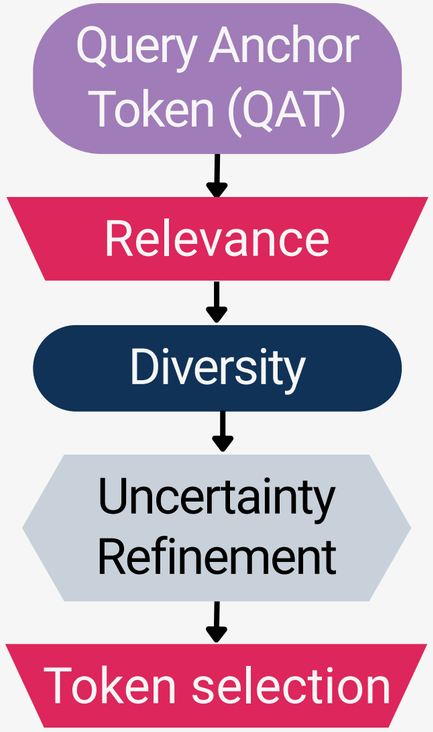
RobustPruner: From “Top-k Importance” to “Relevance–Diversity–Reliability Selection”

Why RobustPruner?

Conventional visual-token pruning ranks tokens using a single relevance score.

Relevance alone cannot tell whether a pruning decision is stable.

How it works?



QAT produces query-conditioned relevance.
DPP preserves relevant but non-redundant candidate tokens.
Uncertainty refinement keeps the most stable evidence within the candidate set.

Robust: Preserves high accuracy across MLLMs, token budgets, and prompt rewrites.
Efficient: cuts many prefill FLOPs and KV cache.
Practical: One-shot pruning at an intermediate decoder layer with a frozen backbone.

Relevance tells us WHAT matters; diversity avoids redundancy; uncertainty tells us WHICH evidence is reliable. RobustPruner preserves stable, answer-critical visual evidence under aggressive compression.

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Information Processing Systems 2026

THANK YOU!

Hanwei Zhu*, Junhan Fu*, Xi Zhang, Jiamang Wang, Weisi Lin

NeurIPS 2026 (Poster)

Presenter: Junhan Fu
Email: junhanfu@link.cuhk.edu.cn

September 2026