RADIO 1D: Elastic Representations for Condensed Vision Modeling

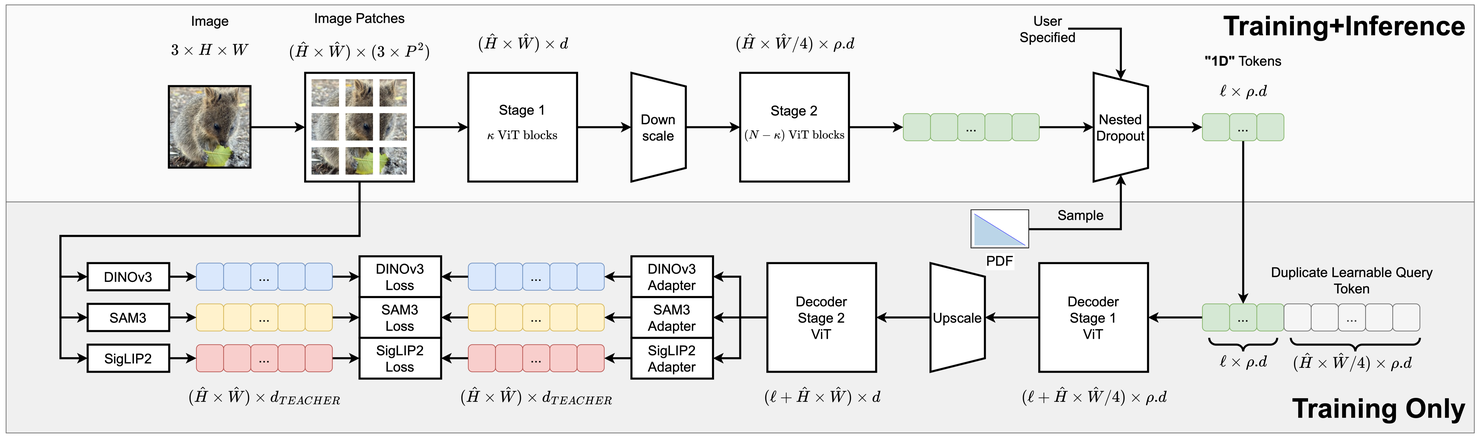
Greg Heinrich*, Mike Ranzinger*, Collin McCarthy*, Natan Bagrov, Eugene Khvedchenya, Bryan Catanzaro, Jan Kautz, Andrew Tao, Pavlo Molchanov

* Equal Contribution

Github

Hugging Face

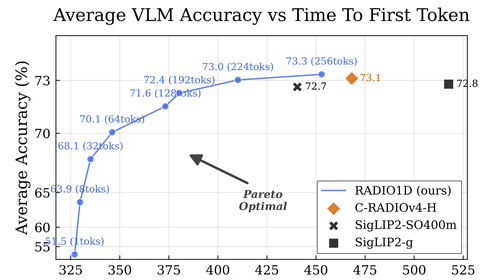
ICML International Conference On Machine Learning



Reduce All Domain Into One 1-Dimensional representation: Auto-Encoder Multi-Teacher Distillation in Latent Space

Applications & Results

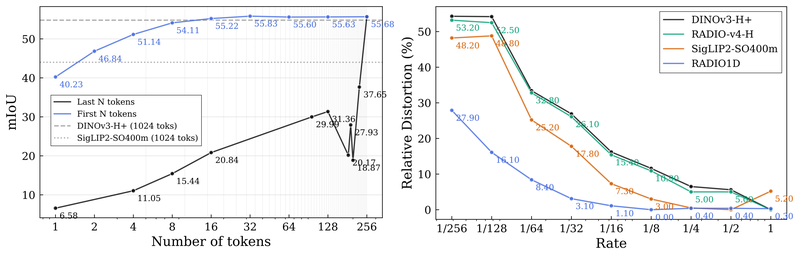
Vision Encoder (VLM)



Vision Encoder	Tokens per tile	TextVQA	DocVQA	InfoVQA	OCRBench	OCRBenchv2-EN	OCRBenchv2-CN	AI2D	ChartQA	MMMU	SeedBench	LVBench	Average
C-RADIOv4+ToMe	128	82.2	90.8	69.0	81.4	58.3	36.3	84.7	87.7	49.2	76.7	57.2	70.31
C-RADIOv4+CDPruner	128	82.8	89.2	69.2	74.7	52.9	33.9	83.5	87.6	50.3	77.0	**58.4**	69.05
RADIO1D(256)+CDPruner	128	82.2	87.2	67.6	76.5	54.6	31.9	84.2	86.8	**50.4**	76.2	56.3	68.56
RADIO1D (ours)	128	**84.0**	**92.5**	**72.6**	**82.8**	**61.5**	**38.6**	**85.1**	**88.6**	47.8	**77.6**	57.0	**71.64**
C-RADIOv4+ToMe	192	83.7	91.7	71.9	82.9	59.6	40.1	84.5	88.5	49.8	77.3	56.4	71.49
C-RADIOv4+CDPruner	192	84.1	92.5	**75.8**	82.1	58.2	39.4	85.3	**89.4**	**51.1**	**77.8**	**58.9**	72.23
RADIO1D(256)+CDPruner	192	83.8	92.3	75.3	82.2	59.7	39.4	**86.0**	88.7	50.6	77.2	57.1	72.02
RADIO1D (ours)	192	**84.5**	**93.2**	75.6	**83.5**	**60.9**	**40.2**	85.7	89.2	48.2	77.6	57.3	**72.36**

Table 1: Comparison of RADIO1D against post-hoc inference-time token reduction methods (Token Merging and CDPruner) at fixed budgets of 128 and 192 tokens. RADIO1D preserves spatial grounding better for reading-heavy tasks, resulting in higher average accuracy.

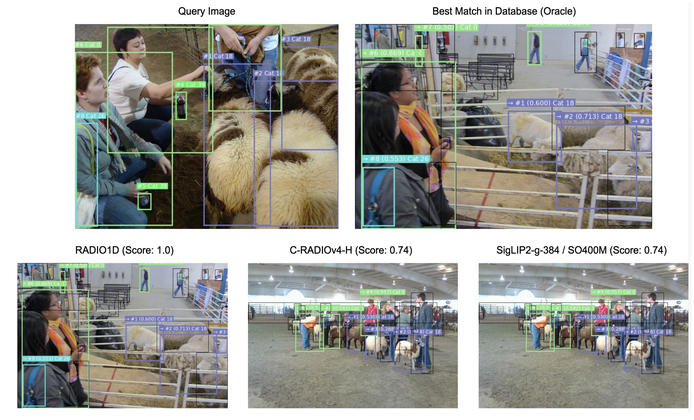
Semantic Segmentation



Truncating the sequence confirms that early tokens capture foundational scene semantics (~40 mIoU with just 1 token).

Best-in-class relative distortion (measuring ADE20k degradation as a function of token compression ratio).

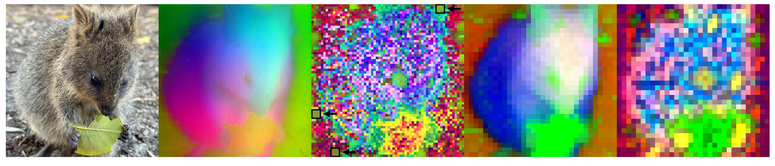
Composition Retrieval



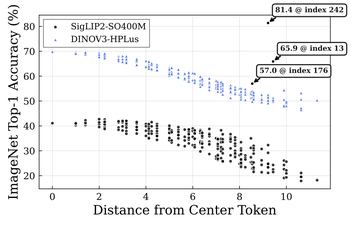
Do We Need 2D Features?

Question: Why are we forcing VLMs to process rigid 2D patches?

Analysis of Vision Encoder features reveals that VLM encoders (e.g. SigLIP) do not exhibit high spatial correlation.



Input **DINOv3** **SigLIP2** **C-RADIOv4** **C-RADIOv4 (VLM Finetuned)**



SigLIP2 has ”outlier” tokens that appear like noise but are very good summarizers.

Waste: Fixed 2D grids carry immense spatial redundancy.

Method

- 1D Tokenization: Flattens standard 2D image patches into an unstructured, compressible 1D token sequence.
- Learnable Downscaling: Integrates Patch Merging directly into the encoder backbone.
- Elastic Bottleneck: Stochastic slicing (nested dropout) during training, forcing earlier tokens to master global semantics while later tokens capture fine details.
- Asymmetric Decoder: Lightweight ViT decoder exclusively during training to reconstruct the 2D grid.
- Multi-Teacher Distillation: Distills features from DINOv3, SigLIP2, etc.