CoIn: Coverage and Informativeness-Guided Token Reduction for Efficient Large Multimodal Models

Supplementary Material

6. Fast Inference Algorithm for CoIn

6.1. Near-optimality Guarantee

Directly optimizing this objective is combinatorially hard, but it is submodular and monotone, which guarantees that a simple greedy procedure can efficiently produce a near-optimal solution. Specifically, if S_{greedy} denotes the subset selected by the greedy algorithm and S^* denotes the global optimum, the following bound holds [29]:

$$f(S_{\text{greedy}}) \geq (1 - 1/e) \cdot f(S^*), \tag{8}$$

where $f(S) = (1-\alpha)\sum_{i \in S} s_{\text{info},i} + \alpha \log\det\left(\mathbf{F}_S^\top \mathbf{F}_S\right)$.

6.2. Algorithm Design

The challenge of selecting the optimal subset of tokens is an NP-hard problem. Therefore, we use a fast greedy algorithm to efficiently find an optimal solution. The core of this acceleration lies in an incremental update method for the QR decomposition in each selection round. As detailed in Algorithm 1, we select k tokens iteratively and in each round, we obtain the token that maximizes a hybrid score, which balances coverage and informativeness.

Algorithm 1 Fast Greedy Selection Algorithm

Require: Token features $\mathbf{X} \in \mathbb{R}^{N \times D}$, informativeness scores $\mathbf{p} \in \mathbb{R}^N$, target size k, trade-off Λ.
Ensure: Selected token indices $\mathcal{S}$.

Normalize each row of $\mathbf{X}$ to unit norm
$\mathcal{S} \leftarrow \emptyset, \mathcal{U} \leftarrow \{1, \ldots, N\}$
$\mathbf{C} \leftarrow \mathbf{0}^{N \times k}$ {Projection coefficients cache}
$\mathbf{Q} \leftarrow [], s \leftarrow 0$
$j \leftarrow \arg\max_{j \in \mathcal{U}} p_j$
for $i = 1$ to k **do**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{j\}, \mathcal{U} \leftarrow \mathcal{U} \setminus \{j\}$
  $\mathbf{x}_{new} \leftarrow \mathbf{X}_j$
  $\mathbf{q}_{new} \leftarrow \text{GramSchmidt}(\mathbf{x}_{new}, \mathbf{Q})$
  $\mathbf{Q} \leftarrow [\mathbf{Q} \mid \mathbf{q}_{new}]$
  $s \leftarrow s + 1$
  $\mathbf{C}_{:,s} \leftarrow \mathbf{X} \cdot \mathbf{q}_{new}$
  if $i < k$ **then**
    $\mathbf{d}_j^2 \leftarrow 1 - \sum_{m=1}^{s} \mathbf{C}_{j,m}^2$, for $j \in \mathcal{U}$
    $\mathbf{S}_j \leftarrow \Lambda \cdot \mathbf{d}_j + (1 - \Lambda) \cdot \mathbf{p}_j$, for $j \in \mathcal{U}$
    $j \leftarrow \arg\max_{j \in \mathcal{U}} \mathbf{S}_j$
  end if
end for
return $\mathcal{S}$

7. Detailed Experiment Settings

7.1. Benchmarks

GQA. A large-scale visual question answering benchmark designed to evaluate compositional reasoning and visual understanding. It provides detailed question-answer pairs covering objects, attributes, and relationships. We follow the standard test-dev balanced split for evaluation.

MMBench. A comprehensive benchmark designed to evaluate the multi-modal understanding capabilities of large language models. It consists of a diverse set of multiple-choice questions that cover a wide range of tasks, from basic perception and object recognition to complex cognitive reasoning and world knowledge.

MME. A comprehensive benchmark measures a model's performance across 14 distinct subtasks, which are divided into two main categories: perception and cognition. Perception tasks test a model's ability to recognize and understand basic visual elements like objects, text, or a scene's context. In contrast, cognition tasks evaluate its higher-level reasoning skills, such as applying common sense, performing logical inference, or solving math and science problems based on an image. The MME benchmark is designed to provide a detailed and fair comparison of MLLMs by using carefully designed instruction-answer pairs and covering a wide range of domains to identify a model's strengths and weaknesses.

POPE. A benchmark designed to rigorously assess object hallucination in large vision-language models. It systematically presents a model with a series of "Yes or No" questions about the existence of specific objects in an image. By strategically sampling objects that are not present, POPE effectively measures a model's tendency to incorrectly confirm or generate objects, providing a robust method for evaluating a model's honesty and accuracy.

TextVQA. A benchmark dataset for Visual Question Answering (VQA) that requires models to answer questions by reading and reasoning about text within images. To succeed, a model must first perform accurate Optical Character Recognition (OCR) to extract the text and then combine this information with the visual context of the image. This

makes TextVQA a crucial test for a model's ability to integrate visual perception with linguistic understanding.

VizWiz. A unique Visual Question Answering (VQA) dataset that focuses on questions asked by people who are visually impaired. Its images are often of poor quality, including blurriness, suboptimal framing, or occlusions, because they were captured by users seeking assistance. The questions are also grounded in real-world needs and are often much more open-ended. The purpose of the VizWiz benchmark is to push the development of VQA models that are robust to real-world visual imperfections and can provide practical, useful information to assist people with vision loss.

OCRBench. A comprehensive evaluation benchmark designed to assess the optical character recognition (OCR) capabilities of large multi-modal models. It tests a model's ability to handle a wide variety of tasks, including text localization, understanding handwritten content, and performing logical reasoning based on the text found in an image. The benchmark includes diverse scenarios such as receipts, documents, and street scenes to provide a robust evaluation of a model's visual and linguistic understanding in real-world, text-rich environments.

ScienceQA. A large-scale benchmark designed to evaluate multi-modal reasoning by presenting models with complex science questions. It includes not only images and text but also detailed rationales—step-by-step explanations for the correct answers. This feature allows researchers to assess a model's underlying reasoning process, rather than just its final output. The questions cover a diverse range of science topics from elementary to high school levels.

RealWorldQA. A benchmark designed to evaluate a multi-modal model's real-world spatial understanding and common sense reasoning. The dataset consists of high-resolution images, often captured from vehicles or other real-world scenarios, each paired with a question and a verifiable answer. Unlike many other benchmarks, RealWorldQA focuses on challenging models to recognize subtle details and perform complex reasoning based on their visual perception. This allows for a robust assessment of a model's ability to comprehend our physical world and act as a practical assistant.

AI2D. A widely used benchmark for evaluating a model's ability to understand scientific diagrams and visually grounded explanations. The dataset contains thousands of annotated diagrams, each accompanied by multiple-choice questions that probe a model's capacity for diagram parsing, structural understanding, and scientific reasoning. Unlike datasets focused on natural images, AI2D requires models to interpret abstract visual elements—such as arrows, labels, and schematic structures—and connect them to underlying scientific concepts. This makes it an important benchmark for assessing a model's fine-grained visual comprehension and domain-specific reasoning abilities.

MVBench. A comprehensive benchmark for evaluating multimodal video understanding. MVBench consists of over 20 sub-tasks covering a wide range of perceptual and reasoning skills — from temporal localization and motion comprehension to high-level event reasoning. It adopts a static-to-dynamic design philosophy, converting conventional image-based multimodal tasks into temporally grounded video tasks. MVBench thus provides a holistic view of how well multimodal large language models (MLLMs) can capture temporal dependencies and integrate visual and textual information across frames.

VSIBench. A benchmark focuses on evaluating visual–spatial reasoning in multimodal large language models. It includes hundreds of real indoor scene videos and thousands of corresponding question–answer pairs that require understanding object relations, spatial layouts, and viewpoint changes. Rather than assessing simple recognition, VSI-Bench probes whether models can "think in space", remembering object locations, reasoning about occlusions, and interpreting 3D spatial arrangements from video input. It provides a fine-grained assessment of models' spatial memory and reasoning abilities in dynamic environments.

VideoMME. A benchmark provides the first large-scale, unified framework for evaluating MLLMs on diverse video understanding tasks. Each video is accompanied by multimodal information including subtitles and audio, enabling comprehensive evaluation across visual, auditory, and linguistic modalities. Video-MME is particularly suitable for assessing models' ability to process long-form and multimodal video inputs, making it an ideal benchmark for testing efficiency-oriented methods such as feature caching, temporal skipping, and token reduction strategies.

7.2. Models

LLaVA-1.5. LLaVA-1.5 is a vision-language model built upon the Vicuna language backbone and CLIP visual encoder, designed to align visual representations with large language models through multimodal instruction tuning. Compared with the original LLaVA, version 1.5 uses an MLP to connect the two modalities, expands the training corpus and adopts improved visual instruction data, significantly enhancing the model's visual grounding and

question-answering capability. It supports image with 336×336 resolution.

LLaVA-NeXT. LLaVA-NeXT represents a significant evolution of the LLaVA series, achieving notable improvements in reasoning, OCR, and world knowledge. Compared to LLaVA-1.5, it supports higher-resolution visual inputs (up to 1344×336 or 672×672), providing four times more pixels and better visual detail comprehension. The model also enhances visual conversation quality across diverse scenarios, enabling more coherent multi-turn dialogues and stronger logical reasoning capabilities grounded in real-world knowledge.

LLaVA-OneVision. LLaVA-OneVision unifies image, video, and document understanding under a single vision-language interface. It leverages a stronger visual encoder and fine-grained multi-resolution tokenization to process inputs of arbitrary modalities and aspect ratios efficiently. Through large-scale multimodal alignment, OneVision achieves state-of-the-art performance across a wide spectrum of benchmarks, including video reasoning, spatial understanding, and image QA. It demonstrates the feasibility of a truly universal multimodal model capable of robust perception and reasoning across diverse real-world inputs.

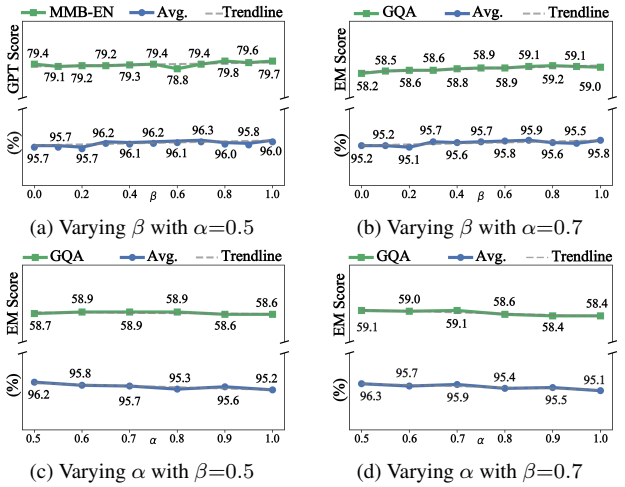
Qwen2.5-VL. Qwen2.5-VL features substantial upgrades in perception, reasoning, and multilingual understanding. Built upon a stronger Qwen2.5 language backbone and an improved ViT-based visual encoder, the model supports high-resolution inputs and dense visual tokenization, enabling fine-grained grounding and detailed scene comprehension. By incorporating large-scale instruction tuning and diverse multimodal corpora, Qwen2.5-VL achieves robust performance on various tasks. The model also exhibits strong generalization across real-world scenarios, making it a competitive and versatile foundation for downstream multimodal applications.

8. More Results for Hyperparameter Setting

We provide additional results under various hyperparameter configurations to complement Section 4.6. As shown in Figure 7, these experiments, conducted on Qwen2.5-VL-7B with 256 retained tokens, further demonstrate the robustness of our method. The extended evaluations confirm that CoIn maintains stable performance across a broad range of parameter settings.

9. Additional Experimental Results

Performance on LLaVA-1.5-7B. When retaining 128 tokens, our method achieves **96.7%** of the original full-token



(a) Varying β with α=0.5 (b) Varying β with α=0.7

(c) Varying α with β=0.5 (d) Varying α with β=0.7

Figure 7. Performance under different hyperparameter settings of α and β. Avg. denotes the average performance over 7 benchmarks in the Qwen experiment.

performance, surpassing the best-performing baseline (DivPrune). As the retention drops to 64, the gap between methods becomes more pronounced. PDrop suffers noticeable degradation, losing over **20%** of their original score. In contrast, our method maintains a robust **94.5%** of the original performance, outperforming VisionZip and DivPrune by **1.7%** and **1.4%**, respectively. This confirms the advantage of jointly considering both token importance and diversity during selection. In the extreme case of only 32 tokens retained (*i.e.*, 94.4% pruned), most baselines experience a drastic performance collapse due to the loss of essential semantic information and excessive redundancy in retained tokens. Our method, however, still preserves **91.3%** of the original score, significantly outperforming DivPrune by **2.4%**. We observe particularly strong results on hallucination-sensitive tasks such as POPE and real-world QA datasets, indicating our method's ability to retain critical visual cues under aggressive pruning.

Results on LLaVA-NeXT-13B. Table 9 reports the performance under two token budgets. When retaining 640 tokens (77.8% reduction), CoIn achieves 92.6% of the original model's accuracy, outperforming PDrop and DivPrune across most benchmarks. Under the more aggressive 320-token setting (88.9% reduction), CoIn remains notably robust, reaching 88.4% average performance and surpassing prior methods by a clear margin. These results demonstrate that jointly modeling informativeness and diversity enables CoIn to preserve essential visual evidence even under extreme compression, consistently outperforming existing pruning strategies.

Table 8. **Performance on LLaVA-1.5-7B**. "Avg." indicates average performance relative to original model across 9 benchmarks.

Method	GQA	MMB-EN	MME	POPE	VQA^{Text}	VizWiz	OCRBench	SQA^{IMG}	RealWorldQA	Avg.
Upper Bound, 576 Tokens										
LLaVA-1.5-7B	62.0	64.1	1508	85.9	46.1	54.3	31.3	69.5	55.8	100%
Retained 128 tokens (↓**77.8%**)										
PDrop (CVPR25)	57.1	61.7	1445	77.4	43.9	53.7	29.1	69.0	51.1	94.7%
PruMerge (ICCV25)	57.6	60.1	1381	81.0	39.2	56.0	27.9	69.5	49.9	93.3%
VisionZip (CVPR25)	57.6	62.2	1445	82.9	43.6	54.1	29.8	68.6	51.9	95.9%
DivPrune (CVPR25)	59.2	62.3	1403	86.6	42.0	56.4	28.8	68.6	49.7	95.7%
CoIn (α=0.9, β=0.6)	59.3	62.4	1406	87.3	43.1	56.0	30.0	69.2	50.7	**96.7%**
Retained 64 tokens (↓**88.9%**)										
PDrop (CVPR25)	46.3	48.1	984	41.3	39.6	50.4	27.0	68.7	49.2	79.4%
PruMerge (ICCV25)	55.1	58.7	1295	75.5	37.7	56.7	26.6	69.5	48.2	90.2%
VisionZip (CVPR25)	55.2	60.1	1373	77.0	42.0	54.7	28.0	68.9	50.9	92.8%
DivPrune (CVPR25)	57.6	59.5	1348	85.8	39.1	57.5	27.1	68.0	49.2	93.1%
CoIn (α=0.9, β=0.7)	57.8	59.8	1378	86.2	41.0	57.6	28.1	68.2	49.8	**94.5%**
Retained 32 tokens (↓**94.4%**)										
PruMerge (ICCV25)	52.6	55.0	1201	70.4	33.3	56.7	23.7	68.8	45.1	84.9%
VisionZip (CVPR25)	51.7	57.0	1249	68.8	36.9	55.3	25.1	68.2	48.1	86.9%
DivPrune (CVPR25)	54.6	57.6	1268	81.2	34.9	56.8	25.3	67.6	47.2	88.9%
CoIn (α=0.9, β=0.8)	55.7	58.3	1326	84.0	37.4	57.5	25.8	69.0	48.2	**91.3%**

Table 9. **Performance on LLaVA-NeXT-13B**. "Avg." indicates average performance relative to original model across 9 benchmarks.

Method	GQA	MMB-EN	MME	POPE	VQA^{Text}	VizWiz	OCRBench	SQA^{IMG}	RealWorldQA	Avg.
Upper Bound, 2880 Tokens										
LLaVA-NeXT-13B	65.4	69.1	1575	86.3	67.0	63.3	550	73.6	59.0	100%
Retained 640 tokens (↓**77.8%**)										
PDrop (CVPR25)	62.1	66.1	1535	85.1	57.2	61.7	359	71.6	52.4	91.2%
DivPrune (CVPR25)	62.9	66.9	1510	86.2	56.1	61.5	356	72.0	52.9	91.3%
CoIn (α=0.9, β=0.4)	63.3	67.1	1580	85.8	56.3	62.1	362	72.7	54.5	92.6%
Retained 320 tokens (↓**88.9%**)										
PDrop (CVPR25)	57.8	62.4	1389	78.6	48.9	57.2	298	71.7	50.6	84.3%
DivPrune (CVPR25)	61.3	65.1	1465	84.1	50.1	59.4	310	72.0	51.4	87.6%
CoIn (α=0.5, β=0.6)	61.5	65.8	1486	84.2	51.0	59.6	318	72.2	51.7	**88.4%**