Uncertainty-Guided Graph Formulation via MWIS for Token Pruning in LVLMs

Supplementary Material

Algorithm 1 Entropy-Adaptive MWIS Token Selection (Summary)

Input: $w^{(b)}$: token importance, $D^{(b)}$: pairwise distances, $H^{(b)}$: entropy, target K
Output: Selected token set $S^{(b)}$, with $|S^{(b)}| = K$
// 1. Compute entropy-based high-confidence rank bound
Estimate entropy statistics μ_H, σ_H from training data
$\tilde{H}^{(b)} \leftarrow (H^{(b)} - \mu_H)/\sigma_H$ $d_s^{(b)} \leftarrow K \cdot (1 - \sigma(\tilde{H}^{(b)}))$ // sigmoid $\sigma(\cdot)$
// 2. Determine distance thresholds and graph masks
τ_{sparse} τ_{dense} Construct sparse/dense masks $E_{\text{sparse}}, E_{\text{dense}}$
// 3. Rank tokens by importance
Sort $w^{(b)}$ in descending order; obtain ranks $r^{(b)}(i)$ for all i
// 4. Greedy MWIS-style token selection
$S^{(b)} \leftarrow \emptyset$, $U^{(b)} \leftarrow \{1, \ldots, N\}$ **while** $|S^{(b)}| < K$ ***and*** $U^{(b)} \neq \emptyset$ **do**
  $i^\star \leftarrow \arg\max_{i \in U^{(b)}} w_i^{(b)}$ $S^{(b)} \leftarrow S^{(b)} \cup \{i^\star\}$ **if** $r^{(b)}(i^\star) > d_s^{(b)}$ **then**
    $N^{(b)} \leftarrow \{j \mid E_{\text{dense}}(i^\star, j) = 1\}$ // uncertain → aggressive
  else
    $N^{(b)} \leftarrow \{j \mid E_{\text{sparse}}(i^\star, j) = 1\}$ // confident → conservative
  $U^{(b)} \leftarrow U^{(b)} \setminus (\{i^\star\} \cup N^{(b)})$
// 5. Diversity-based fill-in
if $|S^{(b)}| < K$ **then**
  Add tokens from $U^{(b)}$ maximizing $\min_{s \in S^{(b)}} D_{is}^{(b)}$ until $|S^{(b)}| = K$
return $S^{(b)}$

6. Implementation Details and Pseudocode

This section provides additional implementation details for the entropy-adaptive token selection module introduced in Section 3.3 of the main paper. The key components are: (1) a function F that determines the size of the *high-confidence region* based on the entropy of token importance, and (2) a function T that defines distance-based graph threshold using the *statistical scale* of the token-pair distance distribution. Both functions are derived from the empirical data distribution and do not rely on fixed heuristic thresholds.

High-confidence region size: function F. The entropy H of the token-importance distribution reflects how confidently the model distinguishes among visual tokens. Because the absolute entropy values vary across datasets and models, we compute the mean μ_H and standard deviation σ_H over the entire LLaVA training set and use these statistics as a reference scale. For each input, entropy is standardized as

$$\tilde{H} = \frac{H - \mu_H}{\sigma_H}.$$

We then define a scaling factor

$$s = 1 - \sigma(\tilde{H}),$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Lower entropy (higher confidence) yields larger s, expanding the high-confidence region; higher entropy results in a smaller region. Given a target token budget K, the upper rank boundary becomes

$$d_s = K \cdot s.$$

Distance-based graph thresholding: function T. The pairwise distance matrix D captures spatial and semantic redundancy among visual tokens. We first select the top-L visual tokens based on attention magnitude (L is the maximum token budget). From these L tokens, we compute all pairwise similarities and extract the top-10 largest values.

The sparse threshold τ_{sparse} is defined as the mean of these top-10 similarities:

$$\tau_{\text{sparse}} = \frac{1}{10} \sum_{i=1}^{10} s_{(i)},$$

where $s_{(i)}$ denotes the i-th largest similarity among the selected L tokens.

The dense threshold is obtained by scaling:

$$\tau_{\text{dense}} = 10 * \cdot \tau_{\text{sparse}}.$$

These thresholds form the sparse and dense graph masks $(E_{\text{sparse}}, E_{\text{dense}})$ used in our adaptive MWIS selection.

Summary and fallback selection. In summary, F determines the high-confidence region using the standardized entropy, and T defines graph connectivity based on the median-scaled distance distribution. Both components

Method	TextVQA	ChartQA	AI2D	OCRBench	MME (Perc.)	MME (Cogn.)	MMB-EN	MMB-CN	Rel.
Vanilla 1225 Tokens (100%)									
Qwen2.5-VL-7B	82.1	77.5	83.0	84.1	1691	642	82.8	83.2	100.0%
Retain 512 Tokens (↓ 60.5%)									
DivPrune(CVPR25)	79.6	72.0	81.9	78.5	1704	620	82.3	81.7	96.5%
CDPruner(NeurIPS25)	79.5	68.2	82.5	77.3	1710	594	81.8	80.9	95.4%
Ours	80.5	72.6	82.1	79.5	1709	616	82.1	81.3	**97.1%**

Table 7. **Performance comparison on Qwen2.5-VL-7B at 512 tokens.** *Rel.* is the relative performance normalized to the full-token Qwen2.5-VL-7B (set to 100%). MME scores are reported as separate Perception (Perc.) and Cognition (Cogn.).

MMMU Acc. (%)	Full	DivPrune	CDPruner	Ours
LLaVA-Critic-R1	43.21	42.11	42.35	**42.89**

Table 8. **Additional results on a GRPO-trained LVLM.** MMMU comparison on LLaVA-Critic-R1.

Benchmark	DivPrune	CDPruner	Ours
MVBench	44.18	44.12	**45.23**

Table 9. **Additional results on video understanding.** Comparison on MVBench.

adapt to the difficulty of each input and the structure of its token layout.

After constructing the adaptive graph, we execute a greedy MWIS-style selection. If fewer than K tokens remain due to graph suppression, we fill the remaining slots with tokens that maximize their minimum distance to the already selected set, ensuring diversity.

Algorithm 1 provides the simplified pseudocode for the entire procedure.

7. Additional Results on Advanced LVLMs and Video Benchmarks

7.1. Additional Results on Qwen2.5-VL (512 Tokens)

For completeness, we provide the extended results of Qwen2.5-VL-7B including the 512-token configuration, which was omitted from the main paper due to space constraints. Table 7 reports the performance of our method and prior pruning baselines (CDPruner, DivPrune) on TextVQA, ChartQA, AI2D, OCRBench, and MME. Our method achieves consistently higher accuracy than the baselines under the same token budget, demonstrating that the proposed approach remains effective across a wide range of pruning ratios and applies robustly to advanced VLM architectures such as Qwen2.5-VL-7B.

7.2. Additional Results on GRPO-trained and Video Benchmarks

To further examine the generalizability of the proposed method, we extend evaluation beyond the settings reported in the main paper. First, In addition to the Qwen2.5-VL results in Sec. 4.4, we evaluate a GRPO-trained model, LLaVA-Critic-R1, on the MMMU benchmark. As reported in Table 8, our method attains 42.89% accuracy with only 10% of the visual tokens, which remains comparable to the full-model accuracy of 43.21%. This indicates that the proposed pruning strategy transfers effectively even to reinforcement-aligned LVLMs. Second, we evaluate on the MVBench video benchmark. As shown in Table 9, our method achieves 45.23%, surpassing DivPrune and CDPruner. Taken together, these results suggest that the proposed uncertainty-guided MWIS pruning generalizes across both post-trained model variants and video understanding settings.

Method	MME	POPE	SQA^{IMG}	TextVQA	GQA
(a) DA (Top-k)	1390	86.3	68.86	55.11	57.99
(b) V-CLS score + DMWIS	1376	78.6	68.86	55.6	55.49
(c) V-Text score + DMWIS	1394	**87.5**	69.16	52.9	58.2
(d) **DA + DMWIS (proposed)**	**1432**	87.3	**69.26**	**55.7**	**58.43**

Table 10. **Ablation study on dual attention fusion and the proposed MWIS-based token selection.** (a) uses dual attention scores with a naive Top-k strategy, (b) and (c) use V-CLS and V-Text attention as node weights in MWIS, respectively, and (d) denotes our full method (dual attention + MWIS).

8. Additional Ablation Studies

8.1. Analysis of Node Weighting Strategies

Our MWIS framework depends critically on the quality of the node weights (w), which determine the importance of each visual–text token. Table 10 reports four node-weighting strategies: (a) dual-attention with a naïve Top-k, (b) V-CLS (global) attention, (c) V-Text (local) attention, and (d) our fused importance score.

Single-modality weighting metrics show complementary yet biased behavior across tasks:

- **V-Text score (c)** performs well on tasks requiring fine-grained visual–text grounding such as **GQA** and **POPE**, where local alignment cues are essential.
- **V-CLS score (b)** shows stronger results on **TextVQA**. Since TextVQA includes not only the question but also multiple OCR tokens, text–vision alignment becomes

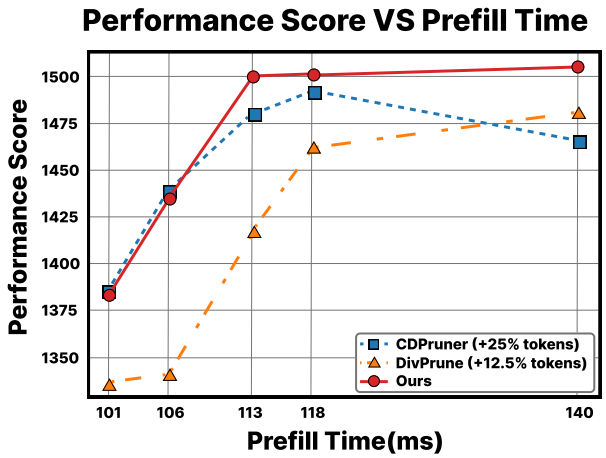

Figure 3. **Performance-time trade-off comparison.** To ensure a fair evaluation under identical time constraints, the retained token budgets of the baselines were deliberately increased (+25% for CDPruner and +12.5% for DivPrune) to match the prefill time of our method. Despite the baselines processing more tokens within the equivalent time budget, our method achieves a strictly superior efficiency curve and the highest peak performance.

ambiguous. In this scenario, the global visual context provided by V-CLS is more reliable than the local V-Text attention.

Using a single metric thus biases the MWIS optimization toward particular task types, reducing overall robustness. In contrast, the fused score (d) integrates both global and local attention cues and provides the **most stable performance across all benchmarks**, indicating that it yields more reliable node weights for MWIS.

8.2. Effect of the MWIS Formulation

To evaluate the contribution of the MWIS formulation itself, we compare Table 10 (a) the naïve Top-k method and (d) our full method, both using the *same fused importance score*. The difference lies solely in whether token similarity (graph edges) is incorporated during selection.

While Top-k (a) improves performance over single-weight baselines, it considers only the importance of each token and cannot address redundancy among highly activated tokens. Our method (d), which optimizes token selection under the MWIS objective, yields clear gains on benchmarks such as **MME** (1390 $\rightarrow$ 1432) and **SQA**^{IMG} (68.86 $\rightarrow$ 69.26). These results demonstrate that MWIS provides a substantial advantage by jointly modeling both **importance** and **diversity**, enabling more effective removal of redundant tokens and leading to consistently stronger performance across tasks.

9. Additional Efficiency Analysis

To rigorously evaluate the performance-efficiency trade-off, we compare the benchmark performance of our method against the baselines under identical time constraints. For a fair comparison, we intentionally increased the retained token budgets of the baseline methods—specifically by 25% for CDPruner and 12.5% for DivPrune—so that their prefill times exactly match ours at each operating point. As illustrated in Fig. 3, our method demonstrates a highly superior trade-off curve. Even though the baselines are granted the advantage of processing a larger number of tokens within the same time budget, our approach consistently yields more robust and higher overall performance. While CDPruner shows highly competitive scores at the most constrained lower time bounds (101–106 ms), our method exhibits much steeper performance scaling as the allowable time increases ($\geq$ 113 ms). Furthermore, at higher time budgets, CDPruner suffers from performance degradation and DivPrune plateaus, whereas our method effectively translates the available computational time into the highest peak performance. This confirms that our approach utilizes the given time budget significantly more effectively than merely increasing the token retention ratio.

10. MWIS Consistency under Diversity-Based Fill-in

To examine whether the diversity-based fill-in step affects the consistency of the MWIS formulation, we analyze the proportion of tokens added after the greedy MWIS-style selection. In our method, fill-in is applied only when graph suppression leaves fewer than the target budget K tokens. These additional tokens are selected to maximize their minimum distance to the already chosen subset, thereby preserving diversity. Although this procedure does not strictly satisfy the MWIS constraint, its impact is minimal in practice: only about 4% of the final selected tokens are introduced by the fill-in step. Hence, the final token subset remains predominantly determined by MWIS-based pruning, confirming that the proposed formulation remains the principal driver of the selection behavior.

11. Robustness under Extreme Pruning Ratios

We further study the behavior of our method under extreme pruning settings by progressively increasing the pruning ratio beyond 94%. As shown in Fig. 4, when only a very small fraction of visual tokens is retained, the overall performance degrades gracefully rather than abruptly. In particular, most benchmarks maintain relatively high normalized scores up to 96–97% pruning, while a clearer drop appears only at 99%. These trends suggest that the proposed uncertainty-guided MWIS pruning remains effective even in highly aggressive compression regimes, and that the selected token

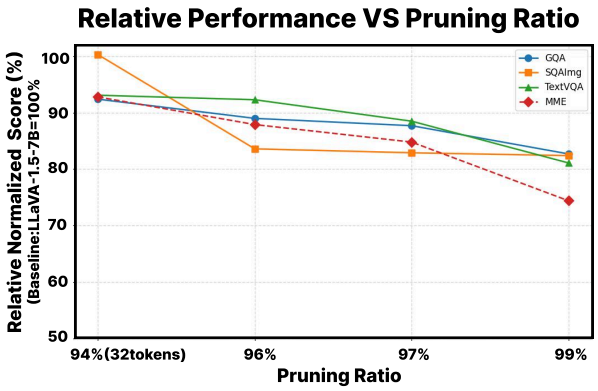

Figure 4. **Robustness under extreme pruning ratios.** As the pruning ratio increases beyond 94%, the performance degrades gracefully rather than collapsing abruptly.

subset continues to preserve informative visual content despite the severely reduced token budget. Taken together, these results indicate that our method degrades gracefully as the pruning ratio increases, while still maintaining strong performance under very limited token budgets. This demonstrates the robustness of the proposed approach in highly constrained inference scenarios.

12. Discussion on Finetuning and GNN Extensions

Both finetuning LVLMs for pruned inputs and incorporating GNN-based processing over the constructed token graphs are feasible extensions of the proposed framework. Finetuning may help the model better adapt to reduced visual token budgets, while GNN-based message passing could provide a more expressive way to exploit token relationships beyond the current graph formulation. However, the focus of the present work is a training-free, plug-and-play pruning framework that improves efficiency without introducing additional optimization or architectural complexity. We therefore leave these directions as future extensions.