[Paper Review] Video Captioning and Summarization, and Their Evaluation Methods

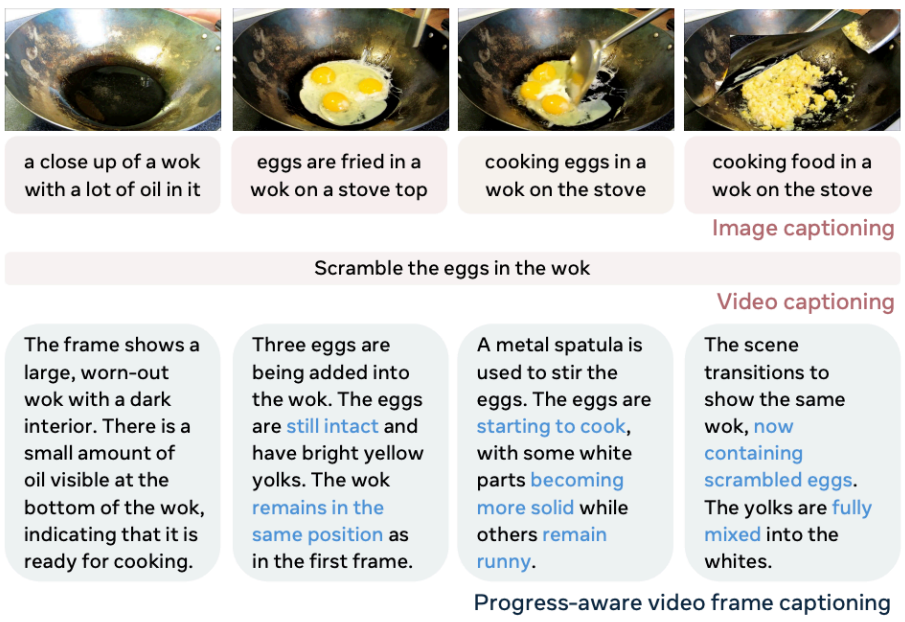
- Video scene 기반 분석을 위해, **비디오 캡셔닝 및 요약에 활용되는 주요 평가 지표와 방법을 조사**하였습니다
- 평가 방법에 대한 이해를 높이기 위해 **최신 비디오 캡셔닝 및 요약 관련 논문들도 함께 리뷰**하였습니다

1. Progress-Aware Video Frame Captioning (CVPR 2025) [link]

▼ **Review**

1. 기존 방식의 문제점

기존 영상 캡셔닝 방식은 다음의 한계를 가진다.



(1) 이미지 캡셔닝 모델의 한계

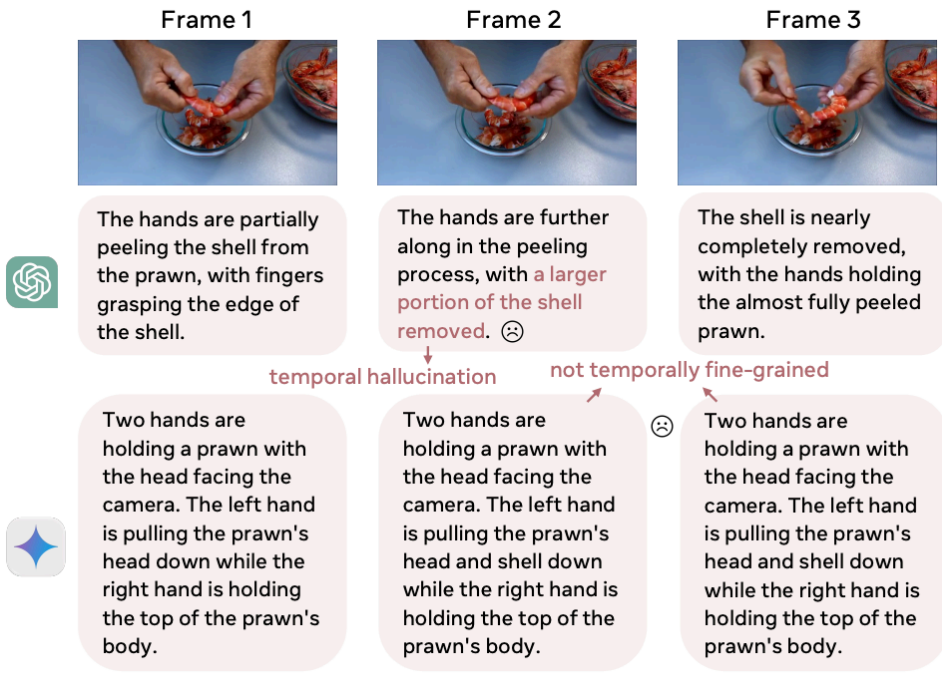
- 각 프레임을 독립적으로 처리하여 프레임 간 미세한 변화(temporal granularity)를 반영하지 못함.

(2) 비디오 캡셔닝 모델의 한계

- 전체 영상을 요약한 **단일 캡션**만 생성.
- 행동이 어떻게 진행되는지(progress)를 설명하지 못함

(3) 비디오 프레임 캡셔닝 모델의 한계

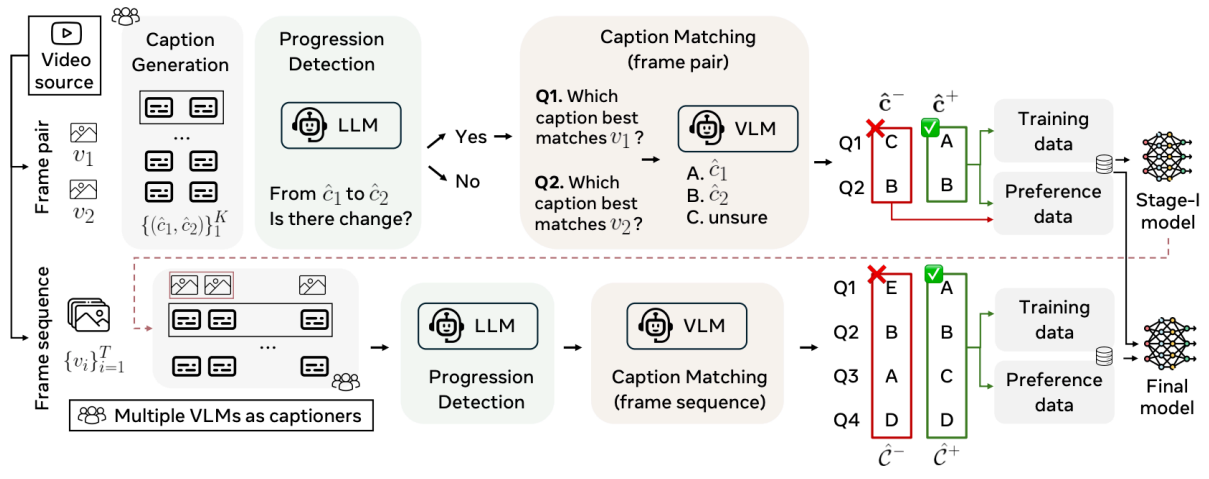
- **Temporal hallucination 발생:** 프레임 간 실제로는 변화가 거의 없음에도 불구하고, 모델이 존재하지 않는 진행 상태(progress)를 지어내는 오류
- **Temporal granularity 부족:** 프레임 간 실제로 존재하는 미세한 변화를 캡션이 충분히 반영하지 못하고 두 프레임을 구분하지 못하는 문제



2. 기여

프레임 간 변화(progress)를 정확하게 포착하면서, 전체 시퀀스에서도 시간적 일관성(temporal coherence)을 유지하는 **Progress-Aware Video Frame Captioning 모델 ProgressCaptioner**를 제안하였다. 이를 위해 pseudo-label 기반 **two-stage training pipeline**을 도입하였다.

3. 방법



3.1 Stage I (Frame pair based)

▼ progress

(a) Caption Generation

- K개의 VLM을 사용해 두 프레임 (v_1, v_2)에 대해 caption pair 생성 → 총 K×2 문장

(b) Progress Detection (caption-based)

- 각 caption pair ($\hat{c}_1$, $\hat{c}_2$)를 LLM에 입력하여 두 문장이 **동일 상태인지 / 진행이 있는지** 판단
- 여러 LLM을 활용해 **majority voting**

(c) Caption Matching (frame–caption alignment)

- progress 있음으로 판단된 pair에 대해 평가용 VLM이 multi-choice로
  - $v_1 \leftrightarrow \hat{c}_1$, $v_2 \leftrightarrow \hat{c}_2$매칭을 수행
  - 두 프레임 모두 올바르게 매칭되면 **$\hat{C}$+**, 하나라도 틀리면 **$\hat{C}$-**

(d) Training

- **$\hat{C}$+** → SFT, **$\hat{C}$-**→ → DPO negative

(e) 가정

LLM은 두 문장이 명확히 다르다고 판단했으나 VLM 매칭이 실패한 경우, 해당 캡션은:

1. **프레임의 미세 변화가 충분히 드러나지 않은 경우**
   - 예:

- Frame1 실제 = 생선을 '토막내는' 중 / Frame2 실제 = 생선을 '회 뜨는' 중
- 생성 캡션 = "칼질한다 / 회를 뜬다"

  → 변화는 있지만 Frame1 표현이 포괄적이어서 매칭 실패 가능

2. **Temporal hallucination**

(f) 기대효과

- 두 프레임의 섬세한 차이를 표현하는 **fine-grained frame pair captioner** 확보

3.2 Stage II (Frame sequence based)

▼ **progress**

(a) Caption Generation

- Stage I 모델 + 추가 K개의 VLM이 T 프레임 전체에 대해 caption sequence 생성 → 총 K′×T 문장

(b) Progression Detection (caption-driven keyframe selection)

- 여러 caption sequence의 문장 변화량(caption differences)을 기반으로 LLM이 progress가 뚜렷한 M개의 key frame을 선택, LLM의 majority voting 사용

(c) Caption Matching

- 선택된 M개의 key frame을 대상으로 각 caption sequence의 M개 문장을 multi-choice 매칭
- M개 모두 맞으면 → SFT, 절반 초과 실패하면 → C→ DPO negative, 중간 영역(모호한 시퀀스)은 제외

(d) 가정

- caption sequence 중간에 hallucination이나 불일치가 있으면 (c) 과정에서 매칭 실패

(e) 기대효과

- 긴 시퀀스에서도 진행 흐름이 자연스럽고 일관된 caption sequence 생성 가능

3.3 전체 과정 요약

Stage I

- **목표:** 프레임 간 미세 변화(progress)를 정확히 서술
- **방법:** 2-frame 중심 학습, LLM 기반 progression 판단 + VLM matching으로 고품질 pair 선별

Stage II

- **목표:** 전체 시퀀스의 시간적 일관성(coherence) 확보
- **핵심:** caption differences 기반 keyframe 선택 + sequence-level matching

▼ **Evaluation**

1. Benchmarking Video Frame Captioning

Model	Size	HTC		COIN		Penn&K	
		Cap	Prog	Cap	Prog	Cap	Prog
Proprietary models							
Gemini-1.5-Pro [53] (img)	-	28.4	59.7	*24.3*	58.6	15.3	51.2
Gemini-1.5-Pro [53]	-	*31.4*	63.8	**25.0**	*63.8*	17.6	*60.3*
GPT-4o [2]	-	**32.4**	*64.2*	21.3	58.4	*18.2*	**63.2**
Open-source models							
Idefics2 [31]	8B	2.0	54.4	2.9	52.2	12.5	50.9
VILA [40]	8B	6.9	53.6	5.1	48.2	15.9	51.4
Qwen2-VL [62]	7B	13.7	**69.6**	11.0	**70.8**	8.5	58.8
LLAVA-Video [83]	7B	3.9	59.3	8.8	53.0	9.7	51.8
LLAVA-OV [33] (img)	7B	5.9	56.3	17.6	55.4	11.9	55.5
LLAVA-OV [33]	7B	7.8	59.0	5.9	57.3	5.1	50.8
PL (VLM ensemble)	-	18.6	62.5	17.6	60.1	**19.3**	52.4
ProgressCaptioner (ours)	7B	**37.3**	**73.6**	**32.3**	**66.1**	**31.3**	**63.7**

Table 1. Results on the FrameCapEval Benchmark, composed of video from four public datasets. Cap and Prog denote caption matching and progression detection accuracy, respectively. PL denotes the pseudo labeling baseline adopting filtered captions from multiple VLMs. ProgressCaptioner greatly outperforms SOTA open-source VLMs and even the leading proprietary models, de-

평가 데이터셋

- FrameCapEval은 프레임 간 실제 progression이 존재하는 영상만 선별하여 구성함.
- HowToChange, COIN, Penn Action, Kinetics Temporal 등 4개 데이터셋에서 총 684개 영상을 추출함.
- 모든 영상은 1FPS로 샘플링된 T-frame 시퀀스로 변환함.
- 기존 비디오 캡션 벤치마크와 달리 프레임별 상태 변화를 평가할 수 있도록 설계됨.
- fine-grained frame captioning 성능 비교를 위해 새롭게 구축된 벤치마크임.

평가 방법

- 다양한 캡셔닝 VLM이 동일 프레임 시퀀스에 대해 프레임별 캡션을 생성함.
  - **Cap(Caption Matching Accuracy)**은 Gemini-1.5-Pro가 프레임–캡션 정합성을 multi-choice로 판정함. → 모델이 **생성한 문장이 실제 프레임의 시각적 상태를 정확하게 기술했는지**를 객관적으로 검증하는 지표
  - **Prog(Progression Accuracy)**은 Llama-3.1-70B가 인접 캡션 간 차이 여부를 판단하고 GT progression과 비교함. → **프레임 간 상태 변화가 캡션에 올바르게 반영되었는지**를 측정하여 temporal granularity를 평가하는 지표
- 모든 모델은 동일한 평가자(Gemini/Llama)를 사용하여 공정하게 비교됨.

실험 결과

- ProgressCaptioner는 Table 1에서 Cap·Prog 두 지표 모두 최고 성능을 기록함.
- Cap은 기존 오픈소스 VLM 대비 3~10배 높고, GPT-4o·Gemini보다도 우수함.
- Prog은 HowToChange 기준 73.6%로 가장 뛰어난 progression 인식 능력을 보임.
- 모든 데이터셋에서 일관된 우수 성능을 달성함.

2. User Study

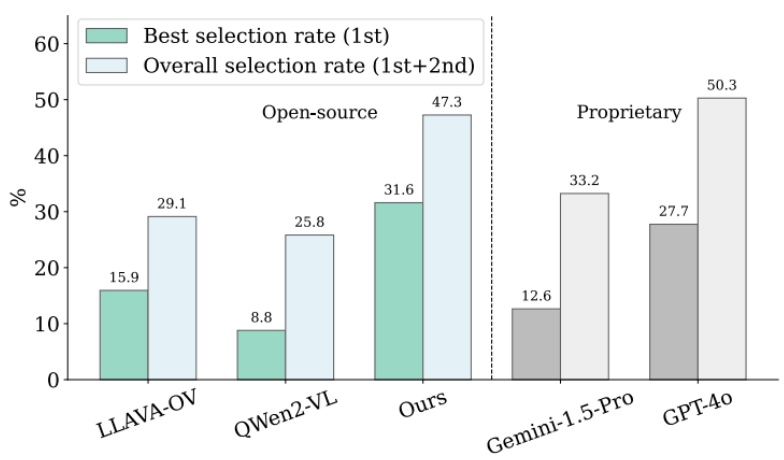


Figure 7. User study results comparing ProgressCaptioner with top competitors show it as the most preferred model (see text).

평가 데이터셋

- User Study는 FrameCapEval과 별도로 사람이 직접 비교 평가할 샘플을 구성함.
- 여러 실제 영상 프레임을 선택하여 progression이 드러나는 장면을 포함시킴.
- 정답 캡션은 존재하지 않고 사람의 직접 선택이 평가 기준이 됨.
- 목적은 **“사람이 보기에 어떤 모델의 캡션이 더 정확·자연스러운가”를 평가**하는 것임.

평가 방법

- 각 프레임마다 5개 모델이 생성한 캡션을 나란히 제시함.
- **선호도 조사: 평가자는 가장 좋은 캡션(top-1)과 상위 두 개(top-2)를 선택하도록 함.**
- 필요 시 “None”을 선택해 어떤 캡션도 적합하지 않다고 표시할 수 있게 설계함.
- 총 15명의 평가자가 동일한 방식으로 독립적으로 평가를 수행함.

실험 결과

- ProgressCaptioner는 top-1 선호도 31.6%로 전체 모델 중 가장 높은 선택률을 기록함.
- top-2에서도 47.3%를 기록하여 GPT-4o 바로 뒤에 위치하며 높은 안정성을 보임.
- 오픈소스 모델(LLAVA-OV, Qwen2-VL) 대비 2~3배 이상 높은 사람 선호도를 보임.
- Gemini·GPT-4o 같은 대형 모델보다도 사람 평가에서 우위를 보인다는 점이 특징임

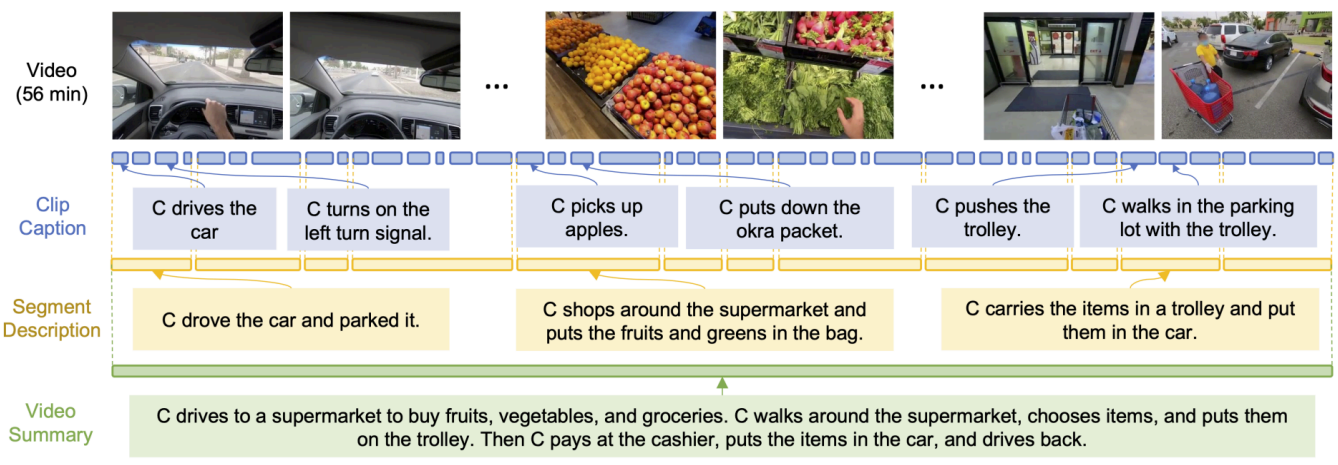
2. Video ReCap: Recursive Captioning of Hour-Long Videos (CVPR 2024) [link]

▼ **Review**

1. 기존 방식의 문제점

기존 비디오 캡셔닝 모델은 다음과 같은 한계를 가진다.

1. **짧은 영상(5~15초)에 최적화:** 영상 전체의 흐름이나 스토리를 연결해서 이해하지 못하고, 매우 단편적 캡션만 생성함
2. **낮은 수준의 행동만 묘사:** 사람의 목표, 의도, 단계적 활동(step), 전체 스토리 같은 상위 수준 의미를 포착하지 못함
4. **시간적 계층 구조를 고려하지 않음:** atomic action → step → goal과 같은 인간 행동의 구조를 반영하지 못해, 긴 영상의 의미를 제대로 구성할 수 없음



예시: Clip caption 수준의 정보(예: “C가 사과를 집는다”, “C가 카트를 민다”, “C가 오크라 팩을 내려놓는다”)만으로는 “이 사람이 어디에 와 있는지”, “무엇을 위해 장을 보고 있는지”와 같은 상위 목적(goal)을 파악할 수 없다.

그러나 인간 행동은 계층적이다: atomic action: 사과를 집는다 → intermediate step: 과일 코너를 돈다 → goal: 장을 본다

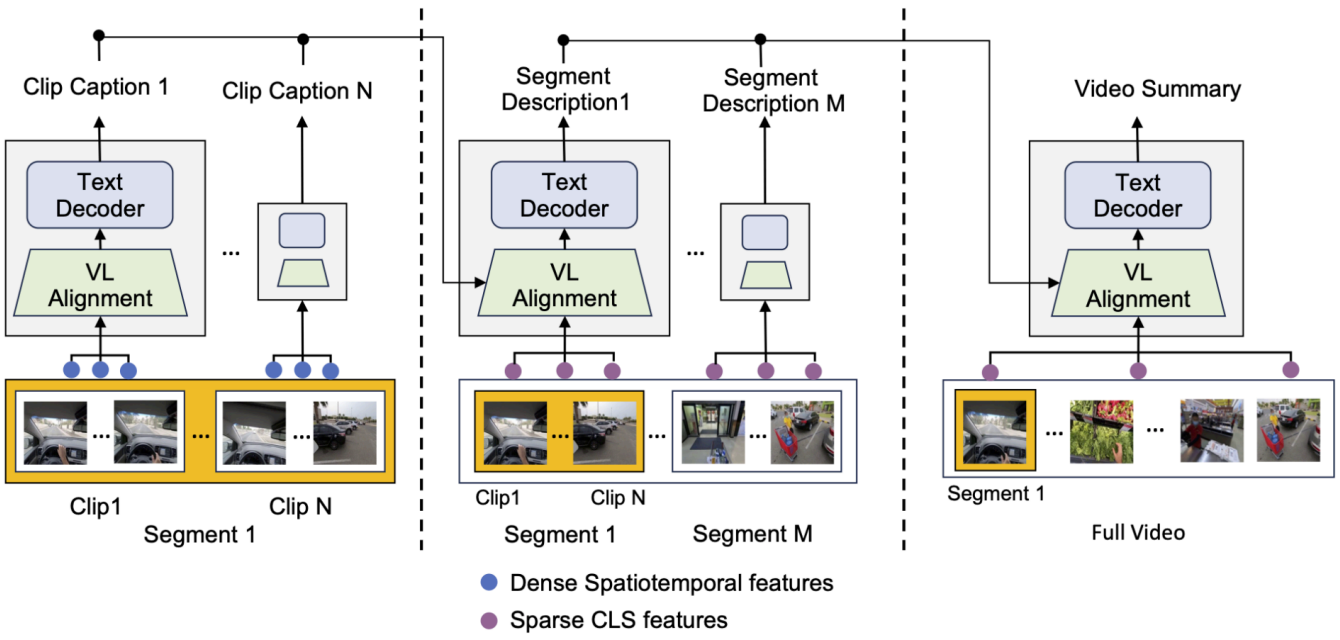
Hierarchy를 구분해야 전체 문맥이 잡히는데, 기존 모델은 모든 장면을 “동등한 중요도”의 짧은 클립으로 보므로, 전체 스토리를 구성할 능력이 없다.

---

2. 기여

장시간 영상에서도 (클립→세그먼트→전체) 요약으로 이어지는 다중 계층 캡션을 효율적으로 생성하는 재귀적 비디오 캡셔닝 모델 Video ReCap과 이를 위한 Ego4D-HCap 데이터셋을 제안한다.

3. 방법



▼ 전체 모델 구성

- Video ReCap은 **Video Encoder + VL Alignment + Text Decoder**로 구성됨
- 긴 영상을 clip 단위로 나눠 TimeSformer로 feature를 추출함
  - 짧은 clip에서는 dense feature를, 긴 구간에서는 CLS feature를 사용함
- 이렇게 얻은 video feature와 이전 계층 caption을 VL Alignment 모듈에 넣음
  - VL Alignment는 video·text를 256개의 토큰으로 고정 길이 압축함
- GPT2 기반 Text Decoder는 이 joint embedding을 받아 해당 계층의 caption을 생성함
- 이 과정이 clip → segment → full video 순으로 재귀적으로 반복됨
- 몇 초부터 몇 시간짜리 영상까지 일관된 구조로 처리함

▼ 계층별 차이

- **계층 1**은 몇 초 clip을 dense feature로 입력받아 세밀한 동작을 묘사함, 텍스트 입력 없이 오직 영상 feature만을 사용함
- **계층 2**는 세그먼트 구간의 clip CLS들을 sparse하게 뽑아 중간 단계 활동을 묘사함, 계층 1의 clip captions도 함께 입력받아 더 긴 행동 단위를 설명함
- **계층 3**은 전체 영상 범위에서 CLS feature를 더 sparsely sampling하여 전체 스토리를 요약함, segment descriptions를 함께 사용해 장기적 목표와 사건 흐름을 요약함
- 계층이 올라갈수록 입력 feature는 더 적어지고, caption은 더 추상적이 됨

▼ VL Alignment 역할

- TimeSformer 기반 비디오 feature와 이전 계층의 텍스트 feature를 하나의 시각-언어 공간으로 정렬함
- **작업**
  1. DistilBERT 내부에 cross-attention 레이어를 삽입하여, 비디오 토큰은 Key/Value로, 텍스트 토큰은 Query로 사용함.
  2. DistilBERT와 텍스트 디코더는 업데이트하지 않고 동결하며, 비디오 토큰이 cross-attention 레이어를 통해 텍스트 표현 공간으로 projection되도록 학습함.
  3. cross-attention 레이어가 텍스트 토큰과 의미적으로 관련된 비디오 토큰을 선택적으로 참조하도록 하여, 멀티모달 토큰을 생성함.
- **효과**
  - 영상·텍스트에서 얻은 수천 개의 토큰을 256개의 멀티모달 토큰으로 압축함으로써 연산 비용을 줄임.
  - Text Decoder가 영상·텍스트 정보를 동시에 참조해 multimodal reasoning을 할 수 있게 함
  - clip caption → segment → summary 로 이어지는 계층 간 정보를 연결하는 허브 역할을 함

- **멀티모달 정보의 중요성**
  - 텍스트 정보만으로는 행동의 의도와 상호작용을 구분하기 힘듦
  - low-level visual cue(예: 속도·방향 변화, 신체 자세)는 fine-grained action(e.g., 달리기 vs 달리는 중 누군가를 쫓음)을 구분하는 데 필요한 핵심 단서
  - high-level visual cue(예: 배경 장면, 장소, 물체 배치)는 coarse-grained summary(e.g., 산책 vs 쇼핑 vs 청소)에서 전반적 활동(goal)을 결정하는 단서

▼ 학습 방법

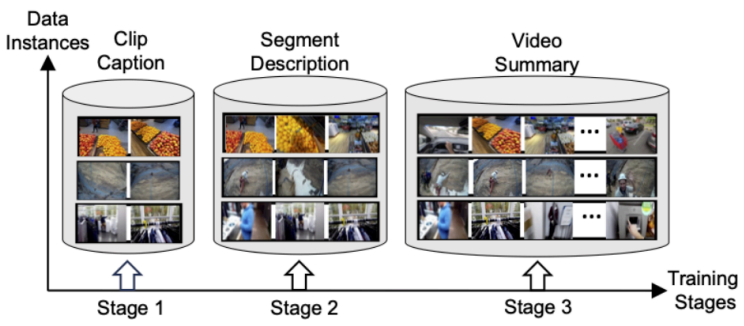


Figure 3. **Hierarchical Curriculum Learning.** We gradually learn the hierarchical structure of the video, starting from short low-level captions to long high-level video summaries.

- 데이터의 계층적 구조를 반영하기 위해 **Curriculum Learning**을 사용함
- 먼저 clip caption만으로 학습하며 가장 풍부한 low-level 인식을 습득함
- 이후 segment descriptions로 확장해 중간 단계 reasoning을 배움
- 마지막으로 video summary를 학습해 긴 스토리 요약 능력을 습득함

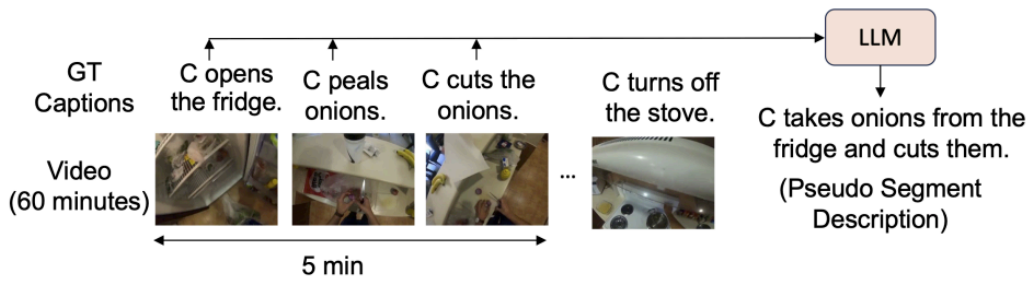


Figure 4. **Large Language Model Supervision.** Given short-term ground truth captions, we use an LLM to generate pseudo-ground truth annotations for medium-length segment descriptions and long-range video summaries to augment our training data.

- segment·summary 데이터 부족 문제를 해결하기 위해 LLM(FLAN-T5)으로 pseudo label 생성함

▼ Evaluation

▼ 평가 지표 설명

1. **CIDEr (Consensus-based Image Description Evaluation)**
   - Reference caption들과 후보 caption 간의 *n-gram* ***TF-IDF 기반 유사도***를 측정하여 평균
   - 사람 주석(reference)에 얼마나 잘 부합하는지 평가함 (사람 주석의 중요 키워드를 포함하는가)
   - **점수가 높을수록 더 인간적인/정확한 캡션**
   - **TF-IDF**
     - **TF (Term Frequency):** 문장에서 자주 등장하는 단어를 중요한 단어로 봄 → 해당 단어는 그 문장을 설명하는 데 중요한 역할
     - **IDF (Inverse Document Frequency):** 단어가 여러 reference 문장에 많이 나타나면 덜 중요한 단어로 봄 → 문장에서 자주 등장하면 전체적으로 중요하지는 않음 (예: a, the)
     - 두 값을 곱해서 **문장에서 자주 등장 + 전체 데이터에서 드문 단어**를 가장 중요한 단어로 봄
     - **TF-IDF 기반 유사도**
       - n-gram으로 만든 키워드 각각에 대해 TF-IDF 값을 계산 → TF-IDF 벡터 구성

        - candidate 벡터와 reference 벡터의 내적 계산 → 중요한 단어가 두 문장 모두에 존재하는가를 판단

2. **ROUGE-L (Longest Common Subsequence)**
    - 두 문장 사이에서 “순서를 유지한 채” 가장 긴 부분 문자열(LCS)을 찾아 유사도를 측정
    - LCS가 없으면, 단어만 같고 순서가 엉망인 문장도 좋은 문장으로 평가될 수 있음
    - **문장 구조·순서 평가하는 지표**
    - **LCS**
        - 두 문장에서 순서를 유지하면서 공통으로 나타나는 가장 긴 subsequence
            - **subsequence:** 원래 시퀀스에서 일부 단어를 순서를 유지한 채 선택하여 만든 시퀀스. 단, 연속일 필요는 없음.
        - Query: **he parks the car** in the lot
        - Reference: **he parks the car** and walks away
        - he → parks → the → car, 이 네 개는 순서가 유지되며 공통임, LCS의 길이는 4
        - 유사도는 정밀도 / 재현율/ F1 점수를 활용함
            - Precision = 모델 문장에서 LCS가 차지하는 비율
            - Recall = 정답 문장에서 LCS가 차지하는 비율
            - 둘 다 분자에 LCS의 길이가 들어감 → 순서가 보존된 겹치는 단어 수가 곧 유사도

3. **METEOR**
    - **단순한 n-gram 일치가 아니라, 단어 의미까지 고려해 정교하게 평가**
    - 매칭 4단계(Exact/Stem/Synonym/Paraphrase)를 수행함
        - unigram을 활용하고 이전 매칭된 단어는 이용하지 않음
        - Query: **The man** driving **his** automobile **to the store and** exited **it quickly.**
        - Reference: **The man** drove **his** car **to the store and** got out of **it quickly.**
            - **Exact:** 단어가 형태까지 완전히 동일한 경우 매칭
                - matches += 9
            - **Stem:** 어근 기반 매칭
                - driving → stem = *drive*, drove → stem = *drive* → matches += 1
            - **Synonynm:** 동의어 기반 매칭
                - automobile ↔ car → matches += 1
            - **Paraphrase:** 문맥적 대체 매칭
                - exited ↔ got out of → matches += 1
    - 이후 Fragmentation Penalty를 계산함 (의미가 역전되는 경우 고려)
        - Query: The man **entered** some items and **quickly bought** the store.
        - Reference: The man **quickly entered** the store and **bought** some items.
        - matches = 10
        - 쿼리 내 단어들을 레퍼런스 단어 위치와 매핑 → 1, 2, 4, 9, 10, 7, 3, 8, 5, 6
        - 증가하는 연속 구간으로 나눠 chunks 계산 → [[1,2,4,9,10], [7], [3,8], [5,6]] → 4

$$Penalty = 0.5 \times (chunks/matches)^3$$
$$METEOR = (1 - Penalty) \times Fmean$$
$$F_{mean} = (10PR)/(R + 9P)$$

        - chunks 증가 → penalty 증가 → 유사도(F_mean) 감소

1. Main Results on Ego4D-HCap

Model	Visual Encoder	Text Decoder	Train Params	Clip Caption		
				CIDEr	ROUGE-L	METEOR
Zero-Shot						
BLIP2 [27]	VIT-G	FT5-XL	0	8.1	7.4	12.7
Finetuned						
LaViLa [64]	TSF-B	GPT2	258M	88.56	47.64	28.03
HierVidCap	TSF-B	GPT2	339M	**98.35**	**48.77**	**28.28**
HierVidCap-U	TSF-B	GPT2	113M	92.67	47.90	28.08

(a) Results for short-range clip captioning.

Model	Video Encoder	Text Decoder	Train Params	Pseudo Ann.	Segment Description			Video Summary		
					C	R	M	C	R	M
Zero-Shot										
BLIP2 [27] + GPT3.5 [11]	VIT-G	FT5-XL	0	✗	5.68	16.87	13.47	11.13	22.41	12.10
LaVila [64] + GPT3.5 [11]	TSF-B	GPT2	0	✗	5.79	19.77	13.45	12.16	24.49	12.48
Finetuned										
LaVila [64] + GPT2 [38]	TSF-B	GPT2	336M	✗	38.22	38.10	16.58	17.98	29.48	12.81
LaVila [64] + FLANT5 [14]	TSF-B	FT5-XL	586M	✗	39.13	38.77	16.88	20.12	30.06	13.17
LaViLa [64]	TSF-B	GPT2	258M	✗	24.63	33.31	15.30	6.54	23.97	10.95
HierVidCap	TSF-B	GPT2	339M	✗	41.74	39.04	18.21	28.06	32.27	14.26
HierVidCap	TSF-B	GPT2	339M	✓	**46.88**	**39.73**	**18.55**	29.34	32.64	**14.45**
HierVidCap-U	TSF-B	GPT2	113M	✓	45.60	39.33	18.17	**31.06**	**33.32**	14.16

(b) Results for medium-length segment description and long-range video summary generation.

평가 데이터셋

- Ego4D-HCap은 Ego4D 영상을 기반으로 구축된 계층적 캡셔닝 데이터셋임
- Clip(5.27M, 약 1초), Segment(17.5K, 약 3분), Summary(8.3K, 약 28분)로 구성됨
- 기존 Ego4D에는 장시간 요약이 없어 Summary 데이터 8,267개를 새로 구축함

평가 방법

- 평가 지표로 CIDEr(C), ROUGE-L(R), METEOR(M)을 사용함
- zero-shot baseline(BLIP2 등)과 fully finetuned baseline(LaViLa 등)을 비교함
- pseudo-annotation 추가 여부에 따른 성능 변화를 분석함

실험 결과

- Video ReCap은 모든 계층(Clip/Segment/Summary)에서 모든 baseline 대비 우수함.
- 특히 CIDEr 기준, LaViLa 대비 Clip **+9.79**, Segment **+17.11**, Summary **+21.52** 향상함.
- pseudo-annotation을 추가하면 Segment **+5.14**, Summary **+1.28** CIDEr 상승함.

2. Video-Language Input Ablation

Input	Segment Description			Video Summary		
	C	R	M	C	R	M
Video	40.17	38.65	17.59	25.64	29.61	13.57
Text	40.10	38.02	17.41	23.23	29.17	13.31
Video + Text	**41.74**	**39.04**	**18.21**	**28.06**	**32.27**	**14.26**

Table 4. **Video-Language Input Ablation.** Using both sparse video features and recursive text inputs leads to better performance for both segment description and video summary generation.

평가 방법

- recursive caption이 없는 경우 (Video-only) / 있는 경우를 비교함
- sparse video feature와 text feature가 서로 보완적인지 검증함
- 입력 모달리티만 변경하고 나머지 모델 구조는 동일하게 유지함

실험 결과

- Video+Text 조합이 Segment와 Summary 모두에서 **최고 성능을 기록함**.
- Segment에서 Video+Text는 Video-only 대비 +1.57, Text-only 대비 +1.64 CIDEr 향상함.
- Summary에서는 Video+Text가 Video-only 대비 +2.42, Text-only 대비 +4.83 CIDEr 향상함.

- Text-only는 Segment에서는 준수하지만 Summary에서는 큰 성능 저하가 있음 → Visual 정보는 기본 성능을 만드는 데 매우 중요한 역할

3. Hierarchical Curriculum Learning

Training Scheme	Segment Description			Video Summary		
	C	R	M	C	R	M
Init → Segment	36.81	38.70	17.17	-	-	-
Caption → Segment	**41.74**	**39.04**	**18.21**	-	-	-
Init → Video	-	-	-	8.62	26.33	11.24
Caption → Video	-	-	-	24.84	30.74	13.25
Caption → Segment → Video	-	-	-	**28.06**	**32.27**	**14.26**

Table 5. **Hierarchical Curriculum Learning.** Using the proposed curriculum learning scheme yields a performance boost of +4.93% in segment description and +19.44% in long-range video summary generation compared to training the model from GPT2 pretrained weights (Init).

평가 방법

- 다섯 가지 학습 방식 비교: Init→Segment, Caption→Segment / Init→Video, Caption→Video, Caption→Segment→Video.
- 각 방식은 학습 시작 계층과 학습 진행 순서가 다름.

실험 결과

- Clip→Segment→Summary 순으로 학습하는 방식이 압도적 최고 성능을 기록함.
- curriculum 방식이 장시간 captioning의 핵심임을 명확히 보여줌
  - 모델이 짧은 시간 단위에서 “행동 개념”을 먼저 습득해야 그 위에서 장거리 이야기 구조를 학습할 수 있음
  - 클립은 수백만 개지만 Summary는 8천 개밖에 없음, 처음부터 Summary를 학습하면 모델이 overfit

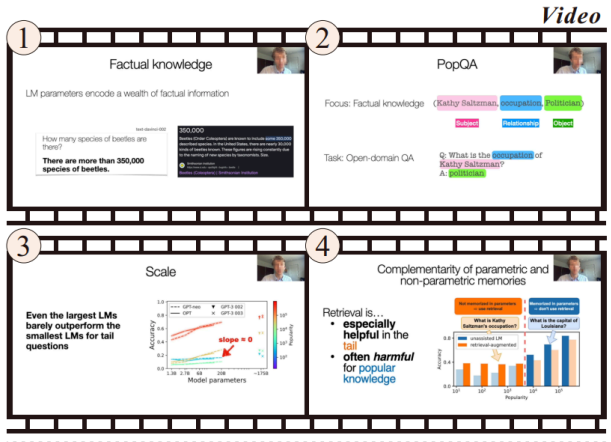
3. What Is That Talk About? A Video-to-Text Summarization Dataset for Scientific Presentations (ACL 2025) [link]

▼ **Review**

1. 기존 방식의 문제점

- 기존 Video-to-Text 모델들은 일반 영상 요약에는 강하지만, **학술 발표 영상처럼 전문적·구조적·시각정보 많은 콘텐츠**에서는 성능이 크게 떨어짐.
- 이유는 전문 용어·도표·논문 구조 이해 부족, 영상·음성·텍스트를 통합적으로 활용할 데이터 부족, 추론 없이 단순한 end-to-end 방식 때문임 .

2. 기여



Summary

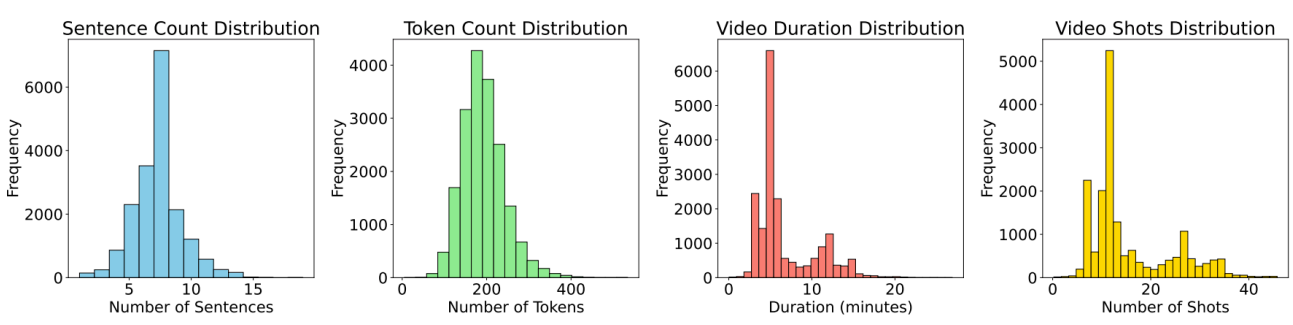
Despite their impressive performance on diverse tasks, large language models (LMs) [...], implying the difficulty of encoding a wealth of world knowledge in their parameters. This paper aims to understand LMs' strengths and [...], by [...]. We find that LMs struggle with less popular factual knowledge, and [...]. Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails [...]. Based on those findings, we devise a new method for retrieval-augmentation[...] memories when necessary.

- 학술 발표 영상 → 논문 초록 수준의 요약을 가능하게 하는 대규모 영상-텍스트 데이터셋 VISTA(18,599개)를 최초로 제공함
- 기존 End-to-End 방식의 한계를 해결하기 위해, **요약 전에 '계획(plan)'을 생성하는 Plan-based Summarization**을 도입해 SOTA 성능을 달성함

3. 방법

Dataset	Language	Domain	#Videos	VideoLen	SumLen
MSS (Li et al., 2017)	English, Chinese	News	50	3.4	—
YouCook2 (Zhou et al., 2018)	English	Cooking	2.0K	5.3	67.8
VideoStorytelling (Li et al., 2019)	English	Open	105	12.6	162.6
VMSMO (Li et al., 2020)	Chinese	Social Media	184.9K	1.0	11.2
MM-AVS (Fu et al., 2021)	English	News	2.2K	1.8	56.8
MLASK (Krubiński and Pecina, 2023)	Czech	News	41.2K	1.4	33.4
VideoXum (Lin et al., 2023)	English	Activities	14.0K	2.1	49.9
Shot2Story20K (Han et al., 2025)	English	Open	20.0K	0.3	201.8
BLiSS (He et al., 2023)	English	Livestream	13.3K	5.0	49.0
SummScreen^{3D} (Papalampidi and Lapata, 2023)	English	Open	4.5K	40.0	290.0
Ego4D-HCap (Islam et al., 2024)	English	Open	8.3K	28.5	25.6
Instruct-V2Xum (Hua et al., 2024)	English	Open	30.0K	3.1	239.0
MMSum (Qiu et al., 2024)	English	Open	5.1K	14.5	21.7
LfVS-T (Argaw et al., 2024)	English	YouTube	1.2K	12.2	—
VISTA (ours)	English	Academic	18.6K	6.8	192.6

Table 1: Comprison of video-to-text summarization datasets. #Videos = the number of videos, whereas VideoLen and SumLen refer to the average of video duration (in minutes) and the average number of summary tokens.



- **데이터셋: VISTA**
  - AI/ML 주요 학회(ACL·EMNLP·NAACL·ICML·NeurIPS)의 **발표 영상 + 논문 초록** 페어로 구성
  - 평균 영상 길이: 약 6.8분
  - 평균 요약 길이: 192.6 토큰
  - 총 18,599개 페어로 기존보다 **길고 구조적**인 데이터셋 제공

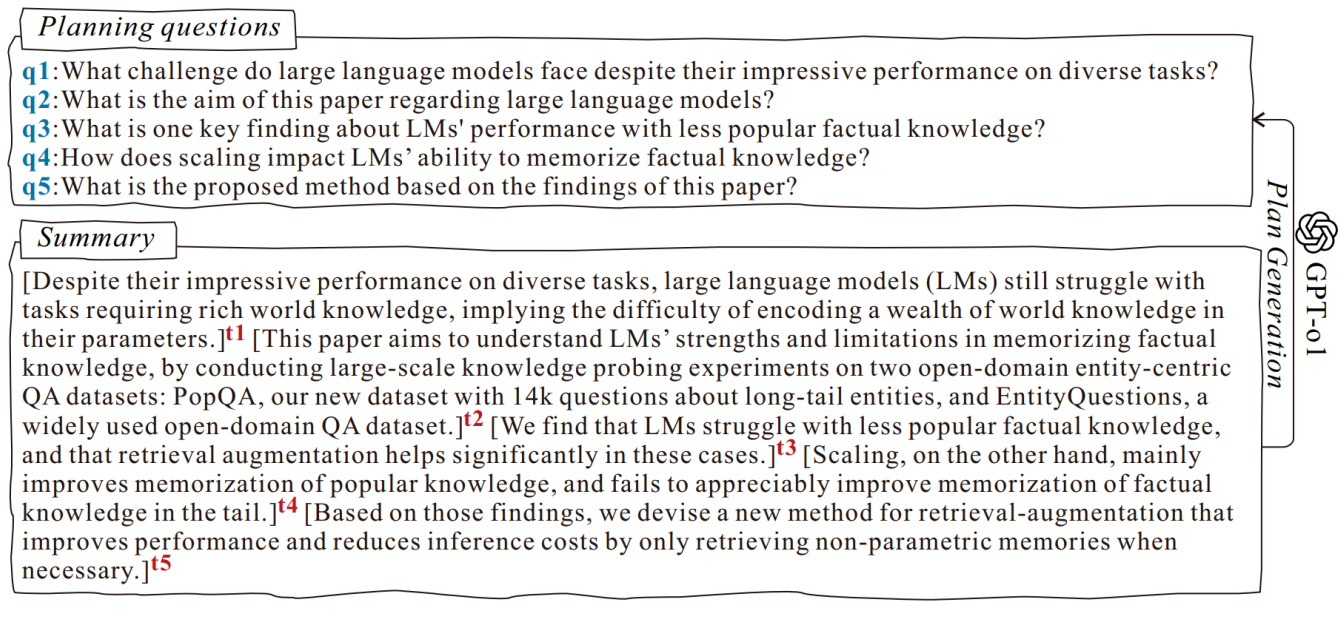


Figure 4: GPT-o1 generates plans based on reference summaries. Each question q_i corresponds to a summary sentence t_i, which we assume constitutes its answer. Index i ranges from 1 to the number of summary sentences.

- **요약 방식: Plan-based Summarization**
  - **(1) Plan Generation (PG)**
    - 영상(v) → 먼저 '계획 p'를 생성
    - 계획 p는 "요약 각 문장에 대응되는 질문 리스트"
    - **QUD(Question Under Discussion) 이론** 기반 → 요약 구조를 안정적으로 맞춤
      - 각 문장이 어떤 질문에 답하는지를 기준으로 글의 논리를 구성하는 이론
      - 과학 발표 영상의 요약은 **문제–방법–결과–결론** 같은 구조적 흐름이 중요한데, 기존 모델은 이를 놓치기 쉬움
      - QUD 기반 질문 리스트(Plan)를 먼저 생성하면 모델이 **자연스러운 구조를 따라 요약**을 생성할 수 있음
    - GPT-o1로 silver-standard plan 생성 후 학습
      - 예:
        - q1: 발표의 핵심 문제는 무엇인가?
        - q2: 제안된 방법은 무엇인가?
        - q3: 어떤 실험적 성능 향상이 있는가?
  - **(2) Summary Generation (SG)**
    - [영상 v + 계획 p]를 입력으로 요약 s를 생성
    - 각 문장이 plan의 각 질문을 순서대로 "답변"하도록 유도
- **학습 방법**
  - 논문 초록을 문장 단위로 QUD 기반의 Plan Questions로 변환하여 (v,p,s)형태의 supervised training triplet 데이터 구성
  - **(PG)** 입력 비디오 v로부터 구조적 질문 시퀀스 p를 예측하도록 학습
  - **(SG)** P(s | v,p)를 최대화하는 방향으로 학습 = 초록 수준의 요약 s를 생성하도록 학습

▼ **Evaluation**

▼ **평가 지표 설명**

1. **Automatic Evaluation Metrics**
   - **main metrics**
     - **ROUGE-2 (R2)**: 생성된 요약과 참조 초록의 2-그램(두 단어 연속) 겹침 비율 (recall 기반)
     - **ROUGE-L:** 요약과 참조 사이의 LCS를 기반으로 한 구조적 유사도
     - **VideoScore:** 영상 입력(video)과 모델이 만든 요약(summary) 간의 시맨틱 정합성(semantic alignment)를 측정하는 지표 [link]

▼ **설명**

- VideoScore 중 TVA(Text-to-Video Alignment)를 사용함
  1. 사람이 37.6K 영상에 1~4점으로 TVA 점수를 직접 부여함 (VIDEOFEEDBACK 데이터셋)
     - 1-Bad, 2-Avg, 3-Good, 4-Perfect
     - 프롬프트에 언급된 객체·인물이 영상에 실제 등장하는지 확인함
     - 프롬프트의 동작·이벤트가 영상에서 재현되는지 확인함
     - 수량·크기·색·상태 등 속성이 텍스트와 일치하는지 점검함

     - 1점(Bad): 핵심 요소 여러 개 불일치 / 전반적으로 정렬 실패.
     - 2점(Average): 일부는 맞지만 중요한 오류 존재.
     - 3점(Good): 대부분 맞고 큰 오류 없음.
     - 4점(Perfect): 거의 완벽한 정렬, 후검수에서 별도로 선정된 사례
  2. 인간 TVA 점수를 정답(label)로 사용해 Mantis-Idefics2-8B 기반 모델을 미세조정하여 VideoScore 모델을 학습함
  3. 학습된 VideoScore 모델은 새로운 영상·프롬프트에 대해 TVA 점수를 자동 산출하는 평가 모델로 사용됨
- 다른 LLM기반 평가 대비 Spearman 상관계수 매우 높다고 논문에서 언급됨 (Gemini-1.5 Pro: 26.7, VideoScore: 59.5)

◦ **FactVC**: 요약이 영상에 나타난 사실과 얼마나 일치하는지 확인하는 사실성 지표 [link]

▼ **설명**

- 영상→텍스트 생성 모델들이 많은 hallucination를 낸다는 것이 관찰됨
- 기존 ROUGE, BLEU 등은 단지 언어 유사성만 측정하고, 요약된 문장이 영상에 실제로 근거가 있는지는 잘 잡아내지 못함
- 영상 프레임 임베딩과 단어 임베딩 간 유사도를 계산해 hallucination 여부를 잡아냄
  - **coarse-grained similarity**

$$S_c = \cos\left(\frac{1}{|V|}\sum_k E_v(f_k),\ E_t(T)\right)$$

    - 영상 프레임 전체 평균 임베딩과 문장 전체 임베딩 간 유사도
    - “영상 전체 분위기와 문장 전체 내용이 일치하는가?” 평가
  - **fine-grained similarity**

$$S_f^p = \frac{1}{|T|}\sum_{x_j \in T} \max_{f_i \in V} \cos(E_v(f_i), E_t(x_j))$$

    - 문장의 **각 단어**가 영상 **어느 프레임에서 근거를 찾는지** 측정
    - hallucination 평가에 획기적임
    - **요약:** A dog is playing with a ball in the snow.
    - **영상:** dog, ball 존재함, snow 없음
    - dog: 0.9, ball: 0.85, snow: 0.1 → 0.6점으로 낮은 점수
◦ **FactVC**

$$FactVC = (1-\alpha)S_c + \alpha S_f^p$$

        - 논문은 기본적으로 α=0.75를 추천
        - fine-grained(단어 근거 확인)을 더 중요하게 반영

- **sub metrics**
    - **SacreBLEU:** 생성된 요약과 참조 초록의 n-gram precision을 평가 (재현성 향상 버전)
    - **METEOR:** 요약과 참조 사이의 정확한 단어, 어간, 동의어, 패러프레이즈 매칭을 기반으로 의미적 유사도를 평가
    - **BERTScore:** BERT 임베딩을 이용해 reference–candidate 토큰 간 cosine similarity를 비교하고, 그 결과를 Precision·Recall·F1로 계산해 의미적 유사도를 평가
        - 분모는 토큰 수, 분자는 각 토큰의 *max similarity* 값을 모두 더한 합
    - **CIDEr-D:** TF-IDF 기반 n-gram 유사도를 기반으로, 생성된 문장이 여러 참조 요약과 얼마나 "정보량 중심으로" 일치하는지 평가 (중복 억제 버전)

2. **LMM-as-Judge Evaluation**

- GPT-o1이 비디오와 요약을 함께 읽고, 인간 평가 기준(정확성·관련성·정보량·간결성·일관성)에 따라 1~5점 점수를 매겨 모델 요약 품질을 대규모로 자동 평가하는 방식

• Faithfulness (정확성)

요약이 원본 영상의 내용을 정확하게 반영하는 정도를 평가합니다.
정확한 요약은 영상의 정보에 충실하며, 모순·오해·근거 없는 내용을 포함하지 않아야 합니다.

• Relevance (관련성)

요약이 영상의 핵심 주제와 내용에 얼마나 잘 집중하는지 평가합니다.
관련성이 높은 요약은 원본 영상에서 가장 중요한 부분에 초점을 맞춰야 합니다.

• Informativeness (정보량/충실성)

요약이 영상의 주요 요점과 중요한 세부 사항을 얼마나 잘 담고 있는지 평가합니다.
좋은 요약은 영상의 핵심 아이디어와 결과를 명확하고 포괄적으로 전달해야 합니다.

• Conciseness (간결성)

정보를 얼마나 효율적으로 전달하는지 평가합니다.
간결한 요약은 불필요한 중복이나 군더더기를 피하며, 중요한 정보는 모두 포함해야 합니다.

• Coherence (일관성/논리적 흐름)

요약의 문장 흐름과 구조가 논리적으로 잘 조직되어 있는지 평가합니다.
좋은 요약은 자연스럽고 이해하기 쉬운 방식으로 정보를 제시하며, 문장과 문장 사이의 연결이 매끄러워야 합니다.

N Human Evaluation Guidelines

Prerequisites To participate in this evaluation, you must meet the following two criteria: (1) be a Master's or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrate English proficiency at C2 level or higher.[a] If you do not meet both criteria, we kindly ask you to refrain from participating in this task. Eligible participants are encouraged to follow the instructions below carefully.

Instructions The following section provides detailed descriptions of the evaluation metrics and criteria used in this study. Please review the accompanying source video and the candidate summaries thoroughly. After evaluating each summary, assign scores based on the five criteria below, using a 1-to-5 Likert scale where higher scores indicate better quality:

- **Faithfulness:** Assess the accuracy of the summary in representing the content of the source video. A faithful summary should adhere closely to the source material, avoiding contradictions, misinterpretations, or unverified information.
- **Relevance:** Measure how well the summary includes the topics and themes central to the source video. A relevant summary should focus on the content that is most pertinent to the original video.
- **Informativeness:** Evaluate the extent to which the summary captures the main points and essential details of the source video. An informative summary should provide a clear and comprehensive understanding of the video's core ideas and findings.
- **Conciseness:** Determine the efficiency of the summary in conveying information. A concise summary should avoid redundancy and extraneous details while retaining all critical information from the source video.
- **Coherence:** Examine the logical flow and overall structure of the summary. A coherent summary should present information in an organized and easy-to-follow manner, ensuring that ideas connect naturally and transitions between points are smooth.

Rating System For each metric, use the following Likert scale:

- 1 (Worst): Does not meet the criteria at all.
- 2 (Poor): Meets the criteria minimally.
- 3 (Fair): Meets the criteria adequately.
- 4 (Good): Meets the criteria well.
- 5 (Best): Fully meets the criteria.

Overall Ranking After assigning scores to each summary for the individual criteria, rank all candidates from best to worst based on their overall quality. Consider the summaries' performance across all criteria when determining the final rankings.

[a] https://en.wikipedia.org/wiki/C2_Proficiency

Figure 19: A snapshot of the experimental instructions provided to human evaluators.

O Prompt for GPT-o1 to Evaluate Summary Quality

Source Video: {Source Video}
Candidate Summary: {Candidate Summary}
You are tasked with evaluating the quality of the candidate summary based on the provided source video. Please adhere strictly to the following evaluation guidelines and scoring criteria to ensure a consistent and objective evaluation.
Evaluation Guidelines: {Guidelines}
Instructions for Output:

- Provide your evaluation using the following format, outputting scores only.
- Assign a score from 1 to 5 for each dimension, with 1 being the lowest and 5 being the highest.

Output Format:

- Faithfulness: [Score]
- Relevance: [Score]
- Informativeness: [Score]
- Conciseness: [Score]
- Coherence: [Score]

If you encounter ambiguity in evaluating any dimension, prioritize adherence to the evaluation guidelines and provide the most accurate score possible based on the provided information. Do not include any additional comments or justifications in your response.

Figure 20: Prompt for GPT-o1 to evaluate summary quality.

1. LLM/VLM Summarization Benchmark

Method	Model	Open-source	R1	R2	RLsum	SacreBLEU	Meteor	BERTscore	CIDEr-D	VideoScore	FactVC
Zero-shot Learning	LLaMA-3.1$_{transcript}$	✓	23.68	4.22	21.39	2.70	14.62	80.93	1.17	1.53	34.32
	LLaMA-3.1$_{OCR}$	✓	24.02	4.37	21.42	2.63	14.59	80.33	1.19	1.50	34.06
	Qwen2-Audio	✓	23.52	4.29	21.53	2.49	14.77	80.62	1.15	1.59	34.31
	Claude 3.5 Sonnet	✗	27.71	5.59	24.14	3.14	17.53	82.57	1.32	1.91	50.11
	Gemini 2.0	✗	27.82	5.66	24.29	4.22	17.83	82.64	1.47	2.02	52.02
	GPT-o1	✗	27.90	5.69	24.37	4.38	17.90	82.63	1.61	2.17	51.36
	Video-LLaMA	✓	20.18	3.19	21.24	1.76	13.73	81.31	1.08	1.63	32.25
	Video-ChatGPT	✓	20.36	3.52	21.43	1.79	14.01	81.35	1.11	1.63	33.21
	Video-LLaVA	✓	25.29	4.50	22.52	2.82	15.13	81.39	1.17	1.65	36.45
	LLaMA-VID	✓	25.31	4.77	22.53	2.88	15.27	81.32	1.14	1.64	36.39
	LLaVA-NeXT-Interleave	✓	25.41	4.82	22.68	2.92	15.25	81.40	1.18	1.73	40.12
	mPLUG-Owl3	✓	25.57	4.82	22.84	2.99	15.33	81.39	1.21	1.77	42.07
	Plan-mPlug-Owl3*	✓	**25.62**†	**4.95**†‡	**22.97**†‡	**3.14**†‡	**15.39**†‡	**81.45**‡	**1.27**†‡	**1.86**†‡	**47.37**†‡
QLoRA Fine-tuning	LLaMA-3.1$_{transcript}$	✓	32.24	11.38	30.39	8.03	21.57	82.39	3.86	2.81	53.22
	LLaMA-3.1$_{OCR}$	✓	33.01	12.11	30.52	8.04	21.55	82.41	3.92	2.77	53.19
	Qwen2-Audio	✓	32.17	12.05	30.77	7.87	21.86	82.36	4.11	2.80	54.27
	Video-LLaMA	✓	30.74	9.44	28.33	6.45	22.49	82.61	3.99	2.77	52.05
	Video-ChatGPT	✓	31.68	10.50	30.40	7.63	23.67	82.62	4.02	2.78	55.02
	Video-LLaVA	✓	33.16	12.64	30.37	8.17	23.92	82.81	4.26	2.83	59.13
	LLaMA-VID	✓	33.31	12.73	30.49	8.22	23.90	83.01	4.31	2.88	62.20
	LLaVA-NeXT-Interleave	✓	33.37	12.77	30.56	8.30	23.95	83.47	4.47	2.93	66.14
	mPLUG-Owl3	✓	33.40	12.82	30.66	8.29	23.97	83.49	4.47	2.92	70.08
	Plan-mPlug-Owl3	✓	**33.52**†‡	**13.01**†‡	**31.10**†‡	**8.33**	**24.11**†‡	**83.53**†	**4.52**	**3.11**†‡	**73.11**†‡
Full Fine-tuning	LLaMA-3.1$_{transcript}$	✓	33.37	11.93	30.86	8.27	25.12	83.71	4.87	3.21	63.38
	LLaMA-3.1$_{OCR}$	✓	34.02	12.42	31.72	8.51	15.11	84.09	4.89	3.32	65.84
	Qwen2-Audio	✓	33.82	12.37	31.63	8.33	25.09	83.62	4.83	3.22	66.62
	Video-LLaMA	✓	32.19	11.86	31.68	8.41	24.99	83.83	4.77	3.04	64.21
	Video-ChatGPT	✓	32.47	12.11	32.21	8.72	25.09	83.91	4.82	3.11	66.09
	Video-LLaVA	✓	33.28	13.39	32.78	9.10	25.42	83.97	4.87	3.13	66.12
	LLaMA-VID	✓	33.47	13.53	32.80	9.21	25.41	84.03	4.91	3.17	68.30
	LLaVA-NeXT-Interleave	✓	33.75	13.61	32.88	9.26	25.63	84.11	5.01	3.23	73.42
	mPLUG-Owl3	✓	34.22	13.62	32.91	9.32	25.72	84.22	5.03	3.28	71.94
	Plan-mPlug-Owl3	✓	**34.53**†‡	**13.74**†‡	**33.25**†‡	**9.56**†‡	**25.88**†‡	**84.37**†‡	**5.15**†‡	**3.33**†‡	**75.41**†‡

Table 3: Model performance on VISTA dataset. In Plan-mPlug-Owl3*, only the PG module is trained. Plans generated by the PG on the test set serve as input to the SG module for zero-shot inference (no training is applied to the SG module). Symbols † and ‡ indicate that the performance of Plan-mPlug-Owl3 is significantly ($p < 0.05$) different from LLaVA-NeXT-Interleave (third best) and mPLUG-Owl3 (second best), when using the paired t-test.

평가 방법

- 다양한 오픈소스/비오픈소스 LMM·VLM 모델을 여러 평가 지표를 통해 비교함
- 입력 방식(transcript / OCR / Audio / Video 등의 modality)에 따라 같은 모델이라도 다른 성능을 보이도록 설계됨
- 모델별로 zero-shot 요약 성능을 직접 비교하여 베이스라인 품질을 평가하는 목적이 있음
- FactVC·VideoScore 같은 영상 기반 정합성 지표도 포함하여 단순 텍스트 매칭이 아닌 멀티모달 품질을 평가함

실험 결과

- Plan-mPLUG-Owl3가 모든 오픈소스 모델 중 전반적으로 가장 높은 요약 품질을 보여줌
- 특히 **FactVC와 VideoScore가 크게 개선**되어 영상 기반 사실성·정합성이 강화됨
- 기본 LLaVA-NeXT나 Qwen2-Audio 등 기존 모델 대비 Plan 적용 모델이 안정적으로 높은 점수를 기록함

2. Modality Ablation

Modality	Zero-shot Learning				QLoRA Fine-tuning				Full Fine-tuning			
	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC
Video only	2.68	20.34	1.55	28.93	8.83	27.51	2.65	50.66	10.78	30.02	2.91	60.87
Audio only	2.14	19.72	1.41	26.84	7.52	26.34	2.48	45.79	9.23	27.93	2.73	58.02
Transcript only	2.02	18.01	1.34	25.53	6.91	24.33	2.39	44.87	8.44	25.81	2.35	54.11
Video + Audio	**3.19**	**21.24**	**1.63**	**32.25**	**9.44**	**28.33**	**2.77**	**52.05**	**11.86**	**31.68**	**3.04**	**64.21**
Video + Transcript	1.87	18.94	1.39	27.76	7.35	24.82	2.51	48.63	9.01	27.19	2.65	58.91
Audio + Transcript	1.64	18.55	1.35	27.48	7.23	24.73	2.38	47.15	8.57	25.82	2.54	55.39
Video + Audio + Transcript	1.92	19.13	1.47	28.60	7.37	25.29	2.52	50.72	9.22	27.21	2.61	59.30

Table 4: Performance comparison of different modality combinations.

평가 방법

- 입력 modality(Video / Audio / Transcript / Video+Audio / Video+Transcript)를 바꿔가며 요약 성능이 어떻게 변하는지 조사함
- Zero-shot / QLoRA / Full fine-tuning 같은 다른 학습 설정을 동일하게 적용해 modality 효과만 분리하여 봄
- 멀티모달 조합(Video+Audio)이 단일 modality 대비 얼마나 개선되는지를 정량적으로 측정하는 목적임

실험 결과

- **Video+Audio 입력이 모든 학습 설정에서 가장 높은 요약 성능**을 달성함
- Fine-tuning(QLoRA → Full)으로 갈수록 전반적인 성능이 증가하며 modality 간 차이도 더 뚜렷해짐
- Transcript only는 내용 일치도는 맞지만 영상과의 사실성(VideoScore, FactVC)에서 한계가 드러남

3. Plan vs No-Plan

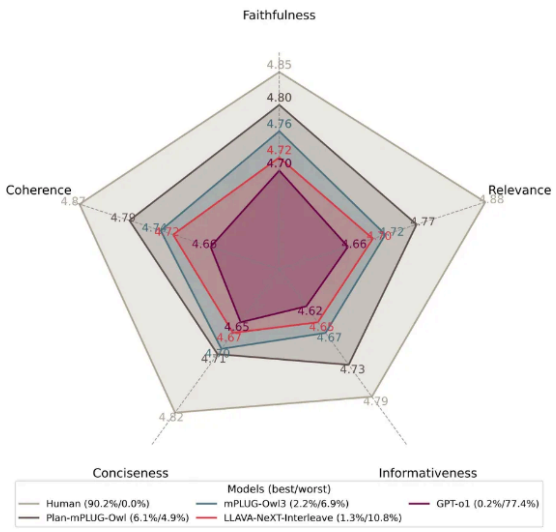


Figure 9: LMM-as-Judge evaluation results showing that human-written summaries consistently outperform neural models.

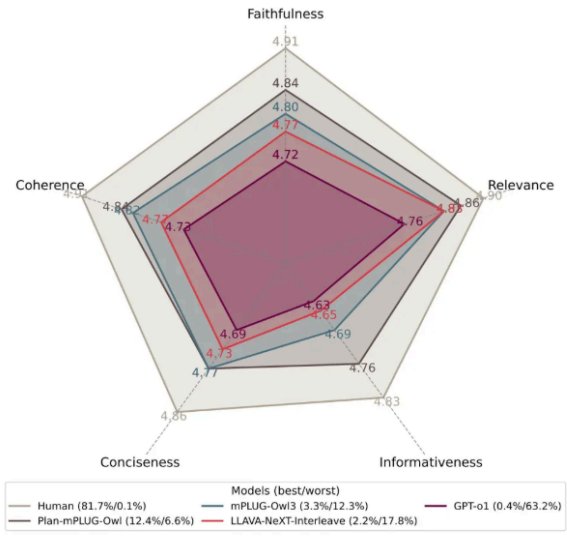


Figure 6: Human evaluation results. Human-written summaries consistently outperform all neural models.

평가 방법

- 동일한 5가지 기준(Faithfulness, Relevance, Informativeness, Conciseness, Coherence)을 사용하여 요약 품질을 평가함
- Figure 9는 GPT-o1을 심사위원으로 사용하는 **LMM-as-Judge 방식**, Figure 6은 실제 사람 평가자에 의한 **Human Evaluation 방식**을 적용함
- 두 평가 방식 모두 인간 요약을 포함한 여러 후보(Plan-mPLUG-Owl, mPLUG-Owl3, LLaVA-NeXT, GPT-o1)를 일관된 조건에서 비교하도록 설계됨
- best/worst(%)는 각 모델이 5개 기준 각각에서 **최고점(1위)** 또는 **최저점(5위)**을 기록한 비율을 나타냄

실험 결과

- 두 평가(Figure 9·6) 모두에서 인간 요약이 모든 기준에서 가장 높은 점수와 가장 넓은 범(≈4.85~4.91)를 기록함
- Plan-mPLUG-Owl은 모든 모델 중 가장 높은 점수대를 형성하며, best 비율도 높고 worst 비율은 낮게 나타남
- GPT-o1 자체 요약은 두 평가 모두에서 가장 낮은 점수와 가장 높은 worst 비율을 보여 요약 품질이 가장 불안정함
- 두 그림 모두 Plan 기반 모델의 일관된 성능 향상을 확인해주며, 인간 평가와 LMM-as-Judge 평가가 동일한 순위 구조를 재현함

4. Integrating Video and Text: A Balanced Approach to Multimodal Summary Generation and Evaluation (AACL 2025) [link]

▼ **Review**

1. 기존 방식의 문제점

- 비디오 + 대본 입력으로 **TV 에피소드 전체를 멀티모달 요약**하는 task를 다룸
  1. 시각·텍스트 모달리티 균형을 맞추지 못하고 **한쪽 정보를 무시**함
  2. 기존 요약 평가지표(ROUGE 등)는 시각 정보 반영 여부를 평가하지 못해 **multimodal summarization 품질을 제대로 측정할 수 없음**

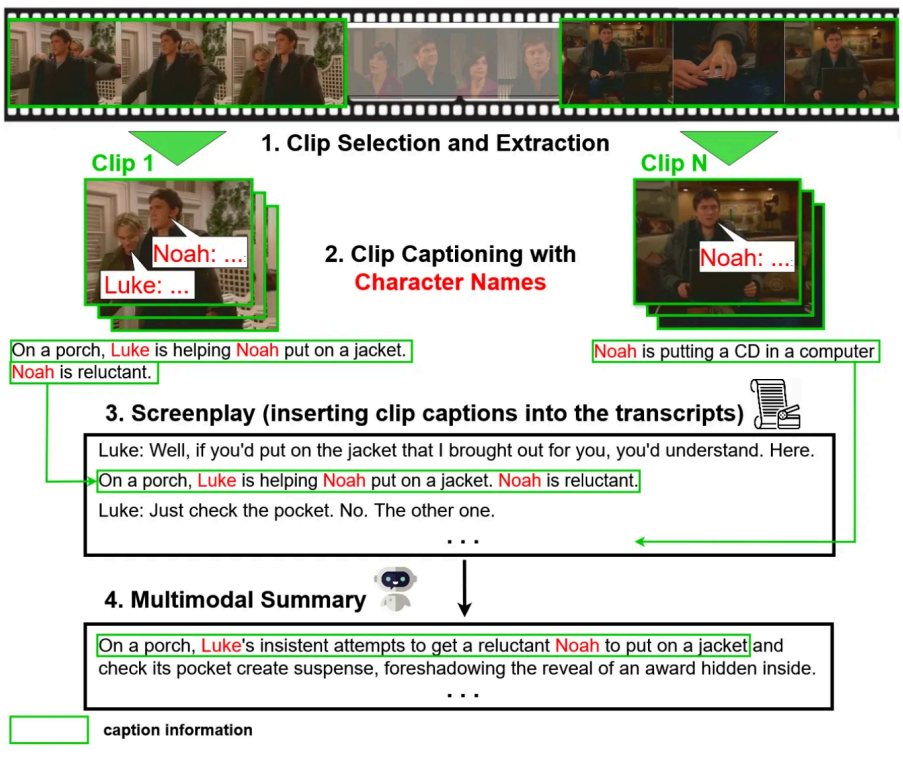
---

2. 기여

- video–text 모달리티 균형 문제를 해결하는 zero-shot 멀티모달 요약 파이프라인인 screenplay 방식을 제안함
- 기존 metric의 한계를 극복하기 위해 visual/textual 균형 평가가 가능한 멀티모달 요약 평가 지표 **MFACTSUM**을 제안함

---

3. 방법

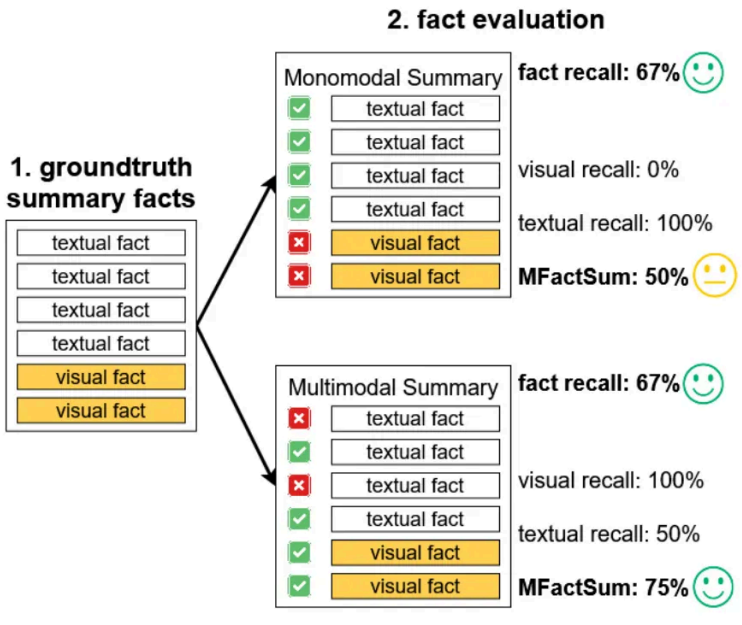


- **Screenplay-based Summarization**
  - 비디오 전체에서 '중요한 구간'을 대사가 없는 구간(silent segments)으로 정의함
    - transcript와 audio를 기준으로 말소리가 일정 시간 동안 감지되지 않는 구간을 Silent Clip으로 판단함
    - 논문은 이러한 silent 구간에서 핵심 시각적 사건(행동·표정·상황 변화)이 자주 발생함을 근거로 삼음

1. 정의된 silent 구간들을 중요한 구간으로 선택하여 video clip으로 추출함
2. 추출된 clip을 VLM에 입력하여, 등장 인물(character)과 행동(action)을 명시하는 clip caption을 생성함
   - 프롬프트에는 transcript를 모델 입력에 함께 제공함
   - 모델이 transcript의 각 대사를 비디오의 오디오와 내부적으로 매칭하여, 각 프레임에 등장하는 캐릭터의 이름을 추론할 수 있게 함

```
<VIDEO CLIP>
Here are the transcripts for the corresponding video:
<CLIP TRANSCRIPTS>
Describe what is happening in the video in all the details.
Explicitly state the names of the characters in your description when possible.
```

3. clip caption을 transcript의 timestamp 위치에 삽입하여, '대사 + 시각 장면 설명'을 결합한 screenplay-like 문서를 구성함
4. LLM이 이 screenplay 문서를 기반으로, 텍스트 정보와 비디오 정보가 균형된 multimodal summary를 생성함



- **MFACTSUM**
  - 비디오 관련 사실(visual facts)과 텍스트 관련 사실(textual facts)을 따로 측정하여 평균을 내는 방식의 metric을 새로 제안 → **시각 정보와 텍스트 정보가 균형 있게 포함되었는지 평가**

- 기존 metric(fact recall)은 두 summary(mono vs multi)를 똑같이 “67%”로 평가하지만, MFACTSUM은 시각 정보(visual fact)를 얼마나 포함했는지를 별도로 계산해서 멀티모달 요약이 더 좋다고 올바르게 평가
- LLM을 기반으로 3단계 과정을 수행함

1. **Fact Identification**
   - groundtruth summary를 문장 → fact(단순 정보 단위)로 쪼갬
   - LLM과 few-shot 예제를 사용
     - “Cassie shakes pills into her hand while Jonathan sleeps.”

       → fact 1: Cassie shakes pills into her hand

       → fact 2: Jonathan sleeps
2. **Visual Fact Classification**
   - 각 fact가 시각 정보 기반인지 또는 대사 기반인지 판단함
   - transcript만 보고 fact를 알 수 있는가? → 알 수 없으면 visual fact
   - few-shot으로 학습된 visual/textual 분류 예시 활용
     - 행동/표정/장면/위치/물리적 움직임 → visual
     - 말한 내용, 관계, 감정 표현 → textual
3. **Fact Evaluation**

> <TRANSCRIPTS> ← generated summary
> Is the Input supported by the above summary?
> Input: <INPUT FACT>. # 1-2과정에서 제작됨
>
> Answer by True or False. Justify your answer.

- 생성된 요약이 fact를 반영했는지 여부를 LLM이 True/False로 결정함
- Recall(정답 요약 속 사실 중에서 예측 요약이 실제로 반영한 비율)을 visual/text 각각 계산 후 평균
- **각 모달리티에서 필요한 fact를 얼마나 빠뜨리지 않았나를 균형있게 평가하는 것이 목표**

$$vis\text{-}rec = \frac{N_{vis\_supported}}{N_{vis\_total}}$$

$$text\text{-}rec = \frac{N_{text\_supported}}{N_{text\_total}}$$

$$MFACTSUM = \frac{vis\text{-}rec + text\text{-}rec}{2}$$

▼ **Evaluation**

1. Evaluation results on SummScreen3D

	vis-rec	text-rec	MFS	fact-rec	r1	r2	rlsum	avg-len
multimodal baselines								
Modular-Kosmos (Mahon and Lapata, 2024)	7.39	19.56	13.48	17.90	44.86	**11.83**	42.97	314.0
VLog	7.77	15.66	11.72	14.62	25.99	3.11	24.69	314.0
Qwen2-VL-72B (no video)	16.16	38.0	27.08	35.0	40.6	8.97	39.28	718.5
Qwen2-VL-72B (video)*	23.69	37.50	30.60	35.61	38.06	8.11	36.95	889.0
Screenplay Summary (Qwen2-VL-72B)	24.43	35.45	29.94	33.93	36.50	7.23	35.51	749.5
Gemini 1.5 Pro (no video)	21.54	42.44	31.99	39.52	41.52	9.04	40.06	573.9
Gemini 1.5 Pro (video)*	27.48	43.00	35.24	40.87	**46.67**	11.77	**44.99**	688.3
Screenplay Summary (Gemini 1.5 Pro)	**33.04**	**45.12**	**39.08**	**43.53**	40.23	8.57	38.82	601.1

Table 1: **Evaluation results on SummScreen3D.** We report the visual recall (`vis-rec`), textual recall (`text-rec`) and MFACTSUM denoted as MFS. For comparison, we also include ROUGE-1 (r1), ROUGE-2 (r2), ROUGE-Lsum (rlsum) and the simple fact recall (`fact-rec`). The average summary word count is denoted by (`avg-len`). Best results are in **bold**. * indicates the VLM is prompted on the full video and transcripts in an end-to-end fashion using the maximum number of frames allowed by the API.

평가 데이터셋

- SummScreen3D 데이터셋의 TV 에피소드 요약을 평가 대상으로 사용함
- 에피소드 길이는 평균 약 40분이며 멀티모달 정보(비디오·대본)를 모두 포함함
- 정답 요약은 인간이 작성한 고품질 요약으로 구성됨
- 평가 대상 모델들은 각 에피소드를 요약해 평균 word 수(avg-len)를 기록함

평가 방법

- **MFACTSUM**을 사용해 visual fact recall(vis-rec)과 textual fact recall(text-rec)을 각각 계산함
- **ROUGE-1, ROUGE-2, ROUGE-Lsum**을 함께 측정해 기존 텍스트 중심 metric과 비교함
- avg-len(평균 요약 길이)을 추가로 보고해 길이 차이로 인한 점수 편향을 점검함
- 모든 baseline은 동일 prompt로 zero-shot 요약을 수행해 공정성을 확보함
- 모델 입력 modality(video / transcript / screenplay)에 따라 성능 차이를 비교함

실험 결과

- video를 직접 입력받은 VLM(Gemini, Qwen2-VL)은 visual fact를 충분히 반영하지 못해 vis-rec이 낮게 나옴
- screenplay 방식은 동일 모델을 사용해도 vis-rec과 text-rec 모두 가장 높은 균형을 보임
- 특히 screenplay 방식의 vis-rec은 Gemini(video) 대비 약 20% 가까이 개선됨
- ROUGE 계열에서는 오히려 전통적 transcript-only 요약이 높음
  - ROUGE는 “정답과 n-gram 단위로 얼마나 똑같이 썼는가?”를 봄
  - 시각적 사실이 포함된 Multimodal summary는 더 abstractive해서 ROUGE에 불리함

2. Ablation results for screenplay

	vis-rec	text-rec	MFS	fact-rec	r1	r2	rlsum	avg-len
w/o handcrafted prompt	20.55	40.66	30.61	37.96	**40.91**	**8.78**	**39.54**	609.9
w/o character ident.	**34.35**	43.32	38.84	42.14	39.97	8.26	38.58	576.5
Screenplay Summary (Gemini 1.5 Pro)	33.04	**45.12**	**39.08**	**43.53**	40.22	8.65	38.90	601.1

Table 2: **Ablation results for our screenplay summarization pipeline using Gemini 1.5 Pro as the based model.** We report the visual recall (`vis-rec`), textual recall (`text-rec`) and MFACTSUM denoted as MFS. For comparison, we also include ROUGE-1 (r1), ROUGE-2 (r2), ROUGE-Lsum (rlsum) and the simple fact recall (`fact-rec`). The average summary word count is denoted by (`avg-len`). Best results are in **bold**.

평가 방법

- 파이프라인의 핵심 구성 요소들이 실제로 multimodal 요약 성능에 기여하는지를 확인하고자 ablation을 수행함
  - **handcrafted prompt**: 스크린플레이 요약 시 LLM이 시각 정보(행동·장소·감정 등)를 적극적으로 포함하도록 유도하기 위한 특별 설계 프롬프트를 사용함
  - **character identification**: 비디오 클립과 해당 대본을 함께 제공하여, 모델이 오디오·대사 매칭을 통해 등장 인물의 이름을 정확히 추론하도록 하는 절차임
- vis-rec, text-rec, MFACTSUM을 사용하여 영상·텍스트 정보 회수율을 분리해 평가함
- ROUGE-1/2/Lsum과 fact-rec도 함께 보고, 텍스트 기반 평가지표에서의 변화도 비교함
- avg-len을 함께 기록하여 요약 길이 차이로 인한 성능 편향을 배제하고자 함

실험 결과

- handcrafted prompt 제거 시 vis-rec이 약 40% 하락하여 LLM이 기본적으로 텍스트 편향적임이 드러남
- handcrafted prompt 제거 시 text-rec도 함께 하락해 multimodal 요약의 품질이 전반적으로 떨어짐
  - 모델은 강한 지시가 없으면 과도하게 압축된 ‘텍스트-only 요약’으로 돌아감
- character identification 제거는 MFACTSUM·ROUGE 모두에 큰 변화가 없음
  - transcripts에 이미 모든 대사마다 화자 이름이 포함되어 있으므로, LLM은 요약 생성 시 등장 인물의 이름과 행동 정보를 자연스럽게 추적·포함할 수 있음
- ROUGE 계열에서는 성능이 감소됨

    - visual 정보를 많이 넣을수록 표현이 추상화되고 단어 overlap이 줄기 때문에 ROUGE가 떨어짐

평가 지표 분류

- **세밀한 캡션 생성 능력 평가:** LLM-VLM based QA

  (Progress-Aware Frame Captioning, CVPR 2025 — Cap/Prog 기반 fine-grained 평가)
- **텍스트 정확도 평가:** ROUGE-L / METEOR / CIDEr

  (Video ReCap, CVPR 2024 — 전통적 캡셔닝 평가 지표)
- **비디오–요약 정합성 평가:** VideoScore / FactVC

  (Video-to-Text Summarization, ACL 2025 — 영상과 텍스트의 사실 일치 여부 평가)
- **멀티모달 품질 종합 평가:** LLM-as-Judge

  (Video-to-Text Summarization, ACL 2025 — 요약의 전반적 품질 및 일관성 평가)
- **텍스트·비디오 균형 종합 평가:** MFACTSUM

  (AACL 2025 — 텍스트와 비디오 정보의 균형 있는 반영 여부 평가)