Non-Linear Consumption of Videos Using a Sequence of Personalized Multimodal Fragments

Gaurav Verma*
gverma@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Trikay Nalamada
Keerti Harpavat
Pranav Goel
Aman Mishra
IIT Guwahati
Guwahati

Balaji Vasan Srinivasan
balsrini@adobe.com
Adobe Research, India
Bangalore

ABSTRACT

As videos progressively take a central role in conveying information on the Web, current *linear-consumption* methods that involve spending time proportional to the duration of the video need to be revisited. In this work, we present **NoVoExp**, a method that enables a **No**n-linear **V**ide**o** Consumption **Exp**erience by generating a sequence of multimodal fragments that represents the content in different segments of the videos in a succinct fashion. These fragments aid understanding the content of the video without watching it in entirely and serve as pointers to different segments of the video, enabling a new mechanism to consume videos. We design several baselines by building on top of video captioning and video summarization works to understand the relative advantages and disadvantages of NoVoExp, and compare the performances across video durations (short, medium, long) and categories (entertainment, lectures, tutorials). We observe that the sequences of multimodal fragments generated by NoVoExp have higher relevance to the video and are more diverse yet coherent. Our extensive evaluation using automated metrics and human studies show that our fragments are not only good at representing the contents of the video, but also align well with targeted viewer preferences.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; *Personalization*; • **Computing methodologies** → **Video summarization**; *Video segmentation.*

KEYWORDS

videos, non-linear consumption/interaction, crossmodal translation

ACM Reference Format:
Gaurav Verma, Trikay Nalamada, Keerti Harpavat, Pranav Goel, Aman Mishra, and Balaji Vasan Srinivasan. 2021. Non-Linear Consumption of Videos Using a Sequence of Personalized Multimodal Fragments. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450672

*This work was done when all the authors were with Adobe Research, India.

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450672

1 INTRODUCTION

> "*[A story] should have a beginning, a middle and an end, but not necessarily in that order.*"
>
> *Jean-Luc Godard* (film director)

Videos have become a central content modality in today's digital world. Content spanning topics like software tutorials, lectures, video blogs, etc. account for a significant fraction of media that is being consumed and delivered on platforms like YouTube, Vimeo, Coursera, Udacity, and Twitch. Besides these usual interactions with videos, the post-pandemic world has witnessed a massive increase in sharing different types of recorded videos – lectures, meetings, and conference talks for offline on-demand viewing [3, 4]. While videos are known to deliver an engaging and immersive experience, some of them, especially instructional and explanatory videos, are often long and require the viewer to spend time in proportion to the duration of the video. This is often sub-optimal for viewers interested in only specific parts of the video as it requires them to spend more-than-required time in skimming through the parts that are not of interest. The current video consumption format does not allow viewers to easily find the specific parts that they are interested in. The '*linearity*' in consumption of videos majorly arises because (a) viewing time is in proportion to the duration of the video, and (b) videos are consumed in their original order. Overcoming the shortcomings of this '*linear viewing*' experience is the major motivation of our work.

Youtube, a popular streaming platform, addresses the shortcomings of current video consumption by incorporating an interface that allows publishers to manually provide section headers and corresponding timestamps in the video. These headers are then used to allow viewers to navigate through different sections to enable more focused consumption [6]. While such an interface alleviates the shortcomings to some extent, it fails on the following fronts: *(a)* providing manual annotations for the videos is a tiresome task, especially for large number of videos of considerable duration each; *(b)* the text-based annotations are not the best representation of the multimodal (visual + auditory) content in the video; and, *(c)* these annotations are not personalized to the needs and preferences of viewers. To this end, we propose a method to provide **No**n-linear **V**ide**o** Consumption **Exp**erience (**NoVoExp**) using a sequence of

Figure 1: ***Sequence of preference-aligned multimodal fragments that represents the contents within an input video generated by NoVoExp. The figure also includes the video description for ease of understanding; timestamps and fragment numbers (out of*** 15 ***target fragments) in the final sequence are also shown.***

automatically generated multimodal (image + text) fragments that are personalized and ordered as per the viewers' preferences. This sequence of multimodal fragments *(i)* provides an understanding of the content within the video *without* watching it in its entirety, and *(ii)* also allows for quick navigation to the parts that the viewers are interested by linking to the corresponding segments of the video.

More concretely, our proposed approach takes a video as input and generates a sequence of image-text (multimodal) fragments that represent various sections in the video. Using multimodal fragments, as opposed to only images or text-based annotations, gives a better representation of the video, which itself relies on multiple modalities to deliver an engaging viewing experience. The images and text in our multimodal fragments are chosen from the frames and transcript of the video, respectively, accounting for the preferences of the viewer. As a final step, we use an information optimization framework to order the selected multimodal fragments to arrive at a sequencing that suits the preference of the viewer, while also covering the information in the original video. Figure 1 shows the final output of our methodology for various user preferences.

The contributions of this work are 3-fold: (1) We propose an automated method that leverages state-of-the-art models to enable efficient consumption of videos. Our method generates a sequence of multimodal fragments that allow viewers to quickly understand and navigate through the content of the video. The fragments and its sequencing are further tailored to the preferences of the viewer. (2) We evaluate our approach using several automated metrics that quantify aspects that are important for an ideal consumption experience – such as coherence of the generated sequence, coverage of the content in the video, alignment to the user preferences, diversity of generated fragments. Although there are no existing methods that address this problem, we design several relevant and competitive baselines to compare their performance against our method. Our quantitative and qualitative evaluation using automated metrics show the efficacy of our proposed approach against the baselines. The evaluation also highlights the variation of performance as the duration and the category of the videos change. (3) Lastly, since the motivation is centered around how humans consume videos, we perform extensive human evaluation to quantify the success of our approach. We find that the proposed approach is perceived as 'moderately good' or 'extremely good' by human evaluators on a number of aspects, with a considerable inter-annotator agreement.

2 RELATED WORK

As mentioned before, Youtube [6] recently allowed manual tagging to enable non-linear consumption of video. VideoKEN [17], a niche AI player for videos, performs a topical decomposition on the video and allows the end-user to browse the video according to the recognized topics. However, the work relies on the textual transcript from the video and is tailored to lecture-style videos only. Further, the work is also agnostic to user personalization.

A line of related work is video summarization [5, 23, 30], where the goal is to summarize a video using a subset of coherent segments. Shemer et al. [23] use an Iterative Local Search to optimize to

search the frames of the video to arrive at the summary in an unsupervised way. Zhou et al. [30] propose a Deep Summarization Network (DSN) using Deep Reinforcement Learning to perform summarization using a novel reward framework. Chen et al. [5] use a reinforcement learning framework with a worker-manager model to achieve the summarization in a hierarchical fashion. The underlying objective in video summarization relies on capturing key information from the original video in the summary video and does not account for aspects like navigation, personalization or reordering which are pertinent to the problem we aim to solve.

Videos are inherently multimodal and any processing of video should understand the content in multiple modalities. Prior work has aimed to understand the semantic relatedness between various modalities like images, text, videos, and audio [8, 11, 16] and have learned to represent multimodal data in common vector spaces [20, 21, 28]. Work on cross-modal translation, i.e., representing videos using other modalities is of active interest – Wang et al. [29] explore the use of cross-modal attention for providing captions to videos. Recently, multimodal understanding related works are on the rise, and in the context of videos, Iashin et al. [14] exploit both visual and audio signals from a video to generate dense captions for the keyframes in the video. The goal of this work is to provide frame-level dense captions that are not necessarily comprehensive and aware of the overall theme in the video. We leverage such methods to gain a holistic understanding of the video to generate multimodal fragments from the input video.

As much as a multimodal understanding of videos is important, it is also important to understand the images and text in the output fragments to present them in a coherent way. For example, Kim et al. [15] generate a coherent story from a sequence of image. However, this would be an overkill in our case, since such unconstrained generation can compromise the story in the input video. We have therefore utilized similarities between content in common visual-lingual embedding space [8] to achieve a coherent presentation of the multimodal fragments.

3 NON-LINEAR VIDEO CONSUMPTION EXPERIENCE (NOVOEXP)

The goal of *NoVoExp* is to enable a '*non-linear*' and *personalized* way to consume videos. This involves *(a)* informing users of the content in the video without having to watch it in its entirety, and *(b)* prioritizing the segments that they are interested in watching. Our proposed approach addresses both these by generating a sequence of multimodal (image + text) fragments, where each fragment points/corresponds to a coherent segment of the video (for *(a)*), and by reordering the fragments as per viewers' preferences without compromising the overall narrative (for *(b)*).

We begin by **extracting visual and textual information** from the video. Given a video, we identify *shots* based on the difference in the color histograms of adjacent frames [18] and pick the median frame of the shot as its *keyframe*. Simultaneously, we also extract the audio transcripts from the video. We use these shots and corresponding transcripts to break the video into coherent segments. Given the transcript sentences, we group them based on their semantic similarity in the BERT embedding [22] space. For each group of sentences, we also assign the keyframes corresponding to the shots that span the duration of the grouped sentences to **form multimodal clusters**. Since the visual content does not change much within a shot, the frames and sentences within the multimodal clusters are a good representation of the content in corresponding video segment. We select **representative multimodal fragment**, composed of an image and an accompanying text, from each of the multimodal clusters. For selecting an image, we pick a representative frame using a semantic importance scoring system accounting for the frame's relevance to video and its alignment to the user preferences. For text, we summarize the sentences of the cluster into a representative text fragment. This yields an image + text (multimodal) fragment for every segment of the video. Finally, we design an *information-gain* based scorer to **select and reorder a subset of fragments**. The information score accounts for relevance of the fragments to the video as well as the alignment with viewer preferences and selects a subset of fragments and reorders them to provide a sequence of multimodal fragments. To summarize, NoVoExp extracts information from the video, segments it into coherent units, generates multimodal fragments for every segment and reorders them based on the user preferences. Figure 2 shows a schematic of our proposed approach.

Information Extraction from Videos. Videos encode information in visual and auditory modality, and we extract information from both these modalities. We then use the information extracted to segment the videos into coherent units. We begin with computing the difference between the color histograms of adjacent frames. If the difference between two consecutive frames is greater than a threshold σ, then the $i+1$th frame is marked as the starting of a new *shot*. We take σ to be the scaled mean histogram difference between adjacent frames of the video, given by, $\sigma = \mathrm{S} \times \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N-1}} \sum_{\mathrm{j}}^{\mathrm{B}} |bin_j^{i+1} - bin_j^i|$, where, N is the number of video frames, B is the number of bins in a color histogram. Identifying shot boundaries helps us break the video into small chunks that comprise very similar visual signals. Once we obtain shot boundaries, we select the median frame from each shot to be that shot's *keyframe*. We extract the *audio transcripts* from the video using automatic speech recognition[1] and break the transcript into sentences along with their starting and ending timestamps in the video. We then extract the sentence embeddings for every sentence using a pretrained BERT model [22].

Potential Preferences. While we extract information from the two modalities, the original video also contains other *preference-related attributes* that can aid in personalizing the generated multimodal fragments to viewers' preferences. Prior studies have shown that viewer preferences relate to prominent sentiments in the content, and hence sentiments be consequently used for personalized delivery of content [9, 26]. Accordingly, we identify a subset of sentiments that lie in similar VAD zones as the three personas and map them to three different personas for our experiments: (i) jolly and fun-lovers, (ii) adventurous and thrill-seekers, and (iii) thoughtful and compassionate individuals. These three personas associate with sentiments that are sufficiently spread over the valence-arousal-dominance (VAD) space [24, 27] thus providing good diversity for

[1] https://cloud.google.com/speech-to-text

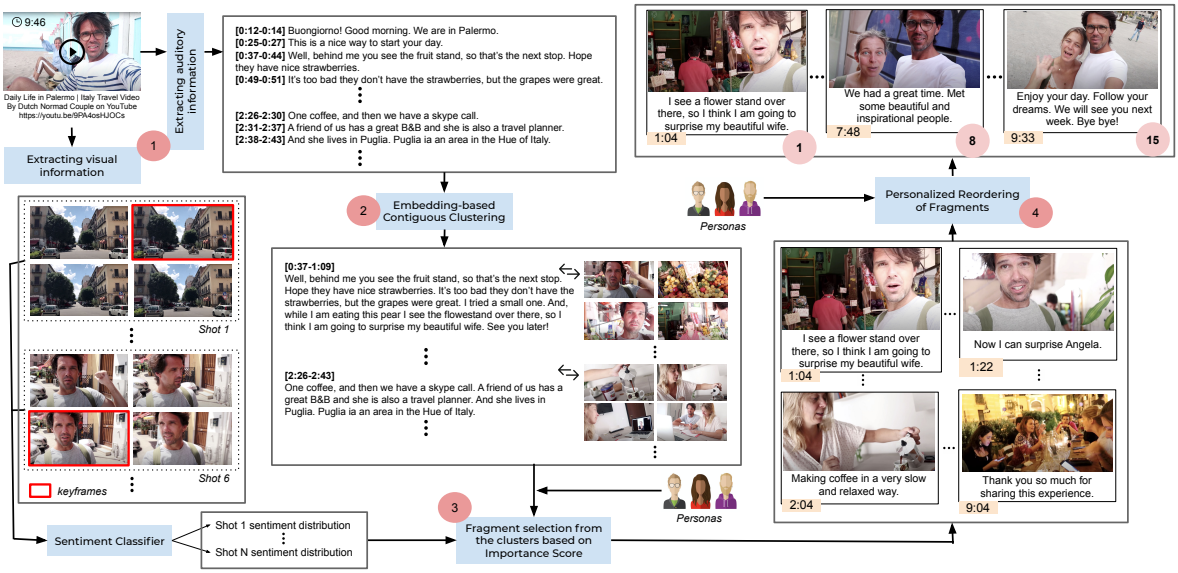

Figure 2: ***Overview of our proposed approach. (1) We extract visual and auditory information from the video and then (2) create contiguous multimodal clusters. (3) Based on the alignment between prominent sentiments of the shots and targeted viewer preference, we select a multimodal fragment from each cluster. This is followed by (4) re-ordering the fragments to obtain a sequence that aligns with the targeted viewer's preference.***

our experiments. We use a corpus of advertisement videos (Pitts Ads Dataset) [13] with around 3,000 videos annotated with 27 overlapping visual sentiment dimensions. Among the 27 sentiment dimensions, we associate high values of 'amused', 'cheerful', 'eager' with jolly and fun-lovers; 'active' and 'amazed' with adventurous and thrill-seekers; 'emotional', 'empathetic', and 'loving' with thoughtful and compassionate individuals. Each of these three personas as represented as multi-hot vectors with 1 assigned to above sentiments and 0 elsewhere. To represent the contents of the video in the same sentiment space as the preferences of the personas, we extract the Resnet-152 embeddings [12] for every frame and train a frame-level sentiment classifier that maps the visual embeddings of a frame to the sentiments present in it by using the Pitts Ads dataset for training. At inference, the inferred vector is averaged over the entire video to obtain the distribution of different sentiments in the video. We use these these video-level, and frame-level, sentiment distributions for identifying segments of the video that are more desirable for the target personas. The same classifier is used to extract the sentiment distributions for every shot and segment in the video, which can similarly be used to compute segment-level preference alignment for a target viewer persona.

Forming Multimodal Clusters. After arranging the sentences in the order as they were uttered in the video, we create clusters of semantically similar sentences. We maintain a running average embedding vector of the current cluster and decide whether the next sentence should be added to this cluster based on its cosine similarity with the running average vector. The sentence is added to the current cluster if the similarity is greater than a threshold, else it is made part of a new cluster. Each cluster thus formed will thus indicate a semantically coherent part of the video. We follow this until all the sentences in the transcript are covered sequentially. We then collate all keyframes within the collective timestamps spanned by the sentences in a cluster and add them the corresponding cluster. These multimodal clusters now comprise contiguous sentences that are semantically similar and keyframes/images from the shots during which they were uttered and represent the information in the corresponding segments in the video.

Selecting Fragments from Clusters. From the segmented video, each segment represented by a multimodal cluster, we choose an image and a short summarized textual description of the transcript of that cluster to arrive at the multimodal fragment that represents the segment. The crucial part, however, is to select the fragment that aligns well with viewers preferences as well. To this end, we select the final fragment by using 3 scores.

- Viewer's preference ($prefScore$) is computed using the average cosine similarity between the viewer's preference vector and the frame-level sentiment score from the classifier.
- Relevance to the overall context of the video ($relScore$) is computed by pretraining a BiLSTM [31] on InceptionV3 representations [25] of frames of the entire video, and computing the relevance of each frame in the cluster to the overall video context using [10].
- Similarity to the sentences in the cluster ($vseScore$) is obtained by averaging cosine similarity of each frame with the sentences in the cluster using a common visual semantic embedding space [8]. This helps in arriving at fragments with semantically aligned image and text.

We use a weighted average of the above scores to arrive at the importance for each fragment, $ImportanceScore = (\alpha \times prefScore) + (\beta \times relScore) + (\gamma \times vseScore)$; such that $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma \in [0, 1]$.

We select the frame with highest Importance Score for the representative multimodal fragment of the segment/cluster. For obtaining the text part of the multimodal fragment, we summarize the sentences in the cluster using a BERT-based Summarizer [19] to arrive at a concise and relevant description of the segment's transcript.

Personalized Reordering of Fragments. We now reorder the multimodal fragments as per target viewer's preference while preserving the overall narrative of the video. We design an information scorer that simultaneously optimizes across various factors (enumerated below) to obtain *a sequencing of the multimodal fragments* that presents viewer-preferred fragments in early parts of the sequence without compromising with the storyline of the video. We initialize an ordered set ($\mathcal{S}$) with the first fragment and add new fragments by accounting for the following factors:

- **Similarity to viewer's preference:** We compute the cosine similarity between the image's sentiment vector for every multimodal fragment and the viewer's preference vector: $prefSim = sim(\mathbf{V}^{\text{sentiment}}_{\text{fragment}}, \mathbf{V}^{\text{sentiment}}_{\text{viewer}})$
- **Follow-up probability of text:** To ensure that the next multimodal fragment *follows* the previous multimodal fragment, we compute the probability of next fragment's textual description (T_{n+1}) following the current one's textual description (T_n) using pre-trained BERT's Next Sentence Prediction model [7]. A high probability value indicates that the transition from the current fragment to the next one will be meaningful and leads to coherent narrative: $textCoherence = P(T_{n+1} \mid Tn)$
- **Relevance of the ordered set to video's context:** To keep a tab of the relevance of fragments in the ordered set to the original video's context, we compute the relevance between the fragment images in the set and the entire video's context by obtaining frame-level representations and the video-level representation using the pre-trained BiLSTM discussed earlier: $relevance(\mathcal{S}) = \sum_{i \in \mathcal{S}} sim(\mathbf{V}_{\text{frame}_i}, \mathbf{V}_{video})$
- **Diversity of ordered set:** We ensure that the fragments in the ordered set are not redundant by explicitly accounting for the diversity of the set. Achieving higher diversity ensures that the fragments cover diverse information in the video, implicitly achieving a good coverage across the video. Starting with the InceptionV3 [25] features for each frame, Coefficient of Variation [1], the metric of dispersion, is calculated across each dimension (D) of the InceptionV3 vector. and then averaged over all dimensions to get a scalar representation of the diversity of a set: $diversity(\mathcal{S}) = \frac{1}{D}\sum_{d}^{D} CV(d)$, where $CV(d) = \frac{\text{standard deviation of dimension } d}{\mid \text{mean of dimension d} \mid}$

The information score is obtained as a linear combination of these 4 scores, $information = w_1 \cdot preSim + w_2 \cdot textCoherence + w_3 \cdot relevance(\mathcal{S}) + w_4 \cdot diversity(\mathcal{S})$, where, $\sum_i^4 w_i = 1$ and $w_i \in [0, 1] \forall i \in \{1, ..., 4\}$. The weights w_i can be tuned as per the desired relative importance to each of the factors. Given this information score, we iteratively loop through the fragments and on each iteration, add the fragment which maximizes information gain of the current ordered set. We continue this process till we either exhaust through all the available fragments or hit an upper limit of fragments needed in the final reordered set, which can be set by viewers or publishers.

4 EXPERIMENTS AND EVALUATION

For evaluating our proposed approach, we curated 2, 700 videos that span three categories – entertainment, lectures, and tutorials), and three durations – short (1-5 minutes), medium (5-10 minutes), and long (15-30 minutes). The resulting nine (3×3) category-duration combinations have 300 videos each corresponding to them. Table 1 gives an idea of the kind of videos that are included in each of the nine combinations. We obtain these $2.7k$ videos by manually identifying YouTube playlists corresponding to these video types (examples are listed in Table 1) and then downloading them for processing. Although there are no other approaches that aim to solve the same problem as ours, we use related approaches from the domains of video summarization, multimodal captioning, and visual storytelling to design competitive baselines and compare the performance of our approach against these baselines using several automated metrics.

4.1 Baselines

The goal of **video summarization** is to summarize a video using a subset of coherent segments from the video. The underlying objective relies on capturing key information from the original video in the summary video and does not account for aspects like efficient navigation, personalization, and reordering which are pertinent to the problem we aim to solve. We start with the series of segments in summarized video from [23], and obtain keyframes for each of these segments. We identify the segment of the original video that corresponds to the summary segment and summarize the corresponding transcripts using the BERT-based summarized discussed above. The keyframe and the summarized transcript are treated as a multimodal fragment and are put together in the original sequence as in the video summary. Given the objective of video summarization, this sequence of fragments is expected to have a good coverage across the video. However, they might not be tailored to user preferences or have the diversity across the fragments.

Work on **audio-visual captioning** uses visual and audio signals from a video to generate dense descriptive captions for different frames of the video [14], that are not necessarily representative of the overall theme in the video. To adapt this work for generating multimodal fragments, we get frame-level captions for the video, and identify frames that have the same caption (this happens because a scene can span multiple frames and a single caption would suffice for the whole duration). We select the frame that is most similar to the common caption by computing the cosine similarity of visual semantic embeddings [8] of the frames and the caption. The frame thus selected along with the caption is treated as a multimodal fragment and are sequenced in the same order as in the video to obtain the final sequence of multimodal fragments. Since audio-visual captioning aims to provide dense captions for each distinct shot in the video, we expect the model to provide fragments that (a) cover most of the video, and (b) are comprised of images and text which are highly relevant to each other. However, since the fragments are not chosen to cater to viewer preferences, they are expected to perform poorly on this front.

Category	Video Types	Example URL
Entertainment	advertisements, news stories, unboxing videos, video blogs (vlogs), etc.	https://youtu.be/dcDW6drsnEo
Lectures	educational videos, classroom lectures, commencement speeches, etc.	https://youtu.be/UF8uR6Z6KLc
Tutorials	how-to videos, software tutorials, recipes, product reviews, etc.	https://youtu.be/sv3TXMSv6Lw

Table 1: Representative types of videos under each category along with an example.

Visual Storytelling [15] aims to generate multiple coherent sentences for a given sequence of images. The sequence of images may or may not belong to a video but are required to have a storyline to them so that the generated sentence can convey a story. We adapt the visual storytelling method [15] by finding the most representative 5 frames (5 is a limitation of the pre-trained model in [15]) from the video by sorting based on the importance score (as discussed in Sec 3). These 5 frames are then passed on to the pre-trained model as input (in the same order as they occurred in the video) to generate 5 corresponding sequential textual pieces that make a larger story (aligned to the story represented by the sequence of frames). We use these image and text pairs as our final multimodal fragments. The expectation from visual storytelling is to have a good storyline independently, but a very poor relevance to the narrative in the video. This can be thought of as a consequence of using a pre-trained model that is trained for a specific type of videos – poor generalization to our context should be reflected in scores that quantify relevance of the fragments to the video. Furthermore, this baseline, like others, does not cater to viewers' preferences.

Finally, we also **randomly sample** the frames from the segments as another baseline. The randomly sampled frames are accompanied by the text from the transcript in the given timestamp. This gives us a naïve baseline to compare our methods against. All the aforementioned baselines work as competitive and relevant adaptations of existing approaches to solve the problem at hand, while the random sampling acting as the lowest performance bar for comparison.

4.2 Evaluation Metrics

- **Image-Video Relevance** measures the closeness of images in the selected fragments to the video context. We take the average ResNet-152 embeddings of the keyframes of the video as the video representation and compute the average cosine similarity of the image representations in the final multimodal fragments with this video representation to quantify image-video relevance.
- **Text-Video Similarity** measures how similar the text in final multimodal fragments is to the transcript of the video. The transcript representation is obtained by averaging the sentence-level BERT embeddings of all sentences in the transcript and computing its average cosine similarity with the text corresponding to every multimodal fragment.
- **Preference Alignment** measures how well the fragments resonate with the viewer's preference and is computed as the average cosine similarity between viewer's preference vector and the average sentiment vector for all the images in final multimodal fragments.
- **Video Coverage** is a measure of the extent to which the video content is covered by the fragments. With an aim to quantify what portion of the original video is covered by our final multimodal fragments, we compute the fraction of segments (identified in the 'Forming Multimodal Clusters' section above) that are included in final set of multimodal fragments. Mathematically, $Coverage = \frac{\text{number of segments covered in the final fragments}}{\text{total number of segments in the video}}$.
- **Image Diversity** quantifies the diversity among the set of selected frames in our final multimodal fragments. To quantify the diversity among the selected images in our final multimodal fragments, we compute the average pair-wise cosine similarity of ResNet-152 embeedings of all images in the final fragments and then subtract it from 1. That is, $1 - \frac{1}{N(N-1)} \sum_{i,j} sim(\text{image}_i, \text{image}_j)$.
- Similar to computing diversity in images, **Text Diversity** is obtained by computing the pairwise cosine similarity of text in the final multimodal fragments and subtract it from 1 (to convert similarity metric into a distance computation).
- **Image-Text Relevance** quantifies the similarity between the output frame and its corresponding sentence. It indicates the semantic relatedness of the text and images and is computed as the fragment-wise cosine similarity between image and text representations in the common visual semantic embeddings [8], averaged across the generated set.
- **Text Coherence** measures the consistency in the semantic flow of the text in final multimodal fragment. We use the BERT Next Sentence Prediction [7] to obtain the likelihood of the text in n + 1th fragment following the text in nth fragment, $P(T_{n+1} \mid T_n)$ and the average across all the values of $n \in 1, ..., N-1$, where N is the total number of multimodal fragments to arrive at the overall text coherence in the fragments.

4.3 Results

Table 2 presents the performance of our proposed approach (NoVoExp) against the different baselines discussed before on the automated metrics.

Comparison against baselines: We start by noting that the performance of NoVoExp is consistently better than all the baselines in terms of (i) relevance of images and text to the video, (ii) preference alignment, (iii) diversity of images and text. We also note that Audio-Visual Captioning consistently [14] shows the best results in terms of the coverage of original video and the relevance between image and text. This can be attributed to the fact that the method *generates* a dense caption to all the major actions in the video and is likely to provide highly relevant captions for almost all the segments in the

Video Type	Duration	Model	Relevance		Preference	Video	Diversity		Image-Text	Text
			Images	Text	Alignment	Coverage	Images	Text	Relevance	Coherence
Entertainment	Short	Random Sampling	0.13	0.21	0.07	0.32	0.39	0.31	0.21	0.33
		Audio-Visual Cap	0.20	**0.57**	0.11	0.72	0.49	0.42	**0.63**	0.42
		VisStorytelling	0.17	0.22	0.13	0.21	0.48	0.39	0.61	**0.82**
		Video Summary	**0.27**	0.26	0.15	**0.76**	0.46	0.40	0.22	0.46
		NoVoExp (Ours)	0.23	0.63	**0.28**	0.67	**0.58**	**0.53**	0.24	0.76
	Medium	Random Sampling	0.11	0.20	0.07	0.28	0.41	0.33	0.20	0.34
		Audio-Visual Cap.	0.17	0.56	0.13	0.69	0.53	0.41	**0.64**	0.47
		VisStorytelling	0.17	0.20	0.15	0.23	0.52	0.42	0.62	**0.83**
		Video Summary	**0.23**	0.25	0.14	**0.72**	0.54	0.44	0.22	0.51
		NoVoExp (Ours)	0.21	**0.59**	**0.31**	0.63	**0.61**	**0.58**	0.26	0.78
	Long	Random Sampling	0.09	0.18	0.08	0.26	0.43	0.47	0.22	0.37
		Audio-Visual Cap.	0.15	**0.58**	0.11	0.63	0.49	0.46	0.63	0.46
		VisStorytelling	0.14	0.21	0.12	0.21	0.47	0.48	0.61	**0.87**
		Video Summary	**0.20**	0.22	0.13	**0.68**	0.48	0.47	0.24	0.47
		NoVoExp (Ours)	0.17	0.53	**0.35**	0.59	**0.65**	**0.62**	0.25	0.81
Lectures	Short	Random Sampling	0.18	0.19	0.06	0.24	0.21	0.37	0.14	0.24
		Audio-Visual Cap	0.26	0.46	0.10	**0.59**	0.27	0.48	0.54	0.39
		VisStorytelling	0.21	0.18	0.12	0.21	**0.28**	0.35	**0.62**	**0.71**
		Video Summary	0.28	0.21	0.14	0.61	**0.28**	0.46	0.15	0.37
		NoVoExp (Ours)	**0.34**	**0.48**	**0.17**	0.48	0.32	**0.59**	0.16	0.64
	Medium	Random Sampling	0.14	0.12	0.05	0.21	0.22	0.42	0.15	0.27
		Audio-Visual Cap	0.23	0.40	0.08	0.54	0.29	0.51	0.57	0.42
		VisStorytelling	0.19	0.15	0.11	0.20	0.31	0.38	**0.64**	**0.72**
		Video Summary	0.27	0.18	0.12	**0.58**	0.30	0.49	0.17	0.40
		NoVoExp (Ours).	**0.31**	**0.42**	**0.20**	0.45	**0.35**	**0.63**	0.18	0.68
	Long	Random Sampling	0.10	0.10	0.07	0.18	0.25	0.45	0.17	0.33
		Audio-Visual Cap	0.19	0.37	0.11	0.51	0.31	0.56	0.58	0.45
		VisStorytelling	0.17	0.14	0.13	0.19	0.35	0.42	**0.65**	**0.74**
		Video Summary	0.24	0.15	0.16	**0.56**	0.33	0.52	0.18	0.45
		NoVoExp (Ours)	**0.27**	**0.39**	**0.23**	0.43	**0.37**	**0.67**	0.19	0.72
Tutorials	Short	Random Sampling	0.12	0.17	0.17	0.27	0.25	0.43	0.09	0.18
		Audio-Visual Cap	0.19	0.42	0.20	0.61	0.29	0.53	0.41	0.29
		VisStorytelling	0.16	0.15	0.23	0.23	0.24	0.40	**0.49**	**0.68**
		Video Summary	0.23	0.18	0.26	**0.68**	0.26	0.51	0.11	0.24
		NoVoExp (Ours)	**0.20**	**0.45**	**0.49**	0.67	**0.59**	**0.57**	0.15	0.63
	Medium	Random Sampling	0.10	0.14	0.20	0.23	0.31	0.50	0.14	0.22
		Audio-Visual Cap	0.16	0.39	0.24	0.63	0.34	0.61	0.44	0.31
		VisStorytelling	0.14	0.13	0.26	0.17	0.29	0.47	**0.52**	**0.70**
		Video Summary	**0.21**	0.12	0.29	**0.66**	0.31	0.59	0.15	0.27
		NoVoExp (Ours)	0.17	**0.40**	**0.51**	0.61	**0.63**	**0.63**	0.18	0.65
	Long	Random Sampling	0.08	0.11	0.25	0.20	0.35	0.58	0.16	0.23
		Audio-Visual Cap	0.15	0.29	0.29	**0.59**	0.41	0.66	0.45	0.35
		VisStorytelling	0.12	0.11	0.32	0.15	0.32	0.52	**0.51**	**0.74**
		Video Summary	**0.17**	0.10	0.33	0.62	0.37	0.64	0.16	0.31
		NoVoExp (Ours)	0.15	**0.37**	**0.54**	0.56	**0.66**	**0.69**	0.17	0.68

Table 2: ***Automated evaluation of multimodal fragments generated by NoVoExp and baselines.***

video. However, it is worth noting that the Audio-Visual Captioning provides fragments that have poor relevance with the video, textual coherence, and preference alignment score. On the other hand, Visual Storytelling [15] as a baseline performs the best in terms of textual coherence and image-text relevance, but loses out on all other metrics. On probing, we find that the fragments generated by this baseline have a generic story that has high coherence, but extremely poor relevance to the video (perhaps a consequence of mismatch of training objectives). Additionally, since the Visual Storytelling baseline can only provide 5 fragments for all videos, we notice a very low value for coverage. The fragments obtained from Video Summary [23] are good in coverage of the video and the relevance of images to the video (as expected), but perform poorly on other metrics. As expected, the naïve baseline using Random Sampling performs poorly on almost all the metrics, except on image-text relevance – that can be attributed to the selection of text from the transcript corresponding to the randomly sampled image. Most importantly, we note that even though NoVoExp is not the best on all the metrics, it is the best overall solution that achieves a reasonable trade-off between important metrics while performing the best in terms of preference alignment, relevance of images and text to the video, and diversity of fragments.

Variation with video duration: Table 2 can also be used to infer how the performance of our proposed method, as well as that of the

baselines, vary as the duration of the videos increase considerably. It can be noted that the average relevance of image and text in the fragments to the original video drops. This is expected because as the duration increases, the total number of segments in the fragments increase and the average relevance of the chosen fragments goes down. This also explains the observed drop in coverage values. However, as the duration of the video increases, NoVoExp has an added advantage of being able to choose from more options – which is also reflected in the increased values of preference alignment, text coherence, and diversity.

Variation across categories: 'Entertainment' Videos have significantly more visual element in them compared to 'lecture' videos. However, lectures have audio that is informative and often takes precedence over visual elements. On the other hand, 'tutorial' videos comprise both the modalities in almost equal proportions with frequent referencing to direct the viewer's attention to specific visual elements ("As you can see, the pizza is now baked."). Given the relative importance of modalities and their interplay varies considerably with the category of the video, we also note the variations in performances of our mode across different categories of videos.

Lectures vs. entertainment videos: We start by noting that for 'lecture' videos, the relevance of images included in the fragments is higher than for other categories, but that of the text is lower. This can be attributed to a lower visual-density and higher auditory-density of lectures. Also, the diversity of images is also lower, while that of the text is higher. Additionally, since lectures do not have a lot of expressive elements, the preference alignment is not as high as those of entertainment videos. We also notice a drop in coherence of text and the relevance between images and text – while the former can be attributed to more domain-specific language (often an academic topic), which may not be familiar to the pre-trained BERT model, the latter can be attributed to a very few visual elements in the video (often lecture videos only comprise of individuals speaking on a topic and illustrations on white/black boards or digital screens).

Tutorials vs. Lectures: Unlike lecture videos, tutorials have significant visual elements in them. Consequently, we note that the relevance of images and text in the fragments to the video decreases. Aligned to this observation, we note that the image-text relevance as well as the text coherence are both lower. However, it's worth noting that the image diversity in fragments is higher than that for the lectures, possibly because of the increased visual density in the videos.

We end this section by noting that across all the durations and categories, NoVoExp consistently remains the best model that can optimize for multiple metrics together. While other baselines perform well on few metrics, they fail as a reasonable alternative solution because of extremely poor performance on the remaining metrics.

5 HUMAN EVALUATION

As discussed earlier, our motivation for addressing this problem is focused around viewers and how they consume videos. Given the centrality of human experience, which is often subjective in the context of this problem, we conduct extensive human studies to evaluate our proposed approach.

Our human evaluation consists of asking several questions using MTurk to assess aspects like informativeness, relevance, preference alignment, diversity, coverage, etc., of the generated multimodal fragments. For all the MTurk surveys, we required the annotators to be 'MTurk Masters' located in the United States having an approval rate $\geq$ 95% and at least 50 annotations approved in the past. We pay all the annotators at the rate of \$12/hour. To get an estimate of the time it would take annotators to complete the surveys, we perform few trial runs on the same platform and use the mean time-to-completion to decide on the annotation cost for the surveys. For the surveys that involve watching a video, we also take the duration of the videos into account.

Our human study consists of 3 parts: *(a)* answering survey questions after viewing just the video (no multimodal fragments), *(b)* answering survey questions after viewing just the multimodal fragments (no video), and *(c)* answering the survey questions after viewing both the video and the multimodal fragments. The purpose of *(a)* and *(b)* is to keep the annotators unaware of a different viewing option (i.e., viewing the video or the multimodal fragments) and assess the similarity of annotations (e.g., the description or the sentimental attributes of the video vs. that of the sequence of multimodal fragments). The purpose of *(c)*, is to let the annotators view both the original video and the generated multimodal fragments and seek more fine-grained responses regarding the quality of generated multimodal fragments. We describe each of these parts in detail below.

- Only viewing the video: The annotators were asked to watch a video and provide a detailed description of the contents of the video in a free-form text field. The annotators were instructed to write detailed descriptions and were provided with both a good (i.e., sufficiently detailed) description and a bad (i.e., superficial and short) description as examples.
- Only viewing the multimodal fragments: The annotators were shown only the generated sequence of multimodal fragments corresponding to the above videos. The annotators were asked to provide a description of the video that the fragments represented (*without* having gone through the video). Additionally, we asked the annotators to choose the most prominent expression from the fragments by selecting one of the following three options: (i) jolly and fun, (ii) adventurous and thrilling, and (iii) thoughtful and compassionate. Since the sequence of multimodal fragments were personalized to cater to the three corresponding viewer preferences, the response of annotators to this particular question will help us understand the alignment between the expression we targeted to cater to vs. the expression that was perceived by the annotators.
- Viewing both the video and the multimodal fragments: In this survey, the annotators were asked to view both the original video as well as the corresponding sequence of multimodal fragments. Following this, they were asked the following questions and were asked to respond on a 5-point Likert scale, where 1 corresponds to 'extremely poor' and 5 corresponds to 'extremely good'.
  - How relevant are the images in the fragments to the original video?

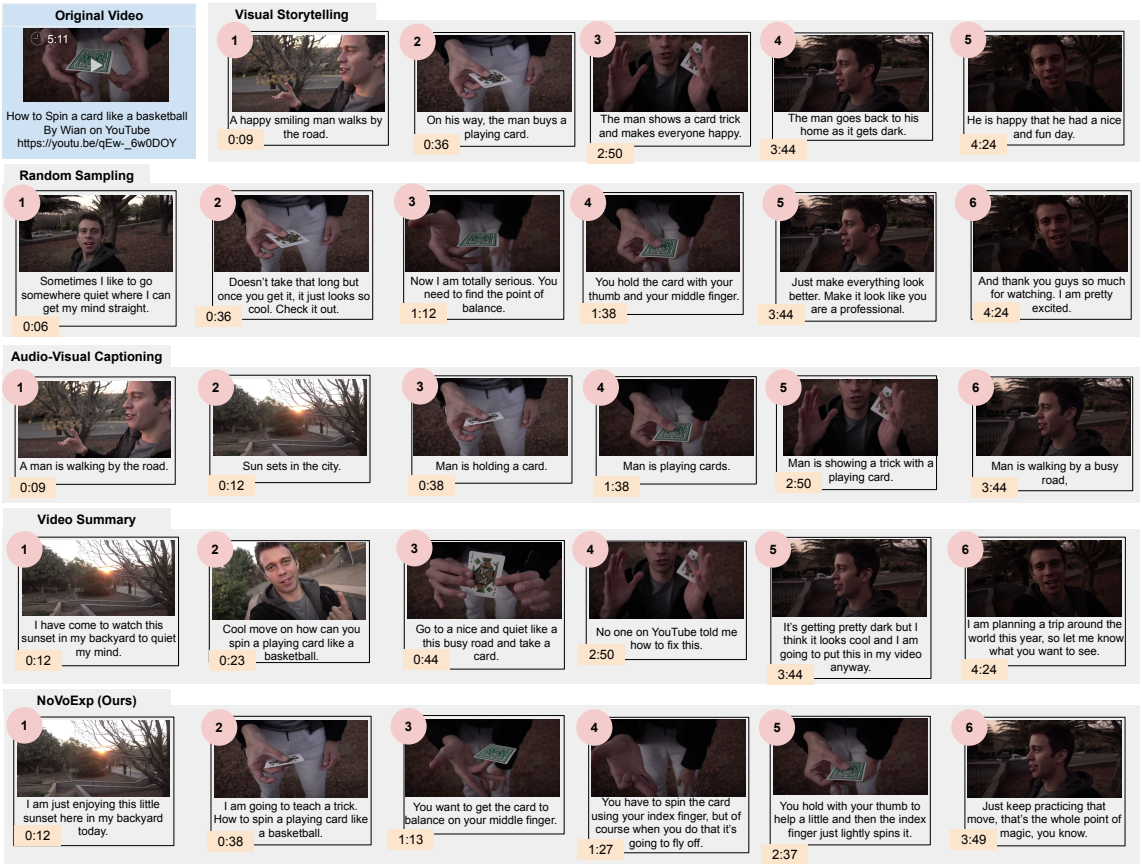

Figure 3: ***Qualitative example comparing NoVoExp against the baselines. Note that the discussed shortcomings of the baselines, as supported by empirical results, are evident here. For instance, the irrelevant yet coherent storyline of the fragments generated using the Visual Storytelling baseline; how Video Summarization achieves the objective of summarizing the video but misses out on the core-parts (i.e., the actual card trick); the generic captioning by Audio-Visual Captioning that remains unaware of the storyline in the video; and lastly, how Random Sampling fails to convey the core of information. In light of these, NoVoExp gives a fairly good understanding of the contents of the video. Refer to Figure 1 for a qualitative example of personalized sequence of multimodal fragments.***

- How relevant are the captions in the fragments to the original video?
- How well do the fragments cater to the the <target preference>? (one of the three)
- How well do the fragments convey the story conveyed in the original video?
- How diverse are all the images with respect to each other in the presented fragments?
- How diverse are all the captions with respect to each other in the presented fragments?
- How coherent (logical and easy to follow) is the story told by the images alone?
- How coherent (logical and easy to follow) is the story told by the text alone?
- How well do the fragments cover important segments of the video?
- How well do the image and text within a fragment relate to each other?

For each of the above surveys, we used 50 videos spanning uniformly across the 3 categories (entertainment, lectures, and tutorials) and time duration (short, medium, and long). Table 3 summarizes the results of our human study, where each question was answered by 5 different annotators. We take the description provided by the annotators after watching either the videos or the multimodal fragments and compute the similarity between BERT-embeddings of these descriptions. As we can note, the similarity between both the descriptions is quite high, indicating that the annotators were able to obtain similar information from the multimodal fragments without even watching the video – thus supporting the claim around effectiveness of the generated fragments in capturing the information in the video. We also note that the similarity between these descriptions decreases with the increase in duration of the videos. For 38 out of 50 videos, majority of the

Video Type	Duration	Description Similarity	Relevance		Preference Alignment	Video Coverage	Diversity		Image-Text Relevance	Coherence	
			Images	Text			Images	Text		Images	Text
Entertainment	Short	0.63	3.68	3.61	3.50	3.92	4.12	3.63	3.32	3.46	3.81
	Medium	0.59	3.26	3.52	3.62	3.95	4.16	3.64	3.33	3.41	3.76
	Long	0.53	3.17	3.41	3.67	3.99	4.17	3.68	3.34	3.38	3.75
Lectures	Short	0.57	3.94	3.59	3.17	3.86	3.13	3.51	3.21	3.17	3.42
	Medium	0.52	3.86	3.51	3.23	3.90	3.17	3.53	3.17	3.16	3.38
	Long	0.48	3.48	3.42	3.29	3.93	3.19	3.54	3.22	3.18	3.41
Tutorials	Short	0.54	3.32	3.57	3.42	3.62	3.91	3.62	3.25	3.20	3.44
	Medium	0.51	3.18	3.41	3.47	3.67	3.92	3.56	3.26	3.27	3.48
	Long	0.46	3.07	3.37	3.49	3.71	3.93	3.58	3.29	3.43	3.52
Overall	Short	0.58	3.65	3.59	3.36	3.80	3.72	3.59	3.26	3.28	3.56
	Medium	0.54	3.43	3.48	3.44	3.84	3.75	3.58	3.25	3.28	3.54
	Long	0.49	3.24	3.40	3.48	3.88	3.76	3.60	3.28	3.33	3.56
	All	0.54	3.44	3.49	3.43	3.84	3.74	3.59	3.26	3.30	3.55

Table 3: ***Human evaluation aiming to assess the quality of multimodal fragments generated by NoVoExp.***

annotations (i.e., >= 3 out of 5) identified the same prominent expression in the generated multimodal fragments as the one we had targeted. In other words, if we targeted one of the three possible preferences while generating the multimodal fragments, 76% times the annotators perceived the fragment to be expressive in a similar manner, independently. This shows the alignment between our targeted preferences while generating the fragments and the viewers' perceptions.

While providing answers to the questions assessing the relevance, coherence, diverseness, coverage, etc. of the multimodal fragments, most annotators chose either 'moderately good' and 'extremely good'. This is indicated by the average values shown in Table 3 where all them are > 3 and some values are even > 4. The Fleiss' kappa score corresponding to these values are in the range of 0.45 to 0.65 indicating moderate to substantial inter-annotator agreement for all the questions. Interestingly, while some of the trends observed in Table 2 persist here as well, e.g., decrease in image/text relevance and increase in preference alignment as duration increases, low image diversity values for 'lecture' videos, etc., there are a few notable trends that emerge from Table 2. Firstly, the existence of more diverse images in 'entertainment' video and more diverse text in 'tutorial' videos is clearly evident (when compared against other categories). Additionally, on observing the average values across all categories ('Overall'), we note how the relevance between image and text in multimodal fragments remain largely the same as the duration increases.

6 DISCUSSION AND LIMITATIONS

As discussed in our motivation, videos make up a fair share of digital media that is consumed by individuals in all walks of life. Recent estimates and forecasts suggest that, on an average, an individual spends over an hour everyday consuming online videos [2]. Given this, and the diverse range of preferences that the viewers of a single video could have, innovations on two fronts are needed: (a) navigation strategies to efficiently consume videos, and (b) personalizing the consumption method to align well with viewer's preferences. NoVoExp, the proposed methodology, is a step in these directions. The generated sequence of multimodal fragments can be skimmed through to understand what the video is about, the topics that it covers, and what is the flow of the narrative. While providing text-based indexing of the video is a simplistic way to achieve similar goals, we argue that multimodal fragments are better representations of the video content. Furthermore, re-ordering the fragments to prioritize the consumption of viewer-preferred segments, while preserving the narrative of the video, helps in achieving greater satisfaction among the viewers. Even though we have focused on three personas in this work, it is a trivial extension to expand this methodology to cater to a larger and wider viewer segments.

However, we do acknowledge some of the limitations of this work. Firstly, not all videos are meant to be consumed in a non-linear fashion; a considerable fraction of videos are consumed for entertainment purposes, like movies and television shows, without worrying about the efficiency of their consumption. Furthermore, some of the entertainment videos are too short, like videos on social media websites and applications like Pinterest, TikTok, and Instagram. It is worth noting that these video consumption behaviors that are largely entertainment-driven are not of interest to the work presented here. Instead, we are interested in videos that need a more-efficient and personalized consumption methodology. Secondly, the sequence of multimodal fragments are not necessarily a *substitute* for the original video. Instead, they *complement* the original video by allowing viewers a quick and efficient navigation and consumption methodology. Lastly, we acknowledge that the video creators and publishers might not always prefer to let their videos be consumed in a way/order that is different from what they intend. This calls for a discussion among publishing platforms, creators, and viewers to decide upon an equitable strategy. Accordingly, the proposed methodology should cater to inputs from the creators while publishing the sequence of multimodal fragments along with the original video.

7 CONCLUSION

In this paper we present NoVoExp, a methodology that uses state-of-the-art computer vision and natural language processing techniques to enable more efficient consumption and navigation of videos. NoVoExp extracts multimodal information from the videos, in form of frames and audio transcripts, and then transforms it to a sequence of fragments that are composed of images and text.

While doing so, NoVoExp personalizes the multimodal fragments by aligning their attributes and sequence with the targeted viewer preference. Our extensive evaluation, using automated metrics and human studies, shows that the sequence of multimodal fragments *(a)* capture the information provided in the video, *(b)* have high relevance to the video while being diverse among each other, *(c)* cover a considerable portion of the video, and most importantly, *(d)* align well with the targeted viewer preference without while preserving the overall coherence of the original narrative. Empirical and qualitative comparison against several competitive and relevant baselines shows that NoVoExp is best in terms of striking a balance between the trade-offs among crucial metrics. We believe that the final sequence of multimodal fragments generated using NoVoExp will enable viewers to efficiently understand the contents of the video without spending time in proportion to the duration of the video and navigate to the specific parts they are interested in.

REFERENCES

[1] Hervé Abdi. 2010. Coefficient of variation. *Encyclopedia of research design* 1 (2010), 169–171.

[2] Online Article. 2019. Online video viewing to reach 100 minutes a day in 2021. https://www.zenithmedia.com/online-video-viewing-to-reach-100-minutes-a-day-in-2021/. Accessed: 2020-10-08.

[3] Online Article. 2020. Microsoft Thinks Coronavirus will Forever Change the Way we Work And Learn. https://www.theverge.com/2020/4/9/21214314/microsoft-teams-usage-coronavirus-pandemic-work-habit-change. Accessed: 2020-10-08.

[4] Online Article. 2020. The Virus Changed the Way We Internet. https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html. Accessed: 2020-10-08.

[5] Yiyan Chen, Li Tao, Xueting Wang, and Toshihiko Yamasaki. 2019. Weakly supervised video summarization by hierarchical reinforcement learning. In *Proceedings of the ACM Multimedia Asia*. 1–6.

[6] Brodie Clark. 2020. How to Make the Most of Video Timestamp Results in Google Search. *Search Engine Journal* (2020). https://www.searchenginejournal.com/video-timestamp-results-google-search/364020/

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[9] Himaanshu Gauba, Pradeep Kumar, Partha Pratim Roy, Priyanka Singh, Debi Prosad Dogra, and Balasubramanian Raman. 2017. Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Networks* 92 (2017), 77–88.

[10] Hongxiang Gu and Viswanathan Swaminathan. 2018. From thumbnails to summaries-a single deep neural network to rule them all. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[11] David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 237–244.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1705–1715.

[14] Vladimir Iashin and Esa Rahtu. 2020. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. *arXiv preprint arXiv:2005.08271* (2020).

[15] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973* (2018).

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[17] Debabrata Mahapatra, Ragunathan Mariappan, Vaibhav Rajan, Kuldeep Yadav, and Sudeshna Roy. 2018. VideoKen: Automatic Video Summarization and Course Curation to Support Learning. In *Companion Proceedings of the The Web Conference 2018*. 239–242.

[18] Jordi Mas and Gabriel Fernandez. 2003. Video Shot Boundary Detection Based on Color Histogram.. In *TRECVID*.

[19] Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165* (2019).

[20] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 19–27.

[21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[22] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3973–3983.

[23] Yair Shemer, Daniel Rotman, and Nahum Shimkin. 2019. ILS-SUMM: Iterated Local Search for Unsupervised Video Summarization. *arXiv preprint arXiv:1912.03650* (2019).

[24] Kai Sun, Junqing Yu, Yue Huang, and Xiaoqiang Hu. 2009. An improved valence-arousal emotion space for video affective content representation and recognition. In *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 566–569.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[26] Xiaoming Tao, Linhao Dong, Yang Li, Jizhe Zhou, Ning Ge, and Jianhua Lu. 2015. Real-time personalized content catering via viewer sentiment feedback: a QoE perspective. *IEEE Network* 29, 6 (2015), 14–19.

[27] Gyanendra K Verma and Uma Shanker Tiwary. 2017. Affect representation and recognition in 3d continuous valence–arousal–dominance space. *Multimedia Tools and Applications* 76, 2 (2017), 2159–2183.

[28] Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. Exploring multimodal video representation for action recognition. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1924–1931.

[29] Xin Wang, Yuan-Fang Wang, and William Yang Wang. 2018. Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 795–801.

[30] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2017. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv preprint arXiv:1801.00054* (2017).

[31] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 207–212.