AD-BTS: Adaptive Dual-Branch Token Sparsification via Spatial Information Density

Xinpei Gao[1,2*], Xin Luo[1,2*], Ming Liu[1,2], Chunjiang Wang[1,2], S Kevin Zhou[1,2,3,4,5†] (* equal contribution, † corresponding author)
[1] School of Biomedical Engineering, University of Science and Technology of China (USTC), China; [2] MIRACLE Lab, USTC Suzhou Institute for Advanced Research, Suzhou, China
[3] State Key Laboratory of Precision and Intelligent Chemistry, USTC, China; [3] Semou; [3] State Key Laboratory of Precision and Intelligent Chemistry, USTC, China
[4] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, USTC, China; [5] Biomedical Basic Research Center of Jiangsu Province, Suzhou, China
email: skevinzhou@ustc.edu.cn

ICML International Conference On Machine Learning

Introduction

Token explosion in high-resolution MLLMs

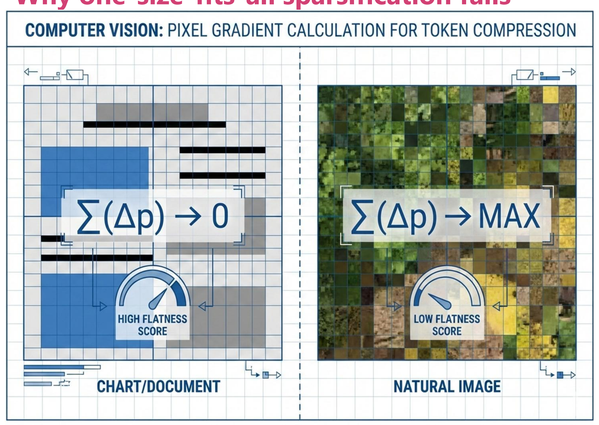
- High-resolution visual encoders generate thousands of visual tokens per image.
- During prefill, attention cost scales as O((N+M)^2), causing latency and memory bottlenecks.
- Static pruning/merging assumes spatial redundancy and fails when key information is sparse.

Core Contributions

- Density-Aware Dual-Branch Sparsification. We introduce AD-BTS to treat token sparsification as dynamic resource allocation, balancing efficiency and structural integrity via a frozen selection branch and a compensatory fusion branch.
- Lightweight Gradient-Based Routing. We propose a parameter-free Gradient-based Routing Gate that identifies structure-sensitive inputs directly from pixel-level gradient statistics.
- Performance Breakthrough at High Compression. AD-BTS sets a new efficiency–accuracy Pareto frontier, consistently outperforming state-of-the-art baselines under 10–30% token retention with substantial prefill speedups.

Method

Why one-size-fits-all sparsification fails



$$\mathcal{S}_{\text{flat}}(I) = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \mathbb{I}\left(\|\nabla I_{u,v}\|_1 < \tau\right)$$

Charts / documents
Dense correlations; many background tokens can be removed safely.

Natural images
Sparse lines, glyphs and axes are structurally essential; losing a few tokens can break reasoning.

Key idea: route by input signal density before deciding how to compress visual tokens.

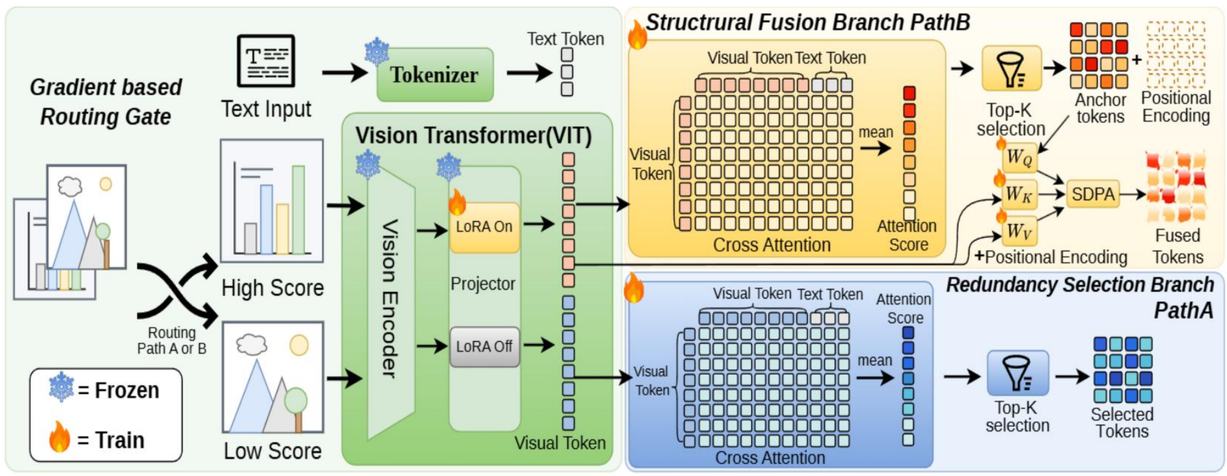
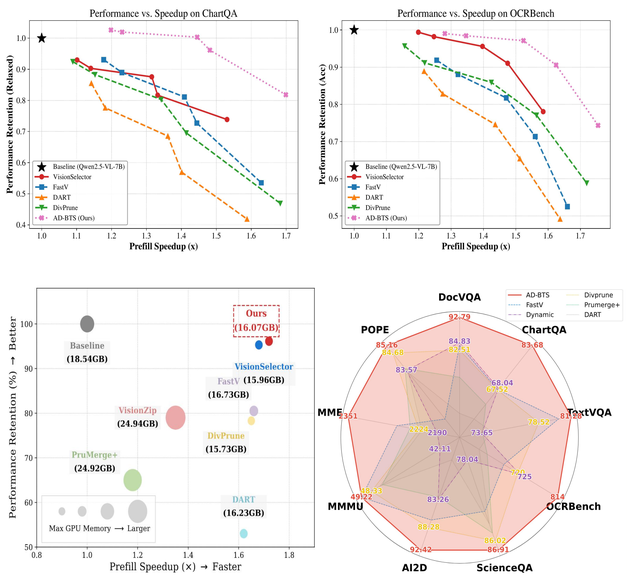


Figure 2. **Overview of the Adaptive Dual-Branch Token Sparsification via Spatial Information Density(AD-BTS) framework.** The architecture consists of three key components: (1) The Gradient-based Routing Gate (GRG) estimates spatial information density (sparsity) to route inputs and configure Dynamic Adapter Modulation (DAM). (2) Path A: Redundancy Selection Branch (RSB) processes redundancy-dominated inputs by keeping LoRA frozen (**Frozen**) and performing hard token selection. (3) Path B: Structural Fusion Branch (SFB) handles spatially sparse structured inputs by enabling LoRA adapters (**Train**) and fusing context into anchor tokens via FlashAttention-based cross-attention. The dynamic routing ensures efficient inference without sacrificing capacity for structural data.

Results

Pareto frontier on structure-sensitive tasks



AD-BTS maintains higher performance as speedup increases, especially for charts and OCR.

Table 1. **Comparison with state-of-the-art methods across varying token retention rates.** We evaluate methods on nine benchmarks under three compression regimes (30%, 20%, and 10% retained tokens). The table reports absolute metrics for individual tasks, while the *Avg Score* column presents the performance normalized relative to the uncompressed Qwen2.5-VL-7B upper bound. **Bold** highlights the best results among compressed models.

Method	Training Free	DocVQA (Anls)	ChartQA (Relaxed)	TextVQA (EM)	OCRBench (Score)	ScienceQA (EM)	AI2D (Acc)	MMMU (Acc)	MME (Score)	POPE (F1)	Avg Score (Norm %)
Dynamic Resolution (MinPix=256x28x28, MaxPix=2048x28x28), Upper Bound (100%)											
Qwen-2.5-VL-7B	–	94.33	83.4	82.84	838	87.26	93.59	50.78	2342.15	86.19	100.00%
Retain 30% Tokens (70% Compression Ratio)											
DART	✓	62.6	56.88	74.45	629	84.33	75.94	47.56	2218.83	83.43	84.72%
Prumerge+	✓	73.95	62.24	73.71	648	85.23	82.77	47.33	2239.64	83.69	87.93%
FastV	✓	84.01	67.64	80.22	687	83.09	86.92	49.22	2263.58	80.47	91.56%
Divprune	✓	82.51	67.52	78.52	720	86.02	88.28	48.33	2224.06	84.68	92.27%
Visionzip	✓	86.11	72.28	77.3	711	86.61	87.86	49.44	2276.04	84.73	93.57%
Dynamic	×	84.83	68.04	73.65	725	78.04	83.26	42.11	2190.78	83.57	88.75%
Visionselector	×	92.61	73.04	**81.28**	801	85.77	91.94	**50.11**	2313.74	85	96.90%
AD-BTS (Ours)	×	**92.79**	**83.68**	**81.28**	**814**	**86.91**	**92.42**	49.22	**2351.76**	**85.16**	**98.72%**
Retain 20% Tokens (80% Compression Ratio)											
DART	✓	73.81	57.88	73.86	648	84.33	82.29	46.33	2198.82	83.55	86.75%
Prumerge+	✓	50.03	47.16	67.53	537	83.34	71.05	46.33	2138.61	80.16	78.02%
FastV	✓	74.75	62.04	72.03	591	84.68	82.32	47	2168.86	83.23	86.43%
Divprune	✓	75.99	60.48	78.01	597	82.75	82.35	**49**	2152.74	76.12	86.45%
Visionzip	✓	61.09	52.56	68.72	562	83.79	77.62	46.11	2219.3	81.74	81.91%
Dynamic	×	78.09	66.20	72.22	675	77.00	81.70	43.78	2105.20	81.95	86.29%
Visionselector	×	89.78	68.12	79.79	763	85.42	90.48	48.78	**2273.94**	84.29	94.42%
AD-BTS (Ours)	×	**90.22**	**80.2**	**80.15**	**766**	**85.97**	**90.67**	48.33	2258.88	**84.3**	**96.09%**
Retain 10% Tokens (90% Compression Ratio)											
DART	✓	33.13	34	53.97	415	81.85	67.1	46.11	1980.7	71.91	68.32%
Prumerge+	✓	42.08	41.56	56.87	417	81.56	71.08	45.22	1948.58	76.52	71.48%
FastV	✓	58.64	44.64	70.83	440	81.95	75.74	45.78	1940.91	65.99	75.35%
Divprune	✓	54.04	39.8	64.65	477	82.15	72.8	46	2015.66	79.27	75.61%
Visionzip	✓	48.29	42.84	55.94	404	82.6	73.09	45.44	1944.04	78.46	72.73%
Dynamic	×	62.34	58.76	69.68	554	76.20	77.36	43.78	2023.10	76.35	79.77%
Visionselector	×	77.01	61.6	**75.11**	654	**83.64**	84.72	46.56	**2146.97**	79.53	87.36%
AD-BTS (Ours)	×	**77.84**	**68.24**	75	**660**	82.6	**85.4**	**46.67**	2134.74	**80.42**	**88.43%**

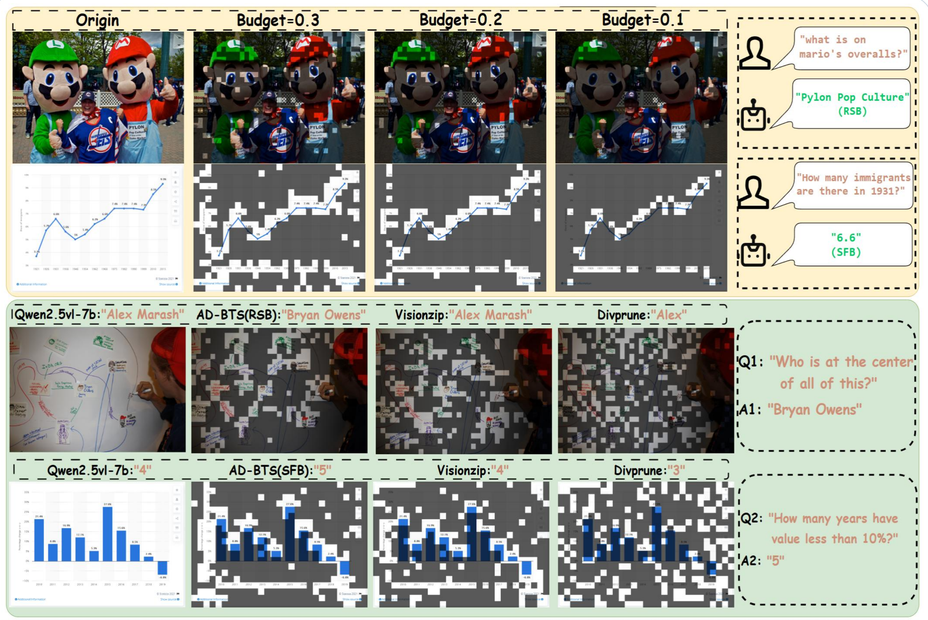


Figure 4. **Qualitative Visualization of Adaptive Token Retention. Top Panel (Budget Sensitivity):** (1) Under Spatial Redundancy (Natural Scene), the **RSB** eliminates background noise and selectively retains semantic regions guided by the query. (2) Under Spatial Sparsity (Chart), the **SFB** preserves the structural skeleton (e.g., trend lines, axes) via feature fusion, maintaining topology even at 10% retention. **Bottom Panel (Baseline Comparison):** Compared to VisionZip and DivPrune, AD-BTS avoids semantic distraction in natural images and prevents structural fragmentation in charts, yielding correct answers for both fine-grained reasoning and counting tasks.