Entropy-based Patchification Creates Semantic Tokens

Suhao Yu[1] Jingjia Peng[1] Yao Tang[1] Jiatao Gu[1]
[1]University of Pennsylvania
{suhaoyu, jgu32}@seas.upenn.edu

Abstract

Vision Transformers (ViTs) partition input images into uniformly sized patches regardless of content, treating all regions identically and ignoring the natural semantic coherence among tokens. We focus on document analysis, where the goal is to localize and accurately recognize semantically meaningful regions. To this end, we propose ***Entropy-Based Patchification****, a method that automatically delineates semantically rich regions by cropping multi-scale patch sizes based on page average entropy, and aggregates them into compact representations via local attention and merge MLP. We call the merged tokens as* ***semantic tokens****. Experiments on OCR tasks demonstrate that semantic tokens capture more precise and information-dense representations compared to uniform patches. Beyond OCR, semantic tokens exhibit natural advantages in reading order detection and broader document understanding tasks, suggesting a promising direction for content-aware tokenization in VLM.*

1. Introduction

Vision Transformers (ViTs) [4] have become the dominant paradigm for visual recognition tasks. Yet at the input stage, images are uniformly partitioned into fixed 16×16 patches, treating every region identically and discarding any prior knowledge of semantic content. However, information in natural images is distributed unevenly: semantically rich regions, such as text, fine-grained objects, or structural boundaries, require much finer representation, whereas background regions contribute comparatively little. Intuitively, computational resources should be allocated in proportion to the density of local information, yet the standard patchification scheme provides no mechanism to do so.

Prior work such as APT [3] relies on a fixed threshold to determine patch granularity, implicitly assuming that a fixed entropy level carries the same semantic significance across all images. Our key insight is that the human visual system responds not to the absolute entropy of a region, but to its entropy *relative* to the global context. To this end, we use the global mean entropy as a dynamic threshold, patches exceeding this threshold are recursively cropped into finer-grained patches to allocate more representational capacity to semantically important regions. However, simply cropping high-entropy regions into smaller patches is insufficient without dedicated computation at each scale. We therefore propose a hierarchical fusion mechanism that applies local attention to fine-grained patches before merging them upward, ensuring that information-dense regions receive proportionally more computation.

As illustrated in Figure 1, compared to the hard-coded threshold of APT [3], our method demonstrates superior robustness across both object-centric images and word-centric inputs. Experimental results further show that replacing the uniform patchification with our entropy-based cropping yields substantial improvements in recognition accuracy.

In summary, we (1) introduce a **dynamic entropy threshold** computed from the global mean entropy of each input image, which replaces fixed scalar thresholds and enables content-aware patch cropping that generalizes robustly across word-centric and object-centric images; (2) propose a **hierarchical fusion mechanism** that applies local attention to fine-grained patches before merging them into compact semantic tokens, allocating more computation to information-dense regions; and (3) demonstrate that our semantic tokens significantly improve OCR accuracy on document images, with gains up to 5.5% on precision.

2. Related Work

Semantic Tokenization The construction of semantically meaningful visual tokens has attracted significant research attention, driven by the limitation that uniform grid-based patchification fails to respect the inherent semantic structure of images. Early efforts focused on clustering-based approaches [9, 11], which group fixed-size patches by semantic similarity to encourage more coherent token representations. A complementary line of work draws on superpixel segmentation [6], bypassing the grid altogether in favor of irregular, content-adaptive regions that align with perceptually meaningful structures. More recently, hierarchical methods [3] have explored entropy-guided, coarse-to-fine

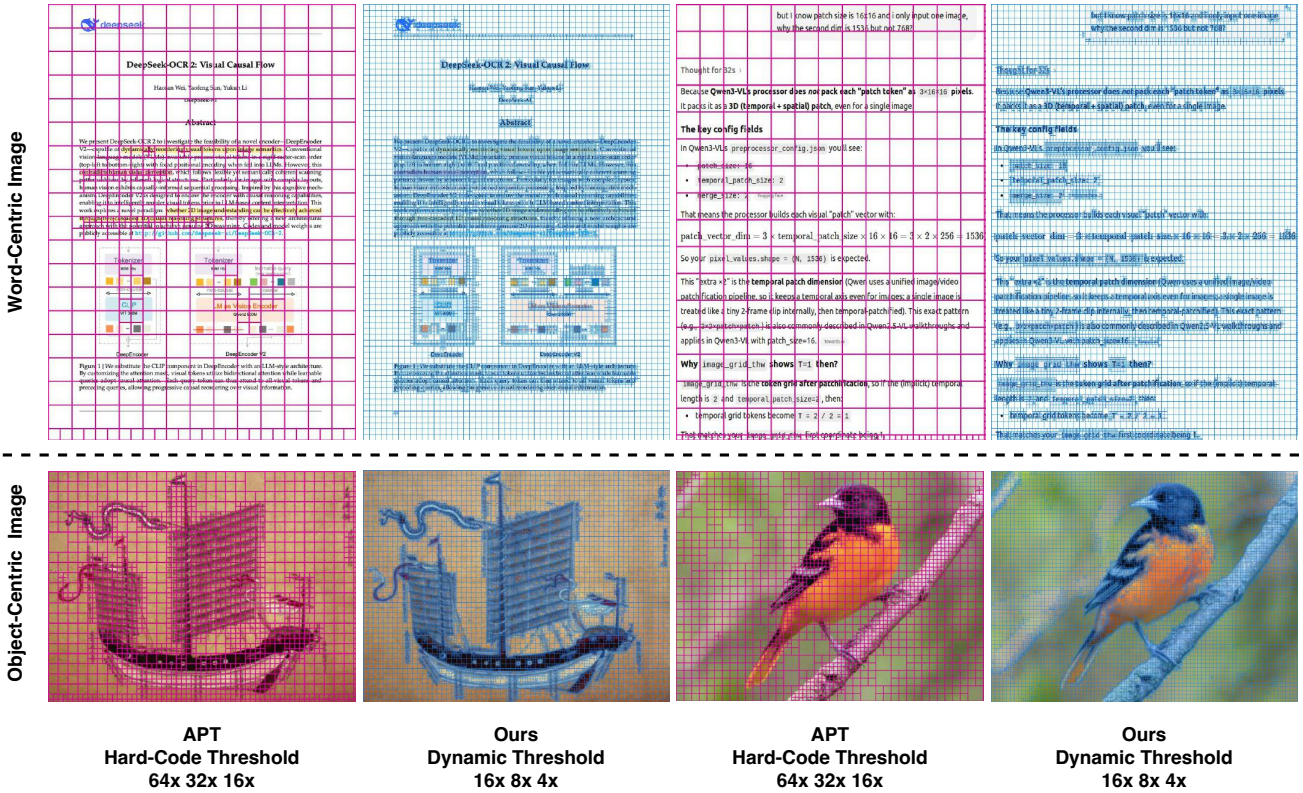

Figure 1. **Comparison of patch cropping strategies between APT [3] and our method on two representative image types.** For APT Following the default setting, patches are hierarchically subdivided across three scales (64×, 32×, 16×): a patch is cropped into the next finer scale if its entropy exceeds a fixed scalar threshold, stopping at 16×. Our method adopts finer scales (16×, 8×, 4×) and replaces the fixed threshold with the mean entropy computed over all maximum-scale patches of the input image. **(Top)** On word-centric document images. APT's hard-coded threshold fails to capture fine-grained text regions. In contrast, our entropy-based dynamic threshold correctly identifies high-information text regions and allocates finer patch granularity accordingly. **(Bottom)** On object-centric images, both methods produce reasonable patch layouts, demonstrating that our adaptive threshold generalizes across diverse image types without manual tuning. Overall, our approach yields more semantically coherent and information-dense patch across domains.

patchification, allocating finer granularity to information-dense regions and consolidating multi-scale representations via bottom-up fusion. These approaches rely on fixed splitting criteria that may not generalize across diverse image types. Our work builds on this foundation, replacing such hard-coded thresholds with a globally adaptive entropy criterion for more robust and consistent semantic tokenization.

Multi-Scale Patch Merging Hierarchical patchification has long been recognized for its ability to capture multi-scale visual information, making it broadly applicable across tasks ranging from fine-grained recognition to high-level semantic understanding. Effectively fusing such multi-scale representations, however, remains a non-trivial challenge. CrossViT [2] addresses this through a dual-branch architecture that independently processes features at different patch scales before cross-attention alignment. In single-branch settings, Quadformer [10] directly resizes inputs to a uniform 16×16 resolution, while APT [3] employs a sequence of resizing and cropping operations followed by fusion, standardizing all patch representations to 16×16 prior to attention computation. In contrast, our approach avoids fixed-resolution preprocessing by adaptively determining patch granularity based on local entropy, enabling a more content-aware merging of multi-scale tokens.

3. Method

Average Entropy as a Dynamic Threshold Prior work such as APT [3] relies on a fixed scalar threshold to determine patch granularity. While this fixed threshold may perform adequately on object-centric images, it generalizes poorly to document analysis tasks, as Figure 1 shows. To address this limitation, we adopt a dynamic thresholding strategy in which the threshold is set as the mean entropy computed over all patches of the input image, allowing it to

adapt to different domains.

Our method operates over three different patch sizes. This multi-scale design is motivated by the observation that text and other fine-grained structures typically occupy a small spatial footprint, requiring finer patch granularity to adequately capture local detail.

Concretely, we begin by computing the entropy of the largest patch. Entropy is estimated by binning pixel intensities into a histogram and computing the resulting empirical distribution. The entropy of a patch P is defined as:

$$H(P) = -\sum_{i=0}^{L-1} p_i \log_2 p_i \tag{1}$$

where L denotes the number of intensity bins and p_i is the probability of the i-th bin. Entropy measures the unpredictability of pixel intensities within a patch, serving as a proxy for local information density and compressibility. The global entropy threshold is then defined as the mean entropy over all N patches of the input image:

$$\tau = \frac{1}{N} \sum_{k=1}^{N} H(P_k) \tag{2}$$

where P_k denotes the k-th patch. Using a constant threshold τ, patches with entropy exceeding τ are recursively cropped into finer scales until a minimum size is reached, while low-entropy patches are retained at their current resolution.

Hierarchical Patch Merging Following top-down cropping, we aggregate multi-scale patches into unified representations via a hierarchical merging module. As illustrated in Figure 2, fine-grained patches at each scale are first processed through a local attention block to capture intra-patch contextual dependencies. The attention block consists of:

$$\mathbf{x}' = \mathbf{x} + \text{MHA}(\text{Norm}(\mathbf{x})) \tag{3}$$

$$\mathbf{z} = \mathbf{x}' + \text{MLP}(\text{Norm}(\mathbf{x}')) \tag{4}$$

where $\mathbf{x}$ denotes the input token sequence, $\text{MHA}(\cdot)$ is multi-head self-attention, $\text{Norm}(\cdot)$ is layer normalization, and $\text{MLP}(\cdot)$ is a two-layer feed-forward network. The output $\mathbf{z}$ is then passed to a MergeMLP, which concatenates 2 $\times$ 2 spatially adjacent patches and projects them into a single compact semantic token that preserves local structure.

To facilitate stable training and preserve pretrained representations, we adopt an initialization strategy inspired by APT [3]. Specifically, for the maximum patch scale (denoted $P_{\max}$), which anchors the coarsest level of spatial context, the output of the MergeMLP is augmented with the original patch embedding of the corresponding $P_{\max}$:

$$\mathbf{t} = \text{MergeMLP}(\mathbf{z}) + \mathbf{e}_{P_{\max}} \tag{5}$$

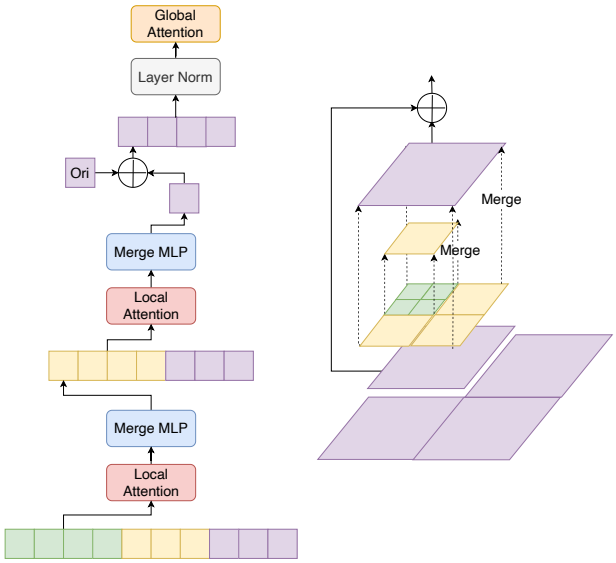

Figure 2. Overview of the hierarchical patch merging module. Fine-grained patches at each scale are processed by a local attention block, then merged by a MergeMLP that merges 2 $\times$ 2 spatially adjacent patches into a single token. For the coarsest scale ($P_{\max}$), the merged token is augmented with the original patch embedding via a residual connection. Before global attention, LayerNorm is applied to all the tokens.

where $\mathbf{e}_{P_{\max}}$ is the standard patch embedding of the $P_{\max}$ region. The final linear layer of this MergeMLP is zero-initialized, ensuring that at the start of training the merged token is identical to the original patch embedding. This residual formulation allows the merging module to learn incrementally without discarding pretrained visual representations in the early stages of optimization. Prior to feeding the token sequence into the decoder, we apply layer normalization across all tokens. This normalization step harmonizes the scale of merged tokens, ensuring a consistent input distribution for the decoder.

4. Experiments

4.1. Experimental Setup

Base Model. We adopt Qwen3-VL-2B-Instruct [1] as our base model. Our modifications are confined to the vision encoder, where we replace the standard uniform patchification module with the proposed entropy-based hierarchical patchification and merging mechanism, leaving the language model backbone unchanged.

Datasets. For training, we use the OCR subsets (OCR_1 through OCR_3) from the nvidia/Llama-Nemotron-VLM-Dataset-v1 [5, 7], which provide diverse document and text recognition samples. For evaluation, we adopt FOX [8] as our benchmark, which covers a broad range of document

Table 1. Precision (%) on the FOX benchmark. N denotes the number of tokens per images, following the split of DeepSeek-OCR [12]. **Red**: best; **blue**: second best per row. †DeepSeek-OCR model. ‡Qwen3-VL with entropy-based patchification (ours).

Token Number	N	Tiny†	Small†	Gundam†	Qwen3-VL	Ours‡
600–700	8	95.5	**97.9**	97.8	95.8	**99.1**
700–800	28	94.1	97.2	**98.0**	94.9	**98.6**
800–900	27	87.7	96.6	**98.0**	93.8	**98.3**
900–1000	14	84.8	96.6	**99.0**	94.9	**98.5**
1000–1100	12	78.9	91.6	**96.5**	92.3	**97.2**
1100–1200	7	79.2	88.3	**94.4**	90.0	**91.1**
1200–1300	4	59.3	86.1	88.4	**89.9**	**95.4**
1300+	12	55.3	85.7	**96.4**	89.7	**94.3**

Table 2. Ablation study on layer normalization before global attention and two stage finetune strategy using Qwen3-VL-2B-Instruct [1]. **Red**: best; **blue**: second best per row.

		New Modules Only		**New Modules + Full Finetune**	
Token Number	N	w/o Norm	w/ Norm	w/o Norm	w/ Norm
600–700	8	89.9	**98.5**	97.3	**99.1**
700–800	28	94.7	**98.0**	93.6	**98.6**
800–900	27	91.5	**97.2**	95.1	**98.3**
900–1000	14	87.4	**98.3**	97.2	**98.5**
1000–1100	12	79.7	**96.5**	88.1	**97.2**
1100–1200	7	87.6	**92.2**	87.1	**91.1**
1200–1300	4	68.5	**93.8**	86.3	**95.4**
1300+	12	78.2	**94.6**	87.6	**94.3**

understanding and fine-grained OCR tasks.

Training. We employ a two-stage training strategy. In the first stage, only the newly introduced modules, patch embedding, local attention, and MergeMLP, are trained, allowing the model to smoothly adapt to the novel multi-scale token representations while keeping the pretrained weights frozen. In the second stage, we perform full end-to-end fine-tuning of the entire model. Each stage is trained for 3 epochs with a learning rate of 1×10^{-5}.

4.2. Main Results

Table 1 reports word-level precision on the FOX benchmark [8], detailed by token numbers. We compare our entropy-based patchification model against DeepSeek-OCR-Tiny, DeepSeek-OCR-Small, DeepSeek-OCR-Gundam [12], and Qwen3-VL-2B-Instruct [1].

Our model consistently outperforms Qwen3-VL-2B-Instruct across all token count (determined using DeepSeek-OCR [12]), achieving gains of up to 5.5%. This improvement is most pronounced in text-heavy images with large token counts. Specifically, for images containing 1200–1300 tokens and 1300+ tokens, our model achieves 95.4% and 94.3%, respectively, compared to 89.9% and 89.7% for Qwen3-VL-2B-Instruct. Furthermore, our method achieves performance on par with the DeepSeek-OCR-Gundam model across all token count.

4.3. Ablation Study

As shown in Table 2, we conduct an ablation study on the FOX benchmark [8] to evaluate two design axes: the necessity of pre-decoder layer normalization (w/ Norm vs. w/o Norm) and the effectiveness of the two stage training (New Modules Only vs. New Modules + Full finetune). The results clearly demonstrate the following:

- **The Necessity of Normalization:** Across both finetune strategies, omitting normalization consistently degrades performance, with the most severe drops occurring at higher token count ($\geq$ 1000). This degradation occurs because, without normalization, the merged patches can exhibit anomalous scale variations in the feature space, which may disrupt or overshadow the representations of other normal tokens. Adding normalization effectively aligns the heterogeneous token distributions and stabilizes the model's performance.
- **Effectiveness of Two-Stage Fine-Tuning:** Compared to the partial "New Modules Only" baseline, incorporating the multi-scale merging module alongside full fine-tuning consistently improves performance on images with varying token counts. Specifically, this second fine-tuning stage is crucial as it allows the encoder and decoder to mutually align their feature representations across the entire network. Furthermore, when combined with normalization, the fully fine-tuned model achieves the best results across most token-count bins (highlighted in red), demonstrating the effectiveness of this two-stage training strategy in handling complex image features.

5. Conclusion

In this work, we proposed Entropy-Based Patchification, a content-aware tokenization method that replaces the uniform grid partitioning of standard ViTs with a dynamic threshold. By identifying high-information regions and fusing multi-scale patches into compact semantic tokens via hierarchical local attention and MergeMLP, our method achieves up to 5.5% improvement in word-level precision on the FOX benchmark over Qwen3-VL-2B-Instruct, reaching performance on par with specialist OCR systems. Beyond recognition accuracy, semantic tokens inherently encode spatial layout structure, conferring natural advantages for reading order detection and broader document understanding tasks, capabilities that uniform patchification fundamentally cannot provide. We further foresee strong potential in GUI understanding, where the entropy mechanism naturally allocates finer tokens to information-dense interactive elements. Future work will explore these features and scale this approach as a drop-in upgrade across larger vision-language backbones.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 3, 4

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2

[3] Rohan Choudhury, JungEun Kim, Jinhyung Park, Eunho Yang, László A Jeni, and Kris M Kitani. Accelerating vision transformers with adaptive patch sizes. *arXiv preprint arXiv:2510.18091*, 2025. 1, 2, 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Ilia Karmanov, Amala Sanjay Deshmukh, Lukas Voegtle, Philipp Fischer, Kateryna Chumachenko, Timo Roman, Jarno Seppänen, Jupinder Parmar, Joseph Jennings, Andrew Tao, et al. \'eclair–extracting content and layout with integrated reading order for documents. *arXiv preprint arXiv:2502.04223*, 2025. 3

[6] Jaihyun Lew, Soohyuk Jang, Jaehoon Lee, Seungryong Yoo, Eunji Kim, Saehyung Lee, Jisoo Mok, Siwon Kim, and Sungroh Yoon. Superpixel tokenization for vision transformers: Preserving semantic integrity in visual tokens. *arXiv preprint arXiv:2412.04680*, 2024. 1

[7] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 3

[8] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024. 3, 4

[9] Yuanbing Ouyang, Yizhuo Liang, Qingpeng Li, Xinfei Guo, Yiming Luo, Di Wu, Hao Wang, and Yushan Pan. Back to fundamentals: Low-level visual features guided progressive token pruning. *Journal of Systems Architecture*, page 103579, 2025. 1

[10] Pratyaksh Prabhav Rao, Feng Qiao, Weide Zhang, Yiliang Xu, Yong Deng, Guangbin Wu, and Qiang Zhang. Quadformer: Quadruple transformer for unsupervised domain adaptation in power line segmentation of aerial images. *arXiv preprint arXiv:2211.16988*, 2022. 2

[11] Giorgos Savathrakis and Antonis Argyros. Enact: Entropy-based clustering of attention input for reducing the computational needs of object detection transformers. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 295–300. IEEE, 2025. 1

[12] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025. 4