LOOKING BEYOND TOKENS: CONTEXT-AWARE VISUAL TOKEN COMPRESSION FOR MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

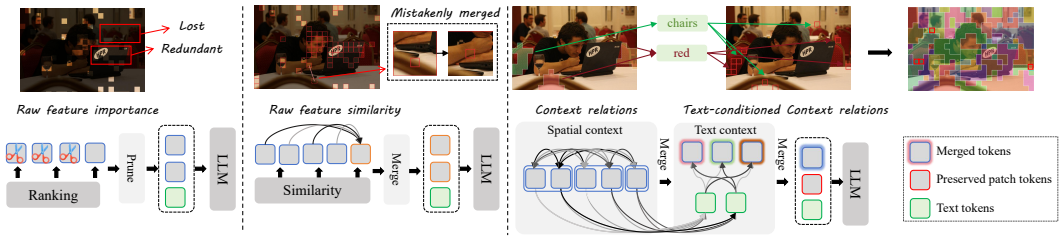
Multimodal large language models (MLLMs) suffer from heavy computational and memory overheads due to the excessive number of visual patch tokens. Existing visual token compression approaches typically rely on raw token importance scores or raw feature similarities, leading to redundant budget allocation or invalid merging across distinct objects. In this paper, we argue that reliable visual token compression should be guided by multimodal context, where token relations are jointly characterized by spatial structure and text-conditioned semantics rather than isolated visual features. Motivated by this insight, we propose Context-aware Visual Token Compression (CVTC), a novel training-free compression paradigm for MLLMs. CVTC employs a cascaded context coarsening mechanism to construct a multimodal context graph, progressively integrating spatial topology and text-conditioned relevance to guide context-aware token merging. Based on the propagated multimodal context, CVTC performs greedy contextual merging to identify tokens sharing coherent contextual relations, while aggregating their original visual features to construct the compressed representation. Extensively evaluated on various benchmarks across different backbones, CVTC achieves a favorable balance among performance, efficiency, and robustness. The codes are available in supplementary materials will be publicly available.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have demonstrated remarkable progress across a wide range of vision-language tasks, including visual question answering, visual reasoning, and visual understanding (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Bai et al., 2023; Liu et al., 2023; 2024a; Chen et al., 2024b). MLLMs typically decompose visual inputs into long sequences of patch-level tokens, which are jointly processed with textual tokens by the language model. However, as visual resolution and input complexity increase, these visual sequences can dominate the multimodal context, substantially increasing the computational and memory costs of inference. Efficient visual token compression therefore requires not only reducing sequence length, but also distinguishing redundant visual information from complementary evidence essential for multimodal reasoning.

Existing visual token compression methods can be roughly grouped into pruning-based and merging-based approaches. Early pruning-based methods design various importance scoring strategies to select salient tokens while discarding less important ones (Chen et al., 2024a; Jiang et al., 2025; Xing et al., 2025; Zhang et al., 2025; Zou et al., 2025; Alvar et al., 2025). However, relying solely on importance metrics inevitably leads to severe information loss from the discarded tokens; furthermore, highly ranked tokens may still carry overlapping or redundant visual information (e.g., Figure 1(a)). To mitigate information loss, recent merging-based methods aggregate tokens based on pairwise feature similarity and have made promising progress in visual token compression (Bolya et al., 2023; Shang et al., 2025). Nevertheless, merging strategies determine which tokens to combine primarily based on raw feature similarity, causing distinct features from different objects or semantic regions to be incorrectly merged together (e.g., the cross-object mismatch in Figure 1(b)). Although recent methods have attempted to combine these two paradigms (Yang et al., 2025; Yu et al., 2026), they still struggle to balance redundancy removal with the preservation of local details due to the above problems. These limitations reveal a common issue in existing compression paradigms: token-wise scoring and pairwise matching provide only partial evidence for determining visual redundancy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Pruning-based method (b) Merging-based method (c) Context-aware visual token compression (ours)

Figure 1: Comparison of different visual token compression paradigms under a token budget of 64. Best viewed when zoomed in. (a) Pruning based on raw feature importance may retain redundant tokens (e.g., the laptop) while discarding other evidence with low importance scores (e.g., the scene), resulting in inefficient allocation of the limited token budget. (b) The red solid boxes denote the cluster centers, and the red dashed box denotes the patch to be merged. Merging by raw feature similarity may mistakenly merge tokens from different objects. (c) Our context-aware visual token compression mechanism shifts visual token compression from isolated token scoring or pairwise matching toward multimodal context-aware relation modeling.

In this work, we argue that visual redundancy is inherently context-dependent and should be assessed in relation to the broader context in which each token resides. Such context is naturally available in MLLMs, encompassing both the structural relations among visual tokens and their relevance to the textual query. The former helps distinguish visually similar tokens arising from different semantic structures, while the latter identifies visual evidence that is particularly relevant to the current task. Therefore, considering multimodal context effectively helps mitigate the trade-off between redundancy removal and local-detail preservation (as presented in Figure 1(c)).

Following this principle, we propose **C**ontext-aware **V**isual **T**oken **C**ompression (CVTC), a generic training-free visual token compressor that makes compression decisions based on multimodal contextual relations. CVTC employs a cascaded context coarsening mechanism that progressively constructs a multimodal context graph, first capturing the spatial organization of visual tokens and then incorporating text-conditioned relevance to the query. Based on the resulting contextual states, CVTC performs greedy contextual matching to determine which tokens to merge, while aggregating their original visual features to construct the compressed representation. Our contributions can be summarized as follows:

- We introduce a novel mechanism for visual token compression that uses multimodal context as the evidence for compression decisions. In this design, merge costs are determined by contextual states that encode the spatial and semantic context of visual tokens, rather than solely by isolated-token importance scores or pairwise feature similarity.
- We introduce CVTC, a generic, training-free, context-aware visual token compressor, which employs the cascaded multimodal context graph to guide visual token merging.
- Extensive experiments on multiple benchmarks demonstrate that CVTC substantially reduces the visual-token sequence length and inference cost while maintaining accuracy competitive with strong existing methods.

2 Related Work

Merging and pruning in visual token compression. Visual token compression mainly follows two routes: pruning and merging. Pruning-based methods rank and select tokens using visual-feature signals (HoloV (Zou et al., 2025) and SpecFlow (Li et al., 2026)), language-model attention (FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2025), and PyramidDrop (Xing et al., 2025)), or diversity-aware criteria (Alvar et al., 2025; Fang et al., 2026; Dong et al., 2026). However, these methods can lose information by discarding tokens, and importance alone does not directly measure redundancy. To mitigate information loss, merging-based methods (such as ToMe (Bolya et al., 2023) and PuMer (Cao et al., 2023)) combine visually similar token pairs. Recently, several works have attempted to combine these two paradigms—selecting important tokens while merging

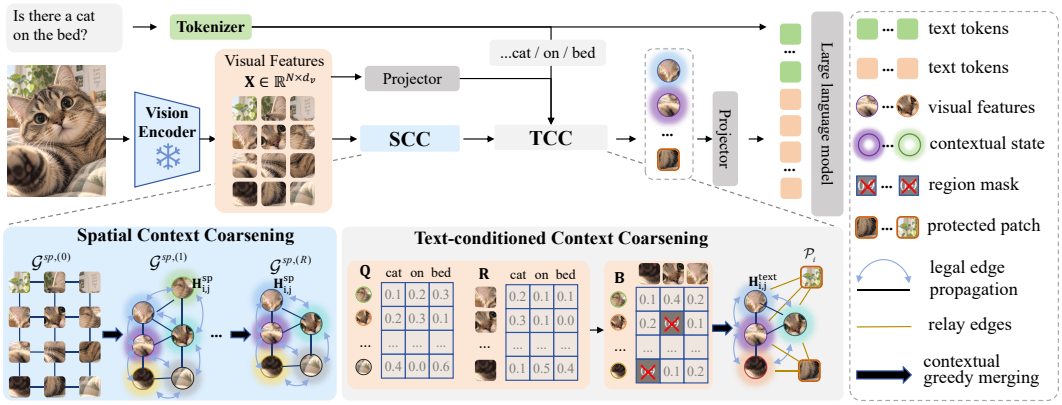


Figure 2: The core idea of CVTC is that the merge decision of a patch is no longer determined only by its appearance, but by the neighborhood content it represents after graph propagation.

the discarded ones (e.g., LLaVA-PruMerge (Shang et al., 2025) and VisionTrim (Yu et al., 2026)). Despite their effectiveness, existing approaches make compression decisions based on isolated raw features or simple pairwise similarity, lacking sufficient contextual awareness to prevent cross-object mismatches or handle the trade-off between redundancy removal and detail preservation.

Context in vision-language models. Context is fundamental to vision-language understanding, as the meaning of a visual token depends on both its position in the spatial topology and the language-conditioned guidance. Vision-language models capture such dependencies through cross-modal attention or query-based visual representations (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023). Recently, Li et al. (2026) found that ViT attention is often not a reliable measure of semantic affinity, as it may connect tokens across inconsistent regions. Moreover, Zhang et al. (2026) found that text-visual attention patterns suffer from attention gravity and attention sink, which undermines the model's perception of spatial context. Consequently, relying solely on implicit cross-modal attention often falls short in preserving the holistic structure and fine-grained spatial-semantic alignments. This motivates explicit modeling of cross-modal context.

3 Method

3.1 Overview

For ease of reading, Appendix A.1 provides a summary of the notations used throughout this section. An overview of CVTC is presented in Figure 2, which builds the merge decision of each patch on its neighborhood content, rather than its isolated appearance. Given an image-question pair, the vision encoder produces a sequence of patch features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d_v}$. CVTC aims to reduce N patch features to a budget of K visual tokens while preserving both the global spatial structure of the image and the fine-grained, task-critical evidence demanded by the question. To achieve this, the entire compression pipeline is structured into two cascaded stages: the spatial context coarsening (SCC) stage and the text-conditioned context coarsening (TCC) stage.

First, in the SCC stage, we consider each patch as an initial node, construct a spatial graph $\mathcal{G}^{sp}$, and propagate contextual states to explore spatial context. This stage aggregates spatially adjacent and contextually similar patches into cluster tokens, avoiding the risk of erroneous merging caused by direct global similarity matching, and forms a compact image-level spatial representation:

$$\mathbf{Y}_{\text{sp}} = \mathcal{F}_{\text{scc}}(\mathbf{X} \mid \mathcal{G}^{\text{sp}}), \quad |\mathbf{Y}_{\text{sp}}| = K_c, \tag{1}$$

where $\mathcal{F}_{\text{scc}}$ models the spatial context coarsening process, using a greedy strategy to reduce N patches into K_c clusters that faithfully preserve the regional appearance and structural layout of the image.

To further use textual context to enhance task-specific local evidence while preserving the established spatial layout, the TCC stage subsequently uses the textual query as a context bridge to iden-

tify K_p original patches that provide supplementary evidence. Specifically, TCC incorporates these selected elements into a hybrid patch-cluster graph, where they influence regional context propagation but are strictly excluded from legal merging:

$$\mathcal{P} = \mathcal{F}_{\text{tcc}}(\mathbf{X}, \mathbf{Y}_{\text{sp}} \mid \text{Text}), \quad |\mathcal{P}| = K_p, \tag{2}$$

where $\mathcal{F}_{\text{tcc}}$ represents the text-conditioned selection mechanism driven by the textual query and spatial clusters. It complements the spatial clusters with question-relevant details to capture fine-grained task-critical evidence. We further merge the $K_c + K_p$ nodes to K to meet the total budget. Ultimately, these K nodes are projected and fed into the language model for subsequent multimodal reasoning.

3.2 Spatial context coarsening

Spatial graph and context propagation. The SCC module constructs a spatial graph to model local image topology before compression, where the graph structure defines both the neighborhoods for contextual propagation and the valid edges for token merging.

Formally, given the visual features $\mathbf{X}$, we construct an undirected spatial graph $\mathcal{G}^{\text{sp}} = (\mathcal{V}^{\text{sp}}, \mathcal{E}^{\text{sp}}, \mathbf{F}^{\text{sp}}, \mathbf{W}^{\text{sp}})$ to capture the local topological layout of the image. Here, $\mathcal{V}^{\text{sp}} = \{1, \ldots, N\}$ represents the original visual patches, with node features initialized as $\mathbf{F}^{\text{sp}} = \mathbf{X}$. The edge set $\mathcal{E}^{\text{sp}}$ contains only edges connecting 8-neighboring patches on the image grid. For any two spatially adjacent nodes i and j, the edge weight is modulated by the cosine similarity of their visual features: $W_{ij}^{\text{sp}} = \exp\left(\cos(\mathbf{x}_i, \mathbf{x}_j)\right)$. To transform isolated visual features into neighborhood-aware contextual states, we apply a random-walk operator:

$$\mathbf{P}^{\text{sp}} = \alpha \mathbf{I} + (1 - \alpha)\, \text{RowNorm}(\mathbf{W}^{\text{sp}}), \tag{3}$$

where the transition matrix $\mathbf{P}^{\text{sp}}$ describes how each patch absorbs contextual cues from its 8-neighborhood, with α controlling the degree to which its original representation is retained.

Then, we update the node states through ℓ_1 steps of graph propagation to aggregate regional context, enriching patch representations:

$$\mathbf{H}^{\text{sp}} = (\mathbf{P}^{\text{sp}})^{\ell_1} \mathbf{F}^{\text{sp}}. \tag{4}$$

$\mathbf{H}^{\text{sp}}$ denotes the spatially contextualized token states, where each patch representation incorporates information propagated from its multi-hop neighborhood rather than describing the patch in isolation (visualized in Appendix A.6).

Contextual greedy merging. The contextual states $\mathbf{H}^{\text{sp}}$ characterize each patch with respect to its surrounding spatial structure, providing a contextual criterion for identifying spatially redundant regions. Based on these states, we design a contextual greedy merging strategy through iterative graph contraction. With the initial spatial context graph $\mathcal{G}^{\text{sp},(0)} = (\mathcal{V}^{(0)}, \mathcal{E}^{(0)}, \mathbf{F}^{(0)}, \mathbf{W}^{(0)}, \mathbf{H}^{\text{sp},(0)})$, SCC performs token reduction through iterative graph contraction:

$$\mathcal{G}^{\text{sp},(r)} = \text{Contract}\left(\mathcal{G}^{\text{sp},(r-1)}, \mathcal{S}^{(r-1)}\right), \quad r = 1, \ldots, R, \quad \text{subject to} \quad |\mathcal{V}^{(R)}| = K_c, \tag{5}$$

where we use the contextual states $\mathbf{H}^{\text{sp}}$ to assess whether neighboring regions exhibit consistent spatial context, rather than determining merges directly from pairwise raw-feature similarity, while aggregating their original visual features to construct the compressed representation. In this way, contextual states determine which tokens to merge, whereas the original features determine what visual information is preserved after merging.

Specifically, at any given round r, for any cluster $u \in \mathcal{V}^{(r)}$, its mass, content feature, and contextual state are dynamically updated from its constituent original patches $\mathcal{R}_u$ as:

$$m_u^{(r)} = \sum_{i \in \mathcal{R}_u} \mu_i^{\text{sp}}, \quad \mathbf{f}_u^{(r)} = \frac{\sum_{i \in \mathcal{R}_u} \mu_i^{\text{sp}} \mathbf{F}_i^{\text{sp}}}{m_u^{(r)}}, \quad \mathbf{h}_u^{(r)} = \frac{\sum_{i \in \mathcal{R}_u} \mu_i^{\text{sp}} \mathbf{H}_i^{\text{sp}}}{m_u^{(r)}}, \tag{6}$$

where $\mu_i^{\text{sp}} = \sum_j W_{ij}^{\text{sp}} / \sum_{u,v} W_{uv}^{\text{sp}}$ is the mass weight for subsequent merging. The merging cost between any two connected neighboring clusters u and v in $\mathcal{E}^{(r)}$ is defined by their contextual dissimilarity: $c(u, v) = 1 - \cos(\mathbf{h}_u^{(r)}, \mathbf{h}_v^{(r)})$. Finally, we iteratively contract $\mathcal{G}^{\text{sp},(r-1)}$ into $\mathcal{G}^{\text{sp},(r)}$ by

Algorithm 1 Contextual Greedy Merging via Iterative Graph Contraction

Require: Initial context graph $\mathcal{G}^{(0)}$ (spatial or hybrid), target budget K_{target}
Ensure: Merged visual tokens or updated graph

```
Set r ← 0
while |V^(r)| > K_target do
    for all cluster u ∈ V^(r) do
        Find the best partner: ν(u) ← arg min_{v:(u,v)∈E^(r)} c(u, v)
    end for
    Collect mutual nearest-neighbor pairs M^(r)
    b_r ← min(|V^(r)| − K_target, ⌈√|V^(r)|⌉)
    S^(r) ← up to b_r lowest-cost pairs from M^(r)
    Contract graph: G^(r) → G^(r+1) by merging legal pairs in S^(r) in parallel and updating E^(r+1)
    Update cluster properties (f_u^(r+1), h_u^(r+1), m_u^(r+1)) via Eq. (6)
    r ← r + 1
end while
return Merged features Y = {f_u^(r) | u ∈ V^(r)} (along with protected patches if in TCC)
```

merging mutually matched adjacent clusters, updating their topological connections, and recomputing their properties until the target budget K_c is met, as presented in Algorithm 1. Upon termination, the remaining nodes in $\mathcal{V}^{(R)}$ constitute the spatially coarsened tokens $\mathbf{Y}_{\text{sp}}$, which encapsulate the holistic image layout and serve as the compact visual representation for the subsequent TCC module.

3.3 TEXT-CONDITIONED CONTEXT COARSENING

Cross-modal context bridging. Beyond the spatial relations captured by SCC, TCC further incorporates query-conditioned evidence to identify fine-grained visual details relevant to the current task. Specifically, it establishes text-conditioned relations between the coarsened regions and original patches according to their shared relevance to the textual query.

Let f be the projection layer, $\mathbf{C} = f(\mathbf{Y}_{\text{sp}})$ be the region tokens, $\mathbf{V} = f(\mathbf{X})$ be the original patch tokens, and $\mathbf{E} \in \mathbb{R}^{M \times d_l}$ be the prompt embeddings. We first measure the text-conditioned relevance of both visual and text representations via cosine similarities normalized with softmax:

$$\mathbf{Q} = \text{softmax}_M\left(\cos(\mathbf{C}, \mathbf{E})\right) \in \mathbb{R}^{K_c \times M}, \quad \mathbf{R} = \text{softmax}_N\left(\cos(\mathbf{E}, \mathbf{V})\right) \in \mathbb{R}^{M \times N}, \tag{7}$$

where $\mathbf{Q}$ and $\mathbf{R}$ capture the relevance of clusters and original patches to the question, respectively. By multiplying $\mathbf{Q}$ and $\mathbf{R}$, we construct a cross-modal bridging matrix $\mathbf{B} = \mathbf{QR} \in \mathbb{R}^{K_c \times N}$ that relates each coarsened region to the original patches according to their shared relevance to the textual query. To avoid trivial associations between a region and its constituent patches, we use the region-assignment indices retained during graph construction to mask each patch's affinity to its parent region in $\mathbf{B}$. The masked relations are then aggregated along the region dimension to derive a text-conditioned contribution score for each original patch. Based on these scores, we select the top K_p most relevant original patches to form the protected set $\mathcal{P}$, preserving query-relevant fine-grained evidence during subsequent compression.

Hybrid graph construction and propagation. To integrate these protected details without disrupting the macro-structure, we construct a hybrid graph $\mathcal{G}^{\text{text}} = (\mathcal{V}^{\text{text}}, \mathcal{E}^{\text{text}}, \mathbf{F}^{\text{text}}, \mathbf{W}^{\text{text}})$ comprising the K_c clusters and the K_p protected patches, which serves as a unified structural bridge to jointly model spatial layouts and task-specific fine-grained evidence. Specifically, we retain the cluster-to-cluster spatial adjacency inherited from the SCC stage. Then, for each protected patch, we establish relay edges to all clusters except its original owner cluster, with edge weights determined by the corresponding masked affinities in $\mathbf{B}$. Through this construction, $\mathbf{W}^{\text{text}}$ organically integrates the regional spatial topology with these cross-modal relay links. These relay edges serve strictly as communication channels for context propagation, whereas legal region merging remains restricted to the inherited spatial edges.

Then, similar to the SCC stage, we perform graph propagation on the hybrid graph to inject text-conditioned guidance:

$$\mathbf{P}^{\text{text}} = \alpha\mathbf{I} + (1-\alpha)\,\text{RowNorm}(\mathbf{W}^{\text{text}}), \tag{8}$$

where $\mathbf{P}^{\text{text}}$ specifies how each cluster and protected patch aggregates visual context from its hybrid-graph neighbors, with text-mediated relay weights controlling cross-type information flow.

Then we initialize the node features as $\mathbf{F}^{\text{text}} = [\mathbf{Y}_{\text{sp}}; \mathbf{X}_{\mathcal{P}}]$. The updated context states are obtained via ℓ_2-step propagation:

$$\mathbf{H}^{\text{text}} = (\mathbf{P}^{\text{text}})^{\ell_2}\mathbf{F}^{\text{text}}, \tag{9}$$

where $\mathbf{H}^{\text{text}}$ denotes the text-conditioned contextual states that jointly guide the final greedy merging.

Text-conditioned context merging. We then reuse Algorithm 1 to contract the hybrid context graph $\mathcal{G}^{\text{text}} = (\mathcal{V}^{\text{text}}, \mathcal{E}^{\text{text}}, \mathbf{F}^{\text{text}}, \mathbf{W}^{\text{text}}, \mathbf{H}^{\text{text}})$, where $\mathbf{H}^{\text{text}}$ determines the contextual similarity while restricting valid merges to the inherited spatial edges. Operating on these text-conditioned representations, the operator further compresses the K_c spatial clusters into K_c' regional tokens, denoted by $\mathbf{Y}_{\text{text}}$, to satisfy the overall token budget allocation ($K = K_c' + K_p$), leaving K_p protected patches unchanged. The final token sequence $\mathbf{Z} \in \mathbb{R}^{K \times d_l}$ is formed by concatenating the K_c' merged clusters and the K_p protected patches, followed by the linear projection layer f to match the input dimension of the language model:

$$\mathbf{Z} = f\left([\mathbf{Y}_{\text{text}}; \mathbf{X}_{\mathcal{P}}]\right), \quad |\mathbf{Y}_{\text{text}}| = K_c', \quad K = K_c' + K_p, \tag{10}$$

which is subsequently fed into the MLLM backbone to facilitate accurate response generation.

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate visual-token compression on eight core benchmarks: GQA (Hudson & Manning, 2019), POPE (Li et al., 2023b), MME (Fu et al., 2025), ScienceQA (SQA) (Lu et al., 2022), TextVQA (Singh et al., 2019), VizWiz (Gurari et al., 2018), MMBench (Liu et al., 2024c), and SEED_{img} (Li et al., 2024). OCRBench (Liu et al., 2024d) is included as an additional OCR-oriented benchmark in the Qwen2.5-VL experiments.

Backbones and compared methods. We study three representative architectures: LLaVA-v1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). We compare against classical pruning-based methods including SparseVLM (Zhang et al., 2025), DivPrune (Alvar et al., 2025), HoloV (Zou et al., 2025), SpecFlow (Li et al., 2026), MMTok (Dong et al., 2026), and recent pruning-merging hybrid methods including VisionZip (Yang et al., 2025) and VisionTrim (Yu et al., 2026). The comparison set is selected according to backbone compatibility and protocol availability. More details are in Appendix A.2 and A.4.

Implementation details. All experiments use the `lmms-eval` framework with matched prompts, answer extraction, splits, and metrics within each table. The uncompressed model is reported as the reference. For hyperparameters, we set $K_c = K$ in the SCC, the self-loop weight to $\alpha = 0.1$, preserve $\rho = \frac{K_p}{K} = \frac{1}{16}$, and use $\ell_1 = 2$ and $\ell_2 = 1$ propagation steps for the spatial and text-conditioned stages, respectively. More details for backbone adaption are in Appendix A.3.

4.2 Main Results

Performance comparison with SOTA methods. Table 1 compares CVTC with recent training-free methods on LLaVA-v1.5-7B at 192, 128, and 64 tokens, using the uncompressed 576-token model as the reference. Avg. is computed by normalizing each benchmark score by its corresponding uncompressed-model score and then averaging across benchmarks. CVTC achieves average retention rates of 98.4%, 97.7%, and 96.0%, exceeding the strongest competitor by 0.1, 0.3, and 1.3 points, respectively. The widening margin shows that its advantage becomes more pronounced as the budget tightens. CVTC ranks first on GQA, POPE, and VizWiz at every budget, while MME and

Table 1: Comparison with training-free methods on the fixed-resolution LLaVA-v1.5-7B backbone. Avg. is the mean relative retention with respect to the uncompressed model.

LLaVA-v1.5-7B	GQA Acc.↑	POPE F1↑	MME P+C↑	TextVQA Acc.↑	SQA Acc.↑	VizWiz Acc.↑	$SEED_{img}$ Acc.↑	Avg. (7) ↑
Uncompressed	61.9	85.9	1862	58.2	69.5	50.0	66.1	100.0%
Retain $K = 192$ *tokens,* 33.3%								
SparseVLM$_{ICML'25}$	59.5	85.3	1787	57.8	68.7	50.5	**64.3**	98.3%
HoloV$_{NeurIPS'25}$	58.6	85.6	1793	55.8	69.1	51.4	63.2	97.8%
VisionTrim$_{ICLR'26}$	58.3	85.5	1731	57.4	70.0	51.5	61.5	97.4%
SpecFlow$_{ICLR'26}$	58.3	85.8	**1827**	**57.9**	69.7	50.5	63.1	98.3%
CVTC (ours)	**59.9**	**86.4**	1756	54.8	**70.1**	**52.9**	63.7	**98.4%**
Retain $K = 128$ *tokens,* 22.2%								
SparseVLM$_{ICML'25}$	58.4	85.0	1746	56.7	68.6	51.4	**63.5**	97.4%
HoloV$_{NeurIPS'25}$	57.4	83.2	1768	55.7	69.1	52.3	61.5	96.8%
VisionTrim$_{ICLR'26}$	56.3	84.2	1693	56.6	**70.1**	51.9	58.9	95.8%
SpecFlow$_{ICLR'26}$	57.6	84.9	**1794**	**56.8**	69.9	51.1	60.8	97.2%
CVTC (ours)	**58.9**	**85.8**	1722	54.0	**70.2**	**54.1**	62.6	**97.7%**
Retain $K = 64$ *tokens,* 11.1%								
SparseVLM$_{ICML'25}$	53.8	77.5	1589	53.4	69.8	50.1	56.9	91.6%
HoloV$_{NeurIPS'25}$	55.1	80.3	**1728**	54.2	69.5	53.1	58.3	94.7%
VisionTrim$_{ICLR'26}$	53.2	81.1	1600	**55.1**	**70.0**	52.4	53.7	92.5%
SpecFlow$_{ICLR'26}$	55.3	80.5	1713	54.9	69.7	52.4	57.3	94.5%
CVTC (ours)	**57.6**	**84.0**	1674	52.6	69.8	**54.7**	**60.2**	**96.0%**

Table 2: Comparison on the LLaVA-NeXT-7B backbone at 320 retained visual tokens (11.1%). Avg. is the benchmark-wise mean relative retention over all benchmarks with reported results in each row.

LLaVA-NeXT-7B	GQA Acc.↑	POPE F1↑	MME P+C↑	SQA Acc.↑	TextVQA Acc.↑	VizWiz Acc.↑	MMB Acc.↑	$SEED_{img}$ Acc.↑	Avg. ↑
Uncompressed	64.2	86.4	1842	70.2	61.3	57.6	67.9	70.2	100.0%
PruMerge$_{ICCV'25}$	53.6	60.8	1534	66.4	50.6	54.0	61.3	–	85.5%
VisionTrim$_{ICLR'26}$	56.3	73.8	1638	71.9	52.6	53.7	**65.3**	60.1	90.7%
SparseVLM$_{ICML'25}$	56.1	82.4	1533	66.7	**55.9**	52.0	60.6	–	90.2%
CVTC (ours)	**60.8**	**86.8**	**1726**	**72.1**	53.8	51.7	62.2	**64.6**	94.1%

TextVQA exhibit task-specific trade-offs. Notably, it surpasses the uncompressed model on POPE at 192 tokens and on VizWiz at all three budgets, suggesting that the original visual sequence contains substantial redundancy and that context-aware compression can suppress noise from redundant tokens. Even at 64 tokens, corresponding to only 11.1% of the original visual sequence, it retains 96.0% of the uncompressed cross-benchmark performance.

The same trend extends to different visual tokenization schemes. On LLaVA-NeXT, CVTC reaches a 94.1% average with 320 of up to 2,880 AnyRes tokens, surpassing VisionTrim by 3.4 points and ranking first on five of eight benchmarks. On dynamic-resolution Qwen2.5-VL, it obtains 92.5%, 86.6%, and 79.7% average retention at the 20%, 10%, and 5% settings, leading the strongest competitor by 0.5, 2.8, and 5.4 points. At 5% retention, CVTC ranks first on six of seven benchmarks. These results show that context-aware compression remains effective across fixed-resolution, multi-scale, and dynamic-resolution visual tokenization, with the largest gains under the tightest budgets.

Efficiency analysis. We measure efficiency on POPE Full with LLaVA-v1.5-7B on a platform equipped with 4 NVIDIA L40-48G GPUs. Table 4 reports total wall-clock time, peak memory, cross-benchmark average performance, and prefill latency for the 192- and 64-token budgets; the prefill is measured from the 576 visual tokens returned by the ViT entering the compressor until the LLM starts generating the first KV cache.

Table 3: Comparison on the Qwen2.5-VL-7B backbone. K denotes the average number of visual tokens for each benchmark. Avg. is the mean relative retention with respect to the uncompressed dynamic-resolution model.

Qwen2.5-VL-7B	GQA Acc.↑	MMB Acc.↑	MME P+C↑	POPE F1↑	TextVQA Acc.↑	SQA Acc.↑	OCRBench Acc.↑	Avg. ↑
Dynamic Resolution (MinPix $= 256 \times 28 \times 28$*, MaxPix* $= 2048 \times 28 \times 28$*)*								
Avg. Tokens K	358.5	276.9	867.6	359.6	976.5	323.0	652.8	545.0
Uncompressed	60.5	83.3	2327	86.2	77.7	87.5	83.8	100.0%
Fixed Resolution (MinPix = MaxPix $= 2048 \times 28 \times 28$*)*								
Uncompressed	58.6	83.6	2339	86.1	76.6	86.9	76.6	99.3%
Retain 20% K	71.7	55.4	173.5	71.9	195.3	64.6	130.6	↓ 80%
VisionZip$_{CVPR'25}$	56.8	**80.3**	2174	83.4	70.4	84.2	59.5	91.2%
DivPrune$_{CVPR'25}$	56.7	77.0	2163	80.6	65.9	80.9	48.1	86.8%
HoloV$_{local}$	56.5	77.4	2122	80.9	63.8	86.7	48.6	87.2%
MMTok$_{ICLR'26}$	58.1	79.3	**2217**	82.4	70.5	81.6	59.6	91.0%
SpecFlow$_{ICLR'26}$	55.2	77.4	2056	**84.0**	66.0	86.7	28.1	83.9%
VisionTrim$_{ICLR'26}$	56.2	79.8	2177	82.3	**74.1**	86.9	**60.0**	92.0%
CVTC (ours)	**59.4**	78.9	2216	83.9	71.3	**87.2**	59.2	**92.5%**
Retain 10% K	35.9	27.7	86.8	36.0	97.7	32.3	65.3	↓ 90%
VisionZip$_{CVPR'25}$	52.5	**75.6**	2003	78.9	63.8	82.3	36.9	82.2%
DivPrune$_{CVPR'25}$	53.4	72.9	1957	75.0	59.6	79.6	37.3	79.9%
HoloV$_{local}$	53.2	71.8	1980	76.5	55.2	85.3	36.6	80.0%
MMTok$_{ICLR'26}$	55.1	74.7	2051	78.8	63.9	80.5	43.6	83.8%
SpecFlow$_{ICLR'26}$	50.6	72.2	1817	79.7	57.6	85.1	22.1	76.9%
VisionTrim$_{ICLR'26}$	52.5	75.4	2001	78.0	63.0	**85.9**	36.3	82.3%
CVTC (ours)	**57.5**	74.2	**2161**	**82.1**	**65.3**	85.5	**43.7**	**86.6%**
Retain 5% K	17.9	13.8	43.4	18.0	48.8	16.2	32.6	↓ 95%
VisionZip$_{CVPR'25}$	46.3	67.5	1677	66.4	54.5	79.6	19.7	70.2%
DivPrune$_{CVPR'25}$	49.0	65.9	1739	68.5	52.0	77.1	24.9	71.3%
HoloV$_{local}$	50.0	66.8	1847	69.2	48.5	83.9	27.2	73.3%
MMTok$_{ICLR'26}$	50.7	65.9	1796	71.4	56.0	77.2	**30.7**	74.3%
SpecFlow$_{ICLR'26}$	46.5	65.8	1664	75.6	49.9	84.4	19.0	71.2%
VisionTrim$_{ICLR'26}$	48.9	68.7	1739	71.2	48.9	84.9	16.3	71.4%
CVTC (ours)	**54.3**	**69.5**	**1953**	**79.7**	**59.8**	**85.2**	28.5	**79.7%**

Table 4: Efficiency comparison on LLaVA-v1.5-7B. Time denotes total wall-clock time over POPE examples. "Avg." is the normalized cross-benchmark average reported in Table 1.

LLaVA-v1.5-7B	Retain $K = 192$				Retain $K = 64$			
	Time	Mem. (GB)	Prefill (ms)	Avg.	Time	Mem. (GB)	Prefill (ms)	Avg.
Uncompressed	0:05:02	14.6	87.1	100.0%	0:05:02	14.6	87.1	100.0%
SparseVLM$_{ICML'25}$	0:04:26	15.0	80.4	98.3%	0:03:38	15.0	76.6	91.6%
HoloV$_{NeurIPS'25}$	0:03:43	14.8	55.7	97.8%	0:03:33	14.9	51.8	94.7%
VisionTrim$_{ICLR'26}$	0:03:57	16.9	53.2	97.4%	0:04:01	17.0	51.2	92.5%
CVTC (ours)	0:03:59	14.8	60.6	**98.4%**	0:03:53	14.8	53.8	**96.0%**

4.3 Ablation Study

Multimodal context mechanism. We evaluate our cascade design on LLaVA-v1.5-7B and LLaVA-NeXT-7B using an 11.1% budget with PruMerge Shang et al. (2025) as the baseline method. Compared to the baseline, combining spatial and text-conditioned context (*+SCC+TCC*) achieves the highest normalized average, reaching 93.6% and 96.3% on the two backbones, respectively, as shown in Table 5. While spatial-only (*+SCC*) variants excel in reasoning (e.g., GQA), the full cascade further enhances hallucination resistance (POPE) and perception (MME), demonstrating the complementary benefits of spatial structure (*SCC*) and text-conditioned refinement (*TCC*).

Table 5: Ablation on LLaVA-v1.5-7B and LLaVA-NeXT-7B. Avg. is the mean relative retention rate with respect to the corresponding uncompressed model.

Method	LLaVA-v1.5-7B, 64 tokens (11.1%) GQA	POPE	MME	Avg.	LLaVA-NeXT-7B, 320 tokens (11.1%) GQA	POPE	MME	Avg.
Uncompressed	61.9	85.9	1862	100.0%	64.2	86.4	1842	100.0%
baseline	51.9	65.3	1549	81.0%	53.6	60.8	1534	79.0%
+SCC	57.6	82.4	1669	92.9%	**61.0**	84.0	1714	95.1%
+TCC	55.9	82.9	1648	91.8%	59.2	85.5	1671	94.0%
+SCC+TCC	**57.6**	**84.0**	**1674**	**93.6%**	60.8	**86.8**	**1726**	**96.3%**

Table 6: Topology ablation on LLaVA-v1.5-7B under different token budgets. Avg. is the mean relative retention with respect to the uncompressed model.

LLaVA-v1.5	192 tokens GQA	POPE	MME	Avg.	128 tokens GQA	POPE	MME	Avg.	64 tokens GQA	POPE	MME	Avg.
Uncompressed	61.9	85.9	1862	100.0%	61.9	85.9	1862	100.0%	61.9	85.9	1862	100.0%
Rewired-local	59.5	83.5	1742	95.6%	58.5	81.1	1691	93.2%	56.2	74.2	1610	87.9%
Global-kNN	58.9	83.5	1727	95.0%	57.0	80.7	1614	90.9%	52.4	67.0	1435	79.9%
Ours	**60.0**	**86.6**	**1755**	**97.3%**	**59.0**	**85.7**	**1731**	**96.0%**	**57.5**	**83.9**	**1640**	**92.9%**

Topology strategy. In Table 6, we evaluate alternative context-graph topologies on LLaVA-v1.5-7B across 192, 128, and 64 tokens. Compared to *rewired-local* (modifying local edge connections while preserving spatial locality) and *global-kNN* (replacing spatial locality with feature-space kNN), our real-local graph consistently achieves higher normalized averages across all budgets (e.g., 97.3% vs. 95.6%/95.0% at 192 tokens). Notably, the performance gap widens under tighter budgets: at 64 tokens, our method maintains 92.9% while *global-kNN* drops sharply to 79.9%, highlighting the necessity of local structural constraints under aggressive compression.

Robustness analysis. We further assess the robustness of CVTC on the LLaVA-v1.5-7B backbone by increasing the strength of Gaussian blur and adding noise to the input images, as shown in Fig. 3. Across these conditions, our method generally outperforms the recent pruning-based method SpecFlow at matched token budgets, with the clearest margins at 64 tokens and on the real-world VizWiz images. The complete GQA, POPE, and VizWiz results are reported in Appendix Table A2; the results show that context-aware merging remains competitive when local patch evidence is less reliable.

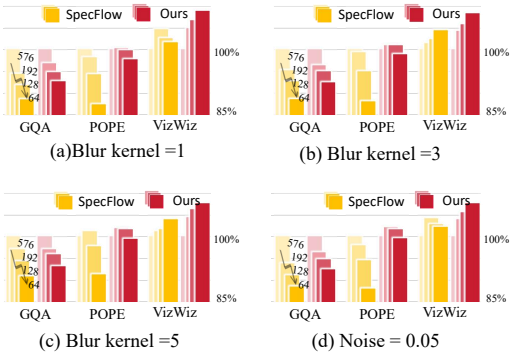


Figure 3: Robustness analysis on corruptions.

Parameter sensitivity. We analyze sensitivity to three main hyperparameters: the spatial propagation depth ℓ_1, the text-conditioned propagation depth ℓ_2, and the preserved raw-patch ratio ρ. CVTC is generally insensitive to moderate changes in these parameters; detailed results and analysis are provided in Appendix A.5.

5 Conclusion

We presented CVTC, a training-free visual token compression framework that unifies spatial context coarsening and text-conditioned guidance. By decoupling matching decisions from raw features, CVTC effectively preserves global layout and task-critical details within a strict token budget. Extensive evaluations demonstrate that CVTC achieves strong cross-benchmark performance while reducing computational overhead.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and analyze experimental data. We did not use generative AI for figure generation, literature retrieval or discovery, research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims. We reviewed all AI-assisted work and take responsibility for the final content, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models, 2025. URL https://arxiv.org/abs/2503.02175.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster, 2023. URL https://arxiv.org/abs/2210.09461.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. PuMer: Pruning and merging tokens for efficient vision language models, 2023. URL https://arxiv.org/abs/2305.17530.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models, 2024a. URL https://arxiv.org/abs/2403.06764.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024b. URL https://arxiv.org/abs/2312.14238.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. MM-Tok: Multimodal coverage maximization for efficient inference of VLMs. In *International Conference on Learning Representations*, pp. 48075–48099, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4f8745c0ca5c951a92844ad7668cfca5-Paper-Conference.pdf.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in VLMs via synergistic importance-diversity. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=i36E5Ezm0H.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://mlanthology.org/neurips/2025/fu2025neurips-mme/.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4075–4083, 2025.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a. URL https://arxiv.org/abs/2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhaoyang Li, Yanjun Li, Wangkai Li, Yujia Chen, and Tianzhu Zhang. SpecFlow: Spectral heat flow for conservative token condensation in vision-language models, 2026. URL https://arxiv.org/abs/2607.10640.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, pp. 216–233. Springer, 2024c. URL https://eccv.ecva.net/papers/eccv_2024/papers_ECCV/papers/00959.pdf.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67, 2024d. doi: 10.1007/s11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0182. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.02122. URL https://arxiv.org/abs/2403.15388.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction, 2025. URL https://arxiv.org/abs/2410.17247.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_VisionZip_Longer_is_Better_but_Not_Necessary_in_Vision_Language_CVPR_2025_paper.html.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. VisionTrim: Unified vision token compression for training-free MLLM acceleration, 2026. URL https://arxiv.org/abs/2601.22674.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 74840–74857, 2025. URL https://mlanthology.org/icml/2025/zhang2025icml-sparsevlm/.

Yuan Zhang, Junpeng Ma, Qizhe Zhang, Chun-Kai Fan, Wenzhao Zheng, Kuan Cheng, Jiwen Lu, and Shanghang Zhang. Sparsevlm+: Visual token sparsification with improved text-visual attention pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3718245.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in MLLMs: Revisiting visual holistic context retention. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39800–39832, 2025. doi: 10.52202/085713-1327. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/38b445e98823a02479582302ee3aa8e4-Paper-Conference.pdf.