Multi modal fusion for video retrieval based on CLIP guide feature alignment

Wu, G., Haider, A., Spence, I., & Wang, H. (2024). Multi modal fusion for video retrieval based on CLIP guide feature alignment. In *MVRMLM '24: Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval* (pp. 45-50). (Proceedings of ACM ICMR Workshop on Multimodal Video Retrieval). Association for Computing Machinery. https://doi.org/10.1145/3664524.3675369

Published in:
MVRMLM '24: Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval

Document Version:
Publisher's PDF, also known as Version of record

Multi Modal Fusion for Video Retrieval based on CLIP Guide Feature Alignment

Guanfeng Wu*
wgf1024@swjtu.edu.cn
Southwest Jiaotong University
Chengdu, China

Abbas Haider
a.haider@qub.ac.uk
Queen's University Belfast
Belfast, Northern Ireland, UK

Ivor Spence
i.spence@qub.ac.uk
Queen's University Belfast
Belfast, Northern Ireland, UK

Hui Wang†
h.wang@qub.ac.uk
Queen's University Belfast
Belfast, Northern Ireland, UK

ABSTRACT

With the rise of short video platforms, a large amount of video data is generated daily. These videos vary in quality and are not well-tagged. How to fully utilize the multimodal information in videos, bridge the differences between modalities, and achieve precise video retrieval is a major challenge currently faced in the field of video retrieval. This paper presents a novel approach to multimodal video retrieval, aiming to boost search precision by incorporating visual, textual, and audio information through the CLIP model and T5. Tackling the issue of retrieving pertinent content from extensive, untagged video repositories, we propose a method that fuses multimodal data through innovative feature extraction and alignment techniques. Our method showcases performance are close to the current state-of-the-art, showcasing its effectiveness in improving search accuracy on MSR-VTT benchmark.

CCS CONCEPTS

• **Computing methodologies → Computer vision tasks**; *Computer vision problems*; Search methodologies.

KEYWORDS

Multi modal fusion, CLIP, Video retrieval, Modal alignment, Video summarizes

ACM Reference Format:
Guanfeng Wu, Abbas Haider, Ivor Spence, and Hui Wang. 2024. Multi Modal Fusion for Video Retrieval based on CLIP Guide Feature Alignment. In *Multimodal Video Retrieval and Multimodal Language Modelling (MVRMLM '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3664524.3675369

1 INTRODUCTION

Video retrieval is a process of finding videos of interest from a vast amount of videos stored in a database. With the explosive growth of digital video content, video retrieval has become a key technology in multiple fields such as media [16], education [22], security monitoring [5], and entertainment. A massive number of videos are produced and uploaded online through multiple platforms, including YouTube and Instagram. These platforms require a significant number of tags during upload. However, there are plethora of online platforms or online video libraries where videos are uploaded daily lacking comprehensive metadata or tags. In this scenario, a text-based search leads to inaccuracies and ineffective video retrieval. Another issue with untagged videos is homogenization, meaning same video clip exists with different names.

Multimodal video retrieval [8] is a technique that uses multiple information modalities within a video for retrieval. This method aims to enhance the accuracy and efficiency of video retrieval by integrating various aspects of video content to the video representation. Multimodal video retrieval typically includes visual (images, scenes, object recognition, and visual features such as color, shape, and motion within the video), audio (speech recognition, music detection, environmental sound recognition, and emotional analysis), textual content (textual information that may be contained in the video, such as subtitles, comments, tags, and descriptive metadata), and social and contextual information (additional data about the video, such as the uploader, view counts, likes, and user feedback).

CLIP (Contrastive Language–Image Pre-training) [19] primarily contributes to multimodal data understanding by training a model that understands the relationships between image content and corresponding textual descriptions through contrastive learning on a massive scale of image-text pairs. This technique significantly enhances the model's generalization capabilities across various datasets and tasks, particularly excelling in zero-shot learning scenarios. Currently, CLIP is widely used in various pioneering efforts, including image classification, object detection, image generation, and image search, especially showing strong capabilities and potential in applications lacking annotated data.

Currently, many projects utilize CLIP for extracting multimodal features from images and text to facilitate video retrieval. For instance, the CLIP4Clip [13] transfers the knowledge of the CLIP model to video-language retrieval in an end-to-end manner. This approach leverages CLIP's capabilities to understand and relate

*This work is a continuation of preliminary efforts while the first author was affiliated with MVSE team. The corresponding author is Hui Wang.

MVRMLM '24, June 10–14, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0684-4/24/06
https://doi.org/10.1145/3664524.3675369

the content of videos and corresponding textual descriptions, enhancing the effectiveness and accuracy of video retrieval tasks. [27] introduced the Open-VCLIP++ approach which makes minimal modifications to the original CLIP to capture spatiotemporal relationships in videos. [7] introduced a model that utilizes Temporal Difference Blocks (TDB) and Temporal Alignment Blocks (TAB) to enhance the cross-modal correlation between video clips and phrases. [29] employed CLIP-guided visual-textual attention for video question answering tasks. [1] conducted a user study to explore methods for multimodal video retrieval using CLIP and introduced 'IMPA', a system that supports video retrieval based on images and textual descriptions. They have all introduced CLIP into their framework, which has improved the performance of the corresponding tasks to a certain extent, and all emphasize the visual and textual alignment capabilities of CLIP.

Despite the superior performance of CLIP, we hold the belief that incorporating additional modalities, such as audio alongside visual and textual elements, could prove beneficial, offering a holistic representation of the video database. The primary concept involves capturing correlations between frames and their textual descriptions, as well as audio content and its corresponding text through cutting-edge techniques like GPT-2(Generative Pre-trained Transformer 2) [20] , T5(Text-to-Text Transfer Transformer) [21], and Automatic Speech Recognition (ASR) technology.Details about CLIP and T5 will be covered in sections 2.3 and 2.4.

The primary contributions of this paper are:

- Proposing a CLIP-guided method for generating video auxiliary captions and a multimodal video retrieval framework.
- Designing and implementing a multimodal video retrieval framework that aligns visual text and audio text.

2 RELATED WORK

2.1 Video Retrieval

Video retrieval can be categorized into three main categories, namely 1) Content-based video retrieval (CBVR), 2) Query-based video retrieval (QBVR), and 3) Interactive video retrieval. CBVR primarily achieves search functionality by analyzing visual and audio features of videos [25]. This includes finding visually similar videos by using visual content such as color, texture, shape, and motion information [6]; conducting audio searches using sound rhythm, pitch, and spectral properties; and utilizing deep learning technologies like convolutional neural networks (CNNs) to recognize specific individuals or objects in the videos [3].

Query-based video retrieval refers to the process where users provide a video clip or sample image, and the system identifies videos containing similar content. Alternatively, users may describe the video content they are searching for in natural language, and the system interprets these queries to find relevant videos [17]. On the other hand, Interactive video retrieval primarily involves collecting user feedback on search results and dynamically adjust retrieval strategies based on this feedback to improve search accuracy [12].

TVR (Text-to-Video Retrieval) involves two main aspects: 1) Searching through video metadata such as titles, descriptions, and tags, and 2) Converting spoken dialogues into text, and then employing standard TVR methods. These were the earliest text-based video retrieval methods. Recent advancements in TVR have been significantly bolstered by end-to-end pre-training on extensive text-video datasets [15] [28] [2]. Efficient training strategies are crucial for end-to-end models such as ClipBERT [10]and Frozen [2], enhancing their effectiveness and performance. The Multi-modal Transformer (MMT) integrates multimodal information from videos, leveraging pre-trained models known as 'experts' for each modality to independently generate embeddings [8]. The Multi-modal Fusion Transformer (MFT) is designed to train the embeddings of vision, audio, and text together in a cohesive manner, achieving a unified embedding representation. This enables the Multi-modal Fusion Transformer to handle inputs of any combination of modalities and any length, focusing on relevant features across different modalities [24]. It then applies these capabilities for video retrieval by comparing similarities.

2.2 Current Challenges

Despite the use of multimodal information as search criteria in the field of multimodal video retrieval improving the retrieval accuracy and achieving higher Recall values, text-based retrieval methods remain the mainstream approach in the video retrieval field. For ordinary users, preparing different modalities of search samples or clips requires specific expertise, such as video and audio editing.

In the vast video databases, performing cross-modal retrieval from text to video on unlabeled videos faces several challenges and issues:

- **Complexity of Semantic Understanding**: Text and video are two entirely different data modalities with significant differences in the way they convey information and details. Text is usually abstract and direct, whereas video contains visual, audio, and temporal sequence information. Accurately matching the abstract descriptions of text with the specific content of videos requires advanced semantic understanding and reasoning capabilities [18].
- **Diversity and Complexity of Video Content**: Videos may include a variety of scenes, objects, actions, and interactions, with these elements constantly changing within the video. Each frame of a video may contain information that is either relevant or irrelevant to the text query, making it a significant challenge to accurately match text queries with video content [11].
- **Lack of Annotated Data**: Unlabeled videos mean there are no prior annotations describing the video content. This makes it more difficult to train models to understand video content and perform accurate retrieval, as the models lack necessary supervision signals.
- **Cross-Modal Feature Fusion**: Designing models that can effectively integrate text features and video features is key to achieving accurate cross-modal retrieval. This typically involves deep learning and machine learning techniques for feature extraction, feature transformation, and feature fusion.

An overview of GPT-2

GPT-2 [20] is a natural language processing and generation model developed by OpenAI. It is primarily based on the Transformer architecture, designed for understanding and generating text. It is the version of GPT currently open-sourced by OpenAI, trained on a massive 40GB dataset, serving as a model for natural language processing and generation. The main features of GPT-2 include:

- **Pre-training and Fine-tuning:** GPT-2 is first pre-trained on a large corpus of text data in an unsupervised manner, and can then be fine-tuned for specific tasks.
- **Autoregressive Properties:** In generating text, GPT-2 predicts the next word sequentially based on all previously generated words.
- **Diverse Applications:** Although initially designed for text generation, GPT-2 is also used for a variety of other natural language processing tasks such as text summarization, translation, and question answering.
- **Large-scale Model and Data:** GPT-2 has multiple versions, ranging from 100 million to 1.5 billion parameters, and is trained on a vast amount of data, including diverse texts from the Internet.

2.3 An overview of CLIP

CLIP is a multimodal model developed by OpenAI [19], capable of understanding images and their associated textual descriptions. CLIP is trained using a large set of image-text pairs through contrastive learning. Specifically, the model is trained to recognize whether pairs of images and textual descriptions match. During training, the model receives paired image and text inputs, such as an image and a sentence describing that image. The goal of the model is to maximize the similarity between matching images and texts, while minimizing the similarity with mismatched images or texts. Through this approach, CLIP learns to connect visual content with language descriptions, allowing it to understand and categorize new images or descriptions without explicit labels. This training strategy enables CLIP to perform exceptionally well across various visual tasks, especially on tasks it has not been directly trained on.

2.4 Why T5

The T5 model [21], developed by the team at Google Research, is based on the core idea that all natural language processing (NLP) tasks can be treated as a "text-to-text" problem. This approach means that tasks such as translation, summarization, classification, or others are accomplished by inputting a piece of text into the model and outputting another piece of text.

By using a unified processing framework, T5 simplifies multitask learning and adaptation through extensive pre-training and fine-tuning for specific tasks, enhancing the model's versatility and flexibility. This methodology allows T5 to excel in a variety of NLP tasks such as text summarization, translation, and question-answering [23]. Researchers have also employed T5 for code-related tasks, demonstrating its exceptional performance and adaptability. These tasks include (i) fixing errors in code, (ii) injecting code mutants, (iii) generating assertion statements, and (iv) producing code comments [14].

Although BERT [4], T5, and GPT-2 all possess the capability for text summarization, GPT-2 is primarily a generative model and requires specific training for text summarization tasks. There is evidence suggesting that T5 outperforms GPT-2 in text summarization tasks, demonstrating superior generalization abilities [9].

3 METHODOLOGY

3.1 Framework

Figure 1 illustrates our CLIP-based multimodal alignment video retrieval framework, which integrates video's visual, audio, and textual data for comprehensive processing. Initially, the CLIP model and automatic speech recognition (ASR) are employed to extract features from images, audio, and text, which are then converted into tokens. Subsequently, self-attention and cross-attention mechanisms are utilized to enhance the correlation of features within and across modalities. Ultimately, these features are pooled and clustered to form a unified feature representation used for similarity calculations, supporting video retrieval tasks. We aggregate these features directly using embedding concatenation.

We achieved feature alignment in two ways: firstly, by using pretrained CLIP to align text and visual features, and secondly, by processing features over time series to achieve alignment of features in the temporal sequence.

3.2 Preprocessing

Before training, we extract the embeddings for each modality offline during the data preprocessing stage and store them in .h5 files. The data preprocessing stage primarily includes the following aspects.

- **Frame Extraction:** During the frame extraction stage, we utilize video processing tools such as FFmpeg or OpenCV to extract independent image frames from video files at set intervals (e.g., one frame per second). This step converts continuous video streams into static images that can be individually analyzed, laying the groundwork for subsequent image processing and analysis tasks.
- **Feature Embedding Extraction with CLIP:** Using the pretrained CLIP model, each frame image is processed to extract visual feature embeddings. The CLIP model, by understanding the relationship between image content and associated textual descriptions, generates feature vectors that capture the core visual information of the images, which is crucial for subsequent image understanding tasks.
- **Subtitle Acquisition:** Using the algorithm from the publication ZeroCap [26], text descriptions are automatically generated for each frame image. This algorithm can generate descriptive text directly from images without the need for fine-tuning for specific tasks. Such descriptions not only enhance the understanding of individual frames of video content but also provide a basis for creating a comprehensive textual overview of the video content.
- **Text Summarization Overview with T5:** The T5 model is used to summarize the descriptive texts of all frames to generate an overview of the entire video's content. By integrating key information from each frame, the T5 model conducts in-depth analysis and synthesis to output a concise summary that encapsulates the main information, helping users quickly grasp the core content and themes of the video.

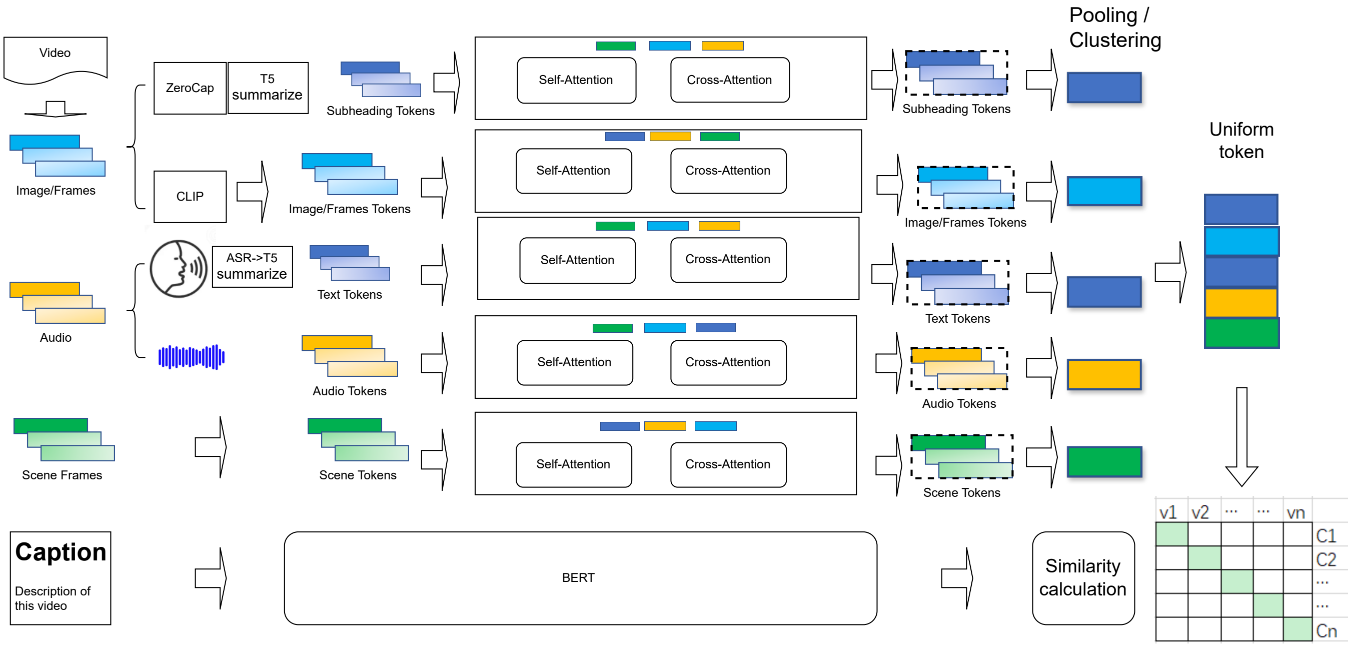

Figure 1: Framework of Multi Modal Fusion Video Retrieval Based on CLIP

3.3 Training objective

Similar to the work with MMT [8], we also adopt a bidirectional margin-maximizing contrastive loss function, which maximizes the similarity between the embeddings processed by the Transformer and the matching text, and minimizes the similarity with non-matching text.

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}\left[\max(0, s_{ij} - s_{ii} + m) + \max(0, s_{ji} - s_{ii} + m)\right] \quad (1)$$

The loss function described in the formula represents a typical structure for learning to distinguish between different classes or categories in a batch of video data, based on descriptions associated with each video. Here's a breakdown of its components and their interpretations:

- **Batch Size** (B): The total number of videos in the batch. Each video has an associated description, and B represents the number of these pairs of videos and descriptions in the computation.
- **Subscript** i **(Video Description Index)**: Refers to the index of a specific video description within the batch. The description at index i is associated with the video at the same index.
- **Subscript** j **(Video Index)**: Refers to the index of any other video in the batch, different from i. The formula includes a summation where j ranges over all videos except the one at index i.
- **Margin** (m): A hyperparameter that defines the minimum difference needed between certain pairwise scores to contribute positively to the loss. The margin helps in driving the separation between the scores of matching and non-matching video-description pairs.

Description of the Loss Function: The loss function computes a sum over all pairs of video descriptions and videos in a batch, except where the indices are the same. For each pair (i, j) where $i \neq j$, it considers two terms:

(1) $\max(0, s_{ij} - s_{ii} + m)$: This term measures how well the description at index i matches the video at index j compared to how well it matches the video at index i itself, adjusted by the margin m. Ideally, the description i should match its corresponding video i better than any other video j, making this term zero or negative (which becomes zero due to the max function).
(2) $\max(0, s_{ji} - s_{ii} + m)$: Similarly, this term assesses how well the video at index i matches the description at index j compared to its own description at index i, again adjusted by m.

The overall loss $\mathcal{L}$ is computed as the average of these maximum values over all video-description pairs in the batch. The goal of this loss function is to ensure that each video is closer to its own description than to any other video's description, effectively helping to learn accurate matching and differentiation across the dataset.

4 EXPERIMENT

4.1 Dateset

The MSR-VTT dataset [28], standing for "A Large Video Description Dataset for Bridging Video and Language," was developed by compiling 257 popular queries from a commercial video search engine, with each query consisting of 118 videos. The latest version of MSR-VTT includes 10,000 web video clips, which cumulatively span 41.2 hours and feature 200,000 clip-sentence pairs. This dataset covers a broad range of categories and showcases a wide variety of visual content, making it the most extensive collection in terms of

Table 1: Evaluation results on MSRVTT: Text to Video

Method	Train Data	R@1	R@5	R@10	MdR	MnR
Everything at once	HT100M	9.6	26.1	36.1	23.0	-
MMT	MSR-VTT	25.1	54.4	68.5	4.5	27.3
MMF-CLIP(ours)	MSR-VTT	31.3	64.0	76.0	3.0	16.8
Cap4Video	MDVM	31.3	74.3	83.8	2.0	12.0

Table 2: Evaluation results on MSRVTT: Video to Text

Method	Train Data	R@1	R@5	R@10	MdR	MnR
Everything at once	HT100M	-	-	-	-	-
MMT	MSR-VTT	25.9	57.3	69.4	4	25.3
MMF-CLIP(ours)	MSR-VTT	32.9	63.9	76.9	3	14.0
Cap4Video	MDVM	47.1	73.7	84.3	2.0	8.7

sentences and vocabulary. Each video segment in the dataset has been meticulously annotated by 1,327 Amazon Mechanical Turk (AMT) workers, who provided around 20 natural sentences per video.

4.2 Evaluation Matrix

In this study, we utilized the MSR-VTT video annotation dataset for model training, aiming to assess the model's performance in video understanding and description. To thoroughly evaluate the model's performance, we employed three metrics: Recall@K, Median Rank (MdR), and Mean Rank (MnR). Recall@K is reported as the model's ability to find the correct video within the top K retrieval results. Median Rank and Mean Rank are used to assess the accuracy of the model in the retrieval results, with lower values indicating better performance, meaning the correct video is ranked higher in the retrieval results. These metrics collectively help us understand the model's effectiveness and accuracy in processing and understanding video content.

4.3 Align with SOTA

We compared the retrieval performance of our method with other algorithms in terms of text-to-video and video-to-text, with the evaluation metrics primarily being Recall, Median Rank, and Mean Rank.The results of the other methods come directly from the corresponding papers. Tables 1 and 2 showcase the text-to-video and video-to-text retrieval results on the MSRVTT dataset. Our method shows significant performance improvements over MMT and Everything at once. We implement our strategy based on the MMT, setting the number of epochs to 50, the same with MMT. The results indicate that using the pretrained CLIP as a feature extraction component can effectively enhance the accuracy of retrieval tasks from text to video and from video to text.

However, there remains a gap between our method and the state-of-the-art model, Cap4Video, which can be attributed to two main reasons. First, our experimental results are preliminary, corresponding only to the addition of CLIP for feature extraction, with experiments involving T5 for summarization still underway. Second, Cap4Video was trained on a larger-scale dataset (**MDAM** is short for MSR-VTT, DiDeMo, VATEX and MSVD datasets), which has positively impacted its generalization capabilities.

5 CONCLUSION

How to extract multimodal semantic information from unlabeled videos, integrate and align multimodal features, and improve video retrieval accuracy are the main research issues in the field of video retrieval today. In this study, we utilize CLIP to align textual semantics and propose the use of T5 to summarize descriptions of extracted video frames, generating video subtitles that enhance video retrieval accuracy. We have designed and implemented a multimodal feature fusion alignment framework. Preliminary experimental results indicate that this framework effectively leverages CLIP's visual-textual alignment capabilities, achieving improved retrieval accuracy on the MSRVTT dataset. Using only pretrained CLIP for feature extraction and text alignment can enhance the capabilities of existing models.

ACKNOWLEDGMENTS

This work is part of the Multimodal Video Search by Examples project, which was funded by the Engineering and Physical Sciences Research Council [EP/V002740/2].

REFERENCES

[1] Tayfun Alpay, Sven Magg, Philipp Broze, and Daniel Speck. 2023. Multimodal video retrieval with CLIP: a user study. *Information Retrieval Journal* 26, 1 (2023), 6.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1728–1738.

[3] Young Rok Choi and Rhee Man Kil. 2020. Face video retrieval based on the deep CNN with RBF loss. *IEEE Transactions on Image Processing* 30 (2020), 1015–1029.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Zuolin Dong, Jiahong Wei, Xiaoyu Chen, and Pengfei Zheng. 2020. Face detection in security monitoring based on artificial intelligence video retrieval technology. *IEEE access* 8 (2020), 63421–63433.

[6] Shiv Ram Dubey. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2021), 2687–2704.

[7] Han Fang, Pengfei Xiong, Luhui Xu, and Wenhan Luo. 2022. Transferring image-clip to video-text retrieval via temporal relations. *IEEE Transactions on Multimedia* (2022).

[8] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16.* Springer, 214–229.

[9] Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. *arXiv preprint arXiv:2005.10433* (2020).

[10] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 7331–7341.

[11] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4953–4963.

[12] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, and George Awad. 2018. On influential trends in interactive video retrieval: video browser showdown 2015–2017. *IEEE Transactions on Multimedia* 20, 12 (2018), 3361–3376.

[13] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[14] Antonio Mastropaolo, Simone Scalabrino, Nathan Cooper, David Nader Palacio, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. 2021. Studying the usage of text-to-text transfer transformer to support code-related tasks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 336–347.

[15] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2630–2640.

[16] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. 2017. Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications* 76 (2017), 22169–22194.

[17] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325* (2020).

[18] Ashish Singh Patel, Ranjana Vyas, OP Vyas, and Muneendra Ojha. 2022. A study on video semantics; overview, challenges, and applications. *Multimedia Tools and Applications* 81, 5 (2022), 6849–6897.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[22] Srihitha Ravi, Shikha Chauhan, Sai Harshini Yadlapallii, K Jagruth, and VM Manikandan. 2021. A Novel Educational Video Retrieval System Based on the Textual Information. In *International Conference on Soft Computing and Pattern Recognition*. Springer, 502–511.

[23] Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Systems with Applications* 208 (2022), 118258.

[24] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 20020–20029.

[25] Newton Spolaôr, Huei Diana Lee, Weber Shoity Resende Takaki, Leandro Augusto Ensina, Claudio Saddy Rodrigues Coy, and Feng Chung Wu. 2020. A systematic review on content-based video retrieval. *Engineering Applications of Artificial Intelligence* 90 (2020), 103557.

[26] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17918–17928.

[27] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. 2024. Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[28] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[29] Shuhong Ye, Weikai Kong, Chenglin Yao, Jianfeng Ren, and Xudong Jiang. 2023. Video question answering using CLIP-guided visual-text attention. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 81–85.