iLLaVA: An Image is Worth Fewer Than 1/3 Input Tokens in Large Multimodal Models

An Efficient Inference Framework for Jointly Optimizing Image Encoder and LLM

Lianyu Hu, Liqing Gao, Fanhua Shang, Liang Wan*, Wei Feng

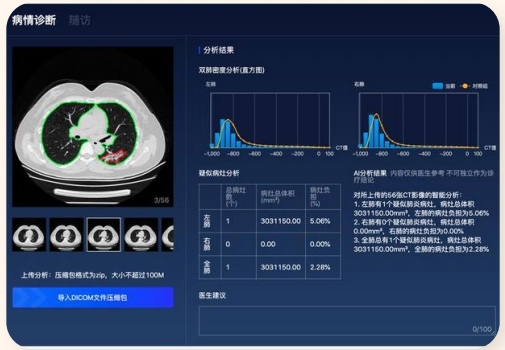
Tianjin University, Tiangong University

ICLR 2026

CONTENTS

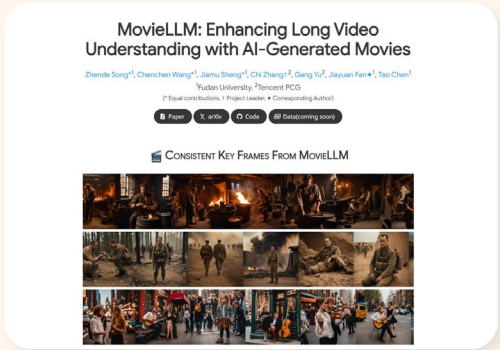
01. Background: Success and Challenges of LVLMs

Great Success: Breakthroughs and Applications



Medical Imaging Analysis

Assisting doctors in precise lesion identification



Long Video Understanding

Deeply analyzing content and generating summaries

Additionally, they demonstrate exceptional understanding of text-image content in cross-modal retrieval.

Severe Challenges: Computational and Efficiency Bottlenecks

High Computational Complexity

Self-attention mechanism grows $O(n^2)$, leading to a computational explosion with surging tokens.

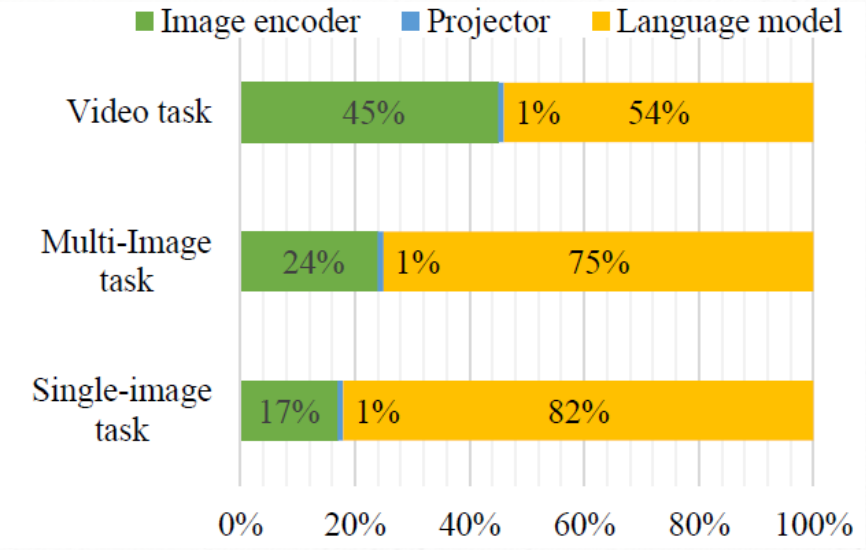
High Resource Requirements

34B parameter models require 80GB VRAM for inference, inaccessible to ordinary devices.

Severe Inference Latency

Slow response times hinder real-time interaction and high-frequency application scenarios.

01. Motivation: Overlooked Bottleneck - Image Encoder



Blind Spot of Existing Acceleration Methods

Mainstream methods focus solely on reducing tokens to the LLM, ignoring the computational cost of the image encoder itself.

Image Encoder: A Key Bottleneck

The image encoder contributes significantly to total inference time (up to 45% in video tasks) and is the primary source of tokens. Optimizing it yields dual benefits.

Core Insight: Optimizing the image encoder is the key to breaking through the efficiency bottleneck of LVLMs.

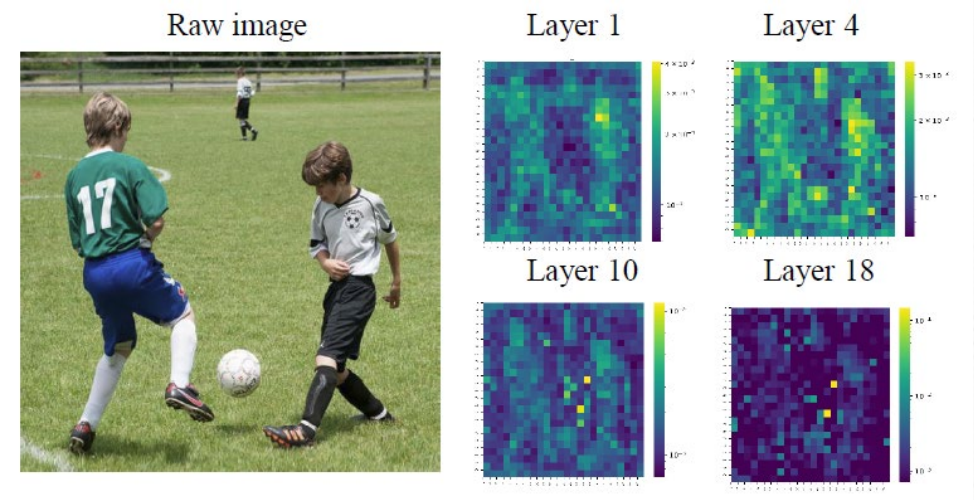
Insight: Visual Redundancy and Acceleration Opportunity

Visual Redundancy: An Intrinsic Feature

- Attention visualization reveals that the model focuses on a small portion of key regions, ignoring extensive background areas. This redundancy is unevenly distributed.

Key Opportunity: Trading Redundancy for Performance

By eliminating this inherent visual redundancy, we can achieve comprehensive end-to-end acceleration of LVLMs.



Experimental Data Visualization (ViT Model)
Left: Attention heatmap. Right: Component time consumption. Clearly shows the model focuses only on local image areas (e.g., the person on the left), with high time cost in the image encoder, confirming that "removing redundancy accelerates speed".

Core Conclusion: Leveraging visual redundancy is the key to achieving full acceleration of LVLMs.

02. Solution: The iLLaVA Framework

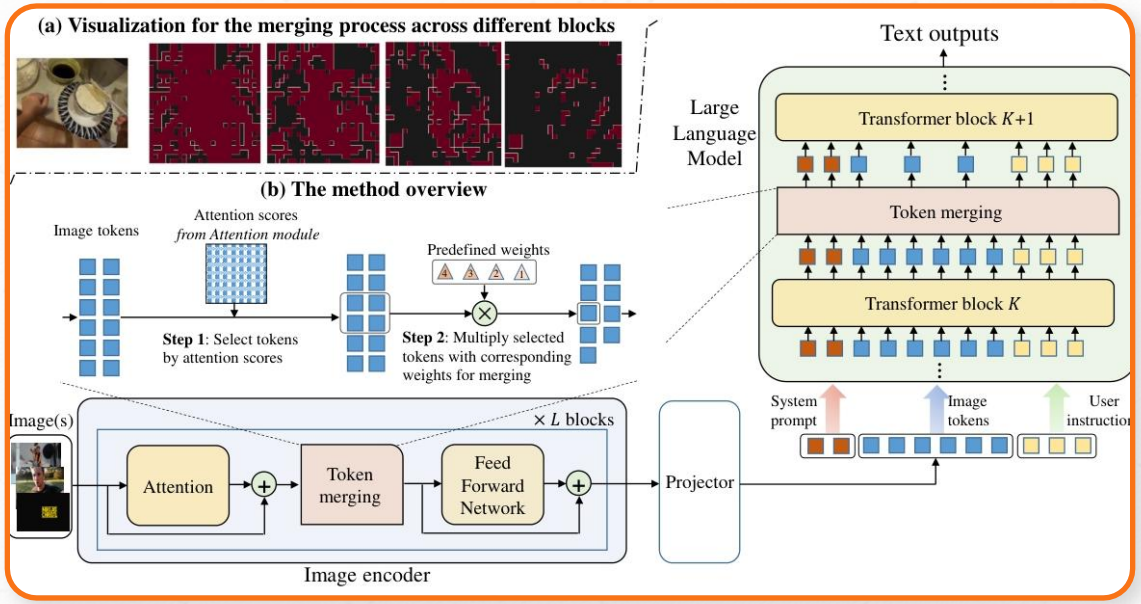


Figure: Overview of the iLLaVA Two-Stage Token Merging Framework

Core Advantage: 40%+ lower computation, >95% info retention

Core Idea: Joint Optimization for Full Acceleration

- Goal: Overcome single-module bottlenecks by jointly optimizing the image encoder and LLM
- Strategy: Intelligently allocate compute budget, reducing tokens without quality loss

Framework Design: Two-Stage Token Merging

Stage 1 (Image Encoder): Early pruning of visual tokens to reduce computational load

Stage 2 (LLM): Further merging in middle layers to optimize inference speed

Summary: Through two-stage token merging, iLLaVA systematically reduces the computational load of the entire model, achieving effi cient inference.

02. Token Merging Strategy: Information Preservation and Recycling

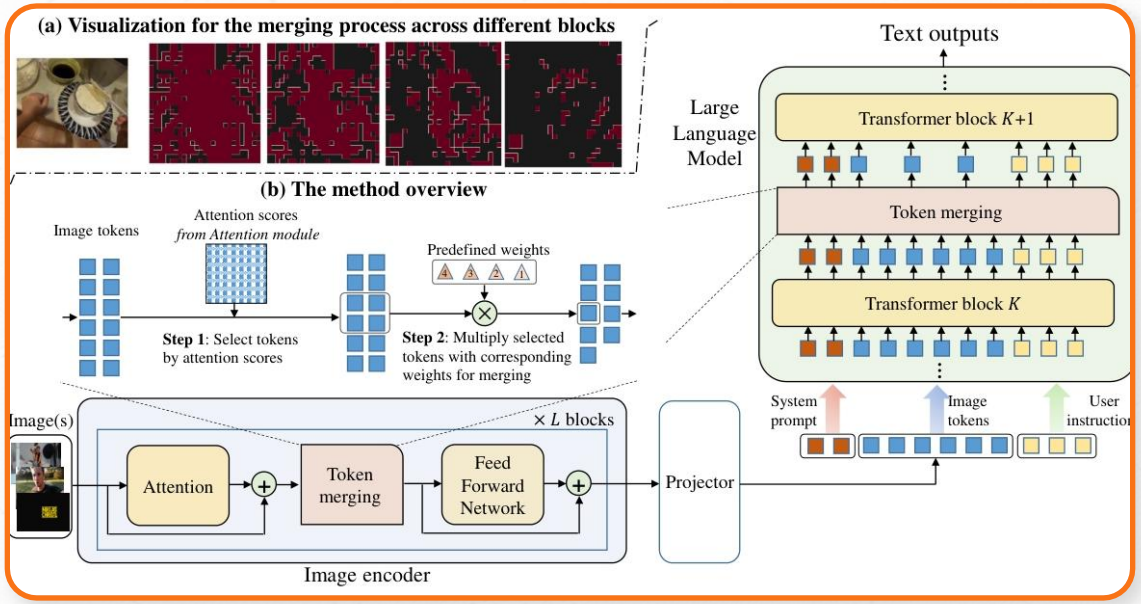


Figure: Visualization of Token selection (red) and discard (black)

Precision Screening: Selecting Informative Tokens

Using attention scores as a metric for token importance, the most critical tokens are directly preserved.

Intelligent Recycling: Reclaiming Useful Information

Information from discarded tokens is aggregated into a few "recycled tokens," maximizing information retention.

Core Value: While drastically reducing tokens, the "dual-mechanism" of preservation and recycling effectively mitigates performance degradation.

03. Experiment

Backbone

Using Qwen2.5-VL 7B by default，validated on multiple backbones。

Qwen2.5-VL

Comparison

Comparison with SOTA methods：

- FasterVLM
- PyramidDrop
- SparseVLM
- VisionZip

Evaluation

Evluation upon 10+ benchmarks：

- MMMU / MMBench
- VideoMME
- MVBench
- MLVU

Effi ciency

- Accuracy
- Throughput
- Memory
- Pefilling time

iLLaVA improves efficiency while maintaining performance of state-of-the-art (SOTA) models

03. Key Result 1: Performance Preservation and Surpassir

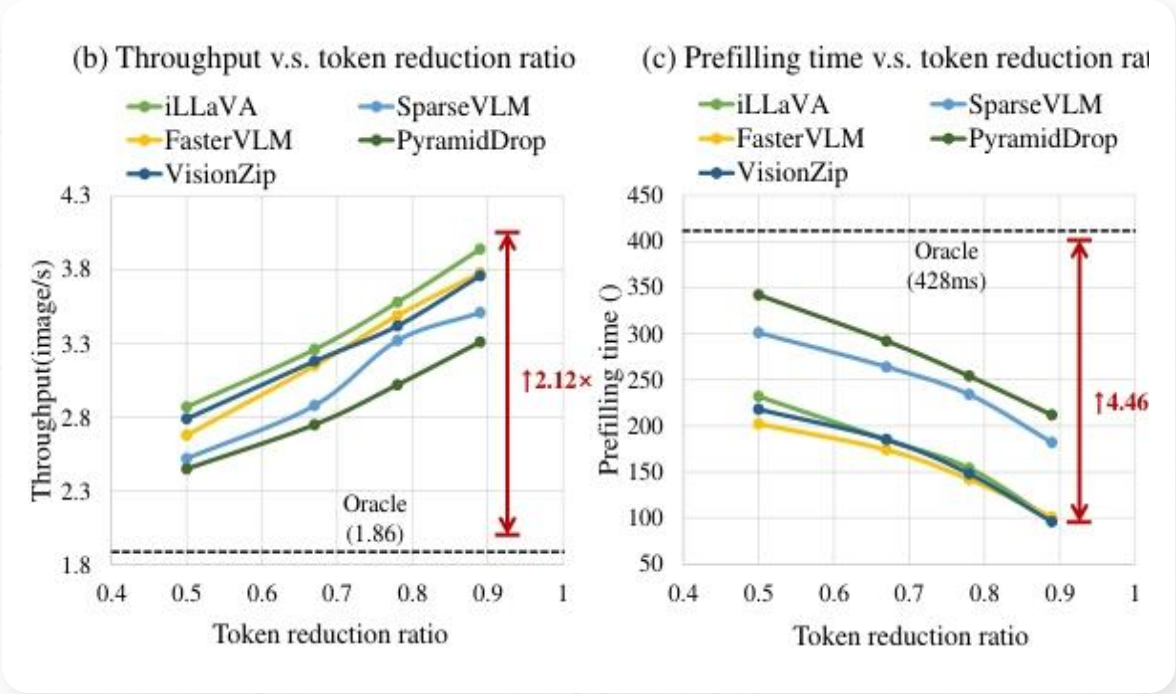


Fig 4d: Comparison of accuracy trends with token reduction (iLLaVA is optimal)

Image Tasks: Drastic Compression

Token reduction 88.9%

Performance retention 95.2%

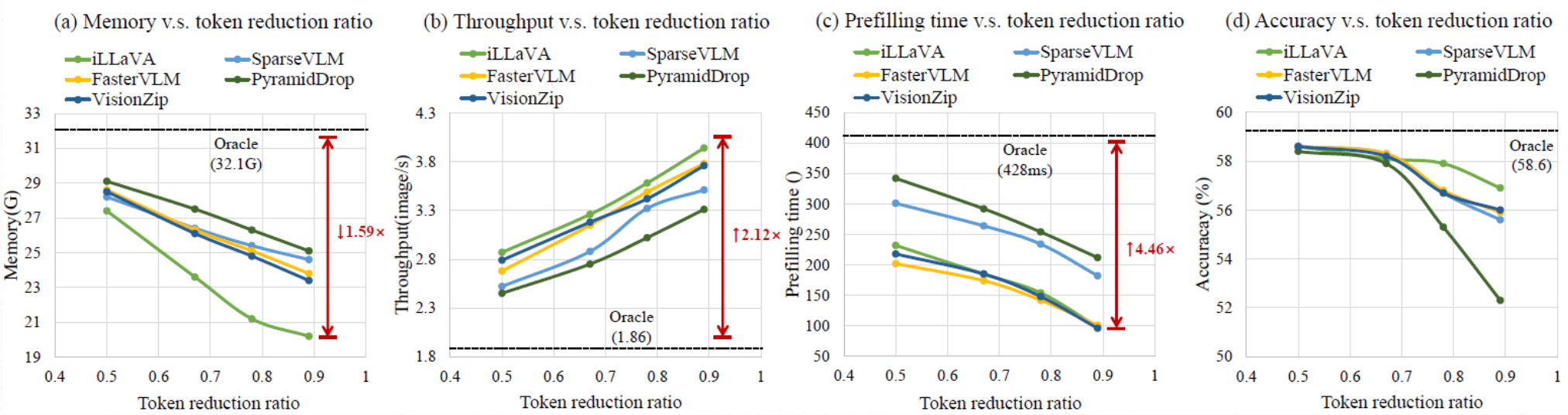
Video Tasks: Performance Leap

Token reduction 95%

Outperforms SOTA +1.7%

Core Conclusion:iLLaVA's token merging strategy successfully preserves the model's core capabilities while significantly reducing computation, demonstrating superior performance in both image and video tasks.

03. Key Result 2: Significant Efficiency Improvement



Throughput 2.12X increase.

Memory 0.64X reduction .

Prefilling 4.46X reduction.

Experimental Data Visualization

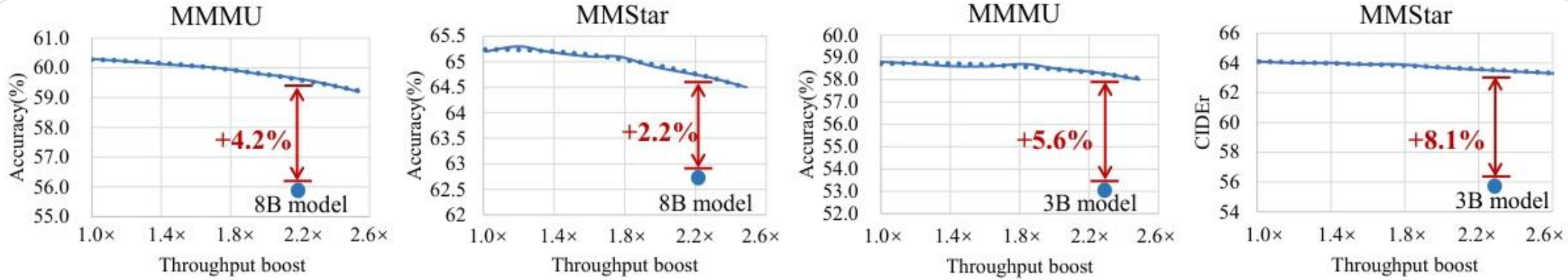
The chart illustrates iLLaVA's (blue line) significant advantages over baselines like SparseLM in memory, throughput, and latency across different token reduction ratios.

Dual breakthroughs in performance and efficiency validate architectural superiority

Core Conclusion: iLLaVA achieves an exceptional balance between performance and efficiency, realizing true end-to-end acceleration.

03. Key Result 3: The 'Dimension Reduction Strike' of Larger Mode

Core Finding: iLLaVA enables larger models tosurpass smaller ones in both precision and effi ciency.



(a) InternVL 2.5 26B (w/ iLLaVA) v.s. InternVL 2.5 8B

(b) Qwen2.5-VL 7B (w/ iLLaVA) v.s. Qwen2.5-VL 3B

InternVL-2.5 26B (iLLaVA)
Throughput > 8B model | Accuracy+4.2%(MMMU)

Qwen2.5-VL 7B (iLLaVA)
Throughput > 3B model | Accuracy+8.1%(MMStar)

Breaks the perception that 'larger models are always slower,' making the deployment of more powerful models on resource-constrained devices a reality.

03. Key Result 4: Flexibility and Generalization

Cross-Model Adaptability

iLLaVA can be successfully applied to LVLMs of different architectures, demonstrating consistent performance advantages.

Experimental Data: Multi-model Average Accuracy Comparison

Model	iLLaVA (Avg. Acc)	Comparison Method (Avg. Acc)
LLaVA-Onevision	68.5	66.2
InternVL-2.5	69.8	67.5
MiniCPM-V	67.1	65.3
Qwen2.5-VL	70.2	68.1

Conclusion: iLLaVA is a versatile, plug-and-play acceleration framework with excellent flexibility and scalability.

03. Visualization Analysis: Validating iLLaVA's Token Selection Strategy

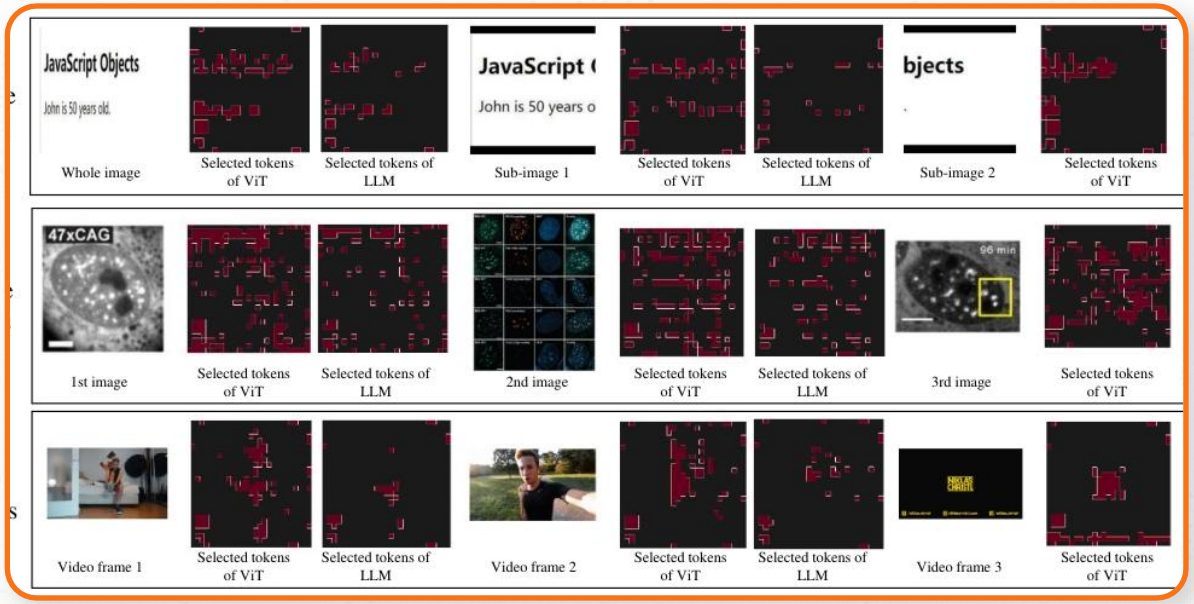


Figure 5: Token Merging Process Visualization (Single/Multi-image/Video Tasks)

Intuitively Insight into Model Attention Flow

Process Visualization

Figure 5 shows the token regions selected by the image encoder and LLM for single-image, multi-image, and video tasks.

Key Observations

The model accurately focuses on key regions. In multi-image tasks, the LLM allocates more tokens to images closer to the output text.

Conclusion

Visualization results intuitively confirm the rationality and effectiveness of iLLaVA's token selection strategy.

THANKS FOR LISTENING

Thanks for watching

Hope have a better future