NEURAL INFORMATION PROCESSING SYSTEMS

Qualcomm

HANYANG UNIVERSITY

# InfiniPot-V: Memory-Constrained KV Cache Compression for Streaming Video Understanding

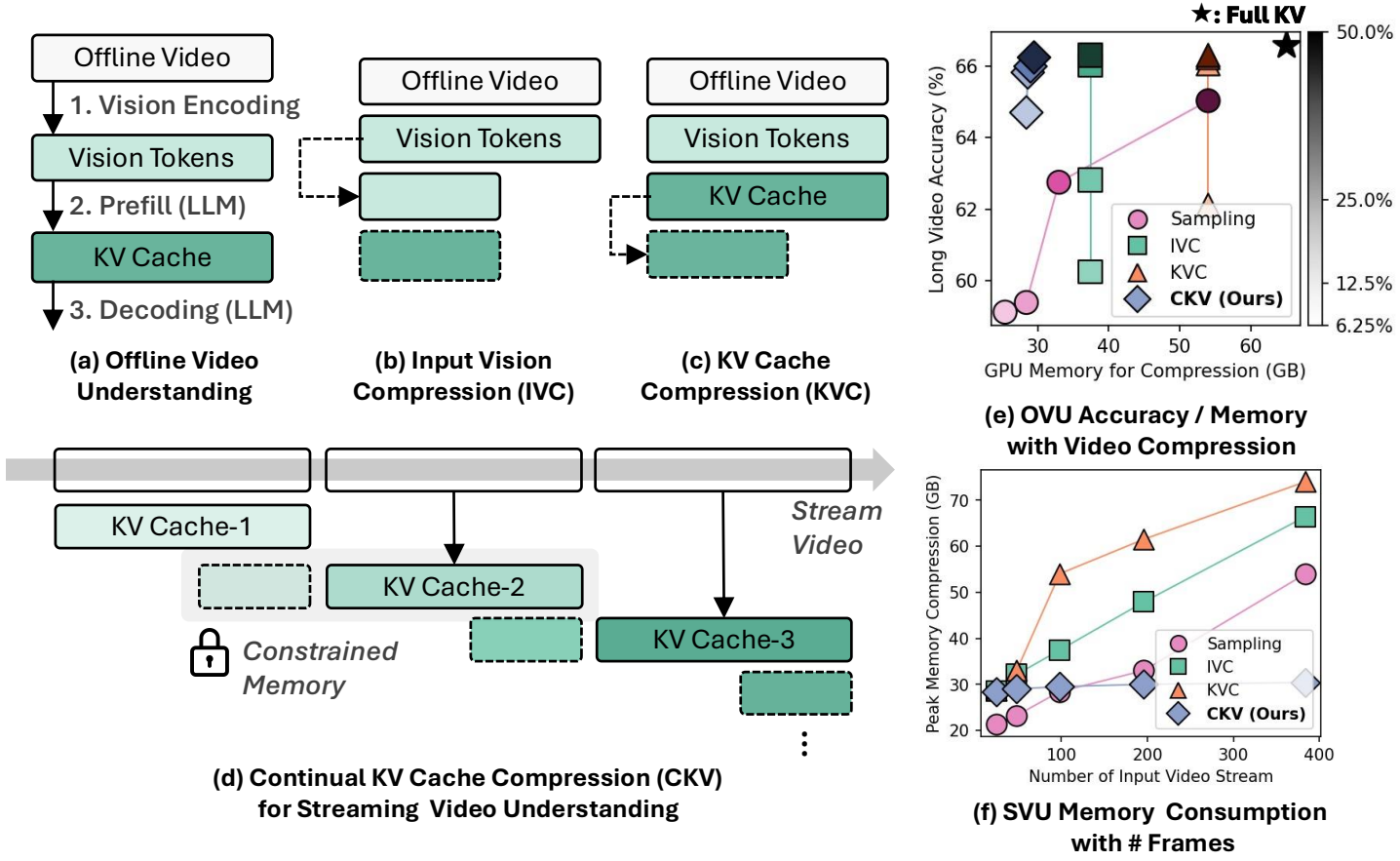
**Minsoo Kim**, Kyuhong Shim, Jungwook Choi, Simyung Chang
Qualcomm AI Research*, Hanyang University

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Paper available at https://arxiv.org/abs/2506.15745

# Motivation

- Memory-constrained continual KV cache compression for processing input video stream







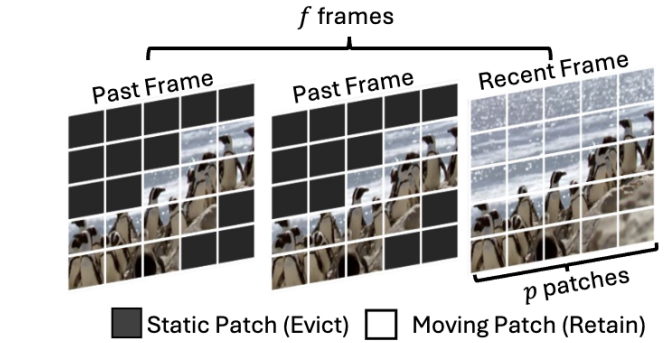
(e) OVU Accuracy / Memory with Video Compression



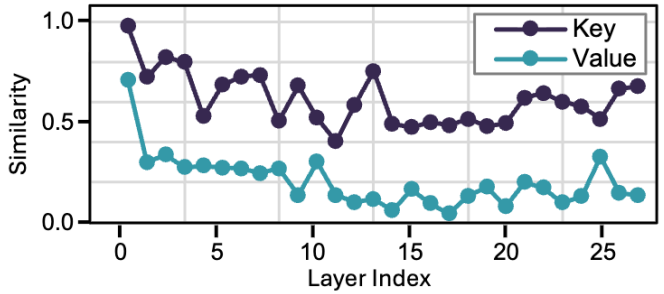

(f) SVU Memory Consumption with # Frames

# Method

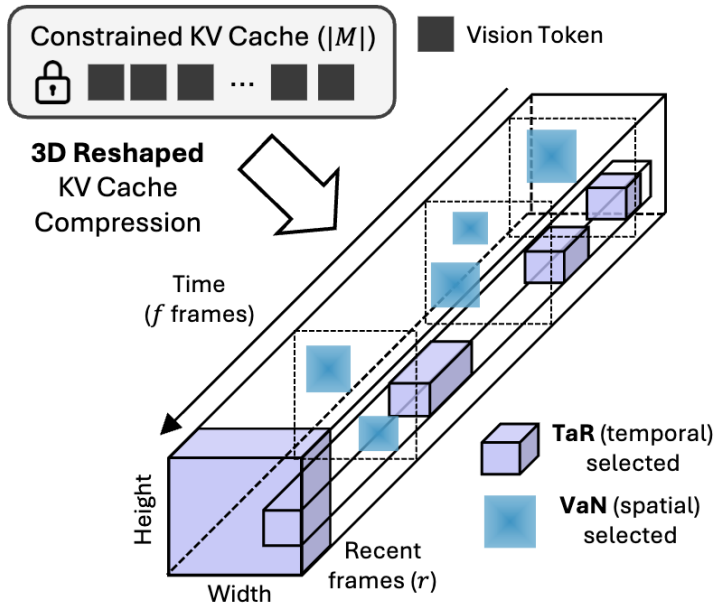
- **Temporal-axis Redundancy Reduction**
- Spatial Semantic Importance Preserving




**(a) Temporal-axis Redundancy Reduction (TaR)**



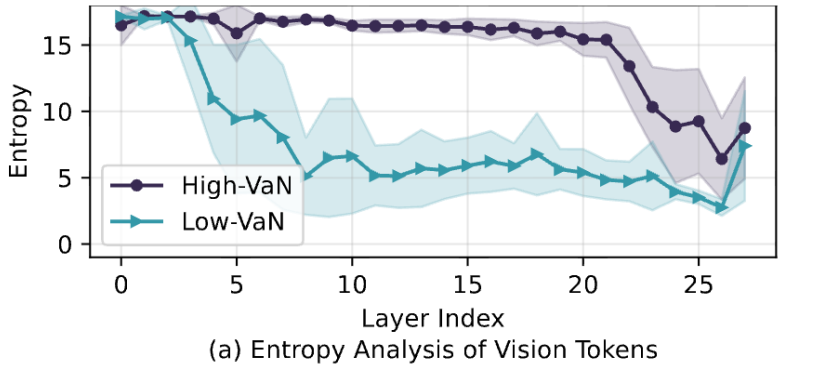

**(b) Static Patch Similarity Comparison**



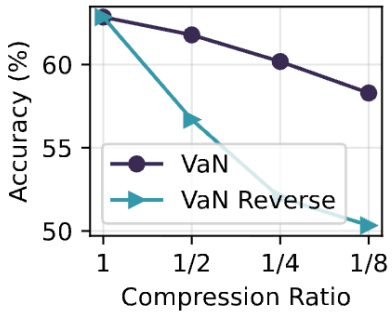

**(c) InfiniPot-V (Ours): Temporal-Spatial Scoring**

# Method

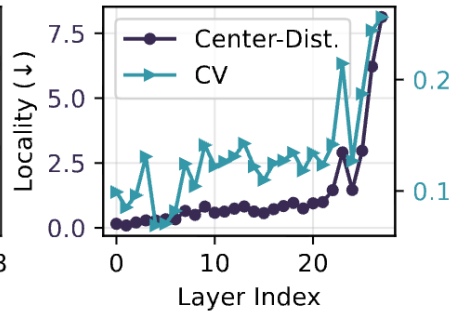
- Temporal-axis Redundancy Reduction
- **Spatial Semantic Importance Preserving**



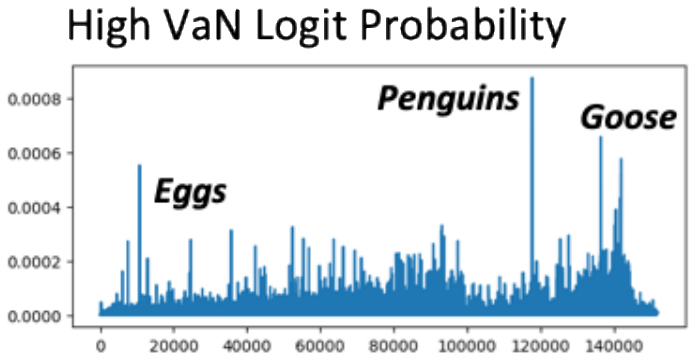
(a) Entropy Analysis of Vision Tokens

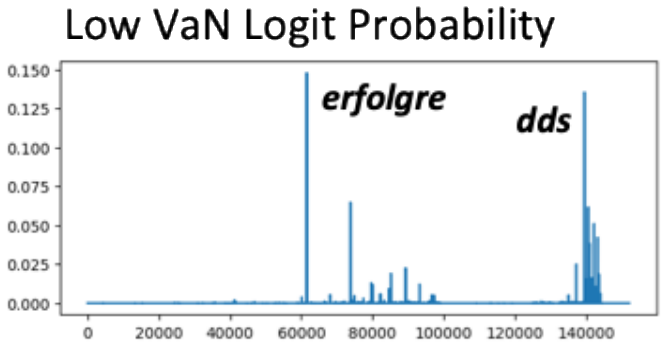


(b) VideoMME Accuracy



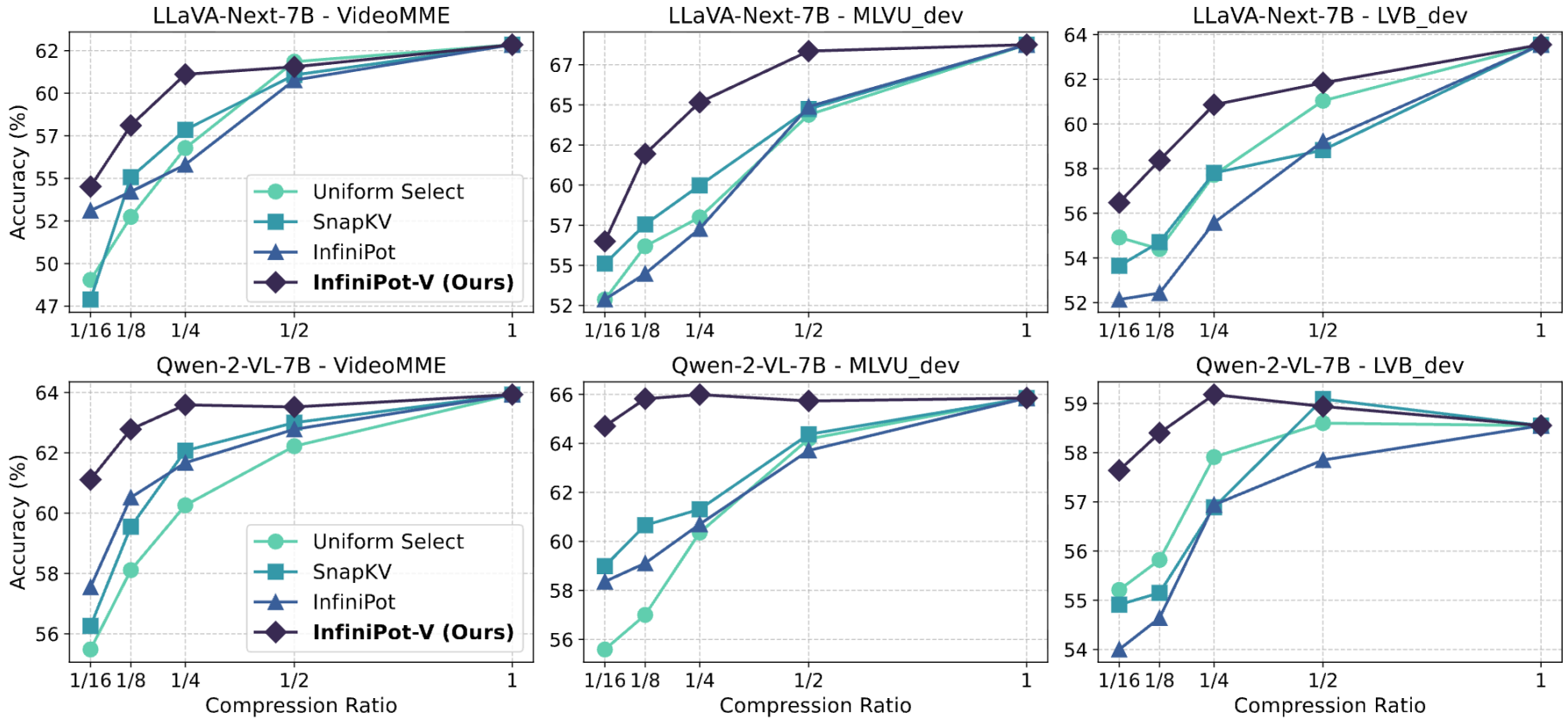
(c) Spatial Locality Analysis



# Offline Video Understanding

- LLaVA-Next-7B, Qwen-2-VL-7B
- VideoMME, MLVU, LongVideoBench



# Streaming Video Understanding

Qualcomm
HANYANG UNIVERSITY

- LLaVA-OV-7B, Qwen-2.5-VL-7B
- RVS-Ego/Movie, OVO-Bench, StreamingBench

| | RVS-Ego | | RVS-Movie | | Execution Time | Total Memory Usage | |
|---|---|---|---|---|---|---|---|
| **LLaVA-OV-7B** | **Acc** | **Score** | **Acc** | **Score** | **Video Enc. (msec/Frame)** | **GPU** | **CPU** |
| ReKV | 60.1 | 3.9 | 53.4 | 3.8 | 285.7 | 37.5 GB | + 18.8GB/h |
| ReKV w/o off. | 55.8 | 3.3 | 50.8 | 3.4 | 74.6 | 27.2 GB | 0 |
| InfiniPot-V | **57.9** | **3.5** | **51.4** | **3.5** | 76.3 | 27.8 GB | 0 |

| **Qwen-2.5-VL** | **Scale** | **OVO-BW** | **OVO-Real** | **OVO-FW** | **OVO-Avg** | **StreamingBench** |
|---|---|---|---|---|---|---|
| Uniform Select | 7B | 44.5 | 61.1 | 47.8 | 51.7 | 75.2 |
| InfiniPot-V | 7B | 47.6 | 65.9 | 47.9 | **53.6** | **76.4** |

# Edge-Device Deployment Results

Qualcomm
HANYANG UNIVERSITY

- NVIDIA Jetson AGX Orin

| Method | Length (sec) | Context Length | VRAM (GB) ↓ | FPS (frame/sec) ↑ | Throughput (tok/sec) ↑ | Power (J) ↓ |
|---|---|---|---|---|---|---|
| Full KV | 100 | 18K | 13.8 | 6.2 | 5.0 | 1.6 |
| InfiniPot-V | | | 9.2 (1.5×) | 6.4 (1.0×) | 9.2 (1.8×) | 1.2 (1.4×) |
| Full KV | 300 | 54K | 26.6 | 5.0 | 2.0 | 4.7 |
| InfiniPot-V | | | 10.1 (2.6×) | 6.4 (1.3×) | 9.2 (4.6×) | 2.4 (2.0×) |
| Full KV | 500 | 90K | 39.0 | 4.1 | 1.2 | 7.5 |
| InfiniPot-V | | | 10.7 (3.6×) | 6.3 (1.5×) | 9.1 (7.3×) | 3.7 (2.0×) |
| Full KV | 600 | 108K | OOM | OOM | OOM | OOM |
| InfiniPot-V | | | 11.3 | 6.4 | 9.2 | 4.3 |

# See you at Poster Session!

Qualcomm
HANYANG UNIVERSITY

- **Thu 4 Dec 11am – 2 pm**
- **Exhibit Hall C,D,E**
- Any Questions?
  - Please contact minsoo2333@hanyag.ac.kr
- Chat?
  - https://neurips.cc/virtual/2025/loc/san-diego/poster/116667