# Vision LLM-based Cross-modal Summarization Framework for Long-range Videos

**Dingrui Tao, Shicheng Peng, Ruiyang Yu**

{taodingr, shchpeng, ruiyangy}@umich.edu

EECS 545 Project Report

## Abstract

Up till now, state-of-the-art LLM-based techniques have made remarkable progress in the area of video summarization. However, there are a number of challenges to be overcome. Specifically, the two most common formats of video summarization are video-based summary and text-based summary. Existing Methods often treat these two formats as separate tasks, thus ignoring the semantic correlation between these two modalities and producing syntactically correct but semantically disjoint summaries. This deficiency becomes severe especially when the video is long-range, which tends to consider more cross-modal information. To address this issue, based on a SOTA vision LLM model called BLIP, we introduce a framework integrated with a local attention module to capture cross-modal information and yield a more readable and understandable video or text summary.

# 1 Introduction

Video summarization is a kind of technique that aims to summarize a long video, producing a concise, accurate, and comprehensive paragraph that describes the main idea of the video [1]. With the rapid growth of online video platforms as well as the rocketing length and volume of videos, there exists an increasing demand for automatic mature video understanding tools [2]. Video summarization has been an important part of machine learning applications. Upon recent remarkable breakthroughs in state-of-the-art LLM models, video summarization has been pushed to another level. Driven by large language models, recent progress in Video-based LLMs has remarkably helped finish video summarization tasks [3].

Video summarization contains a variety of forms. Typical approaches are text-based summary and video-based summary: the first one learns representation from video captions, generating a short paragraph of text; the second format selects a small fraction of frames from the original video, connecting to form a short video summary. However, existing methods often treat either visual or textual summaries as separate tasks, ignoring the semantic correlation between these two modalities. [1]. But caption and video content are often correlated and the summary based on either one solely will lose this correlation and produce an unreadable summary. This issue becomes even worse when the source is long-range, where the probability of the existence of cross-modal semantic relation is higher.

To solve this significant deficiency in the area of video summarization. Based on a popular state-of-the-art vision LLM model BLIP, we proposed a general framework for generating video and text summaries simultaneously. In this framework, we make use of the BLIP encoder, adding a local attention module and redesigning two decoders based on the fine-tuned version of the existing video-summary and text-summary models. The novel attention module enables the framework to learn cross-modal representation and correlate the two decoders to generate semantically meaningful content. Our framework achieves compelling performance on the benchmark, especially beating other popular video summarization models in terms of criteria that evaluate the human readability of the output.

# 2 Related Work

## 2.1 Video Summarization

### Text-based Technique

Text-based Video Summarization aims to summarize a video and generate a short paragraph of text. Some methods rely solely on text input, either directly extract subtitles/captions or convert speech to texts (e.g. TextRank [12], TF-IDF [13], BERT [14], Cap2Sum [15]). These sorts of techniques work well for some well-organized and formal videos such as movies, documentaries, or official instruction videos, but unable to process more diverse and unorganized videos. Vision LLM-based models are also popular in generating text summaries (e.g. Seq2Seq with Attention [16], BART [17], GPT-4 [18]). These methods overcome the dependence on exact text sources in videos, able to generate new sentences outside of videos in the summary, but ignoring the dynamics and visual-spatial features in the videos.

### Video-based Technique

Video-based Summarization aims to summarize a video by extracting key frames from it and then generating a concatenated short video summary. Classical methods use unsupervised methods like clustering to do summarization (e.g. VSUMM [19]). More advanced methods appear with the development of deep learning techniques. (e.g. SUM-GAN [20], DR-DSN [21], PGL-SUM [22]). There is also a special type of technique that focuses on the dynamics of the video content by using motion-based methods (e.g. MeanSum [23], 3D-CNN [24]). Moreover, the SOTA method achieves better performance using Transformer-related methods (e.g. VT-SUM [1], UniVTG [25]). These video-based methods are able to capture the highlight section of the videos, but sometimes yield semantically meaningless summaries due to lack of textual information.

### Multi-modal Technique

Multi-modal video summarization only appeared within a recent decade, as previous methods could not handle multi-modalities very well at the same time. Some early methods using Reinforcement Learning (e.g. HRL [26], Multi-modal RL [27]). Other early techniques join different features together (e.g. VideoBERT [28], UniVL [29]) or use Attention Mechanism (see Section 2.2). In recent years, the multi-modal LLM-based method become more dominant in terms of cross-modal video summarization, which applies to more general topics summarization with promising performance (e.g. Video-LLaMA [30], Flamingo [31]). These methods show promise but still face challenges in grounding fine-grained video semantics.

## 2.2 Local Attention for cross-modal information

The key part we introduce in this novel framework is a local attention module. Much existing literature has shown the effectiveness and power of the local attention module in terms of processing cross-modal information. Local Attention helps the model to learn both local and global representations from videos [8]. Moreover, A Cross-Modal Attention module can enable joint reasoning between texts and frames of a video [9], thus it can help the decoder to learn semantically correlation between two modalities. Beyond that, Local contextual attention can also reduce redundancy by extracting representative information in short video segments [10], this will prevent the decoder from selecting repeat content in a long video with many similar segments such as instructional videos. As a local module, Cross-Modal Attention allows the model to adaptively focus on important video segments and keywords within a sliding attention window [11].

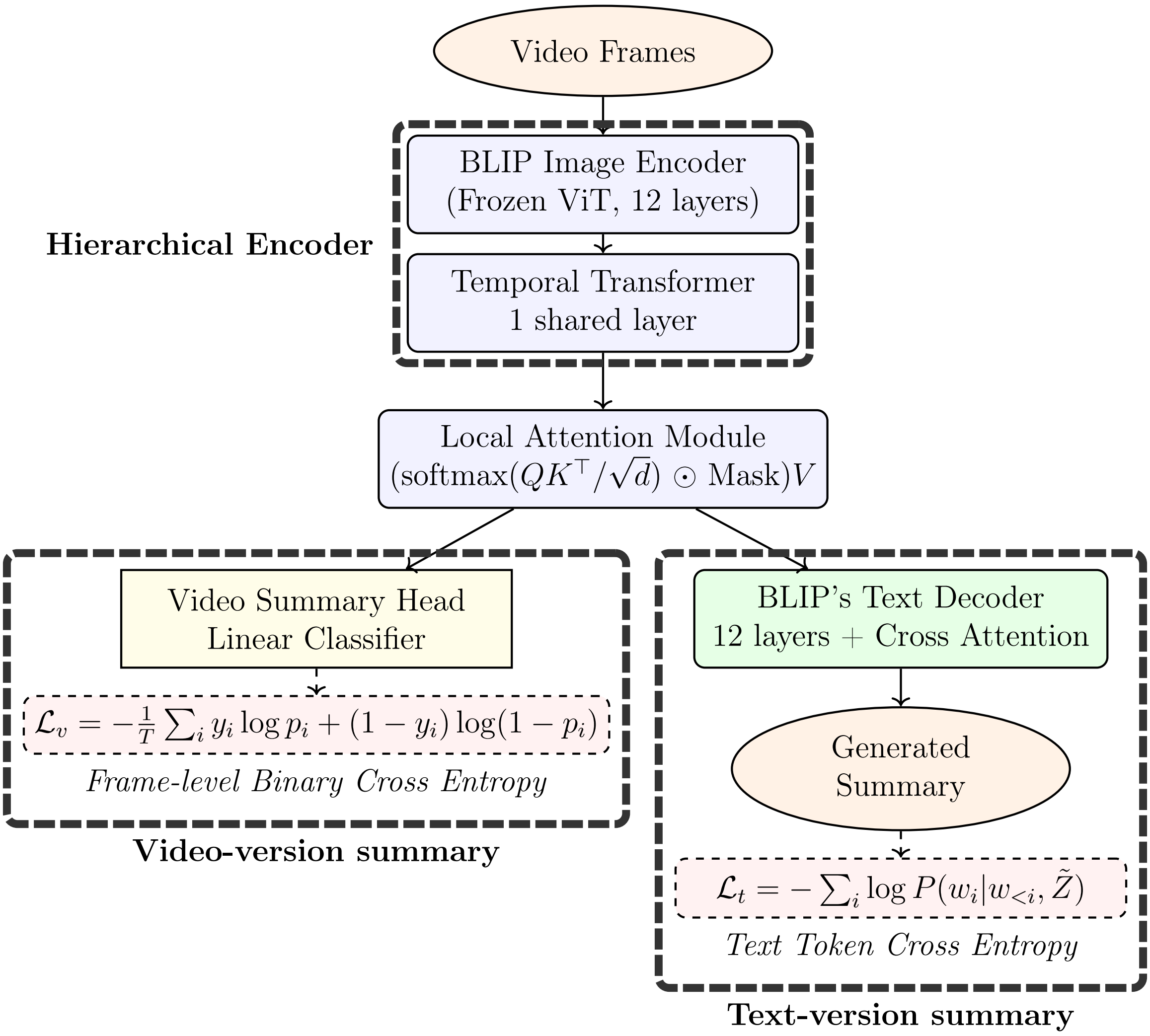
# 3 Method

## 3.1 Dataset

We use the VideoXum dataset proposed by Lin et al. [1], which comprises over 14,000 long-range videos each paired with ten independent video-summary and text-summary annotations. Built on ActivityNet Captions [4]—a high-quality benchmark of 20,000 real-world YouTube videos with dense captions—VideoXum inherits its diverse content and well-annotated structure. To address the subjective nature of summarization, each video is re-annotated by ten human workers with distinct frame-level importance scoring and text summaries, yielding robust multi-reference ground truth for training and evaluation. Video durations range from 10 to 755 seconds (99.9 % under 300 seconds), with frames sampled at 1 fps. In our experiments, we split the data into 8,000 training videos, 2,000 validation videos, and 4,000 test videos.

## 3.2 Overall Framework

To address the cross-modal video summarization task, we learn a shared video encoder followed by two task-specific decoders: a video-sum decoder and a text-sum decoder. We adopt BLIP, a large vision-language pretrained (VLP) model, as our backbone. It is designed to learn joint representations, and is strong at video understanding and language modeling. Inspired by efficient video encoding methods [7] and building on the structure in [1], we integrate BLIP's image encoder into our hierarchical video encoder, enabling efficient extraction of rich features at both the frame and patch levels. Below, we describe our pipeline in detail.



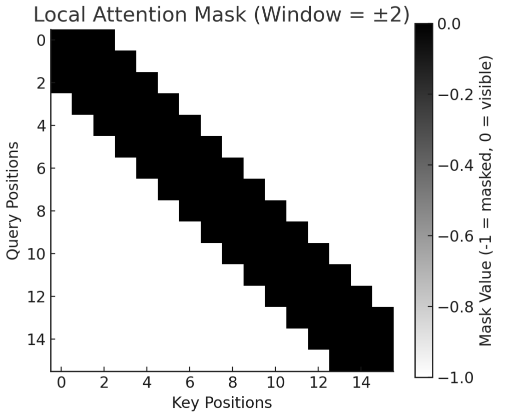


Figure 1: Local Attention Mask with k = 2

The first stage of the **Hierarchical Encoder** leverages BLIP's Image Encoder. Each video frame is split into patches, prepended with a `[CLS]` token, and passed through a 12-layer Vision Transformer (ViT-B/16) [5] whose weights remain frozen during training. The output embeddings for all $T$ frames form the frame-level representations, capturing spatial information.

Next, we process these frame embeddings with the Shared Temporal Transformer (TT). We add learnable *Temporal Position Embeddings* to the spatial features, then process them with a lightweight (one-layer) Temporal Transformer to model temporal dependencies and capture frame-to-frame context, producing spatiotemporal features. Only the TT parameters and *position embeddings* are fine-tuned—the ViT itself stays frozen. They receive gradients from both the video-sum loss $L_v$ and the text-sum loss $L_t$ downstream, and are updated via AdamW under our multi-task training regime.

We introduce the **Local Attention Module** to capture context from neighboring frames, thereby improving the model's ability to understand motion transitions and temporal dynamics. For each temporal position, we define a fixed-size slice window of radius $k$ and construct a binary local attention mask $M^{LA}$ that restricts each frame's attention to its nearby neighbors. In our experiments, we set $k = 2$ (Fig. 1), so each representation attends to five frames—including itself. Next, we compute the locally enhanced attention features via

$$\mathcal{A}_{\text{loc}} = \big(\text{softmax}(QK^\top/\sqrt{d}) \odot M^{LA}\big)\, V,$$

where $Q$, $K$, and $V$ are the queries, keys, and values projected from the spatiotemporal features. The intuition for adding such a module is that the local self-attention can sharpen each frame's awareness of its immediate temporal neighbors, boosting the precision of key-frame selection without the overhead of full global attention. By enhancing features with local context, the model also captures richer motion information, enabling the downstream text decoder to generate more meaningful and detailed summaries.

The parallel **Cross-Modal Decoders** comprise a video-sum decoder and a text-sum decoder, both built on the shared spatiotemporal features to promote semantic alignment between video and text summaries. The **Video-Sum Decoder** predicts each frame's importance score $p_i$ via a linear classifier applied to the locally enhanced attention features

$\mathcal{A}_{\text{loc}}$, and is trained under the binary cross-entropy loss

$$\mathcal{L}_v = -\frac{1}{T}\sum_{i=1}^{T}\big(\hat{y}_i \log p_i + (1-\hat{y}_i)\log(1-p_i)\big),$$

where $\hat{y}_i \in \{0,1\}$ indicates whether the $i$-th frame is a key frame. During inference, we select the top 15% of frames by score to form the video summary.

The **Text-Sum Decoder** uses BLIP's causal, cross-attention-enabled Transformer decoder [5], consisting of 12 layers. At each decoding step, it cross-attends to the spatiotemporal video features, which enables tight vision–language fusion. The decoder weights are initialized from the BLIPCapFilt-L checkpoint [1]. Its training objective is the cross-entropy loss over the ground-truth tokens $w_i$:

$$\mathcal{L}_t = -\sum_{i=1}^{N_{\text{tex}}} \log P\big(w_i \mid w_{<i}, \widetilde{\mathbf{Z}}\big).$$

The overall loss is a weighted sum:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t,$$

We chose $\lambda_v = 15.0$ and $\lambda_t = 1.0$ in our experiments so that the frame-wise training was stabilized and the gradient scales were balanced across modalities.

# 4 Experiment

## 4.1 Training

We implement our framework using PyTorch and train all models on a single NVIDIA RTX 4050 GPU. We optimize the model using the AdamW optimizer with a learning rate of 2e-5, weight decay of 0.01, and gradient clipping at 1.0. The batch size is set to 16, and each training run uses a fixed random seed for reproducibility. The maximum sequence length for text decoding is 32 tokens. No data augmentation or frame interpolation is applied, and video frames are sampled uniformly at 1 fps with spatial resolution fixed at 224 × 224. The total fine-tuning process took 21 hours over 28 epochs, reflecting the complexity of the temporal dynamics and cross-modal objectives.

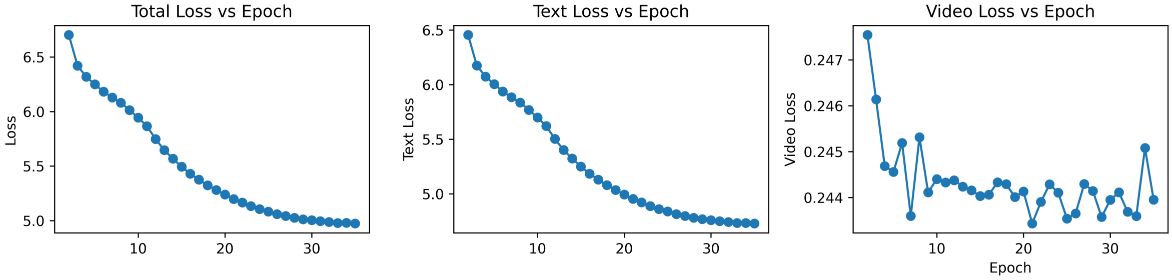


Figure 2: Loss Plot

## 4.2 Inference Protocol

During inference, text summaries are generated using greedy decoding to reduce computational cost and the visual summaries are constructed by selecting the top 15% of frames with the highest predicted saliency scores.

We evaluate model outputs on the held-out test set using standard metrics: F1[1], Kendall's $\tau$[32], and Spearman's $\rho$[33] for video summarization; and BLEU@4[34], METEOR[35], ROUGE-L[36], CIDEr[37], and VT-CLIPScore[1] for text summarization.

# 5 Result and Evaluation

We evaluate our proposed model on the VideoXum benchmark and compare its performance against previous baselines in both video-level and text-level summarization tasks. The evaluation covers three core subtasks: frame-level saliency prediction, textual summarization, and cross-modal alignment.

## 5.1 Metric Comparison

Table 1: Performance Comparison Across Models with Highlighted Improvements

| Metric | Base | TT+CA | (Our) TT+CA+LA |
|---|---|---|---|
| F1 | 21.7 | 23.5 | **23.6** |
| Kendall's $\tau$ | 0.131 | 0.196 | **0.201** |
| Spearman's $\rho$ | 0.207 | 0.258 | **0.269** |
| BLEU@4 | 5.5 | 5.8 | **5.9** |
| METEOR | 11.7 | 12.2 | **12.4** |
| ROUGE-L | 24.9 | 25.1 | **25.1** |
| CIDEr | 18.6 | 23.1 | **25.4** |
| VT-CLIPScore | 28.4 | 29.4 | **29.6** |

*Table 1* presents results for the video-version summary task. Although our model introduces only a marginal improvement in F1 score (from 23.5 to 23.6), it significantly improves both Kendall (from 0.196 to 0.201) and Spearman (from 0.258 to 0.269), which are more sensitive to rank-level alignment between predicted and human-annotated saliency scores. This indicates that while the overall key frames selection remains comparable, the predicted importance distribution better reflects human judgment.

This improvement can be attributed to the Local Attention Module, which enables the model to capture fine-grained temporal transitions, allowing it to differentiate more subtly between informative and redundant segments. This is further supported by the saliency score visualization in *Figure 4*, where the predicted score curve from our model better overlaps with ground-truth annotations than the baseline or TT+CA variants. The comparable F1 score between our model and the previous TT+CA baseline is expected, as the Local Attention Module primarily sharpens the saliency distribution—making the importance scores on key frames more peaked—without significantly altering the top 15% highest-ranked frames selected for summary.

For the text-version summary task, our model achieves consistent gains across multiple evaluation metrics, as reported in *Table 1*. Specifically, BLEU@4 increases from 5.8 to 5.9 and METEOR from 12.2 to 12.4, indicating improvements in both lexical precision and semantic adequacy. ROUGE-L remains constant at 25.1, suggesting that overall sequence-level coverage is maintained. As CIDEr improves from 23.1 to 25.4, our model's ability to capture distinctive and informative content is significantly improved. The alignment between generated summaries and the corresponding visual inputs is much stronger, indicating by the rise in VT-CLIPScore (from 29.4 to 29.6). These results demonstrate that our model produces more fluent, semantically coherent, and visually grounded textual summaries, attributed to the improved local temporal modeling enabled by the proposed Local Attention Module.

These results suggest that our model is not only more accurate in content selection but also more effective at producing fluent and semantically coherent summaries. We attribute this to the local attention enhancement of encoder outputs, which supplies the BERT decoder with representations enriched with short-range contextual cues.

## 5.2 Sample Illustration

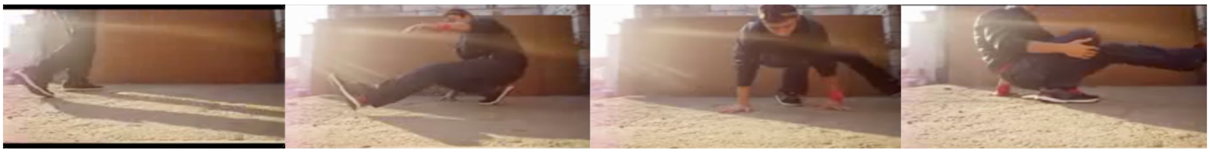


Figure 3: Sample Video

### Video-version Summary

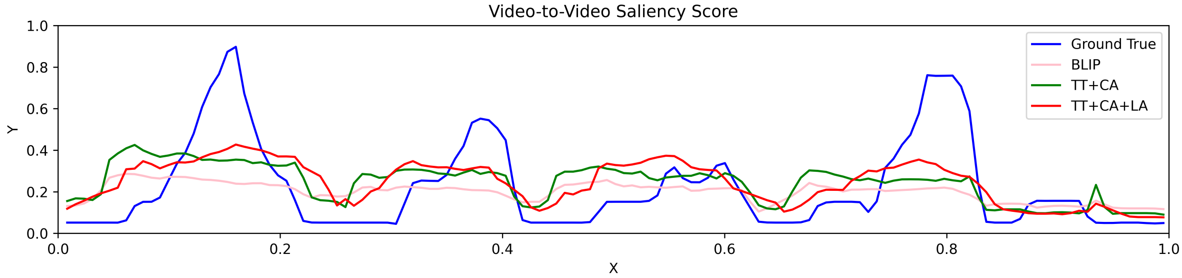


Figure 4: Saliency Score Comparison

### Text-version Summary

**Ground Truth**: several shots of text are shown followed by a **person walking into frame** the person then walks in and out of frame performing various **dance moves on the ground** the person **continues moving around** on the ground and demonstrating **how to perform moves**

**Base: a man is seen standing in front of a camera** and leads into him bending down on the ground the man is then seen bending down on a piece of wood and leads into him bending down on the ground

**TT+CA: a close up of a shoe is shown** followed by a **person stepping up and down** on the shoe the person **continues stepping up and down** on the shoe while **the camera captures his movements**

**(Ours) TT+CA+Local Attention: a close up of a shoe** is shown followed a **person moving around the ground** the person is then **bending down dancing** across the ground on the shoe as **the camera tracks his motions**

*Figure 3*, *Figure 4* and the *block5.2* above provide illustrative examples. The video shows a man playing street dancing on the ground. While the baseline model generates generic statements (e.g., "stepping up and down on the shoe"), our model produces more vivid and temporally grounded descriptions such as "bending down dancing across the ground." This demonstrates improved motion understanding and more human-like narrative structure, which is particularly valuable in instructional or action-dense videos.

## 6 Future Directions

Due to time constraints and limited storage resources, we have not yet been able to evaluate our model on large-scale instructional datasets such as YouCook II (150 GB) and HowTo100M (465 GB). These datasets offer more professional and fine-grained textual annotations compared to VideoXum, and we expect that our model, with minimal additional fine-tuning, could achieve even stronger performance on them. In future work, we also plan to improve the Local Attention Module by replacing the current hard sliding-window mask with a Radial Basis Function (RBF)-based attention weighting scheme. This would enable smoother, distance-aware attention decay, allowing the model to better balance local precision and global context.

## 7 Conclusion

In this work, we present a cross-modal video summarization framework that enhances the pretrained BLIP vision-language model with a lightweight yet effective Local Attention Module. Unlike prior approaches that treat visual and textual summarization independently, our model captures both global sequence context and local temporal dependencies, enabling more coherent and semantically aligned summaries. Experimental results on the VideoXum dataset demonstrate that our method improves cross-modal saliency alignment and yields more expressive textual descriptions, particularly for fine-grained instructional or action-based content. While the frame-level selection performance remains comparable to previous models, our approach achieves higher correlation with human-labeled saliency and consistently outperforms baselines on standard text generation metrics, including BLEU, METEOR, and CIDEr.

# Bibliography

[1] J. Lin et al., “VideoXum: Cross-modal Visual and Textural Summarization of Videos,” Apr. 23, 2024, arXiv: arXiv:2303.12060. doi: 10.48550/arXiv.2303.12060.

[2] Y. Tang et al., “Video Understanding with Large Language Models: A Survey,” Jul. 24, 2024, arXiv: arXiv:2312.17432. doi: 10.48550/arXiv.2312.17432.

[3] Y. Weng, M. Han, H. He, X. Chang, and B. Zhuang, “LongVLM: Efficient Long Video Understanding via Large Language Models,” in Computer Vision – ECCV 2024, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 453–470. doi: 10.1007/978-3-031-73414-4_26.

[4] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, “Dense-Captioning Events in Videos,” May 02, 2017, arXiv: arXiv:1705.00754. doi: 10.48550/arXiv.1705.00754.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” Feb. 15, 2022, arXiv:2201.12086v2 [cs.CV]. doi:10.48550/arXiv.2201.12086.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 11, 2018, arXiv:1810.04805 [cs.CL]. doi:10.48550/arXiv.1810.04805.

[7] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, “HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training,” in Proc. Conf. Empir. Methods Nat. Lang. Process., Nov. 16–20, 2020, pp. 2046–2065.

[8] L. Lan, L. Jiang, T. Yu, X. Liu, and Z. He, “FullTransNet: Full Transformer with Local-Global Attention for Video Summarization,” Jan. 01, 2025, arXiv: arXiv:2501.00882. doi: 10.48550/arXiv.2501.00882.

[9] S. K. Gorti et al., “X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 4996–5005. doi: 10.1109/CVPR52688.2022.00495.

[10] Y. Pan et al., “Exploring Global Diversity and Local Context for Video Summarization,” IEEE Access, vol. 10, pp. 43611–43622, 2022, doi: 10.1109/ACCESS.2022.3163414.

[11] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, “Referring Segmentation in Images and Videos With Cross-Modal Self-Attention Network,” IEEE Transactions

on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3719–3732, Jul. 2022, doi: 10.1109/TPAMI.2021.3054384.

[12] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, “Graph-Based Text Summarization Using Modified TextRank,” in Soft Computing in Data Analytics, J. Nayak, A. Abraham, B. M. Krishna, G. T. Chandra Sekhar, and A. K. Das, Eds., Singapore: Springer, 2019, pp. 137–146. doi: 10.1007/978-981-13-0514-6_14.

[13] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries”.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[15] C. Zhao, C. Wang, Z. Song, G. Hu, H. Chen, and X. Zhai, “Cap2Sum: Learning to Summarize Videos by Generating Captions,” Aug. 23, 2024, arXiv: arXiv:2408.12800. doi: 10.48550/arXiv.2408.12800.

[16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” May 19, 2016, arXiv: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473.

[17] M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” Oct. 29, 2019, arXiv: arXiv:1910.13461. doi: 10.48550/arXiv.1910.13461.

[18] OpenAI et al., “GPT-4 Technical Report,” Mar. 04, 2024, arXiv: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[19] S. E. F. de Avila, A. da_Luz Jr., A. de A. Araújo, and M. Cord, “VSUMM: An Approach for Automatic Video Summarization and Quantitative Evaluation,” in 2008 XXI Brazilian Symposium on Computer Graphics and Image Processing, Oct. 2008, pp. 103–110. doi: 10.1109/SIBGRAPI.2008.31.

[20] B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised Video Summarization with Adversarial LSTM Networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2982–2991. doi: 10.1109/cvpr.2017.318.

[21] K. Zhou, Y. Qiao, and T. Xiang, “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward,” Feb. 13, 2018, arXiv: arXiv:1801.00054. doi: 10.48550/arXiv.1801.00054.

[22] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, “Combining Global and Local Attention with Positional Encoding for Video Summarization,” in 2021 IEEE International Symposium on Multimedia (ISM), Nov. 2021, pp. 226–234. doi: 10.1109/ISM52913.2021.00045.

[23] E. Chu and P. J. Liu, “MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization,” May 22, 2019, arXiv: arXiv:1810.05739. doi: 10.48550/arXiv.1810.05739.

[24] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, “Video Summarization through Reinforcement Learning with a 3D Spatio-Temporal U-Net,” IEEE Trans. on Image Process., vol. 31, pp. 1573–1586, 2022, doi: 10.1109/TIP.2022.3143699.

[25] K. Q. Lin et al., “UniVTG: Towards Unified Video-Language Temporal Grounding,” Aug. 18, 2023, arXiv: arXiv:2307.16715. doi: 10.48550/arXiv.2307.16715.

[26] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, “Weakly Supervised Video Summarization by Hierarchical Reinforcement Learning,” Feb. 29, 2020, arXiv: arXiv:2001.05864. doi: 10.48550/arXiv.2001.05864.

[27] B. Chen, X. Zhao, and Y. Zhu, “Personalized Video Summarization by Multimodal Video Understanding,” Oct. 2024, pp. 4382–4389. doi: 10.1145/3627673.3680011.

[28] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “VideoBERT: A Joint Model for Video and Language Representation Learning,” Sep. 11, 2019, arXiv: arXiv:1904.01766. doi: 10.48550/arXiv.1904.01766.

[29] H. Luo et al., “UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation,” Sep. 15, 2020, arXiv: arXiv:2002.06353. doi: 10.48550/arXiv.2002.06353.

[30] H. Zhang, X. Li, and L. Bing, “Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding,” Oct. 25, 2023, arXiv: arXiv:2306.02858. doi: 10.48550/arXiv.2306.02858.

[31] J.-B. Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” Nov. 15, 2022, arXiv: arXiv:2204.14198. doi: 10.48550/arXiv.2204.14198.

[32] M. G. Kendall, “The treatment of ties in ranking problems,” Biometrika, pp. 239–251, 1945.

[33] D. Zwillinger and S. Kokoska, CRC standard probability and statistics tables and formulae. Crc Press, 1999.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 2002, pp. 311–318.

[35] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proc. Assoc. Comput. Linguist. Workshop, 2005, pp. 65–72.

[36] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81.

[37] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus based image description evaluation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4566–4575.