2026 하계세미나

Efficient Token Compression for Omnimodal LLMs



Pusan National University
Dept. of Electronic Engineering
Computer Vision & Signal Processing Lab.



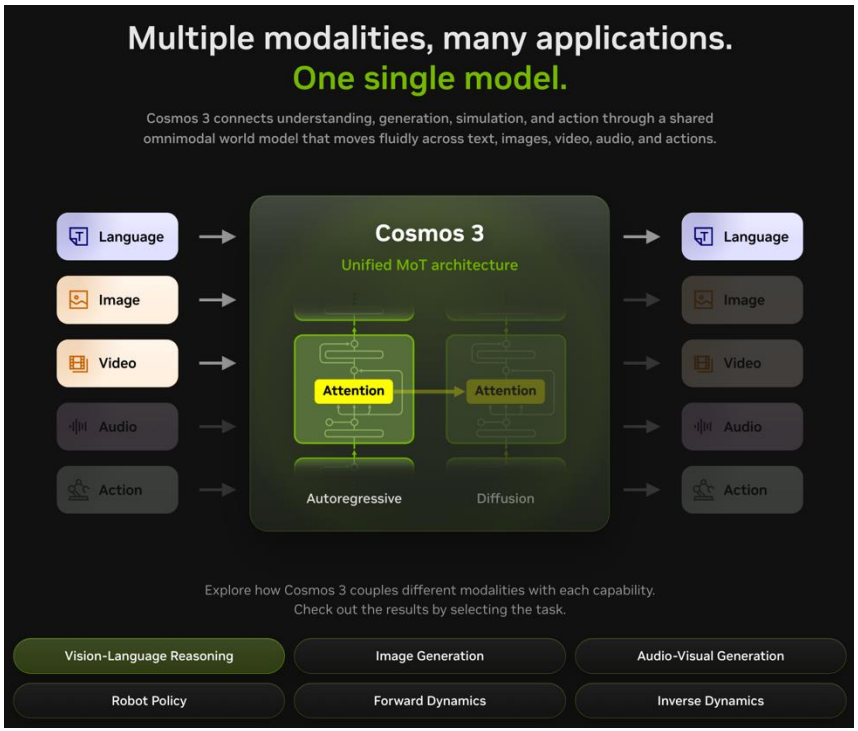
Presented By
조범권

개요

- 주제 선정 이유
- 발표 논문
  - OmniZip: Audio-Guided Dynamic Token Compression for Fast Omnimodal Large Language Models [CVPR’ 26]
  - OmniFit: Bridging Modalities via Layer-Adaptive Token Compression for Omnimodal Large Language Models [ICML’ 26 spotlight]
- 향후 계획

주제 선정 이유

- 최근 Omni Modality가 주목받고 있음
  - 이러한 모델을 실제 로봇에 적용하기 위해서는 '효율성' 필수!!



NVIDIA의 Cosmos 3



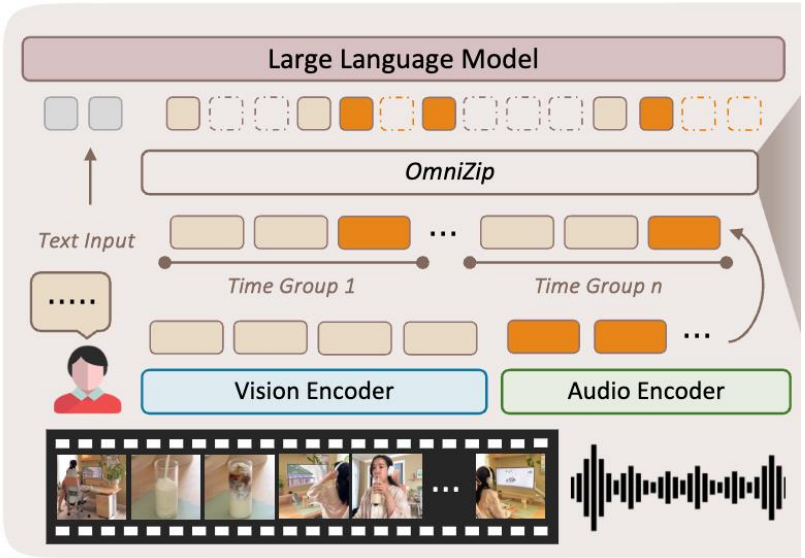
OmniZip: Audio-Guided Dynamic Token Compression for Fast Omnimodal Large Language Models [CVPR'26]

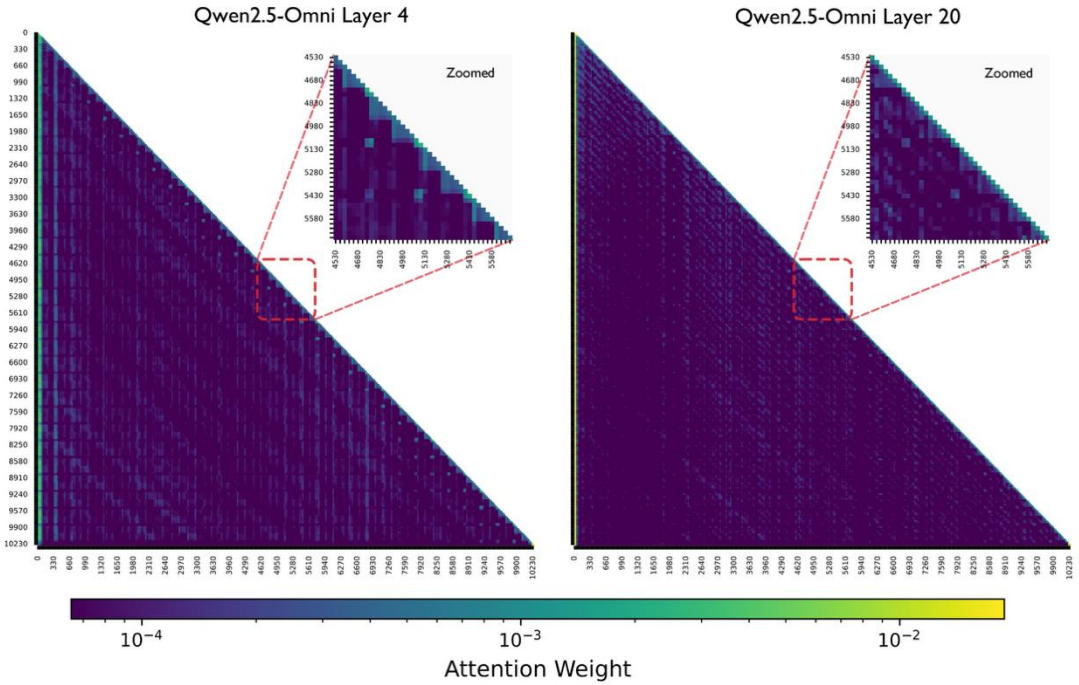
OmniZip (CVPR'26)

- Motivation
  - 기존의 VideoLLM에서 계산 효율을 위해 Video 토큰을 압축하는 연구들이 존재
    - 비디오의 특성상 긴 시퀀스로 인한 수많은 토큰이 발생
  - OmniLLM으로 넘어오면서 Audio 토큰이 추가
    - 추가된 Audio 토큰까지 함께 고려하며 압축 필요

OmniZip (CVPR'26)

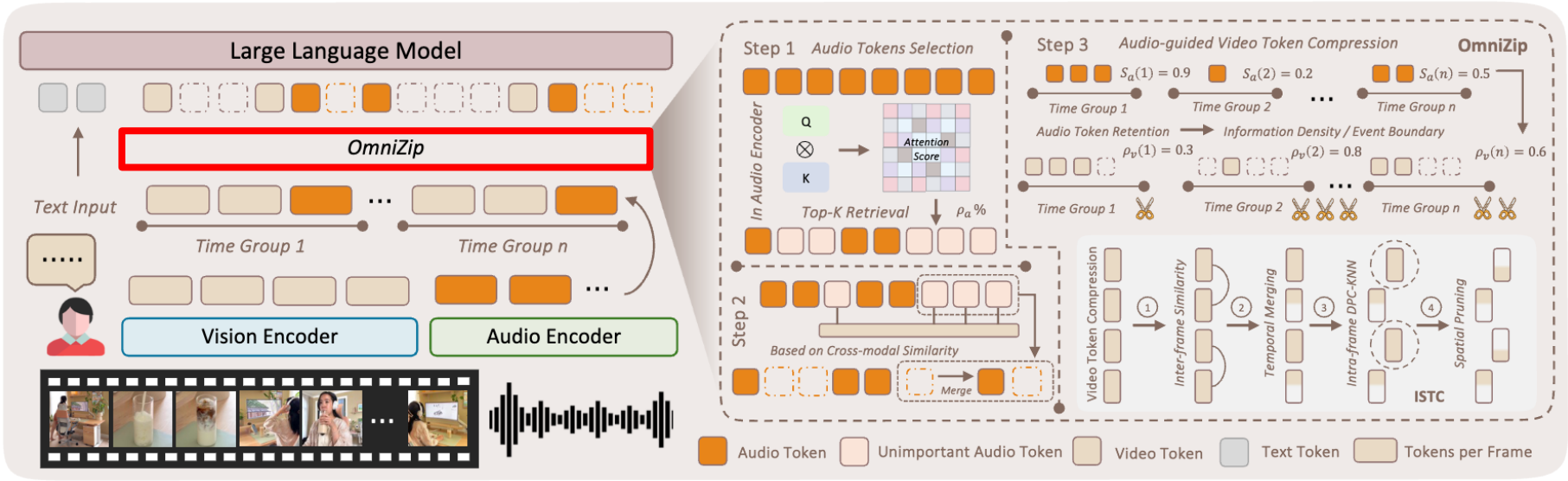
- Finding
  - Audio 토큰이 Video 토큰보다 일관되게 더 큰 attention을 받음
    - Audio 토큰 기준 토큰 압축이 필요
  - Time window 안에서 강한 attention이 주로 일어남
    - Time window 단위 토큰 압축이 필요





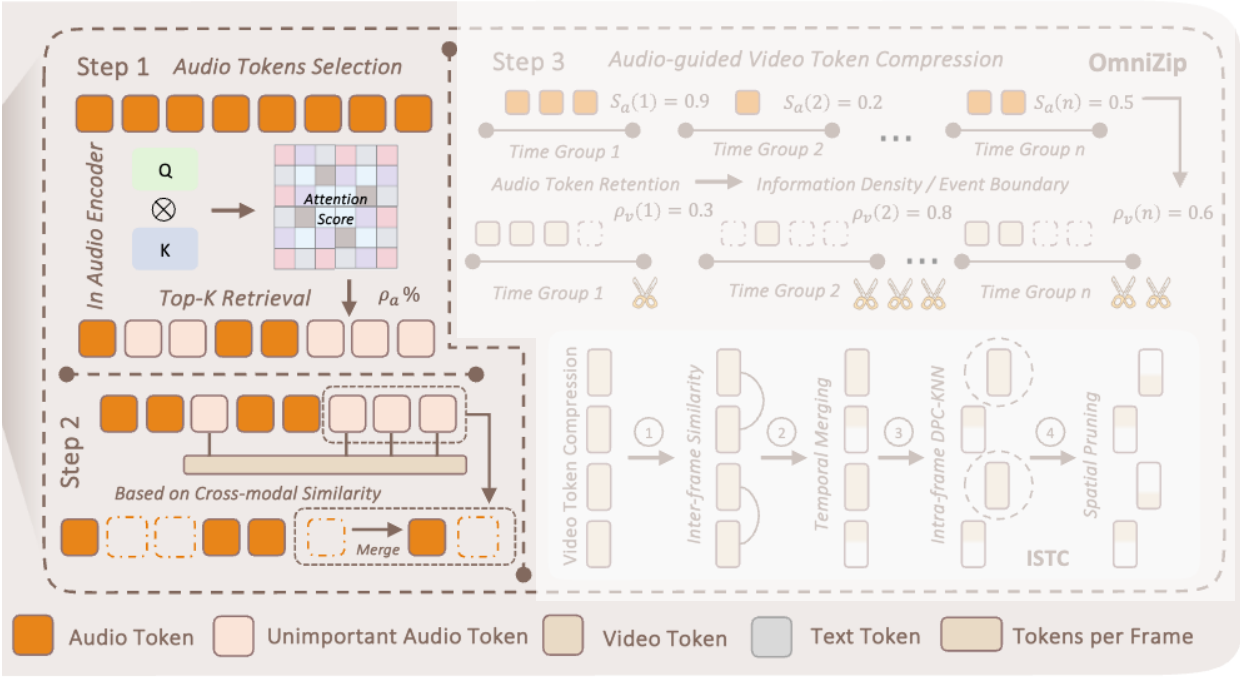
OmniZip (CVPR'26)

- Method
  - 각 모달리티의 입력이 인코더를 거쳐서 나온 토큰에 대해 압축 하는 방식



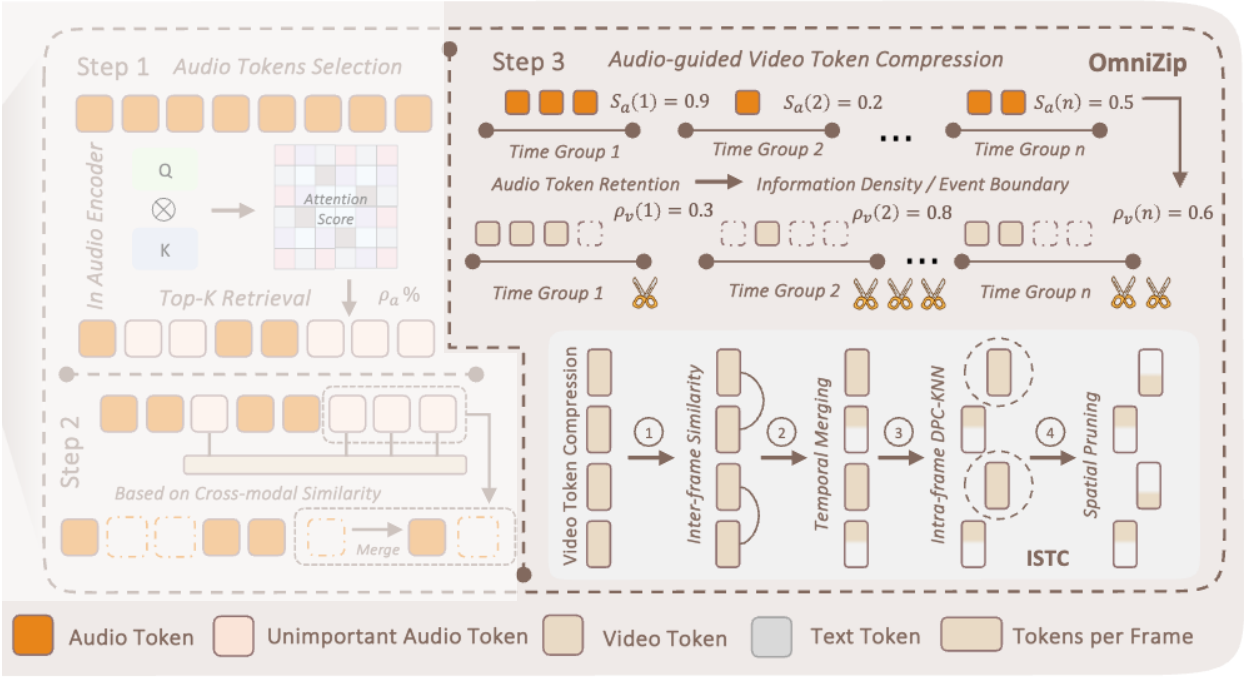
OmniZip (CVPR'26)

- Method
  - 모든 Audio 토큰에 대해 Self-Attention score로 Top-K 중요한 Audio 토큰 선택
  - 남은 Audio 토큰 중에서 균일하게 anchor 선택
    - Cross Modal Similarity로 Top-G merging



OmniZip (CVPR'26)

- Method
  - Time window별 Audio 토큰 보존율로 해당 window의 Video 토큰 보존율 계산
    - Audio 토큰 보존율 높을 수록 Video 토큰도 보존율 높은 방향
  - 계산한 Video 토큰 보존율을 따라 압축
    - 인접 프레임끼리 반복되는 토큰 제거 / 프레임 내에서 반복되는 토큰 제거



OmniZip (CVPR'26)

- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능

Method	Settings		AVUTBench							VideoMME	ShortVid-Bench	Avg.
	Retained Ratio	FLOPs Ratio	EL	OR	OM	IE	CC	CM	Avg.	wo	Avg. Score	
Qwen2.5-Omni-7B												
Full Tokens	100%	100%	38.2	67.8	59.6	85.6	44.1	66.7	64.5	66.0	70.5	100%
Random	55%	48%	38.2	64.9	55.6	80.1	34.7	65.0	61.0	65.4	68.3	96.9%
FastV	50%	54%	34.1	64.3	57.1	77.6	36.4	56.4	58.4	-	68.0	94.3%
DyCoke (V&A)	50%	44%	**38.8**	**67.2**	**58.2**	81.9	39.0	62.4	62.0	65.5	68.5	97.5%
OmniZip (Ours)	45%	39%	38.4	**67.2**	56.9	**85.3**	**42.4**	**66.0**	**63.0**	**66.3**	**69.9**	**99.1%**
Random	40%	34%	31.7	58.5	53.3	74.9	43.2	59.0	56.9	65.0	67.7	94.3%
FastV	35%	42%	24.1	60.7	54.3	81.6	40.7	58.3	57.8	-	67.9	93.8%
DyCoke (V&A)	35%	29%	32.9	62.1	**54.9**	74.5	39.0	58.3	57.4	65.2	68.0	94.7%
OmniZip (Ours)	35%	29%	**34.1**	**67.5**	54.6	**83.7**	**42.4**	**61.2**	**61.0**	**66.1**	**69.0**	**97.6%**
Qwen2.5-Omni-3B												
Full Tokens	100%	100%	32.9	65.3	58.4	85.0	44.1	62.6	62.2	62.6	69.4	100%
Random	55%	45%	31.7	59.2	55.4	77.3	**44.9**	**62.1**	58.7	61.1	67.9	96.6%
FastV	50%	49%	27.1	57.0	56.3	80.5	42.3	60.1	55.9	-	68.0	95.7%
DyCoke (V&A)	50%	40%	31.9	64.3	57.3	82.2	40.7	61.3	60.7	61.6	67.4	97.7%
OmniZip (Ours)	45%	36%	**32.4**	**65.0**	**57.7**	**84.9**	41.5	61.4	**61.3**	**62.8**	**68.5**	**99.2%**
Random	40%	31%	28.2	60.8	54.9	73.1	42.3	**61.6**	57.5	60.6	67.0	95.4%
FastV	35%	37%	24.2	60.8	54.3	81.6	40.7	58.3	57.7	-	67.7	96.9%
DyCoke (V&A)	35%	26%	32.9	62.1	54.9	74.5	38.9	58.3	57.4	61.0	67.5	95.7%
OmniZip (Ours)	35%	26%	**28.8**	**63.1**	**58.2**	**84.0**	**42.4**	60.4	**60.1**	**62.7**	**68.0**	**98.3%**

OmniZip (CVPR'26)

- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능

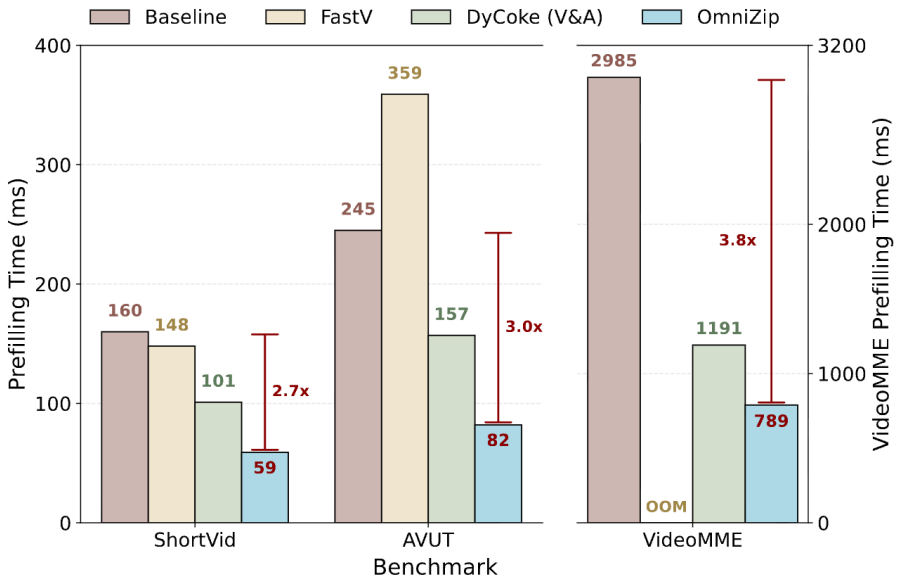
Method	Retained Ratio	FLOPs (T)	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
Qwen2.5-Omni-7B											
Full Tokens	100%	73.2	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	35.5	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	39.3	48.8	47.4	44.2	44.1	**41.2**	38.3	40.0	46.6	44.3
DyCoke (V&A)	50%	31.9	48.4	49.9	46.7	41.4	39.9	**40.8**	40.2	46.5	44.6
OmniZip (Ours)	45%	28.3	**49.4**	**51.1**	**45.6**	**43.9**	40.1	**40.8**	41.9	**46.7**	**45.9**
OmniZip (Ours)	35%	21.4	48.3	49.5	**47.6**	42.5	40.1	40.2	**42.3**	46.3	45.3
Qwen2.5-Omni-3B											
Full Tokens	100%	37.4	51.5	50.8	45.0	45.4	43.8	42.5	44.2	46.1	46.4
Random	55%	17.0	48.2	46.3	40.7	41.4	38.6	40.0	41.8	**43.4**	42.8
FastV	50%	18.2	50.0	**50.5**	**44.1**	43.0	**40.5**	41.6	41.8	42.1	44.4
DyCoke (V&A)	50%	15.1	48.1	48.5	42.3	43.3	39.7	**43.4**	42.1	43.0	44.0
OmniZip (Ours)	45%	13.3	**50.1**	**50.5**	43.9	45.6	**40.5**	40.8	**43.7**	43.1	**45.2**
OmniZip (Ours)	35%	9.9	48.8	48.9	41.8	**46.4**	39.8	42.5	42.6	43.1	44.3

OmniZip (CVPR'26)

- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능

Method	GPU Mem. ↓	Prefilling Time ↓	Acc. ↑	Latency per Example ↓
Qwen2.5-Omni-7B				
Full Tokens	35G	291ms (1.00×)	46.8	4.52s (1.00×)
FastV		OOM		
DyCoke (V&A)	31G	184ms (1.58×)	44.6	3.64s (1.24×)
Ours (45%)	28G	116ms (2.51×)	45.9	3.40s (1.33×)
Ours (35%)	**25G**	**85ms (3.42×)**	45.3	**3.18s (1.42×)**
Qwen2.5-Omni-3B				
Full Tokens	25G	258ms (1.00×)	46.4	3.61s (1.00×)
FastV	45G	222ms (1.16×)	44.4	3.45s (1.05×)
DyCoke (V&A)	20G	171ms (1.51×)	44.0	3.12s (1.16×)
Ours (45%)	17G	104ms (2.48×)	45.2	2.86s (1.26×)
Ours (35%)	**16G**	**79ms (3.27×)**	44.3	**2.75s (1.31×)**

WorldSense 기준 비교



Qwen2.5-Omni-7B 기준 비교

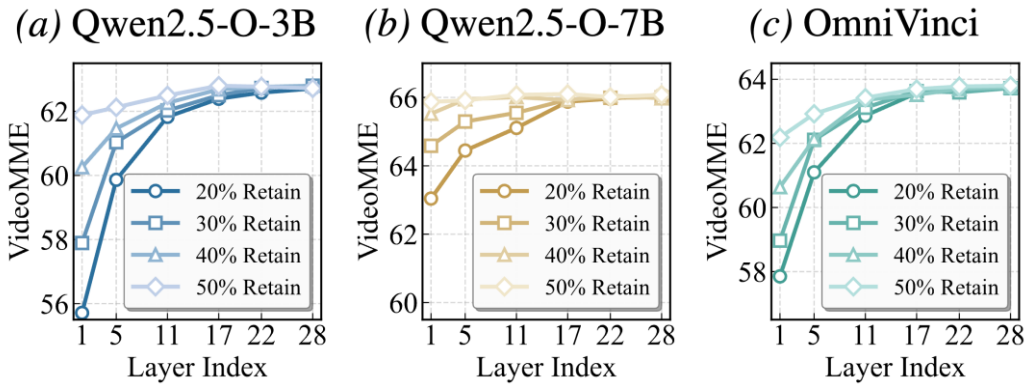
OmniFit: Bridging Modalities via Layer-Adaptive Token Compression for Omnimodal Large Language Models [ICML'26 spotlight]

OmniFit (ICML'26 spotlight)

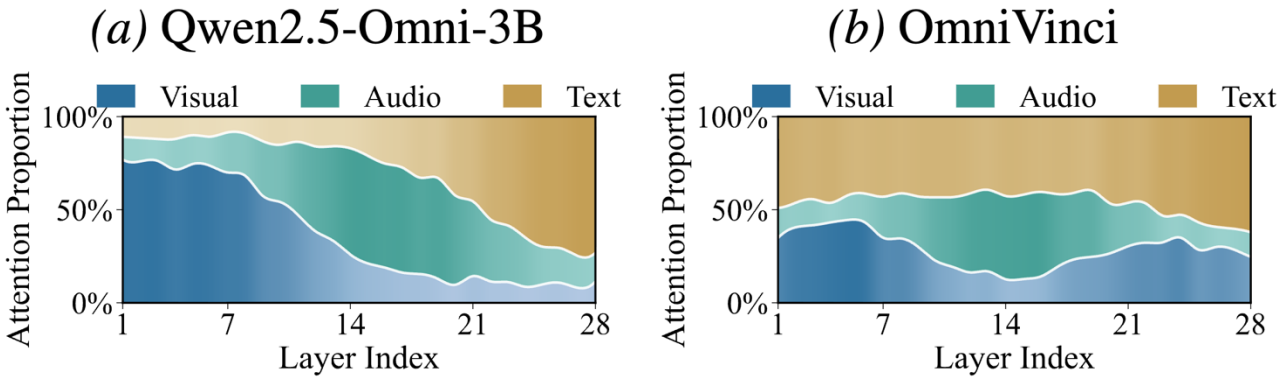
- Motivation
  - 백본 Layer별로 다른 압축 비율 필요
    - OmniZip은 입력 토큰들만 줄이고 백본 내부에서는 안 건듦
  - 단일 모달리티 중심 압축의 한계
    - OmniZip은 오디오의 보존율을 기준으로 비디오의 보존율을 선택하기 때문에 텍스트 Query와 관련된 비디오 단서까지 날릴 위험이 있음
      - ex) 소리 없이 춤 추는 장면

OmniFit (ICML'26 spotlight)

- Finding
  - 얕은 Layer가 깊은 Layer보다 토큰 압축에 취약

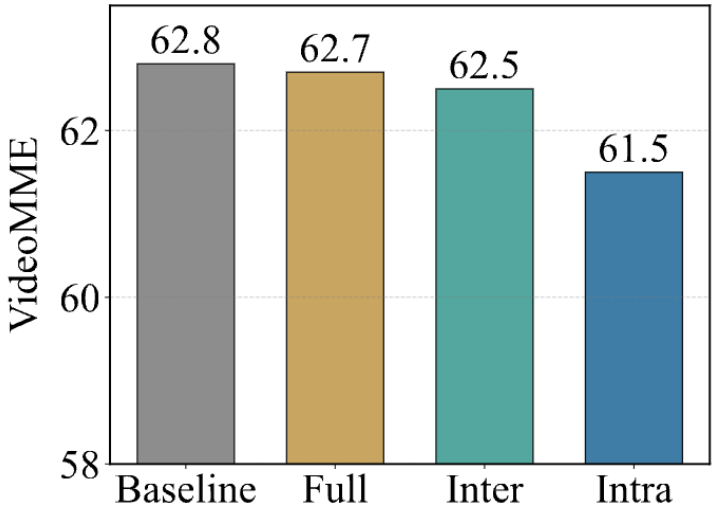


  - Layer마다 Attention을 많이 받는 모달리티가 다름



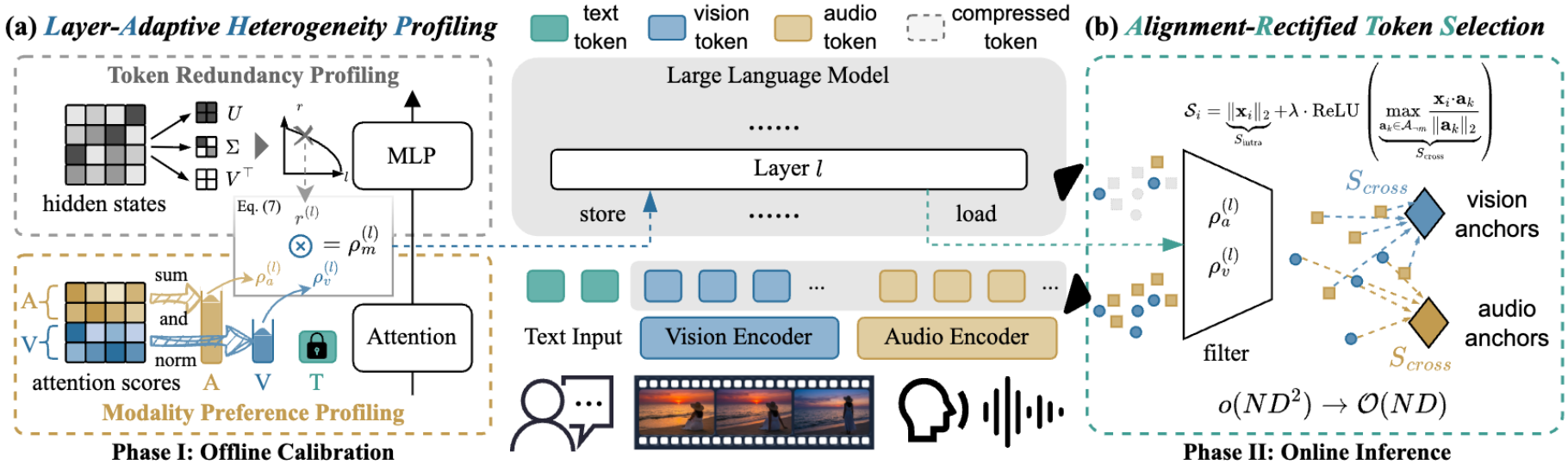
OmniFit (ICML'26 spotlight)

- Finding
  - 모달리티끼리 Attention이 모달리티 내부 Attention보다 중요
    - Intra-modal Attention만 갖고 인코더 직후 토큰 50%를 제거했을때 성능 저하 가장 큼



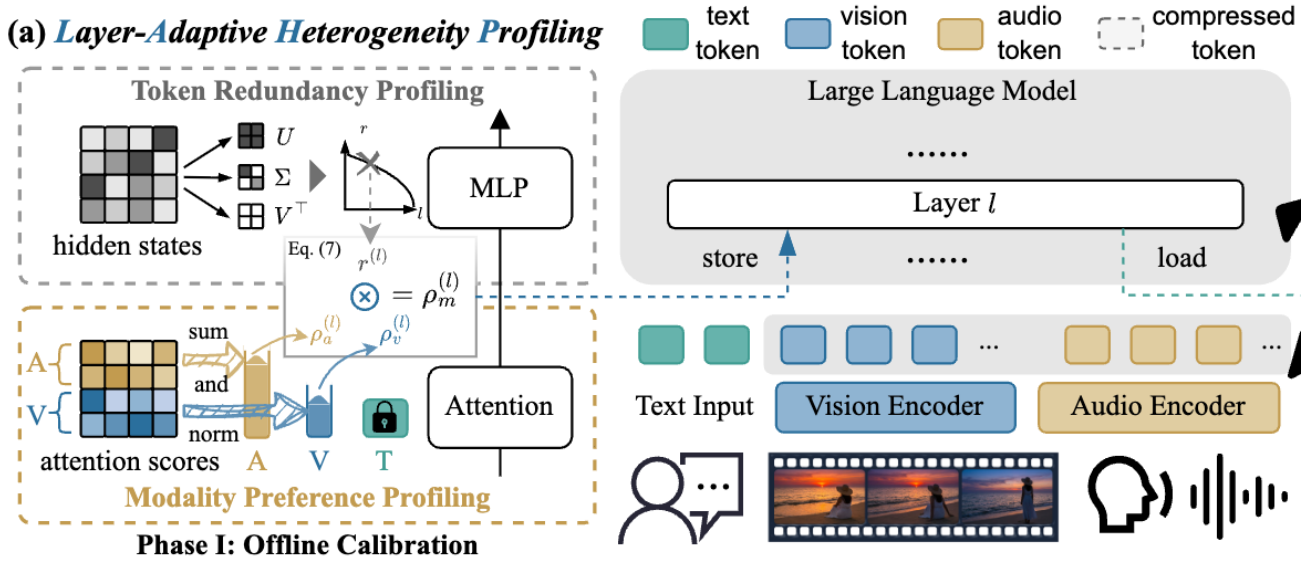
OmniFit (ICML'26 spotlight)

- Method
  - 두가지 Phase로 구성
    - Inference 이전에 미리 보존율을 계산하는 Offline Calibration
    - 미리 계산해둔 보존율을 따라 토큰을 압축하는 Online Inference



OmniFit (ICML'26 spotlight)

- Method
  - LAHP: 얼마나 남길지 선택
    - 샘플들을 넣어서 모델 특성 확인
    - TRP: Layer별로 계산한 중복도로 보존율 계산
    - MPP: 모달리티별 평균 Attention score로 보존율 계산

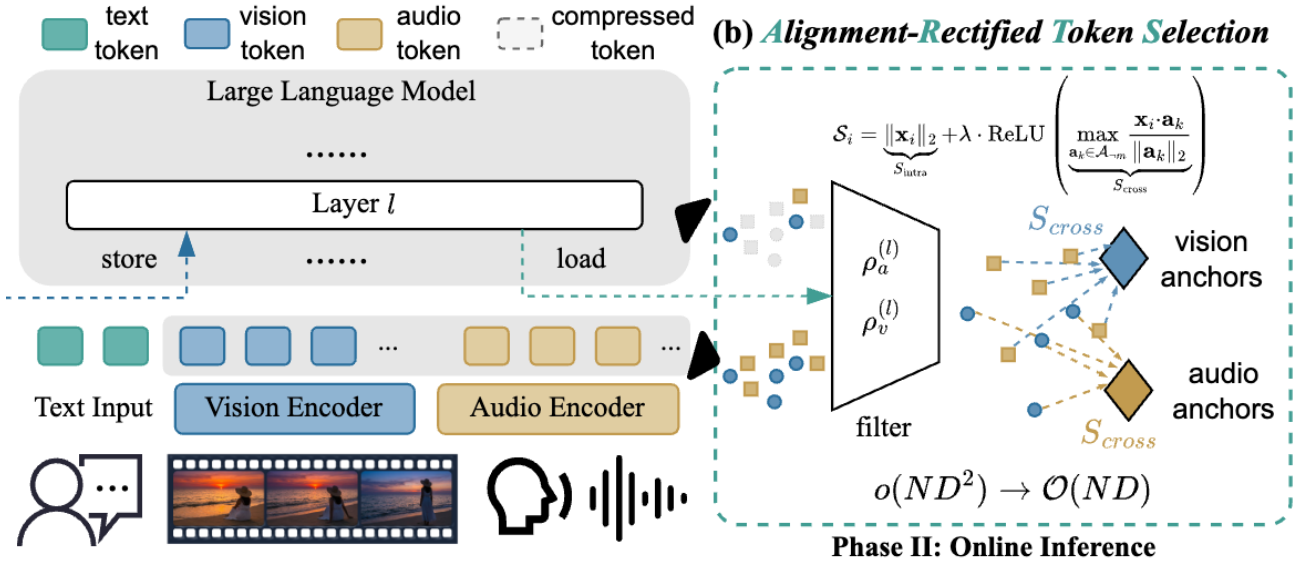


OmniFit (ICML'26 spotlight)

- Method

$$\mathcal{S}_i = \underbrace{\|\mathbf{x}_i\|_2}_{S_{\text{intra}}} + \lambda \cdot \text{ReLU}\left(\underbrace{\max_{\mathbf{a}_k \in \mathcal{A}_{\neg m}} \frac{\mathbf{x}_i \cdot \mathbf{a}_k}{\|\mathbf{a}_k\|_2}}_{S_{\text{cross}}} \right)$$

  - ARTS: 뭘 남길지 선택
    - 인코더 직후 토큰에 DPC-KNN으로 모달리티별 Anchor 토큰 선택
    - 각 토큰의 중요도 점수 계산
      - Intra Score: 자기자신 L2 Norm
      - Inter Score: 반대 모달리티 Anchor 토큰들과의 Cosine Similarity 중 최댓값
    - 앞에서 구한 Layer별 / 모달리티별 보존율에 따라 토큰 merging



OmniFit (ICML'26 spotlight)

- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능

Qwen2.5-Omni-3B 기준 비교

Method	Audio-Video Tasks							Video-Only Tasks			Avg. (%)
	AVUTBench	VideoMME	WorldSense	Daily-Omni	Video-Holmes	UNO-Bench	OmniBench	MVBench	EgoSchema	MLVU	
Full Tokens	62.5	62.8	46.0	59.8	25.8	26.5	52.2	68.7	61.4	68.9	100.00
Retain 40% Tokens											
RandomDrop	57.1	60.9	40.2	52.1	18.8	22.1	49.5	67.5	60.5	65.2	90.54
FastVid (Shen et al., 2025)	58.9	61.5	41.6	55.9	20.1	23.5	50.1	**68.5**	61.4	68.7	93.81
DyCoke (Tao et al., 2025a)	58.1	61.2	42.8	55.1	20.2	22.9	50.0	67.9	61.0	65.4	92.92
OmniZip (Tao et al., 2025b)	60.8	62.7	44.8	59.5	**25.8**	26.1	**52.0**	-	-	98.87	
OmniFit (*Ours*)	**62.3**	**62.9**	**45.7**	**60.1**	**25.8**	**26.6**	51.9	**68.5**	**61.5**	**68.9**	**99.94**
Retain 30% Tokens											
RandomDrop	53.2	59.6	35.3	50.3	17.6	16.5	46.9	63.8	56.5	62.8	83.72
FastVid (Shen et al., 2025)	56.8	61.2	38.9	51.8	20.1	18.4	48.1	66.2	**61.4**	65.1	88.99
DyCoke (Tao et al., 2025a)	55.4	60.5	39.8	51.8	19.6	18.5	47.8	64.1	59.2	65.1	87.97
OmniZip (Tao et al., 2025b)	59.5	**62.5**	42.7	58.6	25.3	25.9	51.4	-	-	-	97.12
OmniFit (*Ours*)	**61.9**	**62.5**	**45.4**	**60.0**	**25.6**	**26.4**	**51.5**	**68.3**	**61.4**	**68.0**	**99.32**
Retain 20% Tokens											
RandomDrop	48.5	50.1	20.9	40.9	10.8	11.2	40.3	59.4	51.5	55.7	68.37
FastVid (Shen et al., 2025)	53.9	60.1	35.2	45.3	19.9	12.0	47.5	65.4	61.0	64.4	83.56
DyCoke (Tao et al., 2025a)	51.6	58.8	35.9	44.5	19.1	16.1	45.6	61.2	58.6	64.9	82.95
OmniZip (Tao et al., 2025b)	57.2	60.5	41.6	57.5	24.6	25.0	50.5	-	-	-	94.41
EchoingPixels (Gong et al., 2025)	-	60.7	45.0	**60.6**	-	-	-	-	-	**68.3**	**98.74**
OmniFit (*Ours*)	**61.5**	**62.0**	**45.1**	59.8	**25.5**	**26.1**	**51.2**	**68.0**	**61.2**	67.2	98.68

OmniFit (ICML'26 spotlight)

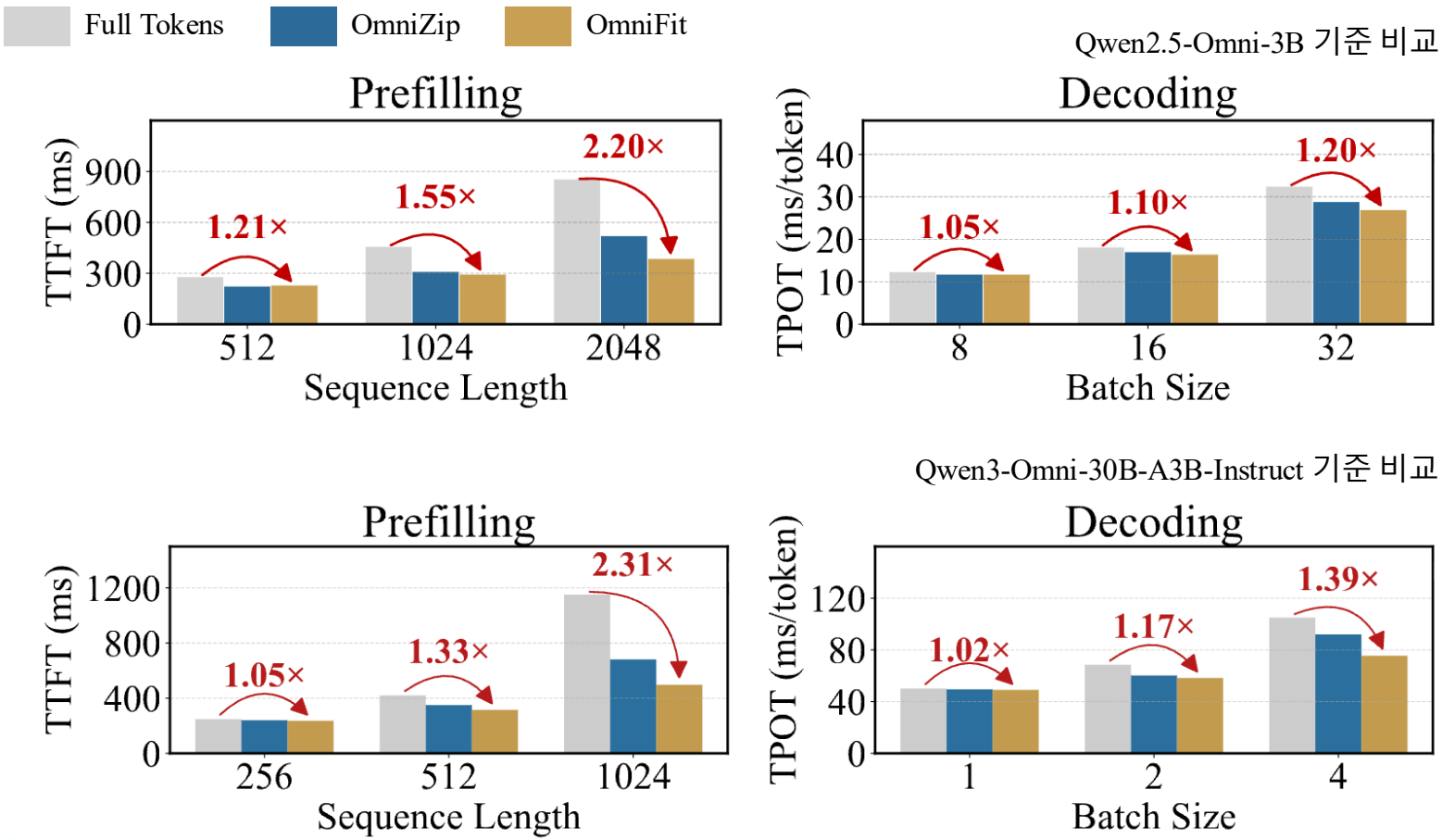
- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능

20% 보존율 기준 비교

Method	Audio-Video Tasks					Video-Only Tasks			Avg. (%)
	VideoMME	WorldSense	Daily-Omni	Video-Holmes	OmniBench	MVBench	EgoSchema	MLVU	
Qwen2.5-Omni-7B (Xu et al., 2025a)									
Full Tokens	66.1	47.5	62.3	27.1	56.1	70.3	68.6	68.5	100.00
RandomDrop	52.3	30.3	40.8	17.6	45.0	59.8	60.2	53.2	75.51
FastVid (Shen et al., 2025)	55.7	33.4	42.6	20.9	48.1	64.3	**65.9**	62.1	83.00
OmniZip (Tao et al., 2025b)	64.4	44.2	59.9	23.2	53.9	-	-	-	93.66
OmniFit (*Ours*)	**65.1**	**46.6**	**61.8**	**25.9**	**55.5**	**68.9**	**65.9**	**64.3**	**97.28**
OmniVinci (Ye et al., 2025)									
Full Tokens	71.3	48.3	66.0	31.2	61.5	71.1	69.2	68.4	100.00
RandomDrop	65.2	29.9	48.3	20.5	47.7	60.9	61.2	51.7	77.43
FastVid (Shen et al., 2025)	67.5	35.5	51.0	23.3	51.1	64.5	68.0	60.2	85.03
OmniZip (Tao et al., 2025b)	70.1	44.2	63.5	26.4	58.9	-	-	-	93.29
OmniFit (*Ours*)	**70.8**	**46.2**	**64.9**	**28.9**	**60.4**	**65.1**	**68.2**	**63.4**	**95.87**
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)									
Full Tokens	73.5	54.0	72.9	56.4	67.5	74.3	72.4	75.2	100.00
RandomDrop	62.7	36.3	51.5	39.2	59.8	65.4	63.8	62.5	80.07
FastVid (Shen et al., 2025)	65.1	40.2	58.8	45.4	62.2	68.9	**66.1**	65.0	85.85
OmniZip (Tao et al., 2025b)	67.8	46.2	66.3	50.3	63.1	-	-	-	90.28
OmniFit (*Ours*)	**70.1**	**51.2**	**70.6**	**52.1**	**65.0**	**70.1**	65.9	**65.1**	**93.46**

OmniFit (ICML'26 spotlight)

- Experiments
  - 여러 벤치마크에서 Random pruning이나 기존 방식들보다 좋은 성능



향후 계획

- 이번 발표에서는 OmniLLM 효율화 중
  Token Compression 관점의 연구들을 살펴봄

- 향후에는 KV-cache, Efficient Attention 등
  Token Compression 외의 다양한 효율화 방법도 함께 검토

- 또한 Cosmos 3와 같이 더 다양한 modality/task를 다루는
  Omni model까지 연구 범위를 확장

Appendix

Appendix

- Qwen2.5-Omni
  - Audio 입력
    - STFT mel 스펙트로그램 (25 ms window / 10 ms hop) → 40 ms 마다 1토큰
  - Video 입력
    - 샘플링 간격에 따라 다름
  - Time window
    - 2 s 간격

Appendix

- Flash Attention
  - 기존 Attention과 동일한 결과를 계산하는 효율적인 구현 방식
  - 전체 $N \times N$ Attention matrix를 저장하지 않고,
    block 단위로 Softmax와 V 연산을 누적하여 동일한 Attention 출력을 계산
  - HBM ↔ SRAM 간 memory I/O를 줄여 속도와 메모리 효율 개선

Appendix

- DPC-KNN
  - feature space에서 대표적인 cluster center token을 선택하는 clustering 방법
  - KNN으로 각 token의 local density ρ_i 계산
  - 자신보다 density가 높은 token까지의 최소 거리 δ_i 계산
  - ρ_i, δ_i가 큰 token → 주변을 잘 대표하면서 다른 cluster와도 구별되는 대표 token

