# PuMer: Pruning and Merging Tokens for Efficient Vision Language Models


**Qingqing Cao**

Bhargavi Paranjape

Hanna Hajishirzi

UWNLP

UNIVERSITY of WASHINGTON


61 ACL 2023

# We Need Faster Vision Language Models

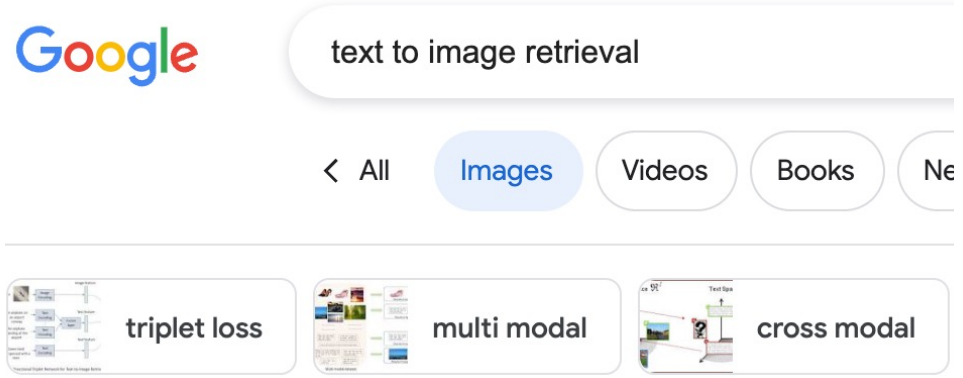
- Deploying to resource-limited devices

  E.g., on-device visual question answering helps the visually-impaired without privacy leaks

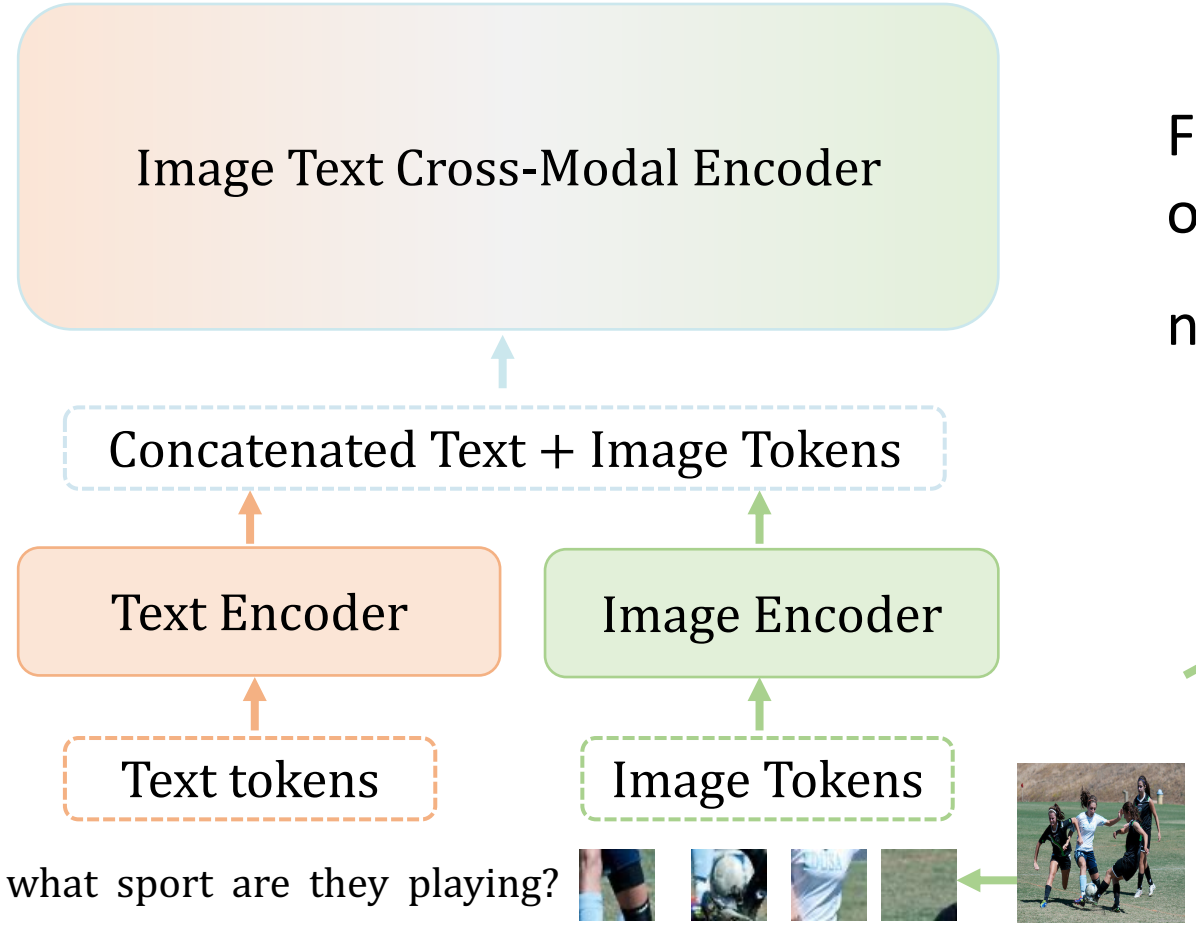

Can you tell me what these pills are?


- Improving throughput and reducing costs for cloud settings

  E.g., text-to-image retrieval, like image search



# VL Models Are Inefficient Due to Processing Many Input Tokens




For example, for an image with a resolution of 384x384 and a patch size of 16, the number of tokens is $\left(\frac{384}{16}\right)^2 = 576$

Hundreds of image tokens

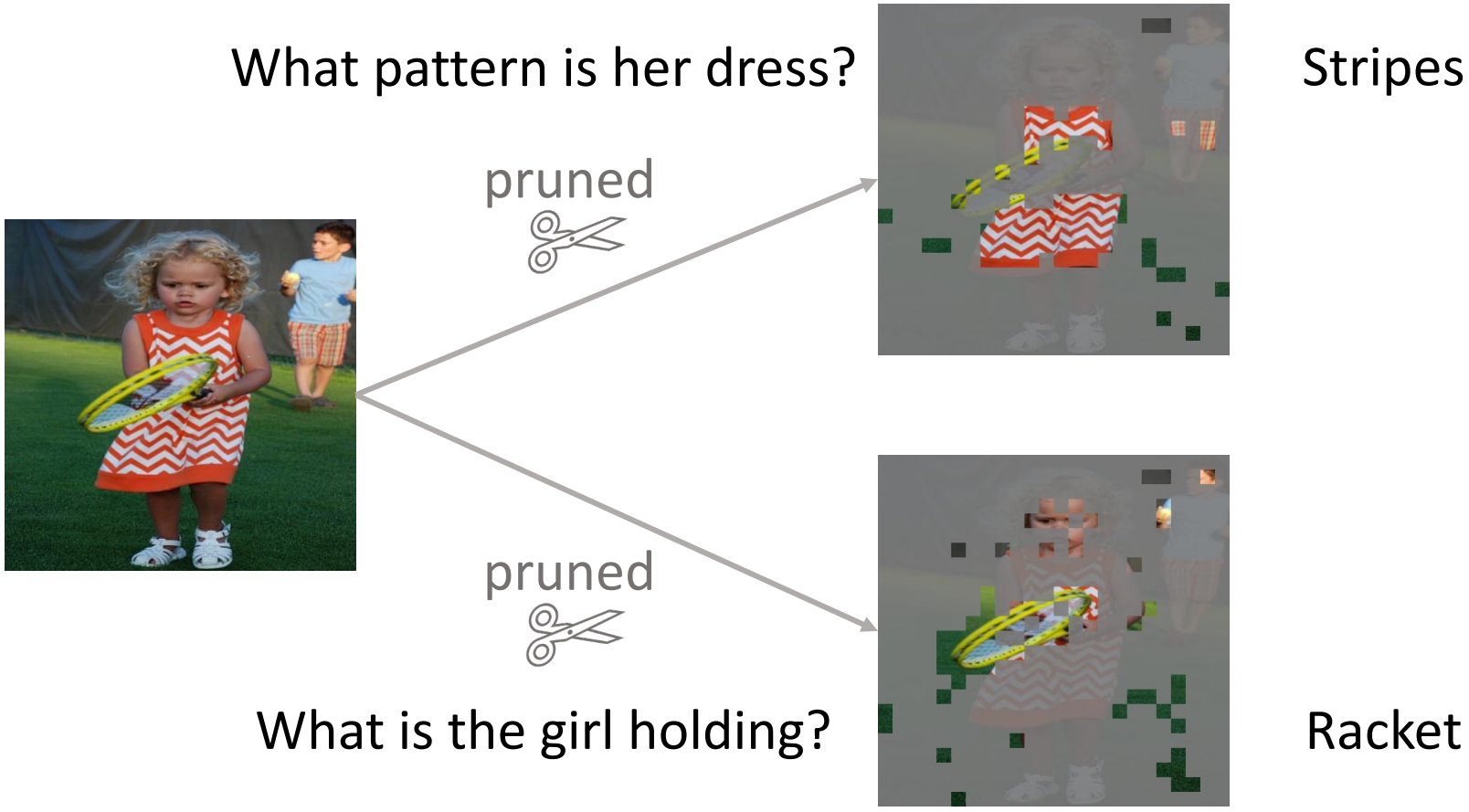
# Intuition 1: The Input Text Queries Different Parts of An Image

What is the girl holding?

How many people are playing?

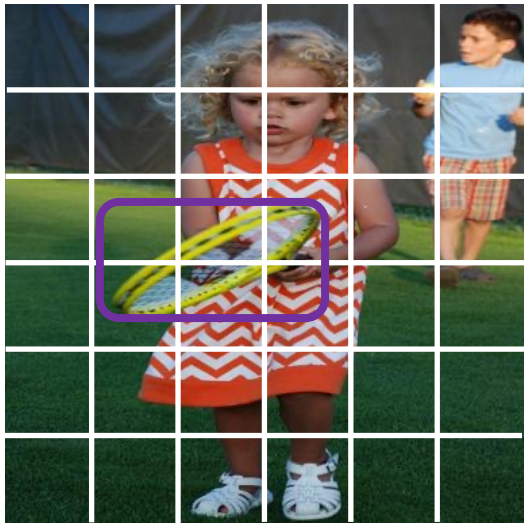
What sport are they playing?

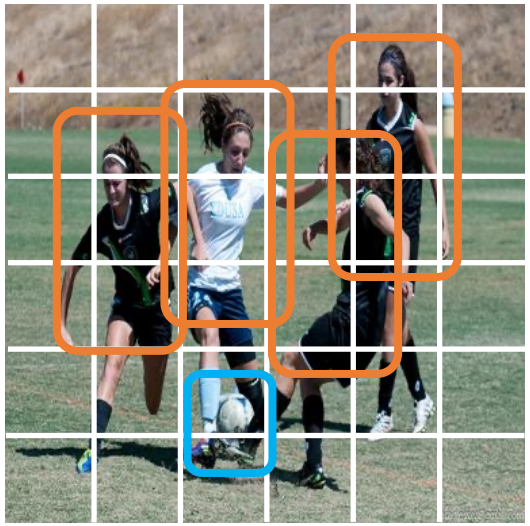
# Technique 1: Text-Informed Image Pruning



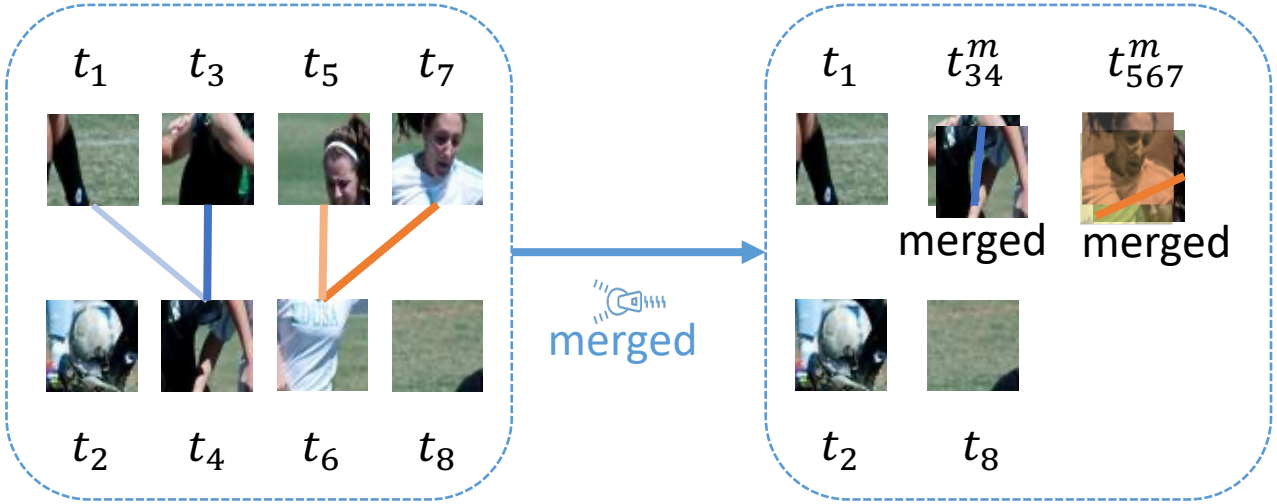
# Intuition 2: Multiple Image tokens Represent Similar Content

Objects span across multiple image tokens





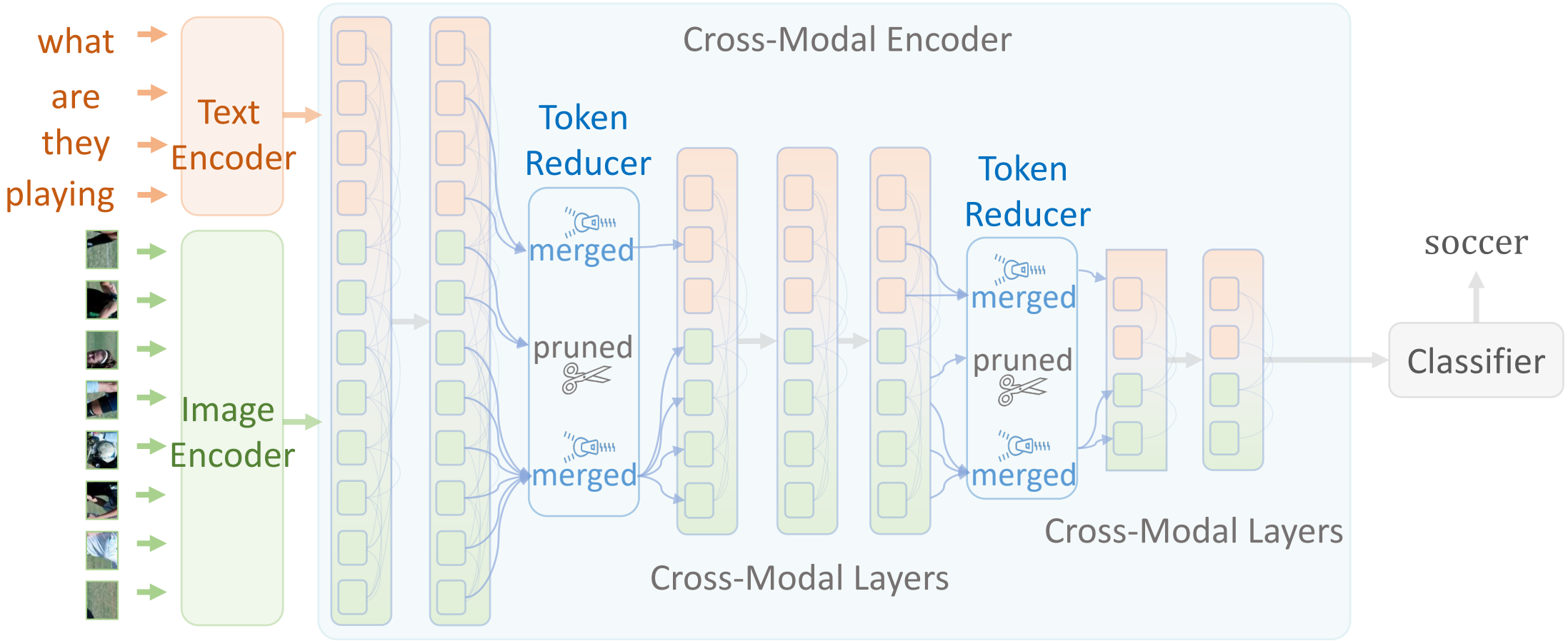
# Technique 2: Modality-Aware Merging



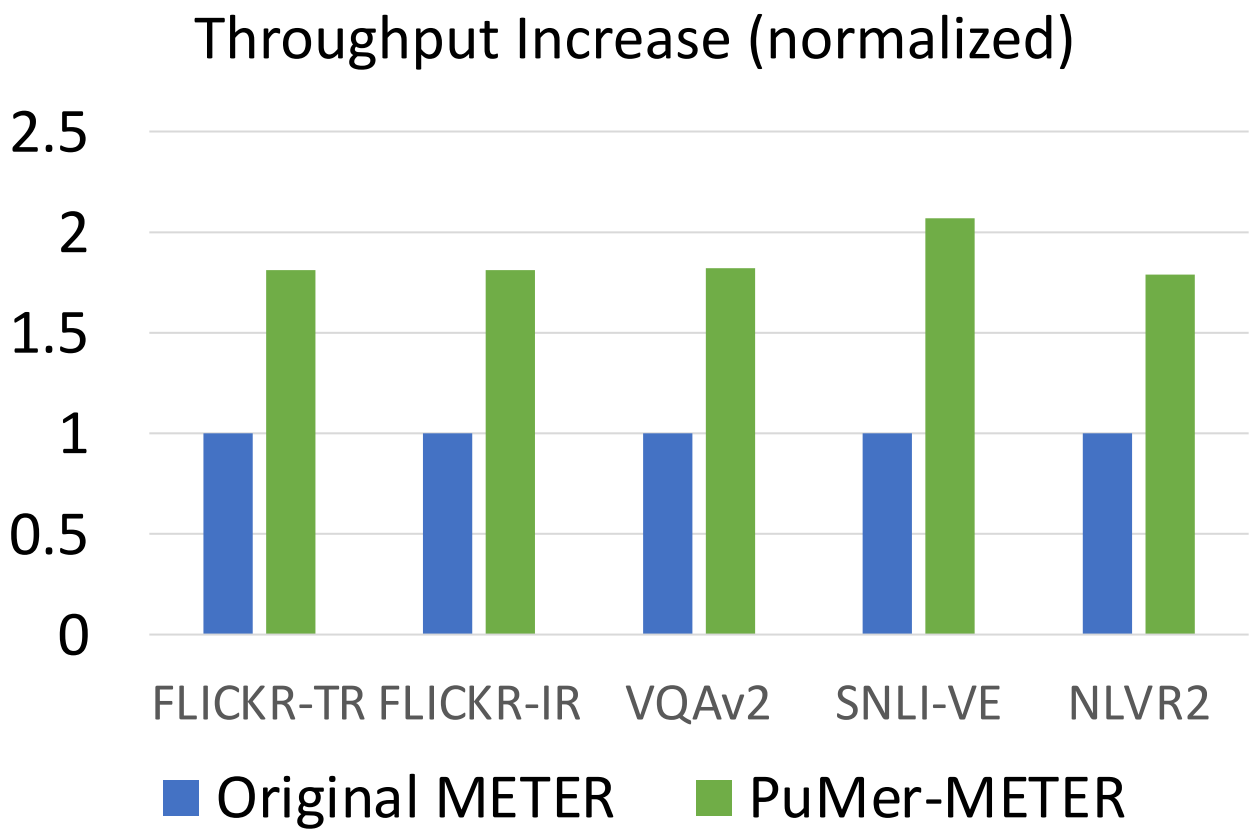

Merging image modality tokens via bipartite matching

Similar process for merging text tokens

# PuMer: Text-Informed Image Pruning and Modality-Aware Merging

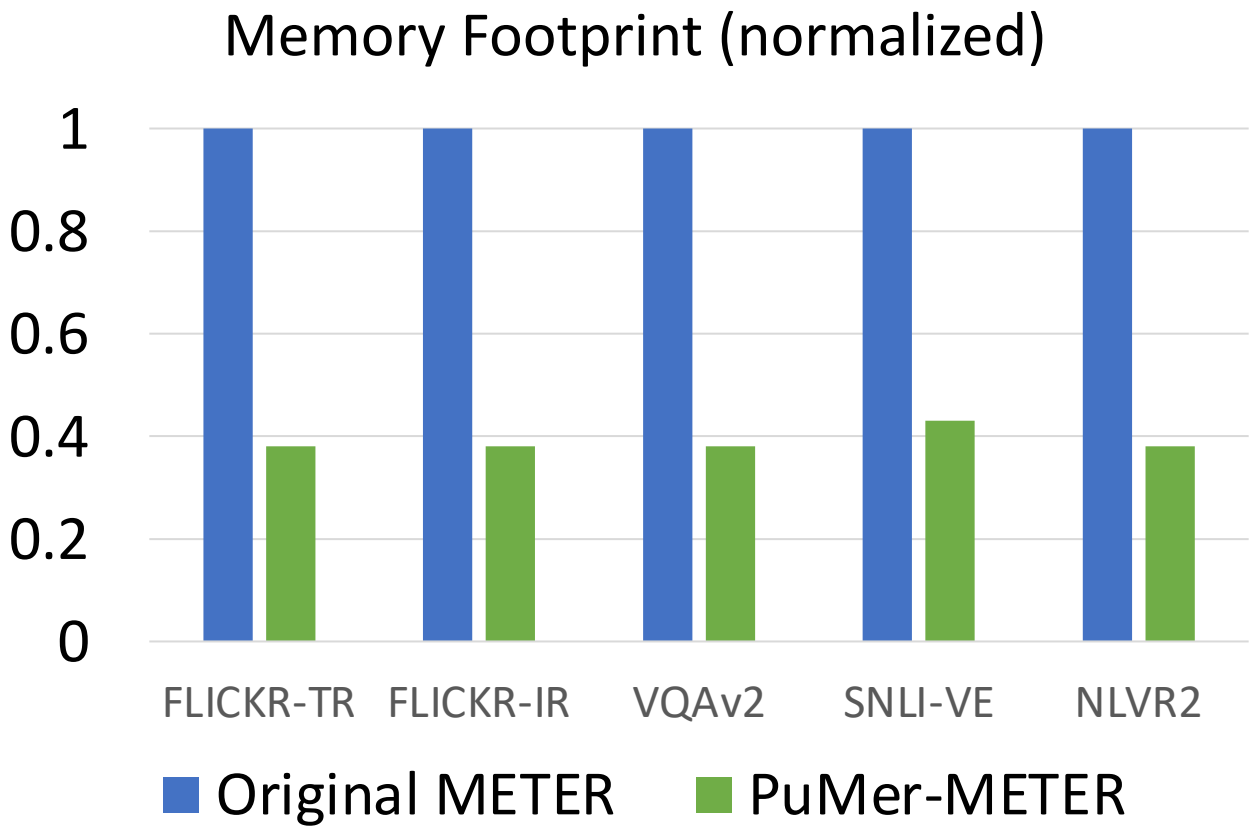


# PuMer Improves Throughput of VL Models



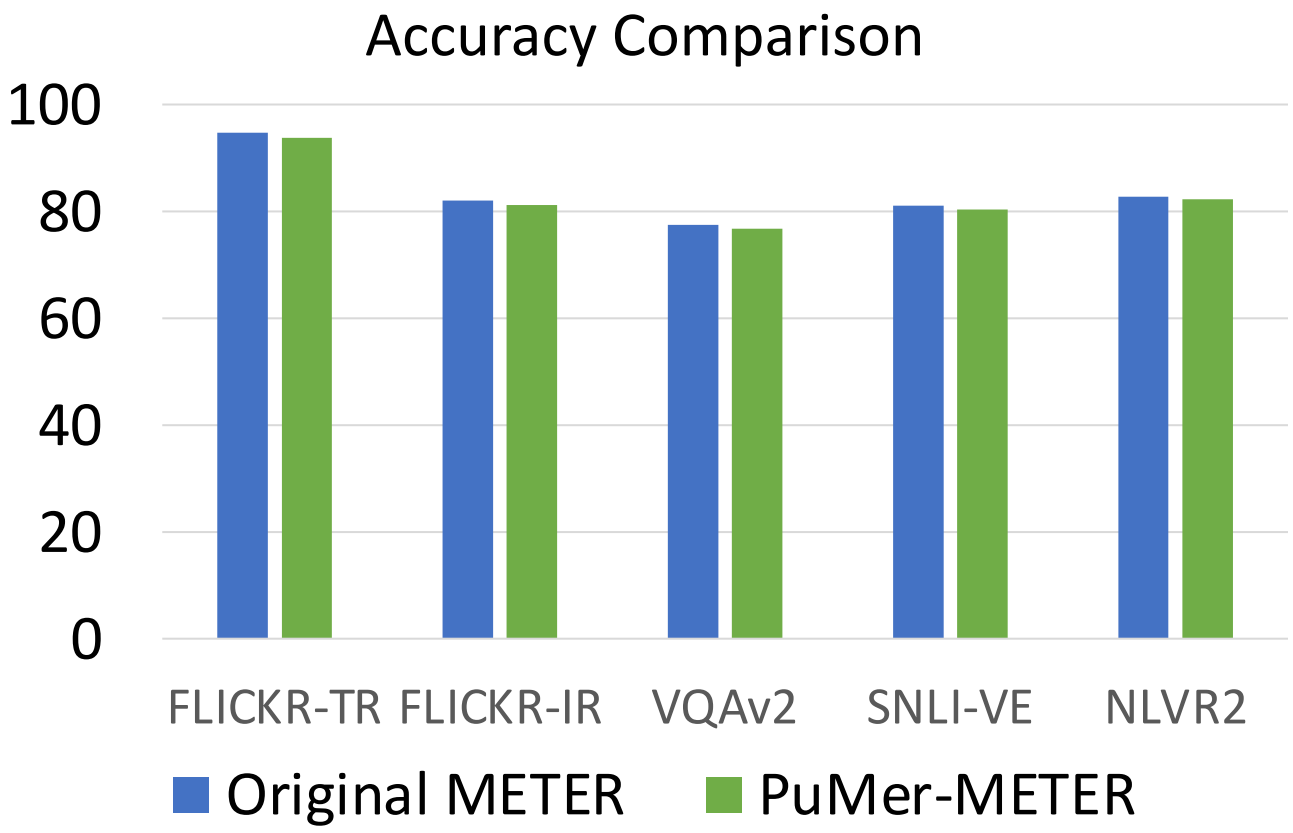

PuMer brings **1.8 ~ 2x** inference throughput increase for SoTA VL models

# PuMer Reduces Inference Memory Footprint




PuMer reduces **38% ~ 43%** inference memory footprint for SoTA VL models

# PuMer Only Incurs Minimal Accuracy Drop




PuMer causes **<1%** accuracy drop for SoTA VL models over all studied VL tasks

# Summary

We present PuMer that uses a set of token reducers to improve the inference efficiency of vision-language models

We design text-informed image pruning and modality-aware token merging in token reducers to remove and merge input tokens

PuMer effectively improves inference throughput and reduces memory footprint of VL models with minimal accuracy drop