

Fast Video LLMs via Token Compression

2026년도 하계 세미나



Sogang University

Vision & Display Systems Lab, Dept. of Electronic Engineering



Presented By

변재현

Outline

- Background
 - Attention-based method
 - Diversity-based method
 - Hybrid-based method
- 논문 선정 이유
- FastVID: Dynamic Density Pruning for Fast Video Large Language Models
 - NeurIPS 2025
- UniComp: Rethinking Video Compression Through Informational Uniqueness
 - CVPR 2026

Background

- Video Token Compression

- Visual token을 무엇을 기준으로 남길 것인가?

- Attention-based

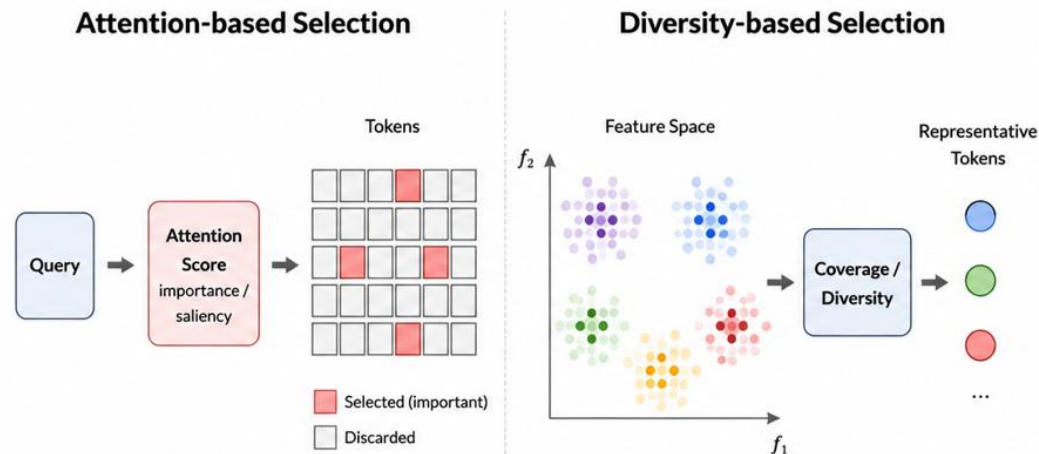
- ※ Attention score를 이용해 중요도가 높은 토큰을 선택

- Diversity-based

- ※ 유사한 토큰 간 중복을 줄이고, feature space의 coverage를 높이도록 토큰을 선택

- Hybrid-based

- ※ 토큰의 attention과 diversity를 둘 다 반영



<Attention-based vs. Diversity-based>

Background

- Video Token Compression

- Attention과 Diversity 사이의 trade-off

- Attention-based (saliency)

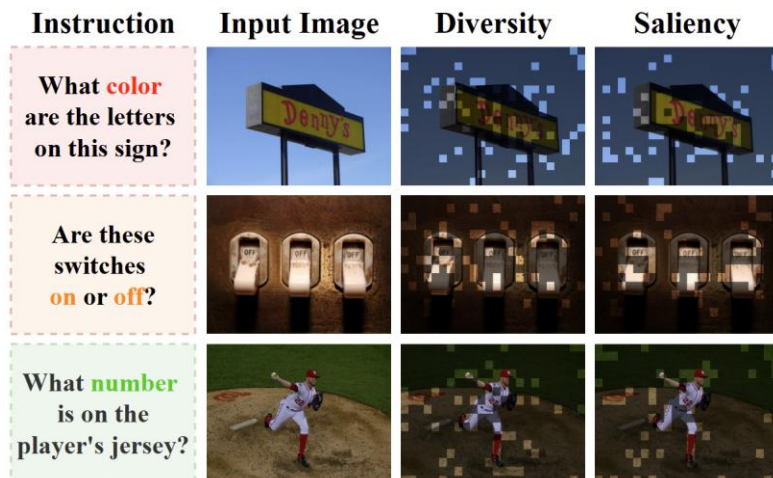
- ※ Attention score가 큰 토큰을 우선 보존하지만, 유사한 salient token이 반복 선택

- Diversity-based (feature-space coverage)

- ※ Feature space를 넓게 대표하는 토큰들을 보존하지만, 중요 단서를 놓칠 수 있음

- Video에서는 동일 객체도 시간축에 따라 위치, 크기, 방향이 달라질 수 있음

- ※ Spatial similarity만으로는 temporal correspondence를 안정적으로 보전하기 어려움



<Attention & Diversity 예시>



<Video에서 위치만 고려>

Background

• Video Token Compression

▪ Hybrid-based compression

- 토큰의 중요도, 다양성을 함께 고려하여 selection과 merging을 결합

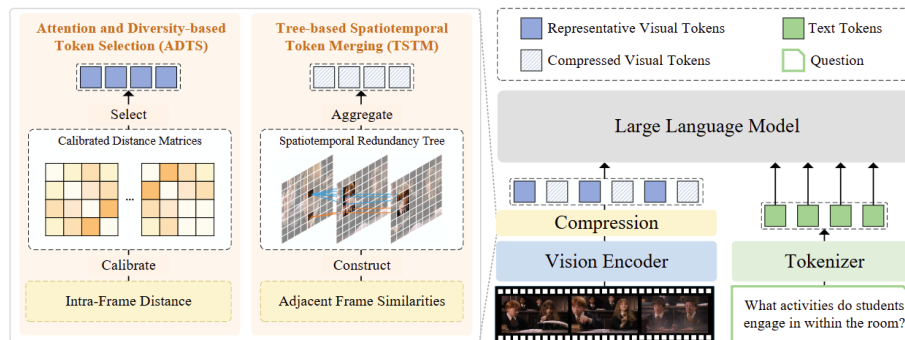
▪ FlashVID¹⁾ [ICLR oral 2026]

- ADTS

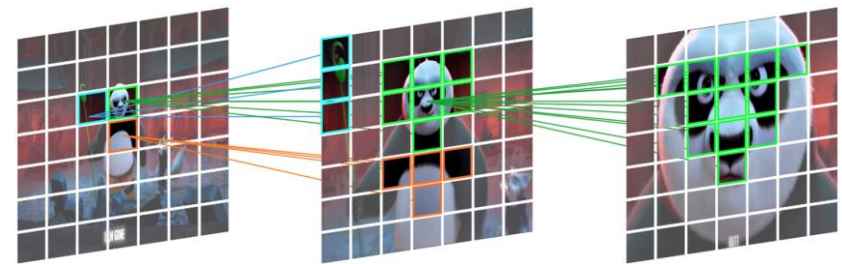
⚙️ Attention으로 보정된 diversity score를 기반으로 중요하면서도 다양한 토큰을 선택

- TSTM

⚙️ 프레임 간 위치 변화를 고려해 시공간적으로 유사한 토큰을 tree 기반으로 병합



<FlashVID pipeline>



<TSTM 예시>

Background

- 논문 선정의 이유
 - 이미지 token compression을 넘어, 비디오의 시공간 중복을 다루는 연구로 확장
 - 동일한 training-free compression 문제를 서로 다른 관점에서 접근한 두 방법
- FastVID: Dynamic Density Pruning for Fast Video Large Language Models
 - Temporal order를 유지하도록 비디오를 동적으로 segment
 - Segment 내부에서 density 기반 merging과 attention 기반 selection을 결합해 시공간 중복을 제거
- UniComp: Rethinking Video Compression Through Informational Uniqueness
 - Attention score 대신 information uniqueness를 토큰 보존 기준으로 제시
 - Frame grouping, token budget allocation, spatial compression을 하나의 기준으로 통합

FastVID: Dynamic Density Pruning for Fast Video Large Language Models [Neurips 2025]

Introduction

- Temporal & Visual Redundancy

- Temporal Context 문제

- Video LLM은 여러 프레임을 순서대로 처리

- 프레임의 연속성이 비디오 이해에 중요

- ⚡ 프레임 순서 변경 → 사건의 전후 관계 손상

- 토큰을 줄이더라도 temporal structure은 유지해야 함

Q: How do the players celebrate scoring the first goal in this video?



Ordered



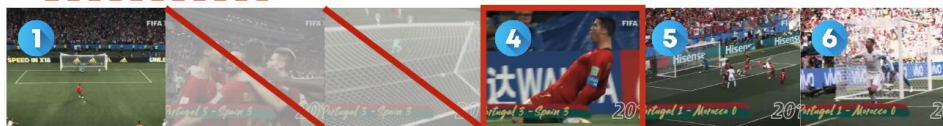
A: High-fiving with the teammates.



Shuffled



A: Running on the field.



Incomplete



A: Kneeling on the field and screaming.

<Temporal Structure 유지>

Introduction

- Temporal & Visual Redundancy

- Visual Context 문제

- 비디오 프레임에는 유사한 visual token이 반복적으로 존재

- Uniform Sampling

- ※ 토큰 내용과 관계없이 일정하게 선택

- ※ 작은 객체 또는 중요한 세부 정보 손실 가능

- 대표성과 다양성을 함께 고려하는 토큰 선택 필요

Original
Image



<Uniform Sampling 예시>

Introduction

- FastVID¹⁾

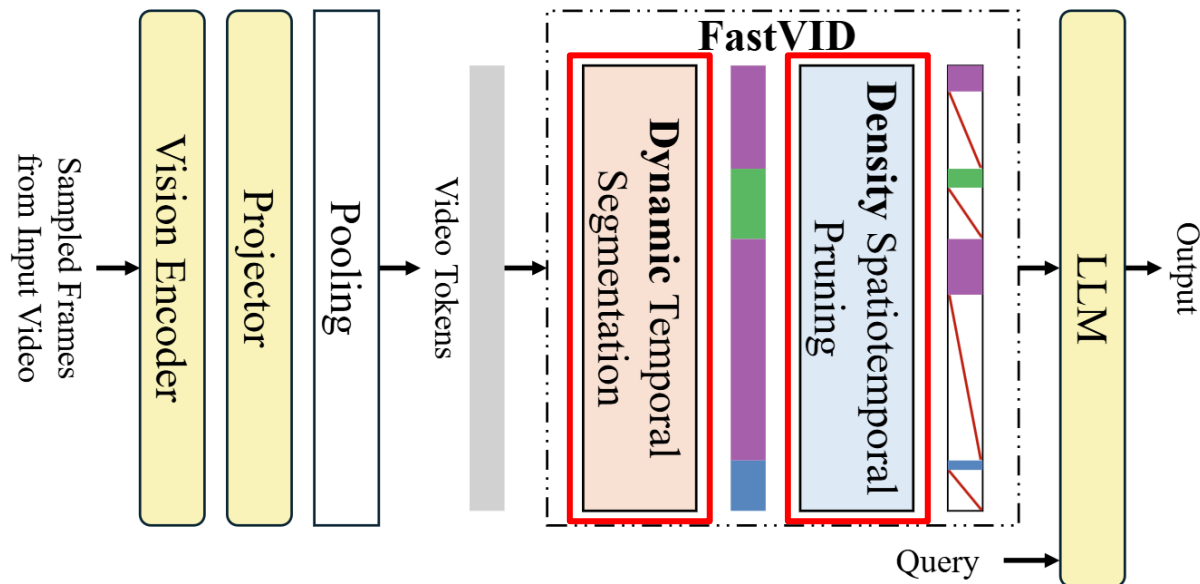
- **Dynamic Temporal Segmentation (DTS)**

- 비슷한 연속 프레임을 시간 순서대로 segment

- **Density Spatiotemporal Pruning (DSP)**

- Segment 내부의 중복 토큰 압축

- 대표적인 시각 정보와 salient detail을 함께 유지



<FastVID pipeline>

Method

- Dynamic Temporal Segmentation (DTS)

- Why Dynamic Temporal Segmentation?

- 균등 샘플링된 프레임들을 시간적으로 유사한 구간으로 분할

- 같은 장면의 연속 프레임에는 redundancy가 높음

- ☼ 유사 토큰을 적극적으로 병합 가능

- ☼ 서로 다른 segment는 토큰을 병합하지 않음



(a) Original Sampled Frames



(b) Fixed-interval Segmentation



(c) Cluster-based Segmentation



(d) Dynamic Temporal Segmentation (ours)

<Temporal Segmentation 예시>

Method

- Dynamic Temporal Segmentation (DTS)

- Frame global feature

- 각 프레임을 하나의 global feature 추출
 - 인접한 두 프레임의 cosine similarity 계산

$$\because t_i = \cos(f_i, f_{i+1}), i = 1, \dots, F - 1$$

$\because t_i$ 가 낮다면 장면 변환 지점

- Segment 경계 선택

- S_1 : 최소 segment 수 보장

$\because$ 전환 유사도가 가장 낮은 $c-1$ 개 지점 선택하여 최소 c 개의 segment 생성

$$\because S_1 = \arg \min_{c-1} T \text{ (} T \text{는 } t_i \text{ 집합)}$$

- S_2 : 장면 큰 변화 탐지

$\because$ 유사도가 threshold보다 낮은 모든 지점 선택

- 최종 segment 경계

$$\because S = S_1 \cup S_2$$



Method

- Density Spatiotemporal Pruning (STPrune)
 - Attention-based Token Selection, ATS
 - 중요한 토큰을 원본 형태로 선택
 - 작은 객체 및 salient detail 보존
 - Attention score가 높은 토큰은 merging하지 않고 유지
 - Density-based Token Merging, DTM
 - 유사하고 반복적인 토큰을 앵커 토큰에 병합
 - Segment의 전체적인 visual context 보존
 - STPrune
 - DTM으로 중복 정보를 병합하고, ATS를 통해 중요 정보는 선택
 - Token merging과 token selection을 함께 사용

Method

- Density-based Token Merging, DTM

- Density-based Anchor Token Selection



- 각 anchor frame 내부에서 대표적인 토큰 선택
 - 동일한 배경 토큰만 반복적으로 선택되는 문제 방지
 - 대표성과 변별성을 동시에 갖는 토큰 선택

- Local Density ρ_i

- 토큰 주변에 유사한 토큰이 얼마나 많이 존재하는지 측정
 - Feature space에서 가까운 k개 이웃 토큰과의 거리를 이용

$$\rho_i = \exp\left(-\frac{1}{k} \sum_{v_j \in k\text{NN}(v_i)} d(v_i, v_j)^2\right)$$

- 이웃 토큰과 거리가 가까움
 - ☞ 유사 토큰이 주변에 많으므로, ρ_i 가 큼
 - 이웃 토큰과 거리가 멀
 - ☞ 해당 토큰이 고립되어 있으므로 ρ_i 가 작음

Method

- Density-based Token Merging, DTM

- Distance to Higher-density Token δ_i

- 토큰 v_i 가 더 밀도 높은 토큰과 얼마나 떨어져 있는지 측정

- $\delta_i = \min_{\rho_j > \rho_i} d(v_i, v_j)$ (더 밀도 높은 토큰이 있는 경우)

- ※ 자신보다 밀도가 높은 토큰 중 가장 가까운 토큰과의 거리

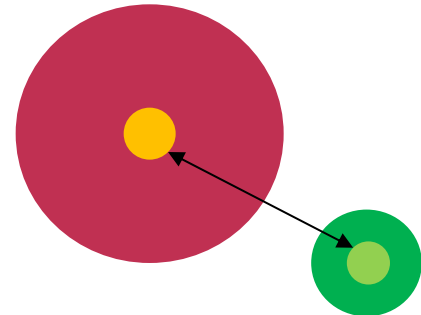
- ※ δ_i 가 작다면 가까운 곳에 더 대표로 뽑힐 토큰이 존재

- ※ δ_i 가 크다면 다른 고밀도 토큰과 멀리 떨어져 있으므로, 새로운 영역의 대표 토큰

- $\delta_i = \max_j d(v_i, v_j)$ (더 밀도 높은 토큰이 없는 경우)

- ※ 전체 토큰 중 가장 높은 밀도를 가지는 토큰

- ※ 큰 δ_i 를 주어 주요 density peak로 선택

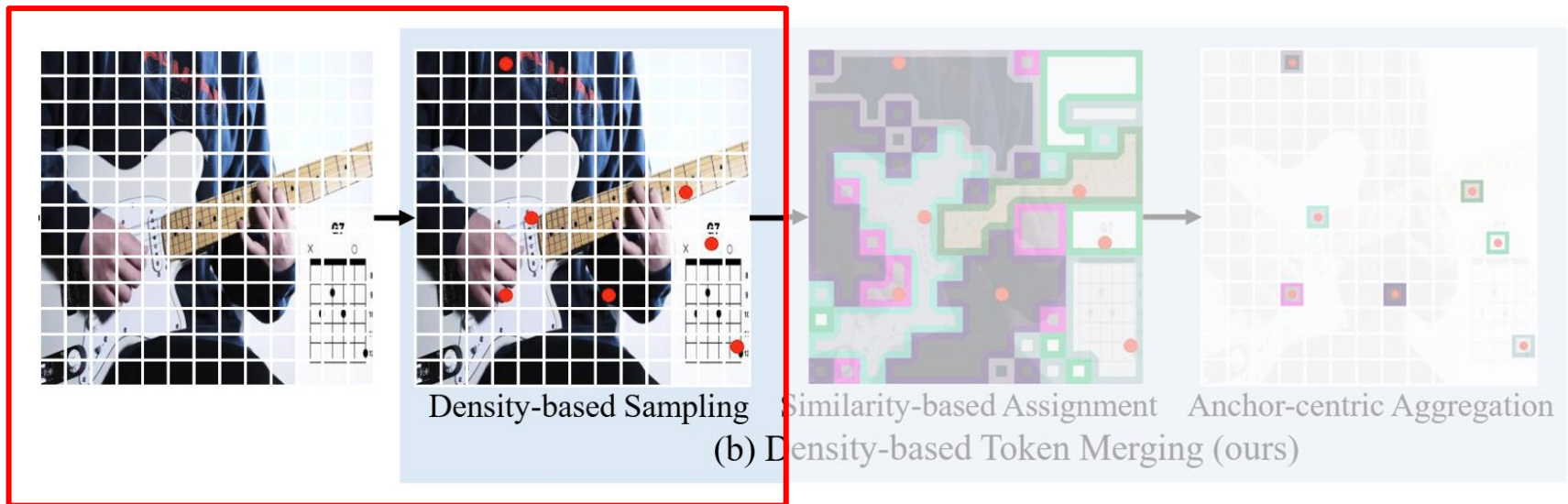


Method

- Density-based Token Merging, DTM

- Anchor Token 선택

- $\text{Score}(v_i) = \rho_i \delta_i$
 - 주변 토큰을 잘 대표 ($\rho_i \uparrow$)
 - 다른 대표 후보와 충분히 구별됨 ($\delta_i \uparrow$)
 - 점수가 높은 토큰을 anchor token으로 선택



<DTM>

Method

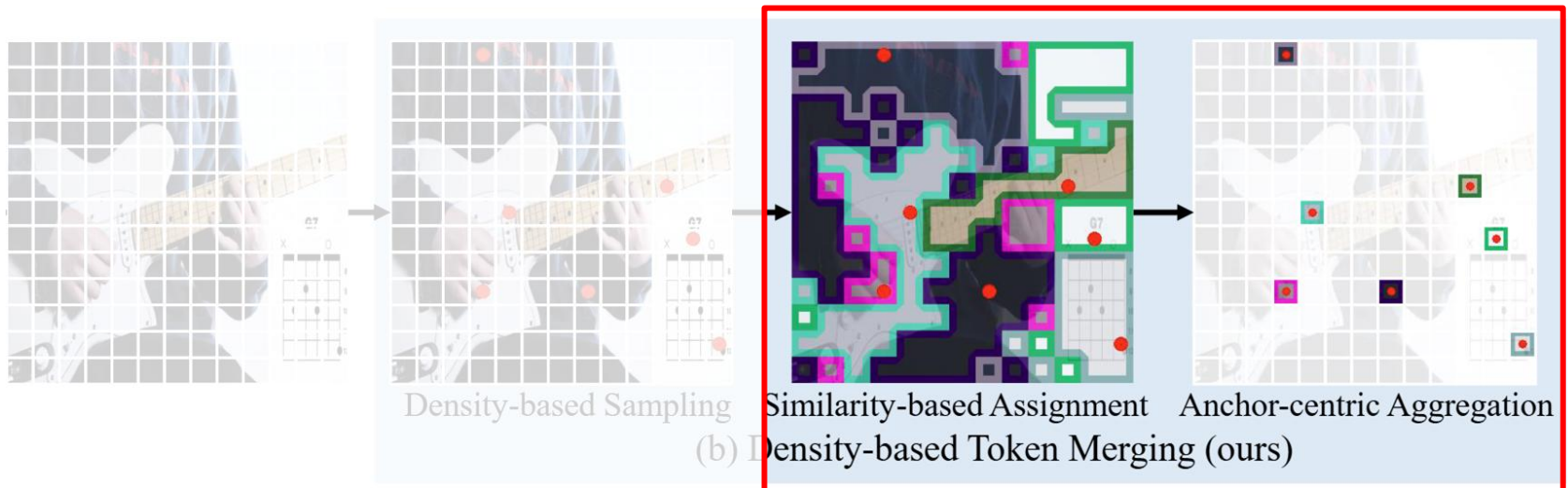
- Similarity-based Assignment & Anchor-centric Aggregation

- Similarity-based Assignment

- Segment에 포함된 모든 프레임의 토큰을 대상으로 수행
 - 각 토큰을 가장 유사한 anchor token에 할당
 - Cosine similarity 기준

- Anchor-centric Aggregation

- Anchor의 위치 정보를 유지하면서 merging 진행



<Similarity-based Assignment & Anchor-centric Aggregation>

Experiments

- Comparison results

- LLaVA-OneVision

- 9.7% token retention에서 원본 성능의 98.0% 유지

- LLaVA-Video

- 9.5% token retention에서 원본 성능의 94.7% 유지

- Qwen2-VL

- 약 24.9% token retention에서 원본 성능의 96.2% 유지

Method	Retention Ratio R	TFLOPs	MVBench	LongVideo Bench	MLVU	VideoMME				Avg. Acc. Score	Avg. Acc. %
						Overall	Short	Medium	Long		
Duration			16s	1~60min	3~120min	1~60min	1~3min	3~30min	30~60min		
Vanilla	100%	48.82	56.9	56.4	65.2	58.6	70.3	56.6	48.8	59.3	100
DyCoke [41] _{CVPR'25}	32.5%	14.13	56.3	56.6	62.1	57.1	68.1	56.7	46.7	58.0	97.8
FastV [6] _{ECCV'24}	100%/25%	13.45	54.7	55.5	61.5	56.2	68.0	54.6	46.0	57.0	96.1
VisionZip [51] _{CVPR'25}	25%	10.73	53.7	51.2	58.5	54.1	61.6	53.4	47.2	54.4	91.7
VisionZip* [51] _{CVPR'25}	25%	10.73	56.6	55.7	64.8	58.0	68.6	57.7	47.7	58.8	99.1
DyCoke [41] _{CVPR'25}	25%	10.73	49.5	48.1	55.8	51.0	61.1	48.6	43.2	51.1	86.2
FastVID	25%	10.73	56.5	56.3	64.1	58.0	69.9	56.6	47.7	58.7	99.0
FastV [6] _{ECCV'24}	100%/20%	11.38	54.1	56.6	61.2	56.2	66.8	54.6	47.2	57.0	96.1
VisionZip [51] _{CVPR'25}	19.9%	8.46	53.0	50.0	57.1	53.0	60.8	51.0	47.1	53.3	90.0
VisionZip* [51] _{CVPR'25}	19.9%	8.46	55.8	55.4	64.2	58.0	68.6	57.0	48.3	58.4	98.5
FastVID	19.9%	8.46	56.3	57.1	63.9	57.9	69.3	56.7	47.7	58.8	99.1
FastV [6] _{ECCV'24}	100%/15%	9.35	53.2	54.9	59.8	54.7	65.1	53.4	45.7	55.7	93.9
VisionZip [51] _{CVPR'25}	14.8%	6.23	50.3	46.9	54.4	49.5	55.8	49.3	43.3	50.3	84.8
VisionZip* [51] _{CVPR'25}	14.8%	6.23	54.3	53.9	63.1	55.5	63.0	54.4	49.1	56.7	95.6
FastVID	14.8%	6.23	56.0	56.2	63.2	57.7	69.3	56.2	47.4	58.3	98.3
FastV [6] _{ECCV'24}	100%/10%	7.36	51.7	52.1	57.7	52.4	60.9	51.4	45.0	53.5	90.2
VisionZip [51] _{CVPR'25}	9.7%	4.04	44.4	43.5	51.5	46.0	50.4	45.8	41.8	46.4	78.3
VisionZip* [51] _{CVPR'25}	9.7%	4.04	51.7	48.3	59.7	52.8	59.4	52.0	46.9	53.1	89.6
PruneVID* [16] _{ACL'25}	10.1%	4.23	54.2	53.8	62.3	55.9	66.4	52.9	48.3	56.6	95.4
FastVID	9.7%	4.04	55.9	56.3	62.7	57.3	67.4	56.0	48.6	58.1	98.0
PruneVID [16] _{ACL'25}	10.1%/3.9%	2.58	54.1	51.8	62.3	55.5	67.1	51.8	47.7	55.9	94.3
FastVID+FastV [6]	9.7%/3.6%	2.47	55.0	53.6	61.9	56.3	66.1	54.8	48.1	56.7	95.6

<LLaVA-OneVision>

Method	Retention Ratio R	# Newline Tokens M	TFLOPs*	MVBench	LongVideo Bench	MLVU	VideoMME			Avg. Acc. Score	Avg. Acc. %
							Overall	Short	Long		
Vanilla	100%	832	103.2	60.4	59.6	70.3	64.1	76.9	53.4	63.6	100
DyCoke [41] _{CVPR'25}	32.1%	256	27.1	59.3	57.9	65.7	61.6	74.6	51.1	61.1	96.1
FastV [6] _{ECCV'24}	100%/25%	832/568.7	29.2	58.0	58.3	63.9	61.0	71.3	51.0	60.3	94.8
VisionZip [51] _{CVPR'25}	24.9%	64	19.5	56.4	54.1	62.1	58.6	66.3	51.2	57.8	90.9
VisionZip* [51] _{CVPR'25}	24.9%	64	19.5	58.3	58.3	66.6	61.9	73.3	52.2	61.5	96.7
DyCoke [41] _{CVPR'25}	25%	208	20.7	50.8	53.0	56.9	56.1	65.8	48.9	54.2	85.2
FastVID*	24.9%	64	19.5	59.3	58.3	67.7	62.6	74.9	52.0	62.0	97.5
FastVID	24.9%	715.5	24.5	59.9	57.4	68.6	63.6	74.9	53.7	62.4	98.1
FastV [6] _{ECCV'24}	100%/10%	832/311.8	16.2	55.8	55.4	58.9	57.9	67.6	48.6	57.0	89.6
VisionZip [51] _{CVPR'25}	9.5%	64	7.3	46.3	46.6	52.2	49.5	54.2	44.3	48.7	76.6
VisionZip* [51] _{CVPR'25}	9.5%	64	7.3	56.6	53.6	61.7	58.7	67.6	50.1	57.7	90.7
FastVID*	9.5%	64	7.3	58.3	56.2	63.9	59.6	70.9	50.7	59.5	93.6
FastVID	9.5%	508.2	10.5	58.5	56.5	64.9	60.7	71.7	51.2	60.2	94.7

<LLaVA-Video>

Method	# Token	TFLOPs			VideoMME			
					Short	Medium	Long	Overall
Vanilla	13447.1	100%	124.0	100%	74.1	60.4	54.3	63.0 100%
FastV [6] _{ECCV'24}	13447.1/3361.8	100%/25%	31.3	25.3%	Out of Memory			
VisionZip* [51] _{CVPR'25}	3349.3	24.9%	24.1	19.4%	70.9	56.3	48.3	58.5 92.9%
PruneVID* [16] _{ACL'25}	3460.6	25.7%	25.0	20.1%	66.7	54.0	48.1	56.3 89.4%
FastVID	3349.3	24.9%	24.1	19.4%	72.7	58.3	50.7	60.6 96.2%

<Qwen2-VL>

UniComp: Rethinking Video Compression Through Informational Uniqueness [CVPR 2026]

Introduction

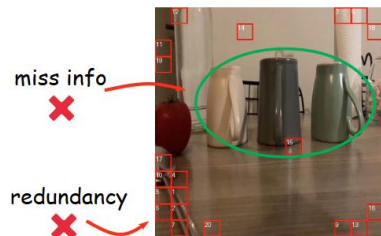
• Information Uniqueness

▪ Attention-based Compression의 한계

- Attention score가 높은 salient token을 우선 보존
- 프레임마다 유사한 salient token이 반복 선택
- 제한된 token budget이 중복 정보에 집중

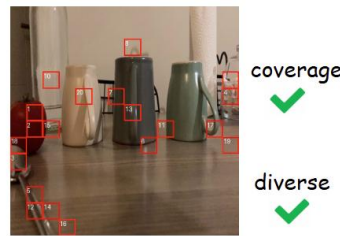
▪ UniComp¹⁾

- 다른 token으로 복원 가능한 정보 → redundant information
- 다른 token으로 대체하기 어려운 정보 → unique information
- 중요도를 추정하기보다, 정보의 대체 가능성을 기준으로 압축



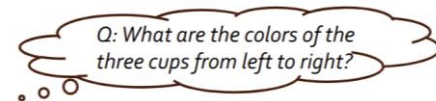
Attention-based

VisionZip: The colors of the three cups are silver, green, and black.



Uniqueness-based

UniComp: The colors of the three cups from left to right are white, beige, and green.



Introduction

- UniComp Overview

- Information Uniqueness-driven Compression

- Frame과 token의 중복을 하나의 uniqueness criterion으로 측정

- Frame Group Fusion (FGF)

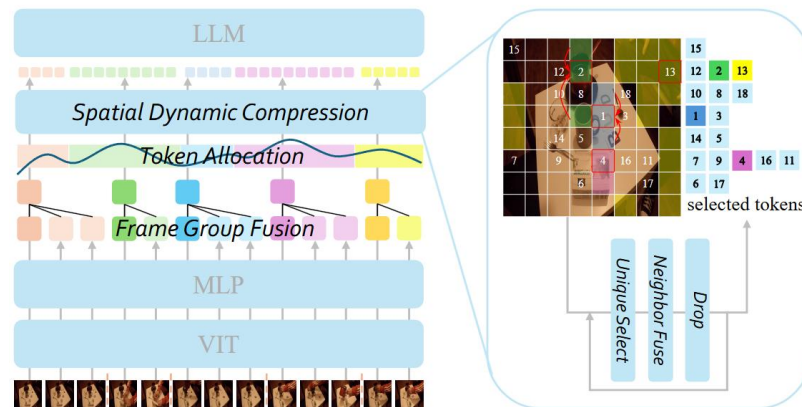
- ⌘ 시간적으로 중복된 frame을 semantic group으로 구성

- Token Allocation (TA)

- ⌘ Frame uniqueness에 따라 전체 token budget을 차등 배분

- Spatial Dynamic Compression (SDC)

- ⌘ Unique token을 선택하고 주변 redundant token을 병합



<UniComp pipeline>

Method

- Information Uniqueness

- Pairwise Information Uniqueness

- 두 token 사이의 cosine distance로 uniqueness 정의
 - 유사도가 높을수록 낮은 uniqueness
 - 다른 token과 구별될수록 높은 uniqueness

$$u_{ij} = 1 - \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2}$$

- Token-level Uniqueness

- Token x_i 와 전체 token 사이의 평균 uniqueness
 - 높은 $U_i \rightarrow$ 다른 token으로 대체하기 어려운 정보
 - 낮은 $U_i \rightarrow$ 주변 token과 중복되는 정보

$$U_i = \frac{1}{N} \sum_{j=1}^N u_{ij} = 1 - \frac{1}{N} \sum_{j=1}^N \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2}$$

Method

- Optimization Objective

- Ideal Objective

- 고정된 token budget K에서 전체 정보를 가장 잘 보존하는 subset S 선택
 - 선택된 token만으로 원본 token 정보를 최대한 복원
 - Information loss를 conditional entropy로 표현

$$\therefore S^* = \arg \min H(X | S)$$

- Practical Approximation

- Conditional entropy를 직접 계산하고 최적화하기는 어렵기에 단순화
 - $\therefore$ 다른 토큰과 비슷하다면 대체 가능하고, 다른 토큰과 다르면 대체하기 어렵다.
 - Token 간 cosine distance를 information uniqueness로 사용
 - Uniqueness가 높은 token은 보존
 - 서로 유사한 token은 redundant token으로 판단하여 병합

$$\therefore u_{ij} = 1 - \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2}$$

Method

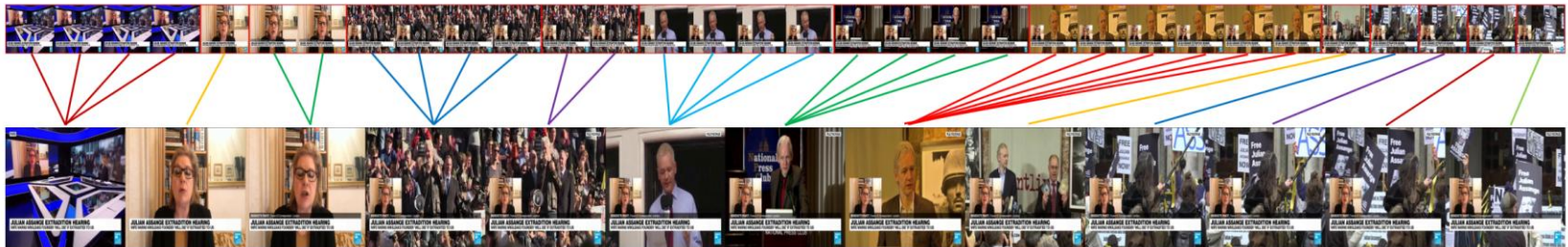
- Frame Group Fusion (FGF)

- Temporal Redundancy Compression

- 각 frame의 visual token을 average pooling하여 global feature 생성
 - 비디오를 시간 순서대로 sequential scan
 - 현재 frame과 현재 group의 첫 representative frame 비교
 - Uniqueness가 threshold보다 낮으면 동일 group에 포함
 - 차이가 크면 새로운 semantic group 생성

32 frames

Frame Group Fusion
13 frames



<FGF>

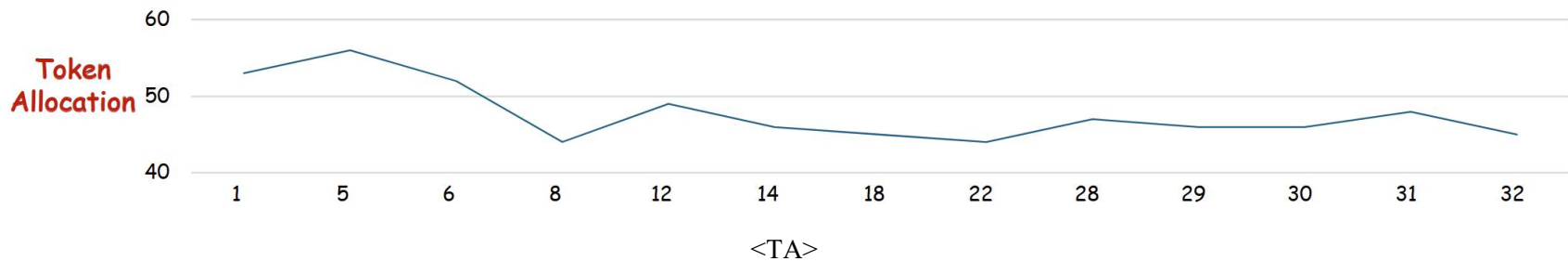
Method

- Token Allocation (TA)

- Uniqueness-aware Token Budget Allocation

- FGF 이후 남은 representative frame 간 global uniqueness 계산
 - 다른 frame과 구별되는 frame에 더 많은 token 할당
 - 중복도가 높은 frame에는 적은 token 할당
 - 전체 token 수는 $TOKEN_{max}$ 로 고정

Frame Group
Fusion
13 frames



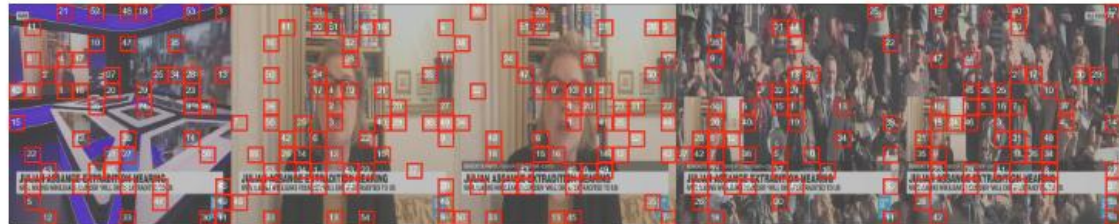
Method

- Spatial Dynamic Compression (SDC)

- Frame-wise Spatial Compression

- 각 frame 내부에서 token-level uniqueness 계산
- ViT 마지막 attention layer의 Key feature 사용
- Uniqueness가 높은 token부터 representative token으로 선택
- 유사한 주변 token은 제거하지 않고 representative token에 merging
- Frame별 budget K_t 에 도달하면 selection 종료

Spatial
Dynamic
Compression



<SDC>

Experiments

- Comparison results
 - LLaVA-OneVision
 - LLaVA-Video
 - Eagle2.5

Method	Frames/Retain Ratio	LongVideoBench	EgoSchema	MLVU	VideoMME	Avg.	Acc (%)
Vanilla	32f / 100%	56.3	60.4	64.7	58.4	59.95	100.0
FastV [5] ECCV'24	32f / 25%	56.4	60.4	61.5	56.1	58.60	97.7
Pdrop [35] CVPR'25	32f / 25%	56.7	58.0	-	56.4	-	-
DyCoke [30] CVPR'25	32f / 25%	49.5	59.9	55.8	57.4	55.65	92.8
VisionZip [37] CVPR'25	32f / 25%	56.5	60.3	64.8	58.2	59.95	100.0
PruneVid [13] ACL'25	32f / 25%	55.7	59.9	-	57.4	-	-
FastVid [25] NeurIPS'25	32f / 25%	56.3	59.3	64.1	58.0	59.43	99.1
HoliTom [24] NeurIPS'25	32f / 25%	56.7	61.2	64.7	58.6	60.30	100.6
UniComp	32f / 25%	57.6	61.6	65.0	58.9	60.78	101.4
VisionZip [37] CVPR'25	32f / 20%	55.2	59.8	64.2	57.9	59.28	98.9
PruneVid [13] ACL'25	32f / 20%	54.7	59.7	-	56.9	-	-
FastVid [25] NeurIPS'25	32f / 20%	57.1	58.3	63.9	57.9	59.30	98.9
HoliTom [24] NeurIPS'25	32f / 20%	57.1	61.0	63.5	58.6	60.05	100.2
UniComp	32f / 20%	57.7	61.1	64.4	58.7	60.48	100.9
VisionZip [37] CVPR'25	32f / 15%	54.4	59.8	63.1	56.1	58.35	97.3
PruneVid [13] ACL'25	32f / 15%	55.4	59.7	-	56.6	-	-
FastVid [25] NeurIPS'25	32f / 15%	56.2	58.7	63.2	57.7	58.95	98.3
HoliTom [24] NeurIPS'25	32f / 15%	56.4	61.2	62.9	57.3	59.45	99.2
UniComp	32f / 15%	57.4	60.9	64.3	58.4	60.25	100.5
VisionZip [37] CVPR'25	32f / 10%	49.3	58.0	59.7	53.4	55.10	91.9
PruneVid [13] ACL'25	32f / 10%	54.5	59.8	62.3	56.0	58.15	97.0
FastVid [25] NeurIPS'25	32f / 10%	56.3	58.3	62.7	57.3	58.65	97.8
HoliTom [24] NeurIPS'25	32f / 10%	56.3	61.2	61.3	56.8	58.90	98.2
UniComp	32f / 10%	57.6	60.5	63.1	58.0	59.80	99.7

<LLaVA-OneVision>

Method	Short Setting			Long Setting		
	F/R	LVB	MLVU	F/R	LVB	MLVU
VisionZip	64f/10%	52.2	62.7	256f/25%	57.2	70.1
HoliTom	64f/10%	54.8	63.3	256f/25%	57.3	69.8
UniComp	64f/10%	56.2	65.5	256f/25%	58.8	71.0
VisionZip	64f/25%	56.3	65.6	320f/20%	56.6	70.2
HoliTom	64f/25%	57.5	66.6	320f/20%	<i>OOM</i>	<i>OOM</i>
UniComp	64f/25%	58.0	66.9	320f/20%	57.9	70.3

<LLaVA-Video>

Method	Frames/Retain	All	Short	Medium	Long
Vanilla	64f / 100%	68.9	80.6	67.1	59.1
VisionZip	128f / 50%	67.8	78.9	65.1	59.4
HoliTom*	128f / 50%	68.4	81.0	65.4	58.9
UniComp	128f / 50%	70.1	81.4	69.0	59.8
VisionZip	256f / 25%	66.3	76.0	64.0	58.9
HoliTom*	256f / 25%	67.9	78.7	65.4	59.4
UniComp	256f / 25%	70.4	80.8	69.7	60.8
VisionZip	512f / 12.5%	65.0	72.9	62.9	59.2
HoliTom*	512f / 12.5%	67.4	77.9	65.4	58.9
UniComp	512f / 12.5%	70.7	80.3	69.3	62.4

<Eagle2.5>

감사합니다.