NEURAL INFORMATION PROCESSING SYSTEMS

Scan for the code

# FrameSelect

## A Unified Library for Video Frame Selection and Evaluation

**Arihant Jain**[1] **Chaosheng Dong**[2]

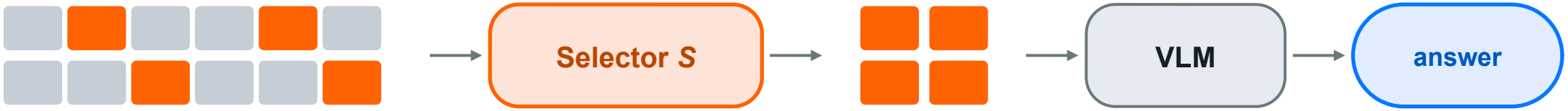
[1]Amazon [2]Ex-Amazon

**NeurIPS 2026** · **Sydney** · Evaluations & Datasets Track

amazon science

BACKGROUND

# Frame selection sits at the front of every video-VLM




candidate pool, $N$ frames | given query $q$ | $\leq K$ frames | answers $q$ | accuracy

$$F_S = S(F_C, q, K) \qquad |F_S| \leq K \ll N$$

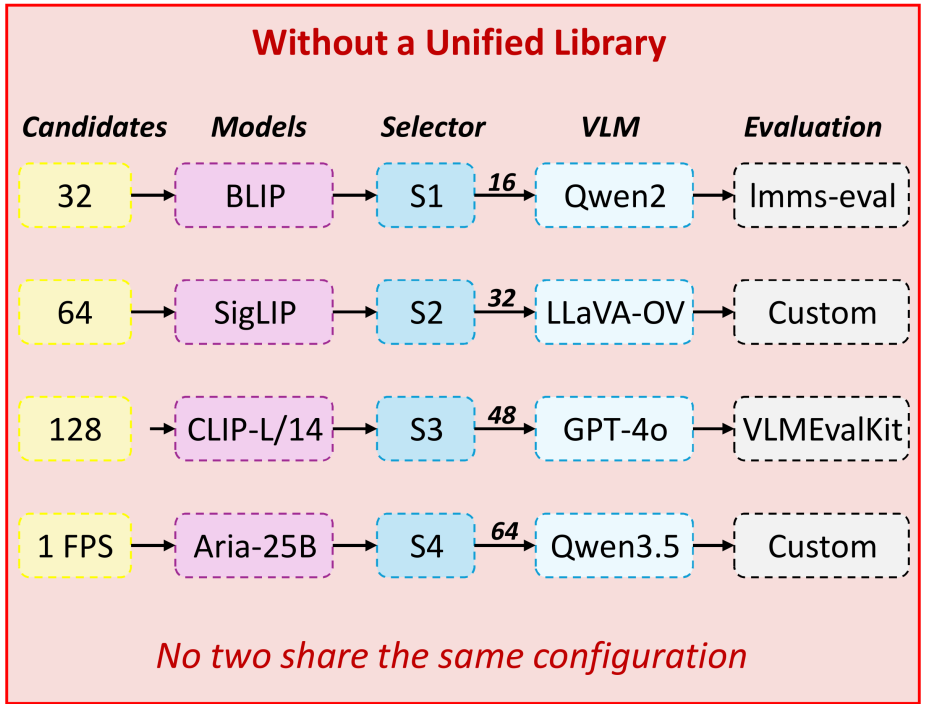
- From a candidate pool $F_C$ of $N$ frames, a **frame selector** keeps at most $K$.
- Good selection targets the key moments, cuts noise and redundancy, fits the context budget and controls cost.

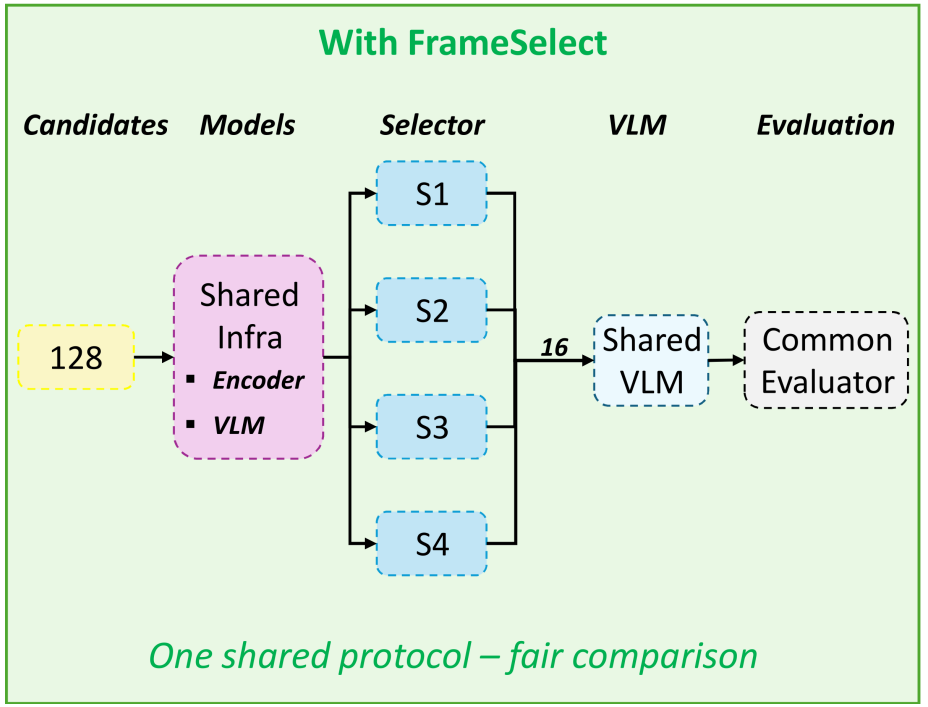
**Answer-level:** does the VLM answer correctly given the selected frames?

**Selection-level:** do the selected frames match human-annotated keyframes or scenes, with no VLM in the loop?

PROBLEM

# The problem: published gains are hard to compare

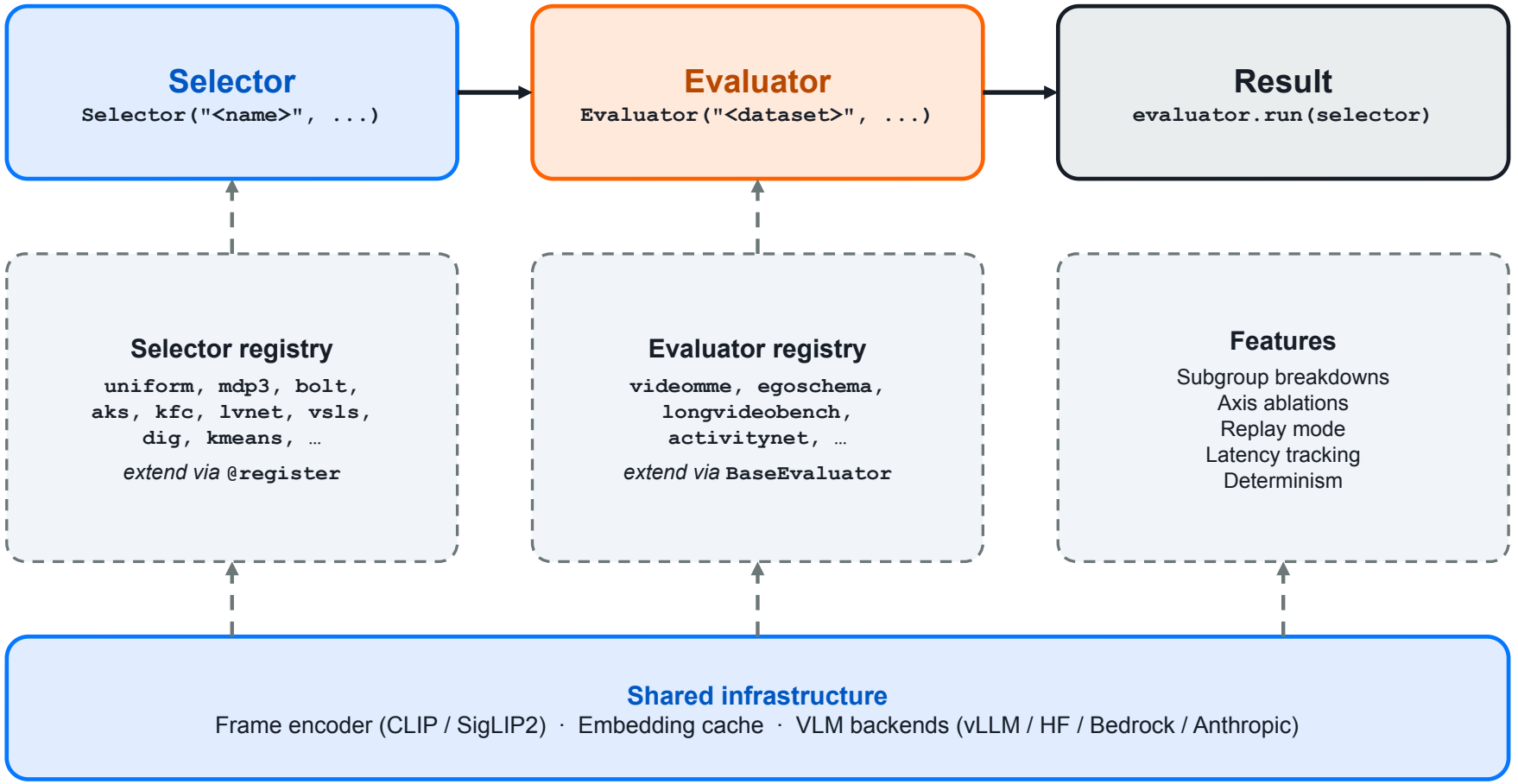



- Each new selector reports gains over uniform sampling, but **most works do not share the same combination of candidate pool, encoder, frame budget and answering VLM**.
- No shared library, API or evaluation protocol exists on which those gains could be compared.

LIBRARY

# FrameSelect at a glance




Selection and answering are decoupled: **two objects and one call** compose any registered selector with any registered evaluator.

LIBRARY

# Two objects, one call

example.py

```python
import frameselect as fs

sel = fs.Selector("kmeans", n_frames=16,
                  candidate_frames=128,
                  encoder="ViT-SO400M-16-SigLIP2-384",
                  encoder_pretrained="webli")
r = sel.select("video.mp4")
r.frames, r.indices, r.all_scores

ev = fs.Evaluator("activitynet",
       video_dir="ActivityNet/videos",
       vlm="openai:Qwen/Qwen2.5-VL-7B-Instruct")
res = ev.run(sel)
res.accuracy, res.subgroup_accuracy
```

- Every axis that matters for a fair comparison is a **keyword argument**.
- **Selector**: candidate pool, encoder, frame budget (and a selection-time VLM for generative methods).
- **Evaluator**: dataset and answering VLM, one prompt contract per evaluator.
- Swapping any one of them is a **one-line change**.

RESULTS

# Answer-level results: 19 selectors under one protocol

| | Method | VMME | LVB | Ego | Hay | ANet | Mean |
|---|---|---|---|---|---|---|---|
| Static | Uniform | 58.07 | 48.92 | **54.60** | 47.95 | **95.80** | **9.20** |
| | Katna | 57.81 | 47.05 | 54.20 | 47.08 | 95.10 | **12.20** |
| Representation | K-Means | 58.74 | 48.02 | 51.60 | 47.66 | 95.30 | **12.20** |
| | MaxInfo | 59.07 | 49.81 | 53.60 | 46.49 | **95.70** | **8.20** |
| | CLIP-TopK | 58.07 | **51.76** | 49.20 | 50.58 | 92.00 | **11.40** |
| | BOLT | 58.93 | 50.41 | 52.60 | 49.71 | 94.10 | **9.20** |
| | AKS | 58.07 | **51.83** | 50.00 | 51.17 | 93.20 | **10.40** |
| | F2C | 58.81 | 50.49 | **55.00** | 50.29 | 94.10 | **6.80** |
| | FOCUS | 58.44 | 49.29 | 54.00 | 48.54 | 94.90 | **9.60** |
| | **MDP³** | **59.52** | 51.46 | 53.00 | **51.75** | 95.20 | **4.60** |
| | Q-Frame | 58.93 | 50.71 | 54.20 | 50.00 | 95.10 | **5.80** |
| | KFC | **59.41** | 50.79 | 53.40 | 50.00 | 94.80 | **6.80** |
| | AdaRD-Key | **59.81** | 50.19 | 52.20 | **51.46** | 94.80 | **7.40** |
| | KTV | 58.22 | 49.81 | 52.40 | 45.03 | 93.80 | **13.00** |
| | HORNet | 52.52 | 45.18 | 52.40 | 40.06 | 92.60 | **17.40** |
| | WFS-SB | 59.04 | **51.68** | 51.60 | **52.05** | 93.70 | **8.00** |
| Generative | LVNet | 58.04 | 49.59 | 51.80 | 49.42 | 93.50 | **13.60** |
| | DIG | 58.00 | 49.29 | **54.40** | 48.25 | 94.90 | **10.40** |
| | VSLS | 57.67 | 49.07 | 53.60 | 48.25 | **95.90** | **10.60** |

## How to read

MCQ accuracy (%) at $K = 16$.

**Mean** = mean rank over the five benchmarks (1 = best).

1st 2nd 3rd per column

**VMME** Video-MME
**LVB** LongVideoBench
**Ego** EgoSchema
**Hay** Haystack-LVBench
**ANet** ActivityNet

**Top mean rank:** MDP³ 4.60, Q-Frame 5.80, F2C and KFC 6.80.

RESULTS

# What the shared protocol reveals

1 **A coherent top tier appears only once the stack is fixed.**

MDP³ (mean rank 4.60) and Q-Frame (5.80) lead; F2C and KFC tie for third (6.80); AdaRD-Key (7.40) is right behind.

2 **Rankings invert across benchmarks.**

AdaRD-Key is 1st on Video-MME but 14th on EgoSchema; AKS is 1st on LongVideoBench but 12th on Video-MME and 17th on ActivityNet.

3 **Uniform is a competitive baseline.**

Ten selectors trail it on mean rank, and it places 2nd on EgoSchema and ActivityNet.

RESULTS

# Rankings are fragile to the setup

**Axis swap (Video-MME)**

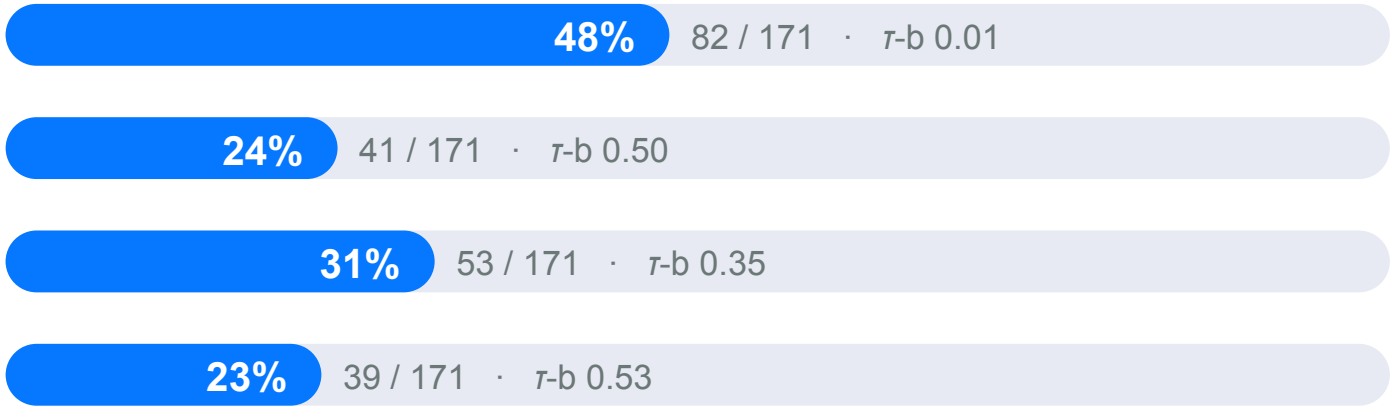
**Candidate pool**
64 ↔ 256

**Encoder**
CLIP-B/32 ↔ SigLIP2

**Output frames**
2 ↔ 16

**Answering VLM**
Haiku 4.5 ↔ Qwen2.5-VL-7B

**Selector pairs whose order flips**




- Across all single-axis swaps in the paper, **16–48% of selector pairs flip** (Kendall $\tau$-b over 171 pairs), so a single-configuration comparison can crown a different winner.
- Each swap is a one-argument change; replay mode re-scores the same selected frames under a new VLM.

SUMMARY

# Takeaways

1. FrameSelect treats the frame selector as a standalone, interchangeable component, separate from the encoder and the answering VLM.

2. Under one shared protocol, **rankings invert across benchmarks and no selector consistently dominates**.

3. Compare selectors only with the candidate pool, encoder, frame budget and answering VLM held fixed; the library makes that a one-line change.



**Try it**

github.com/amazon-science/FrameSelect

Python · CC BY-NC 4.0