# Accelerating Vision-Language Models: L2-Norm Token Pruning for High-Resolution Encoders

UNIVERSITY OF TORONTO

## Introduction

**The Challenge: High-Resolution Efficiency Gap**

- **State-of-the-art VLMs** (e.g., Qwen2.5-VL **[1]**, LLaVA **[4]**) require high-resolution inputs to understand fine-grained visual details.
- **The Bottleneck:** Vision Transformers (ViT) **[2]** process images as sequences of patch tokens.
- **Quadratic Cost $O(N^2)$:** The computational complexity of Self-Attention grows quadratically with the number of tokens (**N**).
  - High Resolution →Thousands of Tokens→ **High Latency & Memory**.

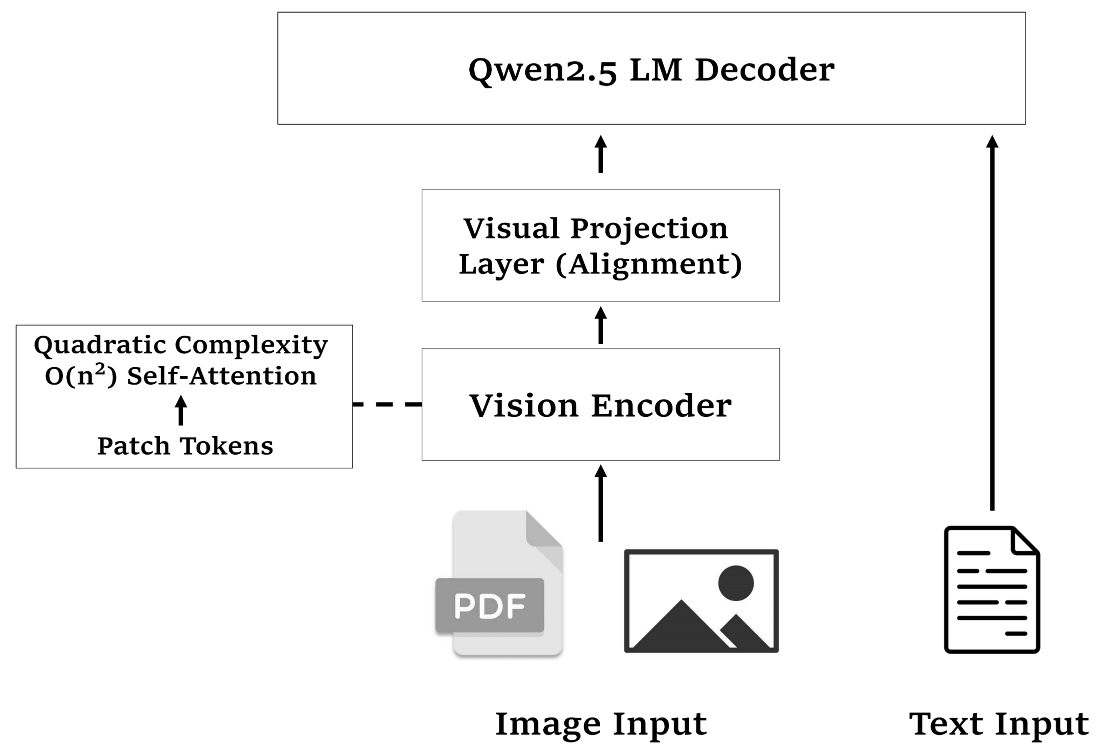


Figure 1. Qwen 2.5 Architecture illustrating Vision Encoder bottleneck

These limitations motivate our research into input-adaptive token reduction. In this experiment, we investigate Vision Token Pruning (VTP) for CLIP-style encoders—the core visual backbone used in modern VLMs like Qwen2.5-VL—to reduce inference cost while preserving semantic performance.

**Motivation & Goals**
To bridge the gap between high performance and deployment efficiency, we propose **Vision Token Pruning (VTP)**.
**Redundancy:** Real-world images contain massive uninformative background patches (e.g., sky, blur).
**Goal 1 (Speed):** Reduce inference latency by removing redundant tokens *before* expensive attention layers.
**Goal 2 (Preservation):** Maintain semantic information using an L2-Norm importance metric (identifying "foreground" vs. "background").

**Contributions:**
**Novelty:** Applied a parameter-free L2-Norm pruning method **[8]** specifically to CLIP-style encoders for VLM efficiency.
**Speed:** Achieved ~1.5x throughput gain on high-res inputs.
Robustness: Verified zero-shot consistency on diverse real-world images.

## Methods

**Methodology: L2-Norm Vision Token Pruning [8]**

1. **Core Hypothesis**
   - **Magnitude ≈ Information:** We hypothesize that the magnitude (L2-Norm) of a patch token correlates with its semantic importance.
     - **High Norm:** Salient objects, edges, textures → **KEEP**
     - **Low Norm:** Flat backgrounds, sky, blur → **PRUNE**
2. **The VTP Pipeline** We introduce a lightweight, parameter-free selection module inserted before the Vision Transformer layers:
   - **Step 1: Importance Scoring**
     Calculate the L2-norm for each input patch token $\boldsymbol{x_i}$:
     $$S_i = ||x_i||_2$$
   - **Step 2: Top-K Selection**
     Sort tokens by score $\boldsymbol{S_i}$ and retain only the top $\boldsymbol{K}$ tokens based on the keep ratio $\boldsymbol{r}$:
     $$K = N \times r$$
   - **Step 3: Spatial Alignment**
     - Unlike standard pooling, simply dropping tokens destroys the 2D spatial structure required by ViTs.
     - **Solution:** We dynamically gather the **Positional Embeddings** corresponding only to the kept indices, ensuring the Transformer understands the relative geometry of the remaining sparse tokens.
3. **Efficiency Gain**

This input-adaptive reduction transforms the computational complexity of Self-Attention from $\mathbf{O(N^2)}$ to $\mathbf{O((rN)^2)}$: yielding quadratic speedups with linear pruning.

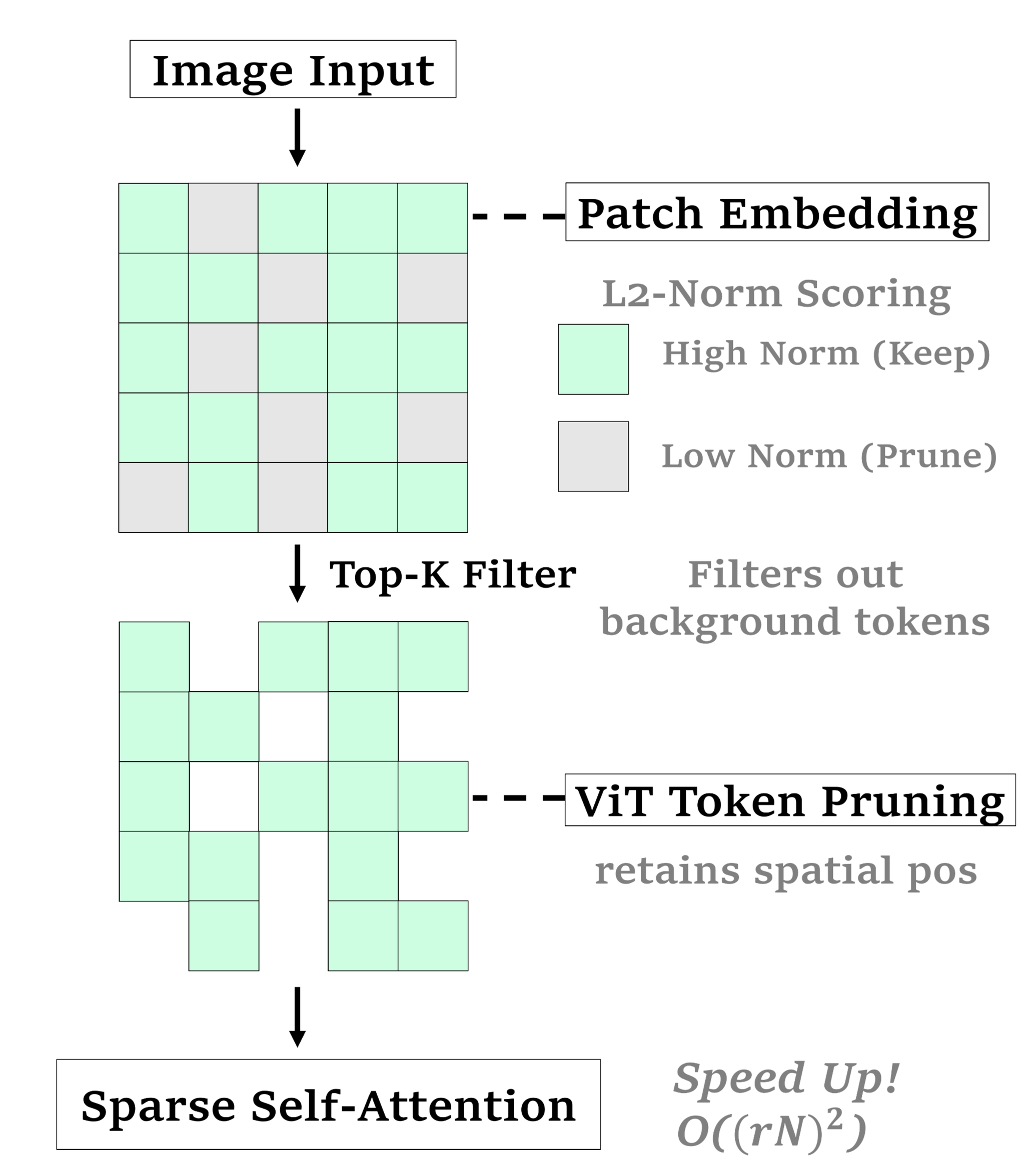


Figure 2. The VTP Pipeline

## Results

**1. Qualitative Analysis: Visualization of Pruning**
We benchmarked the inference latency of the CLIP Vision Encoder on synthetic batches (224 × 224). As shown in **Fig. 3**, reducing the token count directly translates to throughput gains.

- **Latency Reduction:** At a **0.5 keep ratio** (pruning 50% of tokens), average latency drops from **89.6ms to 59.6ms** per batch.
- **Throughput Gain:** This corresponds to a **1.5 times increase in throughput**, jumping from **89 images/s to 134 images/s**.
- **Conclusion:** The quadratic complexity of Self-Attention (**$O(N^2)$**) is effectively mitigated, proving that lighter vision encoders can serve high-speed VLM applications.

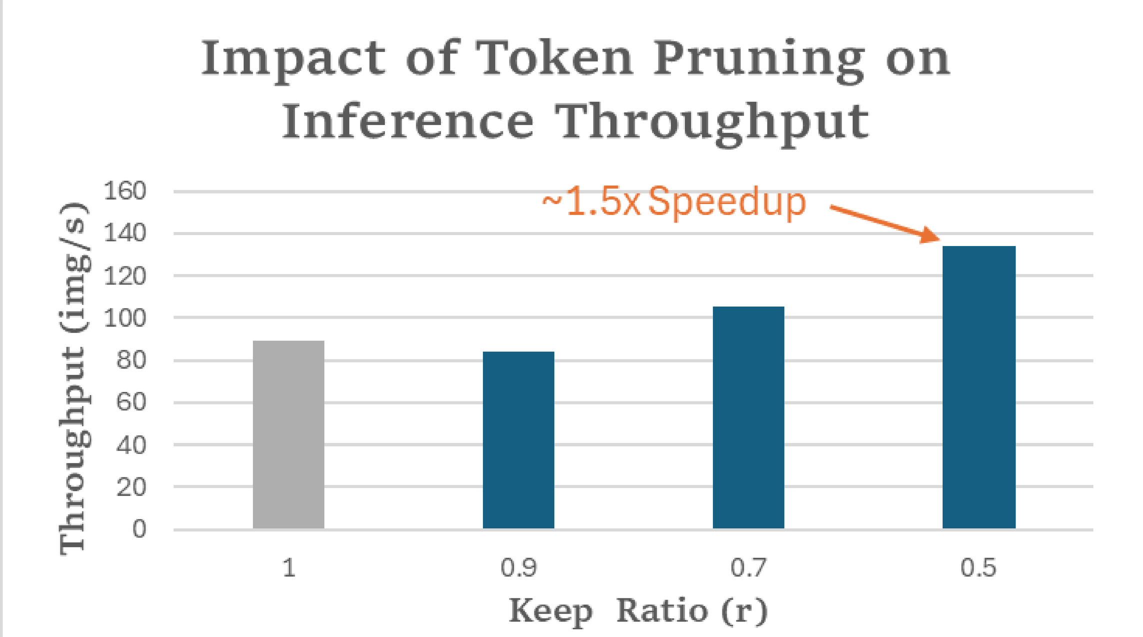


Figure 3. Efficiency Analysis: Throughput vs. Keep Ratio

**2. Qualitative Robustness: Decision Consistency**
High-Resolution Consistency To verify generalization, we extended our validation to 700 high-resolution images (Imagenette).

- **Quantitative Success:** The pruned model achieved **96.9% prediction consistency** with the baseline. **Despite a drop in raw feature similarity (Cosine Sim ≈ 0.47)**, the semantic classification decisions remained robust.
- **Visual Insight:** As visualized in **Fig. 4**, the L2-Norm selector effectively discriminates "foreground" objects (animals) from redundant backgrounds (grass, rug).
- **Sample Predictions: Table 1** shows representative examples where the pruned model correctly aligns with the baseline.

Table 1 Zero-Shot Prediction Consistency $r = 0.5$

| Image ID | Visual Content | Baseline Prediction | Pruned Prediction (r=0.5) | Consistency |
|---|---|---|---|---|
| Image 1 | Two Cats | Cat | Cat | Match |
| Image 2 | Brown Bear | Bear | Bear | Match |
| Image 3 | Weimaraner | Dog | Dog | Match |

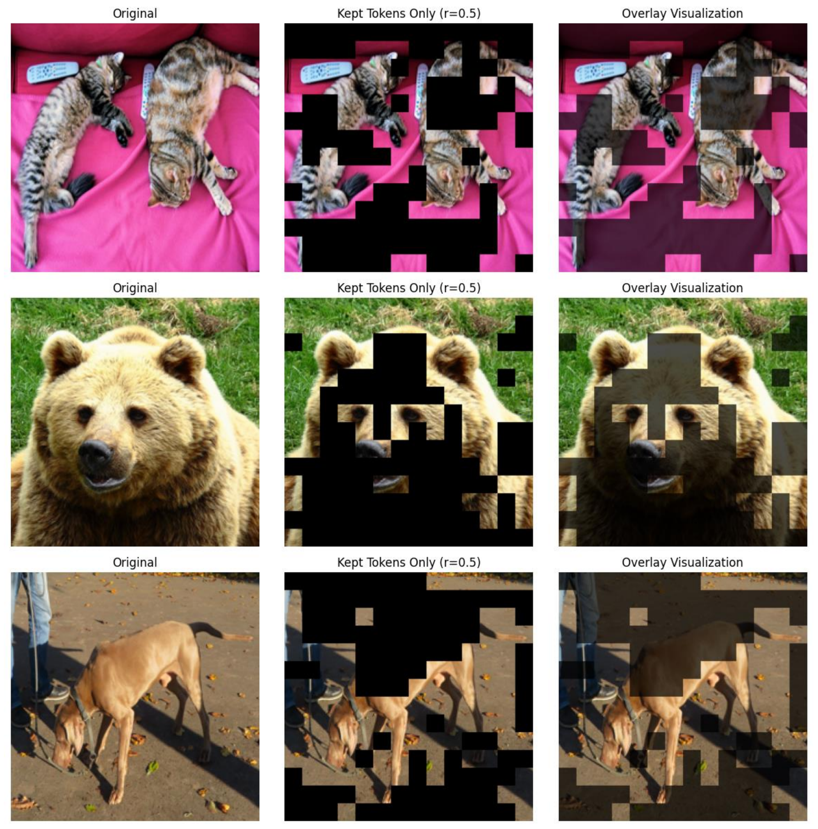


Figure 4. Pruning Visualization

## Conclusions

**Conclusion Summary**: In this work, we addressed the computational bottleneck of high-resolution VLMs (e.g., Qwen2.5-VL) by implementing **Vision Token Pruning (VTP)**. By leveraging the L2-norm of patch embeddings, we identified and removed redundant background tokens before the expensive Self-Attention layers.Key Findings

1. **Validating the Hypothesis:** Background tokens (low L2-norm) contribute minimally to semantic decisions. Removing **50%** of tokens maintained **96.9%** prediction consistency with the baseline on high-resolution validation data.
2. **Efficiency Gains:** We achieved a **1.5 times speedup** in vision encoding throughput, demonstrating a practical path to reduce latency for large multimodal models.

**Future Work** While zero-shot pruning offers speed, the drop in cosine similarity (~0.47) suggests a distributional shift. Future work will explore **fine-tuning (LoRA)** on the pruned features to recover feature alignment without sacrificing efficiency.

## Bibliography

1. Qwen Team. Qwen2.5-VL: A Scalable Vision-Language Model for High-Resolution Perception and Multimodal Understanding. Technical Report, 2024.
2. Zhai, X., et al. SigLIP: Contrastive Vision-Language Pretraining with Sigmoid Loss. NeurIPS 2023.
3. Bai, J., et al. Qwen2 Technical Report: A High-Performance LLM Family. Alibaba Group, 2024.
4. Liu, H., et al. LLaVA: Large Language and Vision Assistant. arXiv 2023.
5. Li, J., et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Q-Formers. ICML 2023.
6. Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.
7. Dosovitskiy, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.
8. Rao, Y., et al. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. NeurIPS 2021.