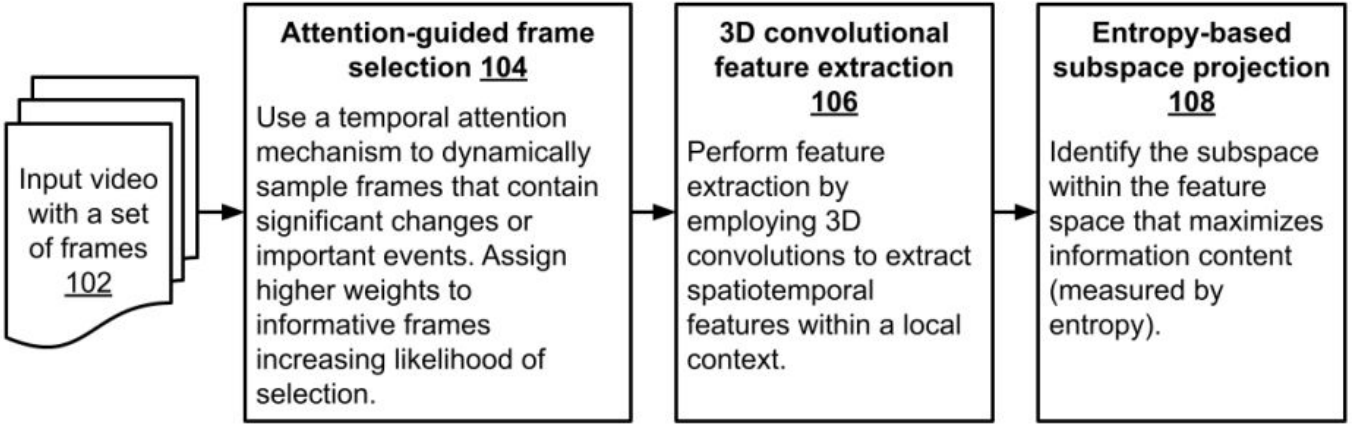
SmolVLM for Dense Video Captioning

Week 4: Internship Update

- Adaptive Thresholding in PySceneDetect
    - tested the default threshold and min_scene_len for Contentdetector and they do not work well always because _calculate_frame_score returns 0.0 for the first frame and it needs at least two frames to detect a shot which might be a problem
    - still used ContectDetector but with observing all the differences and setting the threshold in two ways:
        - threshold = mean + 2std: in the Gaussian case 95.4% of data
        - threshold = 98th percentile of differences: worked okay and extracted at least two frames

- Dynamic Frame Sampling for Multimodal Large Language Model Video Understanding



Input video with a set of frames 102

Attention-guided frame selection 104

Use a temporal attention mechanism to dynamically sample frames that contain significant changes or important events. Assign higher weights to informative frames increasing likelihood of selection.

3D convolutional feature extraction 106

Perform feature extraction by employing 3D convolutions to extract spatiotemporal features within a local context.

Entropy-based subspace projection 108

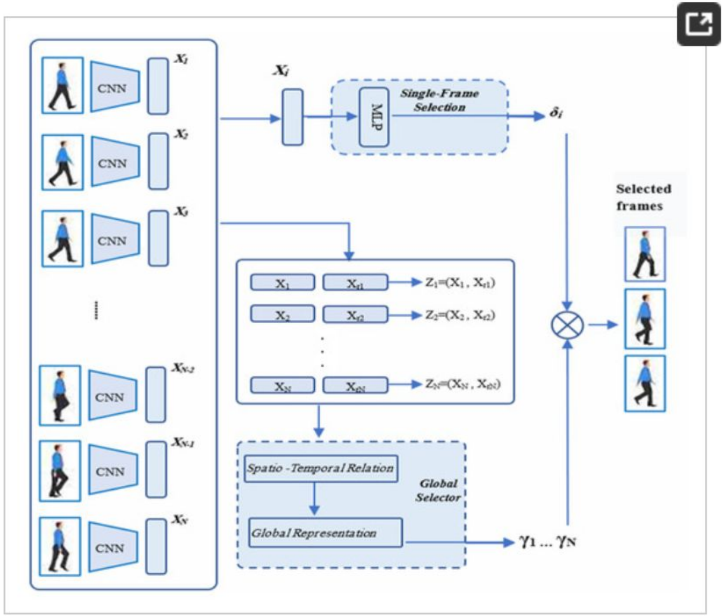
Identify the subspace within the feature space that maximizes information content (measured by entropy).

Fig. 1: Dynamic frame sampling for multimodal large language model (LLM) video understanding

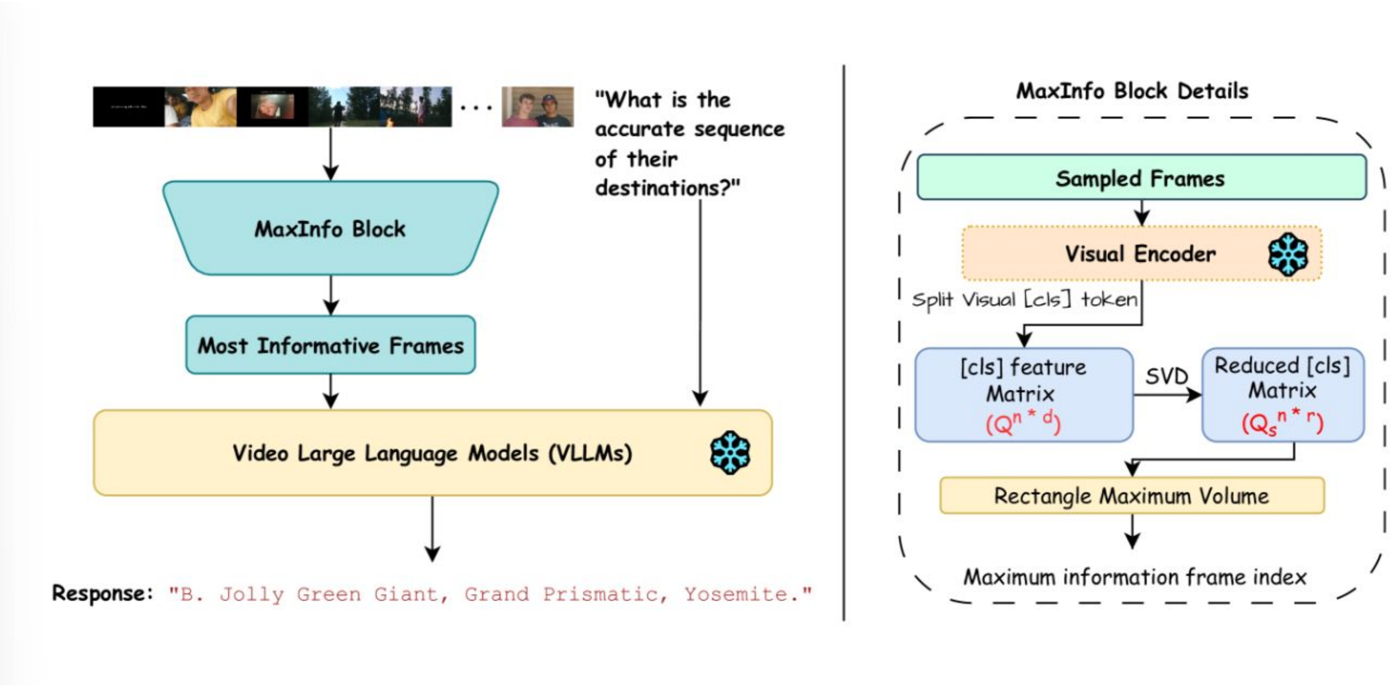
- DFS-QA: Dynamic Frame Selection for Better Video Question Answering
    - mostly for the question answering task, not very applicable

- A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection
    - Human Action Recognition task
    - SMART frame selection
    - two-dimensional convolutional restricted Boltzmann machine for spatial feature extraction
    - LSTM for temporal modeling
    - grayscale images
    - SMART Frame Selection for Action Recognition

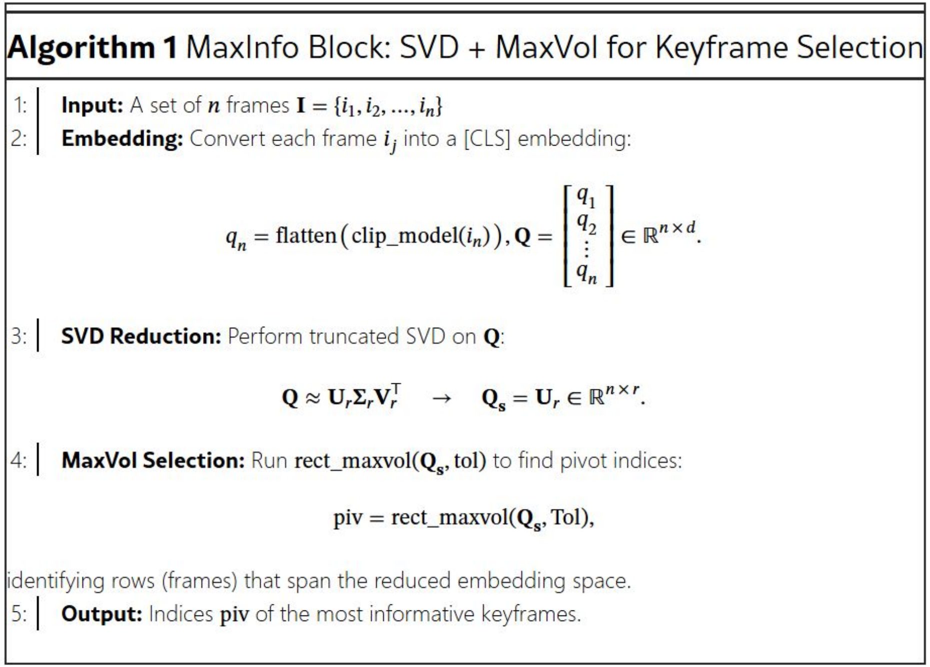
- SMART Frame Selection for Action Recognition
  - CNN: MobileNet trained on ImageNet + averaged representation of 10 most likely classes (text embedded with GloVe)
  - Single-Frame Selector: MLP BUT I would need classes
  - Global Selector: Attention model over the entire video
  - Multiplication of weights gives the final weight



- MaxInfo: A Training-Free Key-Frame Selection Method Using Maximum Volume for Enhanced Video Understanding



- MaxInfo: A Training-Free Key-Frame Selection Method Using Maximum Volume for Enhanced Video Understanding



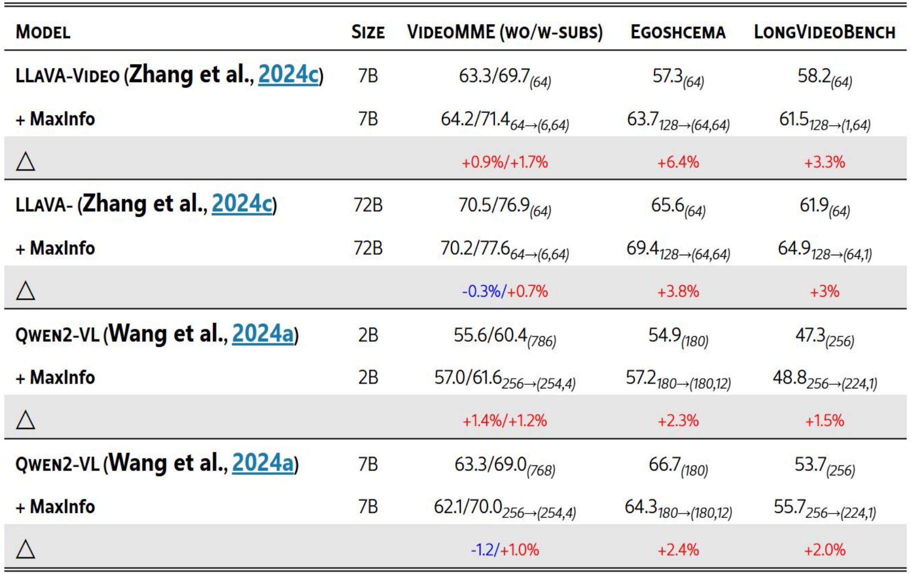
Algorithm 1 MaxInfo Block: SVD + MaxVol for Keyframe Selection

```
Input: A set of $n$ frames $\mathbf{I} = \{i_1, i_2, \ldots, i_n\}$
Embedding: Convert each frame $i_j$ into a [CLS] embedding:
    $q_n = \text{flatten}\big(\text{clip\_model}(i_n)\big), \mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} \in \mathbb{R}^{n\times d}.$
SVD Reduction: Perform truncated SVD on $\mathbf{Q}$:
    $\mathbf{Q} \approx \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^{\top} \quad \rightarrow \quad \mathbf{Q_s} = \mathbf{U}_r \in \mathbb{R}^{n\times r}.$
MaxVol Selection: Run $\text{rect\_maxvol}(\mathbf{Q_s}, \text{tol})$ to find pivot indices:
    $\text{piv} = \text{rect\_maxvol}(\mathbf{Q_s}, \text{Tol}),$
identifying rows (frames) that span the reduced embedding space.
Output: Indices piv of the most informative keyframes.
```

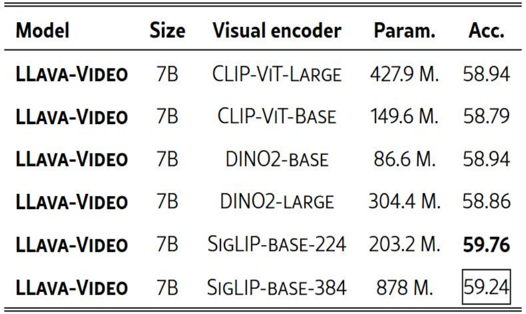
- MaxInfo: A Training-Free Key-Frame Selection Method Using Maximum Volume for Enhanced Video Understanding
    - Truncated SVD(Singular Value Decomposition)
        - Reducing feature dimension
        - Optimal s could be calculated
    - Rectangular MaxVolume Selection
    - Scene-Aware MaxInfo

$$\text{rect-vol}(\mathbf{A}) = \sqrt{\det(\mathbf{A}\,\mathbf{A}^T)} \qquad \mathbf{r} = \arg\max_{\mathbf{r}} \text{rect-vol}\big(\mathbf{Q}_s(\mathbf{r},:)\big)$$

- MaxInfo: A Training-Free Key-Frame Selection Method Using Maximum Volume for Enhanced Video Understanding



MODEL	SIZE	VIDEOMME (WO/W-SUBS)	EGOSHCEMA	LONGVIDEOBENCH
LLAVA-VIDEO (Zhang et al., 2024c)	7B	$63.3/69.7_{(64)}$	$57.3_{(64)}$	$58.2_{(64)}$
+ MaxInfo	7B	$64.2/71.4_{64\rightarrow(6,64)}$	$63.7_{128\rightarrow(64,64)}$	$61.5_{128\rightarrow(1,64)}$
△		+0.9%/+1.7%	+6.4%	+3.3%
LLAVA- (Zhang et al., 2024c)	72B	$70.5/76.9_{(64)}$	$65.6_{(64)}$	$61.9_{(64)}$
+ MaxInfo	72B	$70.2/77.6_{64\rightarrow(6,64)}$	$69.4_{128\rightarrow(64,64)}$	$64.9_{128\rightarrow(64,1)}$
△		-0.3%/+0.7%	+3.8%	+3%
QWEN2-VL (Wang et al., 2024a)	2B	$55.6/60.4_{(786)}$	$54.9_{(180)}$	$47.3_{(256)}$
+ MaxInfo	2B	$57.0/61.6_{256\rightarrow(254,4)}$	$57.2_{180\rightarrow(180,12)}$	$48.8_{256\rightarrow(224,1)}$
△		+1.4%/+1.2%	+2.3%	+1.5%
QWEN2-VL (Wang et al., 2024a)	7B	$63.3/69.0_{(768)}$	$66.7_{(180)}$	$53.7_{(256)}$
+ MaxInfo	7B	$62.1/70.0_{256\rightarrow(254,4)}$	$64.3_{180\rightarrow(180,12)}$	$55.7_{256\rightarrow(224,1)}$
△		-1.2/+1.0%	+2.4%	+2.0%



Model	Size	Visual encoder	Param.	Acc.
LLAVA-VIDEO	7B	CLIP-VIT-LARGE	427.9 M.	58.94
LLAVA-VIDEO	7B	CLIP-VIT-BASE	149.6 M.	58.79
LLAVA-VIDEO	7B	DINO2-BASE	86.6 M.	58.94
LLAVA-VIDEO	7B	DINO2-LARGE	304.4 M.	58.86
LLAVA-VIDEO	7B	SIGLIP-BASE-224	203.2 M.	**59.76**
LLAVA-VIDEO	7B	SIGLIP-BASE-384	878 M.	$\boxed{59.24}$

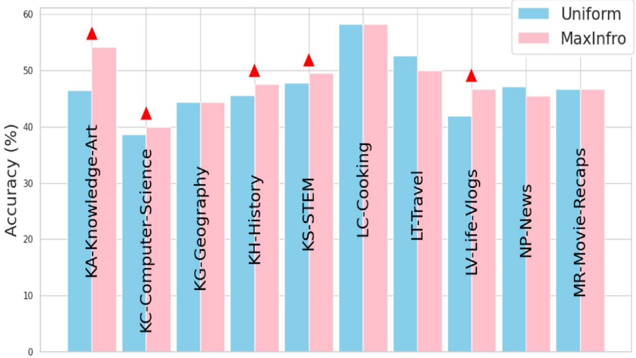


Figure 3: Accuracy comparison between Uniform Sampling and MaxInfo on Video-MME (Qwen2-VL-2B).