# State-of-the-Art Review on Video-to-Text Summarization Techniques

UT UNIVERSITÉ DE TOULOUSE

Alejandro Argüelles Trujillo, Yanis Zerkani, Mira Boudiaf

TIR IAFA IMA

## Multimodality

Combining multiple data streams—visual, auditory, and textual—to enrich the semantic understanding of videos.

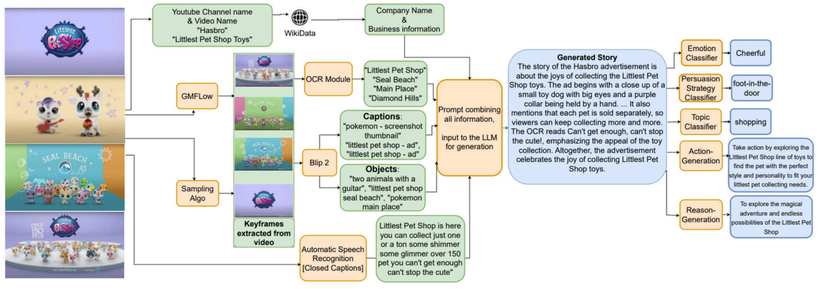



**Fig. 1.** The Process for Converting Video Content into a summary using tokens (Bhat-tacharyya et al., 2023)[1]

## Modular Architectures

Modular pipelines break down video summarization into interpretable stages — scene detection, scene reordering, visual captioning, dialogue summarization... and final summary fusion — enabling flexible upgrades and precise fact-based evaluation.

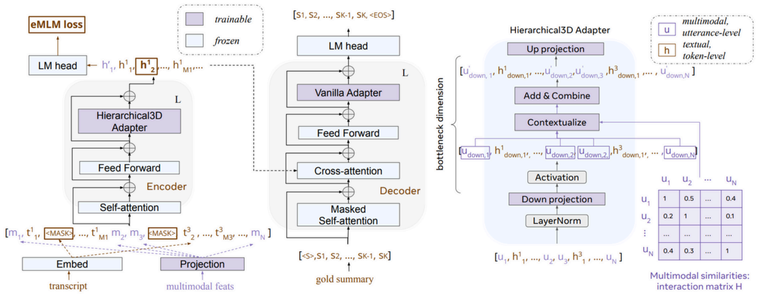



**Fig. 2.** Multimodal Augmentation of BART via Hierarchical3D Adapters (Papalampidie& Lapata, 2023) [2]

## Self-supervision

Train powerful summarizers without human annotations by aligning audio, visual, and text signals through self-supervised learning.

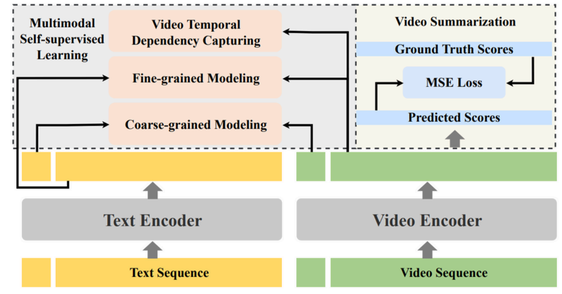
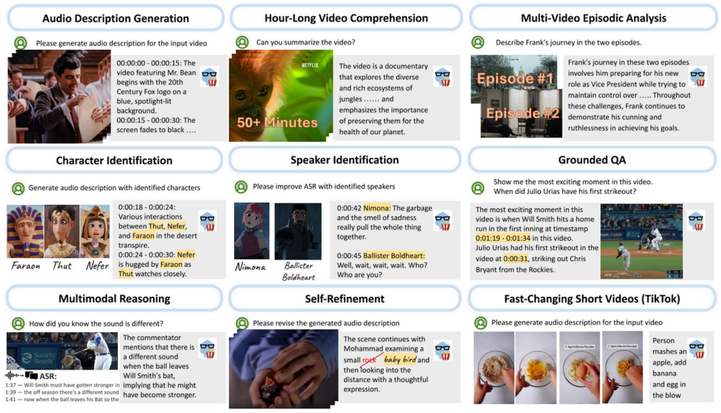



**Fig. 3:** The proposed multimodal self-supervised framework for video summarization[3]

## Specialized Experts




## Challenges

- Content Selection & Coherence
- Domain Adaptability
- Real-Time Processing
- Multimodal Alignment
- Dataset Coverage and Diversity

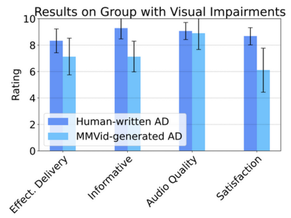
## Future Directions

- Multimodal Fusion
- Long-Form Video Models
- Cross-Domain Adaptation
- Real-Time Efficiency
- Evaluation Metrics
- Annotated Dataset Scarcity

## Benchmarks & User Studies

- Query-Based and User-Centric Evaluation Protocols (F1 - Score based) [7]
- PRISMA: Harmonic mean of fact-precision FP and fact-recall FR [4]
- ROUGE: A set of metrics that evaluates summary quality based on lexical overlap with human references. [6][4]
- LfVS-T [5]
- A user Study with Blind Participants: MM-VID [8]

$$\text{PRISMA} = \frac{2}{\frac{1}{FP} + \frac{1}{FR}}$$




$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

## Datasets

| Dataset | # of Videos | # of Tasks | Avg. Dur. (min) | Annotation |
|---|---|---|---|---|
| TVSum [36] | 50 | 10 | 4.2 | Manual |
| SumMe [7] | 25 | 25 | 2.4 | Manual |
| TL:DW? [26] | 12.1K | 185 | 3.1 | Automatic |
| LfVS-P (Ours) | 250K | 6.7K | 13.3 | Automatic |
| LfVS-T (Ours) | 1.2K | 392 | 12.2 | Manual |

**Fig. 4:** Comparison of different video summarization datasets.[5]

## Bibliography

[1] Bhattacharyya, A., Ju, D., Hariharan, B., Parikh, D., Schwing, A.: A video is worth 4096 tokens (2023)
[2] Papalampidi, P., Lapata, M.: Hierarchical3D adapters for long video-to-text summarization (2023)
[3] Haopeng, L., Qiuhong, K., Mingming, G., Drummond, T.: Progressive video summarization via multimodal self-supervised learning (2022)
[4] Mahon, L., Lapata, M.: A modular approach for multimodal summarization of TV shows (2024)
[5] Argaw, D.M., Yoon, S., Heilbron, F.C., Deilamsalehy, H., Bui, T., Wang, Z., Dernoncourt, F., Chung, J.S.: Scaling up video summarization pretraining with large language models.(2024)
[6] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. En Text Summarization Branches Out (pp. 74–81).
[7] Zhang, Y., Kampffmeyer, M., Liang, X., Tan, M., Xing, E.P.: Query-conditioned three-player adversarial network for video summarization.
[8]Lin, K., Ahmed, F., Li, L., Lin, C.C., Azarnasab, E., Yang, Z., Wang, J., Liang, L., Liu, Z., Lu, Y., Liu, C., Wang, L.:
Mm-vid: Advancing video understanding with gpt-4v(ision) (2023)