UHR-BAT: Budget-Aware Token Compression Vision-Language model for Ultra-High-Resolution Remote Sensing

Yunkai Dang*, Minxin Dai*, Yuekun Yang, Zhangnan Li, Wenbin Li†, Feng Miao†, Yang Gao

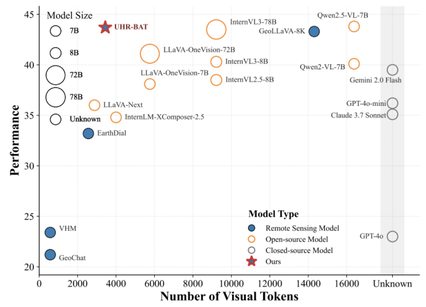
ICML International Conference On Machine Learning

Reasoning and Learning Research Group

code available

1. Background: Ultra-High-Ressolution Remote Sensing

Accuracy–efficiency frontier for Ultra-High-Resolution Remote Sensing MLLMs on XLRS-Benchbenchmark.



MLLM Architecture：

$$E = \{\mathbf{e}_i\}_{i=1}^{N} = \phi(F), \qquad \mathbf{e}_i \in \mathbb{R}^d.$$

Query-aware Token Importance：

$$a_i = \frac{1}{M}\sum_{j=1}^{M} \mathbf{A}_{j,i}.$$

Token Budget

$$\bar{E} = C(E; B, \{a_i\}_{i=1}^{N}), \qquad \text{s.t.} \quad |\bar{E}| \le B.$$

Region Partition

$$\mathcal{P} = \{\mathcal{S}_m\}_{m=1}^{R}, \qquad \bigcup_{m=1}^{R} \mathcal{S}_m = \mathcal{I}, \quad \mathcal{S}_m \cap \mathcal{S}_n = \emptyset.$$

Contribution：

- We propose UHR-BAT, a token compression framework designed for ultra high resolution remote-sensing MLLMs under strict context budgets.
- We propose a query-guided, multi-scale input mechanism to integrate text-derived global priors and capture both holistic context and fine-grained details.
- We propose region-wise preserve and merge strategies to preserve salient local evidence and aggregate redundant background into compact representatives.
- Empirical results across standard benchmarks confirm that UHR-BAT establishes a new state-of-the-art for efficient UHR understanding, outperforming existing methods under strict token budgets.

2. Overview of our method

Cross-scale Importance Alignment.

(General Setting) $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

DF resides within a space spanned by an infinite outer product basis of score and data.

$$\mathbf{A}_{u,v}^{(s)} = \Psi\left(\mathbf{A}^{(1)}, \tau_s(u,v)\right),$$

$$a_i^{(s)} = \frac{1}{|\Omega_i^{(s)}|}\sum_{(u,v)\in\Omega_i^{(s)}} \mathbf{A}_{u,v}^{(s)}.$$

$$\bar{E}^{(s)} = C(E^{(s)}; B_s, \{a_i^{(s)}\}) \text{ with } |\bar{E}^{(s)}| \le B_s$$

Priority-based Serialization and Budgeting

$$\mathcal{P}^{(s)} = \{\mathcal{S}_m^{(s)}\}_{m=1}^{R_s}$$

DF resides within a space spanned by an infinite outer product basis of score and data.

$$\mu_m^{(s)} = \frac{1}{|\mathcal{S}_m^{(s)}|}\sum_{i\in\mathcal{S}_m^{(s)}} a_i^{(s)}.$$

$$\mathcal{K}_m^{(s)} = \left\{ i \in \mathcal{S}_m^{(s)} \mid a_i^{(s)} \ge \mu_m^{(s)} \right\},$$

$$\mathcal{R}_m^{(s)} = \mathcal{S}_m^{(s)} \setminus \mathcal{K}_m^{(s)}$$

$$\mathbf{S}_{\pi(k)}^{(s)} = \{\mathbf{h}_i^{(s)} : i \in \mathcal{K}_{\pi(k)}^{(s)}\}_{\downarrow},$$

$$\tilde{\mathbf{e}}_m^{(s)} = \frac{1}{|\mathcal{R}_m^{(s)}|}\sum_{i\in\mathcal{R}_m^{(s)}} \mathbf{h}_i^{(s)}, \qquad \text{if } |\mathcal{R}_m^{(s)}| > 0.$$

Region-wise Preserve-and-Merge

(General Setting) $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

DF resides within a space spanned by an infinite outer product basis of score and data.

$$\mathbf{S}_{\pi(k)}^{(s)} = \{\mathbf{h}_i^{(s)} : i \in \mathcal{K}_{\pi(k)}^{(s)}\}_{\downarrow},$$

$$E_{\text{cand}}^{(s)} = \text{Concat}_{k=1}^{R_s}\left(\mathbf{S}_{\pi(k)}^{(s)}, \{\tilde{\mathbf{e}}_{\pi(k)}^{(s)} : |\mathcal{R}_{\pi(k)}^{(s)}| > 0\}\right),$$

3. Overview of our method

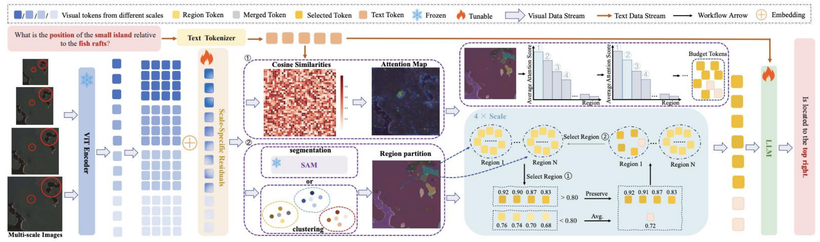


Figure 2. Overview of our method. We encode a high-resolution remote sensing image at multiple scales using a frozen ViT, applying Scale-Specific Positional Embeddings to distinguish visual tokens across scales, and then obtain an anchor-scale query-to-vision attention map from the MLLM interface. A region partition (e.g., induced by SAM or feature+coordinate clustering) enables region-wise preserve-and-merge, while remaining tokens are merged via average pooling to retain coarse context. Finally, the processed sequence is fed into the LLM to generate the final answer.

Algorithm 1 Region-aware Token Pruning

Require: Input feature map $\mathbf{X}$, Number of regions k, Pruning ratio ρ.
Ensure: Pruned tokens $\mathbf{X}_{out}$.
// **Step 1: Region Partition**
Construct clustering embeddings $\mathbf{u}_i$ for all tokens via Eq. (13).
Obtain regions $\{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ by minimizing Eq. (14) (K-means) or via Eq. (15) (SAM).
// **Step 2: Importance Assessment**
Compute the attention score map $A_{u,v}^{(s)}$ for each scale according to Eq. (6).
Derive the individual token importance scores $\{a_i^{(s)}\}_{i=1}^{N_s}$ following Eq. (7).
Calculate the average importance score $\{\mu_m^{(s)}\}_{m=1}^{R_s}$ for each region as defined in Eq. (8).
// **Step 3: Region-wise Preserve, Merge, and Priority Budgeting**
Initialize candidate sequence $E_{\text{cand}}^{(s)} \leftarrow []$
for each region $\mathcal{S}_m^{(s)} \in \mathcal{P}^{(s)}$ **do**
  $\mu_m^{(s)} \leftarrow \frac{1}{|\mathcal{S}_m^{(s)}|}\sum_{i\in\mathcal{S}_m^{(s)}} a_i^{(s)}$
  $\mathcal{K}_m^{(s)} \leftarrow \{i \in \mathcal{S}_m^{(s)} \mid a_i^{(s)} \ge \mu_m^{(s)}\}$
  $\mathcal{R}_m^{(s)} \leftarrow \mathcal{S}_m^{(s)} \setminus \mathcal{K}_m^{(s)}$
  if $|\mathcal{R}_m^{(s)}| > 0$ **then**
    $\tilde{\mathbf{e}}_m^{(s)} \leftarrow \text{MeanPool}(\{\mathbf{h}_i^{(s)} \mid i \in \mathcal{R}_m^{(s)}\})$
  end if
end for
Sort region indices π by descending average importance $\mu_m^{(s)}$
for $r = 1$ **to** R_s **do**
  $m \leftarrow \pi(r)$
  Sort tokens in $\mathcal{K}_m^{(s)}$ by $a_i^{(s)}$ in descending order
  Append sorted tokens in $\mathcal{K}_m^{(s)}$ to $E_{\text{cand}}^{(s)}$
  if $\tilde{\mathbf{e}}_m^{(s)}$ exists **then**
    Append $\tilde{\mathbf{e}}_m^{(s)}$ to $E_{\text{cand}}^{(s)}$
  end if
end for
$\bar{E}^{(s)} \leftarrow E_{\text{cand}}^{(s)}[1 : B_s]$ {Truncate to budget B_s}
return $\bar{E}^{(s)}$

- Region-aware token pruning partitions visual tokens into semantically coherent regions, enabling local compression instead of global top-K selection and preventing dominant background regions from consuming the budget.
- Query-aware attention then estimates token importance: high-importance tokens are preserved, while redundant low-importance tokens are merged into compact representatives.
- Finally, regions are ordered by average importance and truncated to the scale-specific budget B_{s}, yielding a fixed-length sequence that preserves fine details and reduces redundancy.

4. Main Experiment Results

Table 1. Results on XLRS-Bench across perception and reasoning sub-tasks. 'w.Avg.' indicates the sample-count-weighted average.

Method	Size	Perception								Reasoning							Overall
Sub-tasks (L-3 Capability)		OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCD	CCR	w.Avg.	
number of samples		60	100	100	200	800	800	60	500	100	100	100	60	100	3080	
Open-source MLLMs																
InternLM-XComposer-2.5 (Zhang et al., 2024a)	7B	21.7	42.0	7.0	68.0	31.8	27.8	6.7	26.0	72.0	81.0	41.0	36.7	47.0	34.8	
LLaVA-Next (Liu et al., 2024)	7B	26.7	40.0	5.0	67.0	28.8	32.8	**66.7**	30.0	69.0	78.0	27.0	35.0	36.0	36.0	
LLaVA-OneVision-7B (Li et al., 2024a)	7B	25.0	38.0	2.0	69.5	35.9	35.3	65.0	25.2	76.0	**83.0**	24.0	43.3	36.0	38.1	
InternVL2.5-8B (Chen et al., 2024b)	8B	38.3	37.0	10.0	**77.0**	33.4	35.5	65.0	21.6	73.0	**83.0**	34.0	**50.0**	43.0	38.5	
InternVL3-8B (Chen et al., 2024c)	8B	40.0	39.0	10.0	71.5	**44.5**	30.8	65.0	25.2	**77.0**	82.0	36.0	21.7	50.0	40.3	
Qwen2-VL-7B (Wang et al., 2024)	7B	26.7	40.0	11.0	73.0	35.9	34.6	61.7	31.8	70.0	81.0	35.0	46.7	48.0	40.1	
Qwen2.5-VL-7B (Bai et al., 2025)	7B	33.3	40.0	31.0	**77.0**	40.6	**40.5**	**66.7**	36.2	68.0	72.0	27.0	38.3	45.0	43.8	
LLaVA-OneVision-72B (Li et al., 2024a)	72B	33.3	38.0	15.0	72.5	36.3	36.3	**66.7**	35.6	74.0	**83.0**	28.0	36.7	43.0	41.1	
InternVL3-78B (Zhu et al., 2025)	78B	23.3	**49.0**	33.0	74.0	42.5	37.4	**66.7**	30.0	76.0	81.0	40.0	45.0	42.0	43.5	
Closed-source MLLMs																
GPT-4o (Hurst et al., 2024)	–	25.0	32.0	15.0	66.0	9.5	11.3	11.7	24.6	73.0	73.0	35.0	20.0	25.0	23.0	
GPT-4o-mini (Hurst et al., 2024)	–	23.3	25.0	19.0	59.5	40.9	31.0	65.0	23.6	71.0	71.0	29.0	6.7	30.0	36.2	
Claude 3.7 Sonnet (Anthropic, 2025)	–	27.6	22.7	17.4	68.4	30.5	29.9	63.6	27.6	64.8	78.4	34.5	27.8	32.6	35.1	
Gemini 2.0 Flash (Team et al., 2023)	–	**41.7**	45.0	38.0	73.5	34.6	27.6	61.7	32.0	73.0	82.0	43.0	30.0	51.0	39.5	
Remote Sensing MLLMs																
GeoChat (Kuckreja et al., 2024)	7B	16.7	29.0	2.0	23.0	21.1	16.8	35.0	24.2	33.0	43.0	10.0	24.0	21.0	21.2	
GeoLLaVA-8K (Wang et al., 2025a)	7B	26.7	38.0	49.0	69.0	41.6	31.6	65.0	35.0	67.0	78.0	**66.0**	**50.0**	**52.0**	43.3	
EarthDial (Soni et al., 2025)	7B	18.3	42.0	1.0	36.0	31.3	31.0	65.0	24.8	62.0	71.0	43.0	48.3	50.0	33.8	
VHM (Pang et al., 2025)	7B	16.7	30.0	2.0	26.0	21.4	16.8	35.0	25.6	42.0	53.0	46.0	28.3	21.0	23.6	
UHR-BAT (Ours)	7B	21.7	33.0	**50.0**	55.5	43.5	33.8	65.0	**44.8**	62.0	71.0	54.0	46.7	51.0	**44.0**	

Table 2. RSHR-Bench results across Perception and Reasoning tasks.

Model	Size	Perception								Reasoning					Average		
		COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN	AD	FP	MRJC	MRJCS	OSJ	P.Avg.	R.Avg.	Avg.
Open-source VLMs																		
InternVL2.5-8B (Chen et al., 2024b)	8B	25.5	22.0	26.0	26.0	22.5	24.5	30.0	22.5	20.0	26.0	20.0	22.5	34.0	20.0	24.3	24.5	24.4
InternVL3-8B (Wang et al., 2025c)	8B	21.5	28.0	18.0	21.5	29.0	28.5	30.0	26.5	25.0	20.0	16.0	29.0	34.0	26.0	25.3	25.0	25.2
MiniCPM2_6 (Yao et al., 2024)	7B	21.5	28.0	30.0	24.0	19.5	29.5	34.3	22.0	29.0	26.0	30.0	35.0	32.0	30.0	27.4	30.6	28.5
Pix-I-5-Vicuna (Abdin et al., 2024)	7B	25.0	24.0	25.0	25.0	23.5	25.0	22.9	25.0	25.0	24.0	22.0	23.5	30.0	22.0	24.5	24.3	24.4
Qwen2.5-VL-7B (Bai et al., 2025)	7B	29.5	25.0	22.0	28.0	25.0	24.5	24.3	26.5	22.0	26.0	28.0	25.0	10.0	20.0	25.2	21.8	24.0
DeepSeek-VL (Lu et al., 2024)	7B	22.5	22.0	21.0	25.0	20.5	26.0	28.6	20.5	22.0	28.0	24.0	32.0	20.0	22.0	23.1	30.4	25.7
VILA-HD (Shi et al., 2025)	7B	40.0	22.0	22.0	37.0	35.5	26.0	21.4	24.5	24.0	58.0	30.0	50.0	32.0	58.0	28.0	46.6	34.6
Closed-source VLMs																		
GPT-5 (OpenAI, 2025)	–	29.0	10.0	23.0	23.0	37.0	24.5	31.4	20.0	23.0	74.0	58.0	35.0	34.0	66.0	24.5	53.4	34.8
GPT-4o (Hurst et al., 2024)	–	49.5	23.0	15.0	35.5	30.5	28.0	27.1	22.5	41.0	68.0	56.0	30.5	32.0	64.0	30.2	50.1	37.3
GPT-4o-mini (Hurst et al., 2024)	–	41.5	16.0	29.0	31.5	31.5	32.0	28.6	19.5	32.0	54.0	54.0	31.5	48.0	54.0	29.1	48.3	36.0
Remote Sensing VLMs																		
EarthDial (Soni et al., 2025)	7B	41.0	22.0	21.0	30.0	32.5	30.5	27.1	18.0	31.0	42.0	30.0	29.5	32.0	52.0	28.1	37.1	31.3
GeoChat (Kuckreja et al., 2024)	7B	32.5	22.0	24.0	29.5	40.0	25.0	22.9	22.5	29.0	30.0	24.0	25.5	30.0	32.0	25.9	28.3	26.8
GeoLLaVA-8K (Wang et al., 2025a)	7B	25.0	24.0	25.0	25.0	25.0	25.0	21.4	25.0	25.0	0.0	0.0	0.0	34.0	22.0	24.5	16.0	21.5
VHM (Pang et al., 2025)	7B	25.5	25.0	26.0	26.5	55.0	25.0	22.9	25.0	25.0	26.0	24.0	26.5	34.0	28.0	25.7	27.7	26.4
UHR-BAT (Ours)	7B	46.5	32.0	38.0	12.5	38.5	24.0	17.1	17.0	37.0	44.0	34.0	65.0	32.0	50.0	29.2	45.0	34.8

Table 3. Results on the MMERealworld-RS dataset, with methods ranked by Avg

Method	Params	Position	Color	Count	Avg.
GPT-4o (Hurst et al., 2024)	–	33.52	29.83	18.90	27.42
LLaVA-v1.5-7B (Liu et al., 2023)	7B	21.48	22.95	16.31	20.28
LLaVA-v1.6-7B (Liu et al., 2024)	7B	26.49	24.06	20.47	23.70
LLaVA-ov-7B (Li et al., 2024a)	7B	26.81	26.14	27.57	26.83
Qwen2.5-VL-7B (Bai et al., 2025)	7B	22.12	15.54	14.93	17.55
LLaVA-HR (Luo et al., 2024a)	7B	35.56	44.30	7.91	29.26
GeoChat (Kuckreja et al., 2024)	7B	25.06	23.11	15.66	21.32
VHM (Pang et al., 2025)	7B	35.24	20.32	16.80	24.18
GeoLLaVA-8K (Wang et al., 2025a)	7B	34.90	27.92	22.27	28.41
UHR-BAT (Ours)	7B	44.00	42.00	14.00	33.33

Table 4. Performance on XLRS-Bench across multiple resolutions under varying scale-specific token budgets K.

Resolution	Other scales	K	Perception	Reasoning	w.Avg.
672×672	320,600,2400	0	40.76	57.83	43.31
672×672	320,600,2400	40	40.88	58.04	43.44
672×672	320,600,2400	80	40.95	57.83	43.47
672×672	320,600,2400	120	41.15	58.04	43.67
1344×1344	120,600,2400	80	40.69	57.61	43.21
1344×1344	120,600,2400	160	40.88	57.17	43.33
1344×1344	120,600,2400	240	41.07	57.39	43.51
1344×1344	120,600,2400	320	41.15	58.04	43.67
2688×2688	120,320,2400	0	40.80	58.04	43.38
2688×2688	120,320,2400	200	40.84	57.83	43.38
2688×2688	120,320,2400	400	41.07	57.83	43.57
2688×2688	120,320,2400	600	41.15	58.04	43.67
4032×4032	120,320,600	400	40.99	57.39	43.44
4032×4032	120,320,600	1200	40.99	57.61	43.47
4032×4032	120,320,600	2000	40.38	58.04	43.02
4032×4032	120,320,600	2800	41.07	57.83	43.57

Figure 3. Ablation on clustering centers and retained tokens. Left: accuracy versus the number of cluster centers k. Right: accuracy versus the number of retained visual tokens.

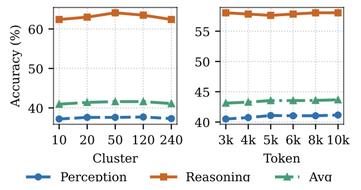


Table 5. Ablation study of each compoment.

Variants	Perception	Reasoning	Overall
w/o Query-aware scoring	39.8	55.9	42.2
w/o Cross-scale alignment	40.4	57.2	42.9
w/o Region partitioning	40.1	57.6	42.7
w/o Preserve-and-merge	40.2	57.6	42.8
Full Model (Ours)	41.5	57.8	**44.0**

Table 7. Performance analysis on XLRS-Bench under varying resolutions and token budgets.

Kept Tokens	Image Resolution	Perception Mean	Reasoning Mean	w.Avg.
3k	8k	40.38	57.83	43.00
4k	8k	40.61	57.52	43.12
3k	4k	40.50	58.04	43.12
4k	4k	40.73	57.83	43.28

Table 9. Evaluation of various region partitioning methods.

Method	4k	5k	6k	8k	1w
BIRCH	43.05	43.28	43.28	43.31	43.51
K-means	43.34	43.57	41.43	43.73	43.31
SAM	43.28	43.54	43.54	43.57	43.67

Figure 4. Multi-scale token selection accuracy on XLRS-Bench. We evaluate vary the number of retained visual tokens at each scale (x-axis).

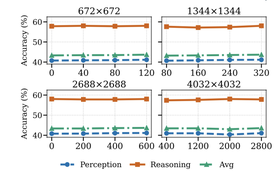


Table 6. Efficiency analysis: Latency and FLOPs per sample under different token budgets (Total input tokens = 131,328).

Compression (×)	Kept Tokens	TFLOPs	Avg. Total (s)	Avg. Gen. (s)
32.83	4k	267.68	20.93	4.34
26.27	5k	274.91	21.58	4.56
21.89	6k	289.10	21.83	4.74
16.42	8k	317.48	21.99	5.11
13.13	10k	345.75	25.90	5.65
11.94	11k	359.94	25.79	5.86
10.94	12k	374.24	22.79	5.99

Table 8. Performance and efficiency analysis on XLRS-Bench under varying token budgets.

Kept Tokens	Latency (s)	Accuracy	TFLOPS	Usage (GB)
131k (No Compression)	180.0*	OOM	8650.0*	128.85*
44k	24.54	44.5	816.0	84.05
11k	14.48	44.0	358.0	29.57
8k	10.38	43.7	315.3	26.73
5k	9.04	43.5	272.7	24.59

Table10. Ablation study: Accuracy (%) across Perception and Reasoning tasks under different cluster counts (k).

Metric	10	50	120	240	360	480	600	720	1000
Perception	37.2	37.6	37.7	37.3	37.4	37.3	37.5	37.1	37.2
Reasoning	62.4	64.1	63.5	62.4	63.3	63.7	63.9	63.3	63.0
Average	40.9	**41.6**	**41.6**	41.1	41.3	41.2	41.4	41.0	41.1

Figure 6. Qualitative examples of coarse segmentation on remote-sensing images produced by our segmentation module, used to form semantically coherent regions for region-aware processing.

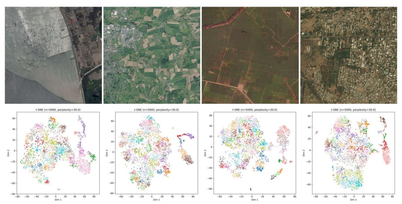
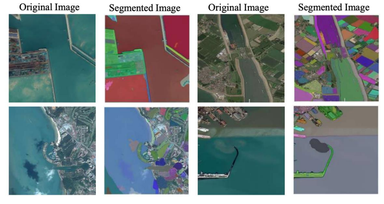
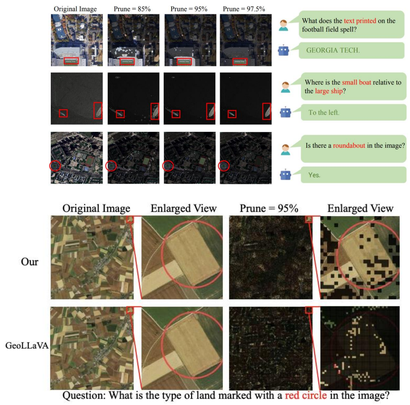
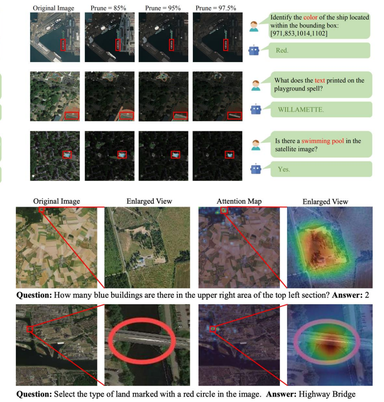


Figure 7. t-SNE visualization of visual token clustering, showing the redundancy of visual tokens and supporting the feasibility of token merging.



5. Visualization of Our Model