

# Multidimensional Text-to-Video Generation: Integrating Sentiment, Pragmatics, and Semantics

Gaganpreet Kaur<sup>1, \*</sup>, Amandeep Kaur<sup>1</sup>, Meenu Khurana<sup>2</sup>

<sup>1</sup>Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India; [gaganpreet.kaur@chitkara.edu.in](mailto:gaganpreet.kaur@chitkara.edu.in), [amandeep@chitkara.edu.in](mailto:amandeep@chitkara.edu.in),

<sup>2</sup>Chitkara School of Engineering & Technology, Chitkara University, Baddi, Himachal Pradesh, India; [meenu.khurana@chitkara.edu.in](mailto:meenu.khurana@chitkara.edu.in)  
Corresponding author : [gaganpreet.kaur@chitkara.edu.in](mailto:gaganpreet.kaur@chitkara.edu.in) (G.K.);

## ARTICLE INFO

## ABSTRACT

Received: 21 Dec 2024

Revised: 27 Jan 2025

Accepted: 12 Feb 2025

In this study, we have developed a new framework for creating videos automatically. Unlike diffusion models, our work focuses on assembly and composition of existing media rather than generation of new content with high overhead. The approach followed in this research work deconstructs the input script into sentences and examines each textual input segment using sentiment, pragmatic, and semantic analysis. The query engine receives entities found in the love letter content—the use case investigated—and gathers pertinent video clips. Our method aligns the segments for meaningful links, creating a seamless video composition that mirrors the written story. The work gives better qualitative outcomes as it incorporates sentiment and pragmatic analysis. We used a panel of five judges to assess the quality of the generated automated movies, and computed an intraclass correlation coefficient to ascertain inter-rater agreement. In terms of the produced videos & quality, we got encouraging results. The successful outcome shows that this is an efficient method (very low overhead as compared to diffusion models) for automating contextual video synthesis based on limited text input, as further confirmed by the panel consensus and our cohesion-focused methodology.

**Keywords:** Automatic Video Generation, NLP, Sentiments, Pragmatics, Narratives

## 1. Introduction

The domain of Automatic Video Generation (AVG) is undergoing strong expansion and dramatic transformation. AVG aims to create videos from text or other media inputs automatically, which provides a useful assistant in education [2], news and entertainment field [3]. For this technology evolution transforms content composition and expands the opportunities for automation beyond conventional uses.

The video production goal is to minimize manual video production work through automatic video generation technology in line with the trend of AI-based content creation. Because video is quickly emerging as the leading content type across digital networks in various verticals including entertainment, news, educational, and advertisement industries, the demand for suitable and cost-effective solutions has increased. AVG methods convert text-based stories into entertaining visual content with the help of AI, Natural Language Processing, and 3D modeling of scenes. Apart from reducing production cost, this progress in automation also brings the ability to create adaptable and personalized media solutions to the market that are dynamic to the audience's need instantly.

Existing literature shows that AVG involves a multi-step process, each crucial for the final output; it includes a media server 'M' that has a diverse media element including images, videos, and audio, forming the basis for video generation[4]. Advanced natural language understanding (NLU) techniques are being used by AVG systems to evaluate and interpret complex scripts, ensuring contextual accuracy and coherence in video output. NLP enhances visual storytelling by uncovering subtleties like tone, sentiment, and meaning. Continuous advancements in NLU and computer vision are crucial for creating realistic, emotionally compelling videos[57][58].

Most of the AVG systems take text script as input and it needs to be preprocessed [5]. It involves utilising NLP and regular expressions; this step involves cleaning, formatting, and transforming text data for further processing. Generally, the input script needs to be converted into a computational scene, for this the text segmentation

algorithm is required [6][7]. Text Segmentations algorithms are used for breaking down text into manageable segments like sentences or paragraphs and help to identify essential elements for constructing video sequences [8]. The creation of the video sequences is equivalent to generation computational narration models. Based on multiple input elements such as atmosphere and settings, mood and tone, action and acts, character/person relationship, location, timeline, and others. The scene needs to be mathematically constructed and sequenced [9]. They include entity identification, which implies extracting and classifying named entities and relevant information from texts to guide the video creation process[56]. In the next step, the queries are constructed to get media elements on the fly based on the identity found in the input text script [10]. Implying a context driven queries - based mechanism comes into play [11]. Sequences are constructed as the algorithm determines the temporal relationships between text segments and media objects. Accurate associations of temporal relationships [2] between text semantic and contextual meaning with the media items are critical for logical video assembly. At the same time the text and media integration occurs with video editing methods or custom rendering engines are employed to produce the end video [12].

Despite so many advancements, AVG faces significant issues, especially in terms of data quality, relevance and bias. Maintaining high-quality, diverse media datasets is imperative to prevent biases and ensure the accuracy of generated videos. Constructing video based on computational models is a challenging task because every person perceives and experiences some video content in different ways. . However, factual video or data driven video can be less subjective [13][50]Therefore, improving accuracy, refining entity identification, query formulation, and timeline analysis is vital for producing videos that align well with the input script and media. In the next sections, we conduct an investigative survey on these aspects with help of the latest high impact scientific literature.

The paper has VI Sections. The subsequent sections seek to enhance comprehension of AVG by offering an in-depth exploration of its technological foundations, methodological methodologies, and prospects for further development. This study specifically aims to show how automation, media composition, and user interaction interact to show how AVG is changing the digital content creation ecosystem. This research highlights the benefits and drawbacks of automation in story development and content customisation by contrasting AVG's capabilities with traditional video production techniques. Sections I and II cover the Introduction and Literature review. Section III and IV show the Problem Structure and Methodology of this research. Section V and VI describe the Results, Discussions and Conclusion and Future directions of the analysis[59].

## 2. Literature

In this section, we consider documentary evidence to understand the contemporary landscape of the Automatic video generation technological stack. It must be noted this domain is also referred to as “Generative AI”, which is considered as a type of computational intelligence that can create content such as articles, blog posts, reports, presentations, images, and videos [14]. These models use complex machine learning algorithms to predict the next word or image based on previous sequences [15]. Generative AI tools can already create most types of written, image, video, audio, and coded content [14]. In the near future, applications that target specific industries and functions are expected to provide more value than those that are more general. One area of interest in generative AI is text-to-video conversion [4] that is attention in terms of quality of videos being generated automatically. While initial endeavours in this area were constrained by certain limitations, recent advancements have enabled the creation of extensive, high-definition video content [16][17].

Text-to-video conversion algorithms are revolutionising the way videos are created. These algorithms use templates, animations, and other elements to automatically convert written text into engaging videos, significantly simplifying video creation for both professionals and beginners alike[18][19]. These algorithms interpret the text, select appropriate visuals, animations, and background music, and compile them into a video. Users can often customise elements such as fonts, colours, and transitions, ensuring the final video aligns with their vision [7]. This level of customization necessitates human intervention, highlighting the semi-automatic nature of these AVG frameworks [20].

Through alignment, verification, and human feedback,[21] AVG ensures relevance, quality and user satisfaction. Alignment is achieved through optimization of objective, relevance is achieved by matching videos to user preferences, and validation ensures standards are met. User feedback allows system adjustments, and machine learning algorithms from interactions and contentment improve user experience [22][48].

Natural Language libraries such as NLTK, TextBlob, SpaCy, and Gensim are used to generate video scripts [23][48]. These libraries perform text preprocessing operations such as text extraction, tokenization, and stemming. Researchers have improved the performance of video production systems but may have reduced the amount of information available to evaluate the emotional meaning of text. Therefore, it is important to carefully consider the potential impact of preprocessing on the emotive meaning of the text before applying these techniques to input text for video generation. Alternative to the issue is to have an intermediary algorithm that can assess emotive meanings in between [49].

For getting a fair idea (Topic Analysis) of the input text for video, the researchers[41] are using algorithms such as LDA, NMF, Word2Vec, GloVe, and BERT and pretrained models such as GPT, BERT, I3D, and T5[24] are frequently used.. AVG also requires text segmentation algorithms to divide the text into logical sequences so that a computation scene can be constructed. Text Tiling, introduced by Hearst in the year 1997[25], remains a foundational unsupervised topic segmentation algorithm that utilises a moving window-based approach. This method effectively identifies transitions between topics by calculating cohesion at potential boundary points, leveraging lexical cohesion between blocks of text to discern topic boundaries and dividing the input text into sequences of relevant tokens. More recently, CoType, proposed by [26], has emerged as a domain-independent framework employing a data-driven text segmentation algorithm. Text segmentation involves extracting entities, placing them in lower ranks, and examining relationships, labels, and types [27]. Continuity modeling is implemented with the CoType method, which uses neural networks to remodel segments. Semantic word embeddings, such as Word2Vec and GloVe, improve text segmentation algorithms. New techniques, such as Hierarchical Topic Segmentation (HTS) and Neural CRF-Based Topic Segmentation (NCTS), enhance the ability of text to capture coarse- and fine-grained topics and pattern changes [28][29].

According to research [30], entity recognition is important for the correct identification and removal of specific entities. This helps improve the accuracy and usability of understanding content and extraction entities in specific domains, especially for applications that automatically generate videos. Identification of specific domains that are not covered by standard NLP, such as object numbers and expressions, can be addressed by generating domain IDs. This feature makes the extracted data more accurate and correct. Moreover, custom entity recognition models excel in contextual understanding by learning the specific context in which entities are likely to appear, enabling distinctions between different types of entities within a given context. Hence, as per researchers they are essential part of frameworks that make AVGs. The flexibility and adaptability of custom entity recognition are highlighted in its ability to accommodate new naming conventions, variations (such as gender identities), and typos, which may be overlooked by generic entity recognition approaches.

Numerous methods and techniques [31] are employed to the dynamic generation of contextual queries, adapting to specific applications and requirements. Text analytics and tokenization are techniques used to extract meaningful content in search engines and databases. Dynamic programming solves optimization problems and takes user input and context into account to create effective searches. GraphQL[32] increases query flexibility and performance. Data-driven queries are built using machine learning algorithms to analyze data and create questions based on patterns. Rule-based questions are built using rules based on user input and context. Time analysis algorithms are essential for automatic video production. These algorithms, [44][45] including time series analysis, delve into techniques such as time series plots, autocorrelation analysis, and seasonal decomposition to unveil patterns and trends in temporal data [34]. Real-time algorithms, crucial for AVG applications involving instantaneous data acquisition and prediction, manage temporal streams, enabling dynamic processing and analysis for the creation of new analysis capabilities [33]. Temporal decomposition algorithms, exemplified by methods like STL (Seasonal and Trend decomposition using LOESS [34]), disassemble time series data into trend, seasonal, and remainder components, contributing to a nuanced understanding of trends and seasonality for the production of dynamic and engaging videos[4]. Additionally, Spatio temporal analysis algorithms, focused on data variance across space and time, provide insights into how variables evolve over these dimensions, proving invaluable for crafting visually compelling and informative videos within the AVG domain [35][60].

The research papers [36] related to time analysis, it is observed that the analysis of the temporal order of sentences and media objects involves a diverse array of methodologies and techniques, as synthesised from the search results. Firstly, the exploration of timeline generation algorithms reveals a methodology centered on evaluating such algorithms based on deep semantic units. This transformative approach focuses on converting

roughly chronological input text into timestamped summary sentences, providing a structured method for timeline generation[1].Secondly, there is an emphasis on deducing the timeline/chronology (order of events) of novel series, involving the division of texts into days and the creation of a relationship database to deduce a best-fit chronology. This process results in a fair degree of complex relationship graph with inherent inconsistencies [37]. Therefore, there is a need for a better approach that helps to compose the video assembly without such issues.

Diffusion Models are quite popular, in fact they dominate the main discourse in this research area of making automatic video generation. Within the domain of Diffusion models, multiple approaches include [61] Denoising Diffusion Probabilistic Models (DDPMs), with subsequent breakthroughs in text-to-image synthesis[62] and AVG. Recent literature published between 2023 and 2025 highlights discussions on the core architectural principles, video-generation frameworks, and efficiency optimizations. Key approaches include cascaded spatial-temporal architectures for high-definition video synthesis, state-space model (SSM)-based efficiency improvements, and hybrid approaches combining reconstruction and density-based anomaly detection[63]. All these approaches are computationally expensive and they are generative in nature, which implies they generate video that never existed earlier, not using the vast media library that humankind has. Yes, these models are trained on existing video, but they are not utilizing the potential of the existing media library of mankind. Researchers have been trying to reduce the memory and computational overhead and the future lies in building distilled and quantized models. Open AI product Sora and Google’s Veo are the most recent developments in this context. These have been trained on large cultures of GPUs and other hardware peripherals[64].

In this study, an attempt to overcome the issues related to the existing methods (see Table 1) and based on the comparative view point of the problem statement.

Table 1: Issues related to Existing Methods

Feature	Existing Methods	Proposed Framework
Text Analysis Approach	Primarily semantic analysis or keyword matching	Integrated sentiment, pragmatic, and semantic analysis
Video Composition	Template-based or rule-based composition	Dynamic video composition based on narrative flow and emotional tone
Contextual Understanding	Limited to surface-level semantics	Deep contextual understanding through pragmatic analysis
Emotional Reflection	Minimal attention to emotional subtext.	Emphasises emotional tone and narrative coherence
Temporal Alignment	Fixed or predefined temporal structures	Adaptive temporal alignment based on script's storyline
Flexibility in Content	Restricted by predefined templates	Flexible composition through dynamic query generation

So technically, in this research, the problem revolves around transforming written narratives into appropriate video content, a structured methodology is proposed to automate the creation of videos ‘Va’ from a written script, specifically a love letter. This approach meticulously breaks down the script into its elemental components, leverages advanced analysis techniques to interpret its content, and dynamically select media components to visually represent and compose the narrative. Here's an organised breakdown of the problem and solution undertaken in this research.

Table 2: List of Abbreviations

Terms	Notation/ Abbreviation
Script	S
Automated Video	Va
Video components	Vc

Universal set of media components	M
Dynamic Queries	q
Computational Scene Construction	CSC
Segments	Seg
Component Selection	CS
Emotions	E
Polarity	P

### 3. Problem Structure

The problem revolves around a holistic approach that uses the combination of sentiment, pragmatic, and semantic analyses to offer a more comprehensive understanding of the text than using these techniques in isolation, which has been undertaken by previous researchers. The framework must have the ability not just the content but also the emotional undertones and contextual nuances of the input script. Further allowing for a more multifaceted representation of the narrative, moving beyond surface-level content matching or single or dual aspect research work .Hence, the elements of the problem include :

- S: The original written narrative (e.g. Love Letter) intended for transformation into video.
- Va : The final video output to be generated from the script.
- Vc : The individual units that make up each automated video ‘Va’.
- M: A comprehensive collection of media elements (images, clips, sounds) from which the specific components ‘Vc’ are selected for composing ‘Va’.
- q: A set of queries dynamically generated to facilitate the extraction of specific media components (mi) from the universal set (M).

#### 3.1. Computational Scene Construction (CSC)

##### 3.1.1. Initial Script Analysis (s):

- Segment the input script (S) into meaningful and manageable sections (Seg).
- For each segment (Seg) in Seg, perform the following analyses and identifications:

##### 3.1.2. Topic Inference

##### Sentiment Analysis

- Compute polarity ‘P’ to gauge the segment’s positive or negative tone.

##### Pragmatic Analysis

- Analyse and identify meaning and context expressed in discourse of ‘S’. Pragmatic analysis enables the algorithm to grasp the intended meaning behind the script’s language, that too in a sentence-wise manner. This helps in capturing aspects such as implied emotions, and the atmospheric setting of each computation scene.

##### Semantic Analysis

- Dive deeper into the meaning of the text to identify:
  - Atmospheric words that set the scene or mood.
  - The timeline, pinpointing when events occur or are mentioned.
  - Actions depicted or described in the narrative.

### 3.1.3. Component Selection (CS)

- Based on the analyses, identify the necessary components for the video 'Va'.

### 3.1.4. Dynamic Media Component Extraction

- For each identified component of CS:
  - Generate a dynamic query (q) tailored to extract a specific media component from the universal set (M).
  - Execute each dynamic query (q) to retrieve the appropriate media components.

Computation Scene 'CSC' = (Polarity (P), Emotions (E), Location, Mood and Tone, Atmospheric, Actions, Persons and Timeline), where, identified through the function of Topic Inference (Semantic, Pragmatic and Sentiment analysis.)

### 3.1.5. Video Assembly Composition

- For each extracted media component 'Vc':
  - Join 'Vc' together in a sequence that aligns with the narrative flow and analysis outcomes.
  - Va = Fully composed video composition.

This structured approach enables the composition and validation of the transformation of written love letters 'S' into automated videos, ensuring that the emotional essence, thematic elements, and narrative structure are accurately represented in visual form. Through the integration of sentiment, pragmatic, and semantic analyses, the process not only selects appropriate media components but also composes them with the narrative's mood, tone, and progression, creating a video that truly encapsulates the original written script's essence.

## 4. Methodology

In this section, the steps are taken to resolve the problem stated. For better understanding of the process, Figure [1] can be considered.

### 4.1. Dataset and Pre-Processing

Two types of dataset involved in this research, the first one [54] was created to store media elements such as sound, movie clips and images. The second type of dataset included the input text scripts. Since, we are developing an AVG for real life specific use case 'visualisation of love letters'. 300 plus odd love letters were collected from various digital platforms such as Quora [47], various social media platforms and some collected by author personally and 87 Love Letters are collected from existing dataset available on kaggle[42]. The dataset was uploaded on Mendeley for public access[52]. The text data was pre-processed and is available publicly. For pre-processing special character removal, stop word elimination, formatting, spell checks and grammar check was done so that no misunderstanding of the context of text happens [39][40].

### 4.2. Text Segmentation Algorithm

In this step, the pre-processed text /cleaned text is processed for the conversion of the text into a tabular format data (Text2Seg) method. In this method, the paragraphs of each unit of text (a letter) are processed to form a row of segments matrix. However, it must be noted that for these steps multiple options were at disposal, including rule-based segmentation, statistical function/metric (that can identify linguistic boundaries) based text segmentation. Next is utilising supervised or unsupervised learning, these algorithms are potent in learning complex data relationships. Their high accuracy comes at the cost of requiring substantial training data and being less interpretable. Some researchers are forced to use hybrid approaches due to the complexity of the content. Combining various segmentation methods, hybrid algorithms offer robust results by leveraging the strengths of different approaches, albeit with increased complexity. In the context of script analysis, segmentation is essential for understanding narrative structures. Techniques like scene segmentation, discourse segmentation, dialog act segmentation, and sentiment analysis are employed, each serving specific purposes like studying plot development, character interactions, and emotional arcs. For monologue content such as love letters, algorithms like topic segmentation, discourse segmentation, argumentative zoning, and emotion segmentation are useful. These



techniques help in dissecting monologues based on themes, discourse shifts, argument structures, and emotional tones, respectively. However, for this research, several experiments were conducted, and a custom algorithm was authored to suit the specific purpose[6][8].

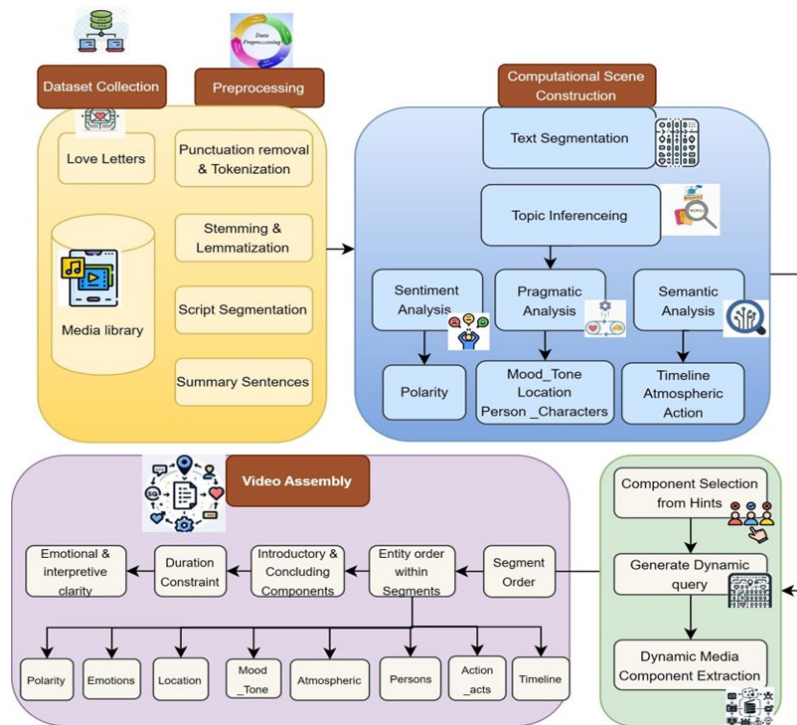


Figure1. Workflow of Automatic Video Composition Model.

#### 4.3. Topic Inferencing Using Sentiment and Pragmatic Analysis:

The new approach followed in this research is to enable the algorithm to infer implicit relationships and situational context. For example, in a love letter, phrases that suggest intimacy or conflict may not explicitly mention emotions but imply them through tone, dialogue, and setting. The algorithm analyses these pragmatic cues from the given script 'S' to identify key elements such as mood, tone, character dynamics, and temporal transitions.

Hence starting from the input script 'S' analysis journey, our initial phase involves the extraction and identification of crucial 'hints,' or 'topics' pivotal elements that form the foundational building blocks for the subsequent video assembly. These nuanced hints/topics are meticulously gathered from the segmented text, a process involving the dissection of sentences within the love letters. The diverse array of hints encompasses entities ranging from individuals and their associated gender, to atmospheric settings, sentiment polarity, distinct acts and actions, prevailing mood and tone, intricate character relationships, timelines, diverse locations, and other entities, including pertinent company names. Each of these discerned 'hints or topics' are systematically catalogued and stored within a structured tabular format, creating a comprehensive dataset that serves as the reservoir for subsequent analytical processes in Table 3. Technically, this amounts to a deep contextual understanding to construct a computational scene 'Cs' that mirrors the narrative's subtleties, there by producing a video composition that resonates more closely with meaning of the letter.

Upon the completion of the exhaustive hint collection and the establishment of contextual relationships within the love letter sentences, our focus seamlessly shifts towards the activation of query engine. The query engine plays a pivotal role in firing queries designed to not only compose the assembly of the video but also to facilitate a structural temporal analysis. This transition from hint/topic collection to the query activation phase signifies a critical juncture where the synthesised data begins to transform into an actionable and composable component of video assembly. The queries to understand the interplay of entities and their contextual significance, creating a narrative-rich video assembly that encapsulates the essence of love letters shown in Figure [2] and [3].

Table 3: Extraction of Topic and Hints

Love Letter Title	Content	Segments	Sentiment	Mood and Tone	locations	Persons	Atmospheric_Words	Time line	Actions
Love Enduring Promise	My Dearest Gudiya,As I sit here, gazing at the distance that separates us, my heart aches with the longing to be by your side once more. Our love story, forged amidst the complexities of an India-Pakistan connection, has faced its fair share of challenges. But through it all, our love remains steadfast, an unwavering flame that continues to burn bright.The moments we spent together in Karachi, in the embrace of love's sweet enchantment, remain etched in my memory.	As I sit here, gazing at the distance .	POSITIVE	love	room	Male , Boy,Female,G irl,Woman	heart	Before separation	Sitting, gazing
		Our love story, forged amidst the	POSITIVE	joy	India	Man, Male , Boy,Female,G irl,Woman	love, connection	Formation of love	Forging a love story
		The moments we spent together	POSITIVE	love	Karachi	Man, Male , Boy,Female,G irl,	embrace, love	Moments in Karachi	Spending moments together
		Every moment, from secret rendezvous	POSITIVE	joy	Sipari	Man, Male , Boy,Female,G irl,Woman	connection	Cherished moments	Being in secret rendezvous
		Since our separation, solace is found	POSITIVE	joy	home	Man, Male , Boy,Female,G irl,Woman	love, love	During separation	Finding solace in correspondence emails
		Education has enriched our understanding ,	POSITIVE	joy	room	Man, Male , Boy,Female,G irl,Woman	connection	Educational phase	Enriching understanding
		Distance is an illusion we understand,	POSITIVE	love	room	Man, Male , Boy,Female,G irl,Woman	calm	Understanding of distance	Understanding distance as an illusion
		As we navigate separation,	POSITIVE	joy	room	Man, Male , Boy,Female,G irl,Woman	love	Navigating	Navigating separation
		Though our countries may be divided, our love	POSITIVE	anger	room	Man, Male , Boy,Female,G irl,Woman	love	boundaries	Defying boundaries
		My beloved, as we continue our journey, let our love	POSITIVE	joy	room	Man, Male , Boy,Female,G irl,	beloved, love, love	Hope for reunion	Eagerly awaiting reunion

The hints/topics to these prompts touch on specific details of the love letter, such as tone, character relationships and context going beyond a literal interpretation of words. An important element of pragmatics is getting beyond the words/sentences to what we communicate with language. Note that ‘contextual relationships’ are identified at the same time. This includes noting how different textual elements work in unison to contribute towards the whole of the text. This fits with the basic concept of pragmatics, which is the study of how language is used in given contexts for specific communicative purposes. It also is crucial to understand that the text describes the assembled information turned into the Narrative rich video assembly ‘Va’. This aligns with the pragmatic goal of using language to achieve specific communicative effects.

This streamlined format facilitates a understanding of the script's narrative nuances, enabling us to make a computational scene ‘CSC’ which is the structure of the video assembly ‘Va’. In Table 4. the logic used for the construction of the computational scene is given.



Table4.Logic for creation of the Computation Scene

Hint/Topic Type		Logic Implemented
Sentiment Analysis	Sentiment Polarity_function	To understand the context of sentences in the love letter in terms of positive , negative , neutral sense a python script was developed . The script utilises various natural language processing (NLP) libraries, including TextBlob, VADER Sentiment, Flair, and Hugging Face Transformers. After loading necessary models and classifiers, the script defines functions for sentiment analysis using TextBlob, VADER Sentiment, Flair, and Hugging Face Transformers. The DataFrame is then processed using a function named 'process_love_letters,' which iterates over each row, applying sentiment analysis and emotion classification functions to extract relevant features. The resulting DataFrame is enriched with additional columns such as 'TextBlob_Polarity,' 'TextBlob_Subjectivity,' 'VADER_Positive,' 'VADER_Neutral,' 'VADER_Negative,' 'VADER_Compound,' 'Flair_Sentiment,' 'Flair_Confidence,' and 'Hugging_Face_Sentiment.' These columns provide insights into the sentiment and emotion context and characteristics of each text segment.
	Mood_Tone_function	The mood_tone_function uses the Natural Language Toolkit (NLTK) library to identify emotions and keywords in the love letter "S" file. The script reads the CSV archive into a Pandas Data Frame, tokenizes each segment, and identify keywords related to mood_tone by matching pre-defined list. The "re.findall" method coupled with tokenization and case normalization to extract keywords. Added comments and keywords to the "Mood_Keywords" and "Tone_Keywords" fields in the updated DataFrame. The modified DataFrame is saved to a new CSV file.
	Location_Function	The location_function is simply responsible for extracting out the locations from the love letters in a CSV file by utilizing Named Entity Recognition (NER) available via spaCy library. This script loads a pre-trained English model, and a new lexicon called "custom_locations" with location-oriented phrases such as nations or landmarks. The function sets up a spaCy Matcher object to match proper nouns in the text, reads the input CSV file into a Pandas DataFrame, and processes each segment of text in an iterative manner for extracting locations. Locations are then located in combination with both the custom lexicon and spaCy model. Otherwise, the script uses your custom lexicon to find possible place references because state_location_places ngram (from a model) fails to recognize locations. With resulting DataFrame, new column "locations" is added, as output.

Semantic Analysis	Person_Characters	The character_relationship_function aims to locate people in a love letter script using named entity recognition (NER) and a proprietary lexicon of names. It uses a Pandas DataFrame to process each text segment and adds a new column called "Persons" to hold identified persons. The algorithm includes an 'indian_names_lexicon' with a carefully selected collection of Indian names. After detecting characters, the script categorizes them into positive and negative features, providing a detailed examination of romantic relationships throughout the story arc.
	Timeline_function	This function creates a time vocabulary, representing different times such as instant and intervals, other abstract concepts. Shown above, it illustrates that writing “romance” involves weaving time references through a broad range. The timeline algorithm then searches through scripts for time-related keywords by utilizing regular expressions, which reveals how and when time is mentioned in the story. This is a relatively high-level way of looking at the chronology. and, it turns out that this script also enables analysts, writers and fans alike to explore the temporal dynamics of love in any narrative material romantic comedies included. The analyze_timeline function categorizes sentences containing explicit time references to demonstrate when love stories take place.
	Actions_function	The action_function implementation aims to analyze nuanced romantic gestures within a script. It uses a large vocabulary of romantic behaviors, including subtle displays of love, affection, and close communication. The vocabulary is a powerful tool for literary analysis, indicating amorous intent or emotional connection depending on the context. The identify_actions function searches a given material using regular expressions to find instances of the specified acts. It filters sentences containing romantic activities by creating a regular expression pattern matching any of the listed actions. This feature allows for the identification and highlightment of romantic activities in a screenplay or other narrative text.
	Atmospheric_Functions	In the initial phase, the custom lexicon, referred to as "La", is intricately constructed to include a complex array of words including romance keywords as well as atmospheric keywords.. Based on this, the lexicon is further enriched by the addition of synonymous , leading to the creation of an expanded lexicon referred to as "Lae". This process represents a significant leap with the identification of atmospheric keywords embedded in the provided love letter script. By application of regular expressions and tokenization techniques this identification is done. Here, 'S' symbolises the script, serving as the input for the subsequent analysis. To enhance the lexicon, the process employs tools such as 'averaged_perceptron_tagger', 'vader_lexicon', and 'wordnet', contributing to the augmentation of the lexicon's richness. This synergistic approach, combining manual construction, synonym integration, and automated lexical expansion,

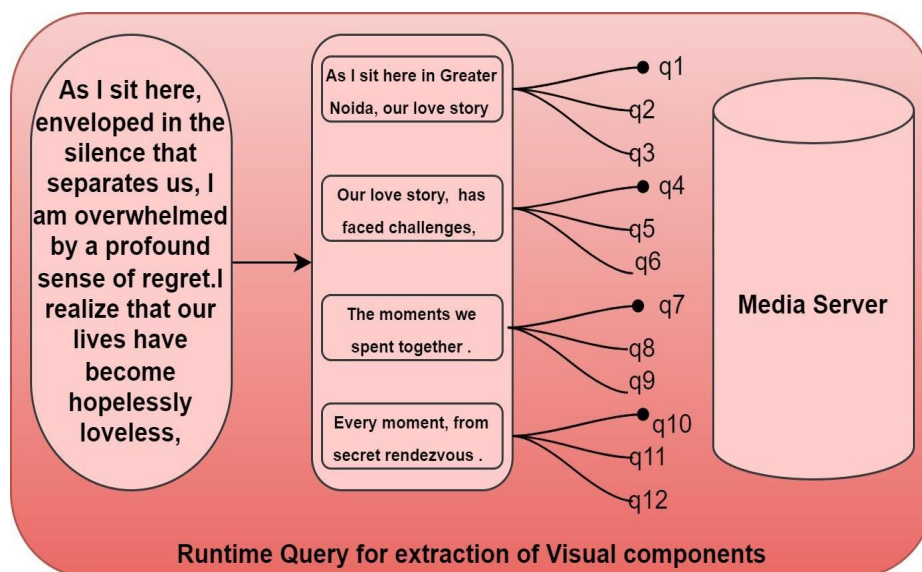
After the Collection of Hints and Topic inference from the Input script, extraction of relevant and corresponding media is required. The next section explains it.

#### 4.4. Dynamic Query Engine:

Dynamic query generation ensures adaptability to contexts (Semantic, Sentiment and Pragmatic) in the love letters. For each element sentiments, emotions, mood and tone, locations, persons involved atmospheric, actions and timeline words of the computation scene 'CSC' to be visualised as corresponding visual elements is searched in the media repository using dynamic query system . Following steps were followed [46].

- 1) **Data Preparation:** The dynamic query engine starts by loading the computational scene CSV file into a Dataframe using Pandas library in python. This step ensures that all the necessary information is structured and accessible.
- 2) **Validation:** It checks for the presence of required columns in the Dataframe. This validation step is crucial to avoid errors down the line and ensure that all necessary data points are available for generating search queries.
- 3) **Generate Query :** For each row in the Dataframe, the script constructs a search query. It combines sentiments and emotions with an "AND" logic, implying that both elements should be present in the search results. The rest of the attributes (locations, persons, and atmospheric words) are combined using an "OR" logic, broadening the search scope to include any of these elements. With the query constructed, the script interacts with Pixabay's API and other media server 'M' for fetching videos that match the criteria. It navigates through the database API's response, extracting video URLs and tags, which are then stored for further processing shown in Figure 2.

Figure2: Dynamic Query Generation



- 4) **Video Component Selection ('Vc'):** Once the component selection keywords are identified using a data-driven dynamic search query form the media server 'M' the next step is to download video assets. For each video, the algorithm initiates a download process, saving the video file to a specified cache directory of the system. Since, the internet is unpredictable, and downloads can fail for numerous reasons. The function is designed with robust error handling to manage such incidents gracefully, ensuring the process continues uninterrupted and logs any issues encountered. Throughout the download process, the algorithm provides real-time feedback, informing the user of each video's download status. This is done for monitoring progress and troubleshooting any issues that may arise. In the next section, we explain how the media resources are composed for making an automated cohesive video generation. In the end a meta-file is generated for each computation scene (CSC) that corresponds to visualisation of a specific love letter.

```

POSITIVE love room OR Man OR heart
POSITIVE joy India OR Man OR love
POSITIVE love Karachi OR Man OR embrace
POSITIVE joy Sipari OR Man OR connection
POSITIVE joy home OR Man OR love
POSITIVE joy room OR Man OR connection
POSITIVE love room OR Man OR calm
POSITIVE joy room OR Man OR love
POSITIVE anger room OR Man OR love
POSITIVE joy room OR Man OR beloved

Segment Sentiment Emotion \
0 As I sit here, gazing at the distance that sep... POSITIVE love
1 Our love story, forged amidst the complexities... POSITIVE joy
2 The moments we spent together in Karachi, in t... POSITIVE love
3 Every moment, from secret rendezvous to stolen... POSITIVE joy
4 Since our separation, solace is found in our c... POSITIVE joy
5 Education has enriched our understanding, teac... POSITIVE joy
6 Distance is an illusion we understand, knowing... POSITIVE love
7 As we navigate separation, let us hold onto th... POSITIVE joy
8 Though our countries may be divided, our love ... POSITIVE anger
9 My beloved, as we continue our journey, let ou... POSITIVE joy

Locations Persons \
0 room Man, Male , Boy, Female, Girl, Woman
1 India Man, Male , Boy, Female, Girl, Woman
2 Karachi Man, Male , Boy, Female, Girl, Woman
3 Sipari Man, Male , Boy, Female, Girl, Woman
4 home Man, Male , Boy, Female, Girl, Woman
5 room Man, Male , Boy, Female, Girl, Woman
6 room Man, Male , Boy, Female, Girl, Woman
7 room Man, Male , Boy, Female, Girl, Woman
8 room Man, Male , Boy, Female, Girl, Woman
9 room Man, Male , Boy, Female, Girl, Woman

Pixabay_URLs
0 [{'url': 'https://cdn.pixabay.com/video/449480...
1 [{'url': 'https://cdn.pixabay.com/video/449480...
2 [{'url': 'https://cdn.pixabay.com/video/449480...
3 [{'url': 'https://cdn.pixabay.com/video/147055...

```

Figure3: Output of Dynamic Queries

```

media_lib
├─ alone-46637.mp4?width=640&hash=587619fa5e99b2af7e65361cdc8b...
├─ blue-1700.mp4?width=640&hash=8433825989ed10b433013ff92a0995...
├─ boat-23838.mp4?width=1280&hash=811371144577e3c36a110b9fda7e...
├─ china-43238.mp4?width=640&hash=8db39b116f4464b7ff536c4128b46...
├─ clouds-146169.mp4?width=1280&hash=4b4f486ca9855c33ca9e08adf2...
├─ clouds-27197.mp4?width=640&hash=c84a671837080c563dbba175da8...
├─ couple-2301.mp4?width=640&hash=270bcdcf122946e541e82ec33068...
├─ cycling-39183.mp4?width=640&hash=5b1437f07aeee129a2ba12c996e...
├─ dog-15305.mp4?width=640&hash=6ea48be00ac930a73562f0208ae50...
├─ guitarist-1651.mp4?width=640&hash=ff2b20426f6586665a3c6b543f79...
├─ hands-552.mp4?width=480&hash=d55bd1d31053db4522e961ec49b12...
├─ hiking-109277.mp4?width=640&hash=0537c055edab49d0ab1d389027...
├─ keyboard-10822.mp4?width=640&hash=e8f6344499c3a6199312a77e3...
├─ love-32021.mp4?width=640&hash=c43fd367173dcb8ff380a73c517b1a...
├─ man-64729.mp4?width=640&hash=c536a1c1f7f16e70e0fcdf279722d6...
├─ meta_info.json
├─ nature-51142.mp4?width=640&hash=fa1621a208e829d6a86c3c72187...
├─ network-14900.mp4?width=1280&hash=5720aa9e12aa40b207555dd40...
├─ ocean-19609.mp4?width=1280&hash=b81a7f85a4ec1a8a03bbbf7b410...
├─ ocean-21175.mp4?width=1280&hash=fe9c59623ee58abb7dca5cd15b1...

```

Figure4: View Media Objects with Meta info

**5) Video Assembly:** The structure of a Video Assembly 'Va' is subject to a set of defined rules and constraints (derived from computation scene construction 'CSC') that govern its composition. These are outlined as follows:

- a) *Segment Order:* The 'Va' is composed of segments that are analogous corresponding to Sentences in love letters. The sequencing of these segments is crucial to the narrative flow of the 'Va', this forms a cohesive 'CSC'.
- b) *Entity Order within Segments:* Within each segment, entities are identified and ordered based on their discovery. These entities include but are not limited to:
  - Sentiment Polarity
  - Key Emotions
  - Location
  - Mood and Tone
  - Atmospheric Settings
  - Persons
  - Actions and Acts
  - Timeline
- c) *Introductory and Concluding Components:* Each 'Va' is book by an introductory (intro) and concluding (outro) component, referred to as 'Vc'.
- d) *Video Component Format:* Each 'Vc' within the 'Va' adheres to a short video format to maintain engagement and clarity. The introductory and the last video are made at runtime using fix format/layout images.
- e) *Duration Constraint:* The entire 'Va' is designed to not exceed a maximum duration of 60 seconds, ensuring conciseness and impact.
- f) *Emotional and Interpretative Clarity:* Each 'Vc' within the 'Va' is crafted to effectively convey the emotions and the intended meaning of the love letter segment it represents.
- g) The clips have attributes including title, description, media\_info, and assembly\_ duration and assembly fps. These are recorded in meta file.

The composition of a 'Va' is expressed as:

$$f(Va) = Vc1 \cup Vc2 \cup \dots \cup Vcn=10, \text{Where}$$

Vc1 = Intro

Vc6 =Atmospheric\_settings

Vc2 = Timeline                      Vc7 = Dominant\_emotion

Vc3 = Mood\_Tone      Vc8 = Key\_actions

Vc4 = Location                      Vc9 = Outro

Vc5 = Persons/Characters

In this formula, "U" symbolises the 'join' operation that combines the video components 'Vc' into a coherent assembly. Each 'Vc' is extracted and refined through a Dynamic Query Engine that operates in real-time. The construction of each query within this engine is informed by a Topic Inferences function, which employs a comprehensive analysis encompassing semantic, sentiment, and pragmatic facets. In essence, the 'Va' is a computationally structured narrative that is dynamically assembled to ensure an expressive and cohesive portrayal of the thematic elements derived from love letters, with a focus on brevity and emotional resonance. Hence, the outcome of this process shown in Figure5.

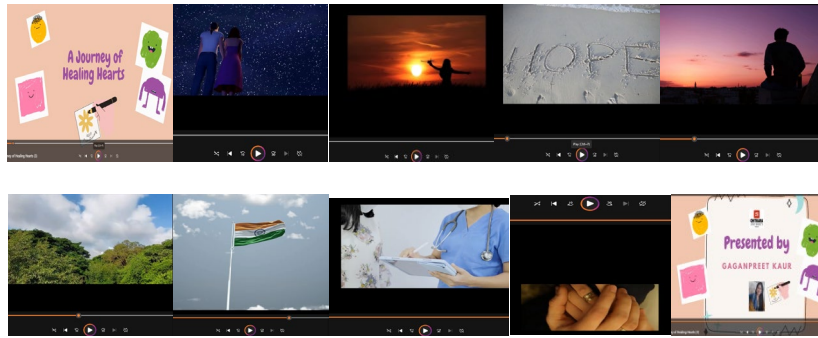


Figure5: Computational Scene Clip by Clip

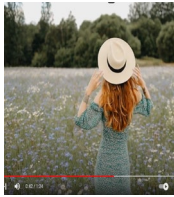

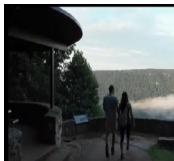
Within each segment of the love letter Script 'S', a multi-faceted analysis uncovers entities in a systematic manner that includes such as sentiment polarity, key motions, emotions, atmospheric settings, characters/persons, actions, mood and tone, and locations. This analytical process aims to capture the meaning of the original content, reflecting the intended message and emotional undertones of the love letters. A distinctive feature of 'Va' is the inclusion of an intro clips and an outro in each assembly, which are composed in a short video format to immediately engage and provide closure to the viewer within the stipulated maximum duration of 60 seconds. These constraints necessitate a concise yet expressive conveyance of the video's core message, emphasising the understanding of emotions and the meaning embedded in the love letters. The composition formula for a 'Va' is an ordered concatenation of its components (Vc1, Vc2, ..., Vcn), signifying the seamless integration of segments to form a unified and coherent narrative. The dynamic selection and ordering of these components are driven by a Dynamic Query Engine that operates at runtime. Queries within this engine are constructed using a Topic Inferences function, which employs semantic, sentiment, and pragmatic logic to draw relevant content for the assembly.



Table 4.Explanation of Video with YouTube Reference

<b>Id</b>	<b>URI</b>	<b>Snippets</b>	<b>Explanation</b>
1	<a href="#">Inspiring your girlfriend to work effectively in her first Job 1 (youtube.com)</a>		The video provides a supportive and motivational message to a loved one facing workplace challenges, emphasizing personal growth, productivity, and emotional support, with warm, comforting visuals.
2	<a href="#">Motivating your girlfriend in hard times (youtube.com)</a>		This video provides emotional support and understanding in the face of hardship and focuses on love, resilience and perseverance. It depicts desire and optimism, reflecting the theme of love and resilience in the face of setbacks and separation.
3	<a href="#">A Journey of Healing Hearts (youtube.com)</a>		The video tells the emotional journey of perseverance, love and overcoming adversity in Indian city, depicting themes of deep love, vulnerability, gratitude and healing.
4	<a href="#">Embracing the Light of Love (youtube.com)</a>		The video highlights a couple's perseverance and bond by following their journey through illness and recovery. An personal, contemplative atmosphere is created by the subdued images and cosy lighting, honouring their strong bond and optimism.
5	<a href="#">An Act of Selfless Love (youtube.com)</a>		The video explores the weight of self love, acceptance and it shows the thoughts and deep emotions of two people, transforming their relationship into something deeper.
6	<a href="#">Love Beyond Borders (youtube.com)</a>		The video depicts a couple's emotional bond despite personal challenges and distance. It highlights resiliency and hope through digital and symbolic connections.
7	<a href="#">Love Enduring Promise (youtube.com)</a>		The video explores the emotional journey of India-Pakistan's long-distance relationship, highlighting themes of positivity, longing and resilience, reflecting separation and hopeful



reunion.		
8	<a href="#">A Promise of Hope and Love (youtube.com)</a>	 <p>The video captures journey of 2 lovers filled with emotions far from each other by careers but still connected by a strong bond. The mood is bittersweet, theme revolves around the sacrifices made for success Visually, it may include soft, nostalgic colors, tender of hopeful atmosphere.</p>
9	<a href="#">Embracing the Simplicity of Love (youtube.com)</a>	 <p>The video soundtracks themes of love, separation and strength, creating a melancholic yet hopeful mood. It is a night that signals the end of their relationship based on individual career paths and yet celebrates love in its purest form.</p>
10	<a href="#">A Journey Back to Love 1 (youtube.com)</a>	 <p>Despite the stress and tension, the film shows the feelings and hopes of two exlovers to reunite against the backdrop of a wintery town, expressing the feelings of distance and sadness between them.</p>

## 5. Results and Discussions

After programmatically assembling individual media components into integrated video output, evaluating the quality of the resulting automated videos was imperative. For this purpose, author conducted a subjective assessment using a judge panel. Specifically, 10 videos were randomly sampled from the total generated collection and provided to 5 evaluators along with a formal parameter rubric as a guideline for video quality validation.

The panel was asked to analyse and rate each video's quality and relevance of selected clips to textual context, appropriate ordering and timing, and success conveying the emotional essence. The graders were also asked for overall impressions, for assessing the quality of videos.

### 5.1. Subjective Evaluation

By leveraging multiple expert judges with a defined evaluation criterion, this study aimed to determine baseline performance and obtain constructive qualitative feedback on the quality of the videos. Aggregate ratings across parameters and inter-rater agreement levels can indicate current effectiveness and improvements needed as this narrative-driven video generation approach advances. Here are the outcomes (ratings) given by the judges.

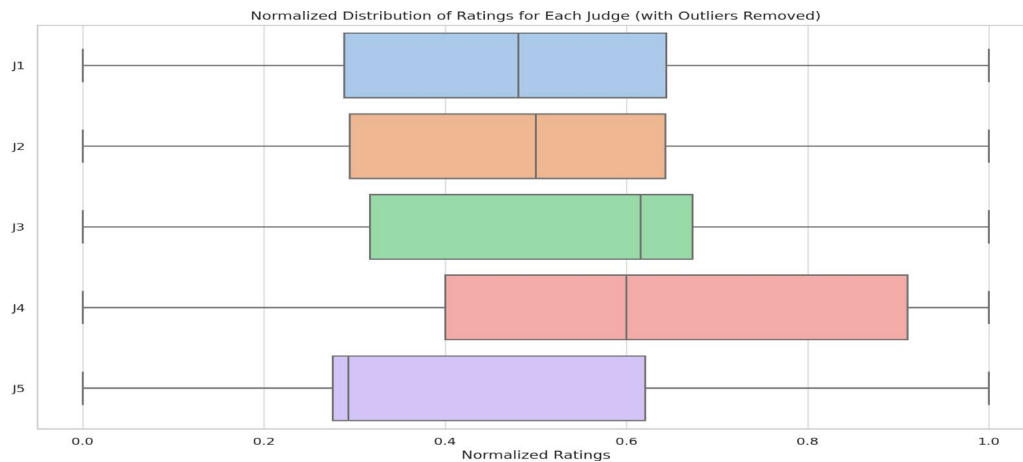


Figure 6: Distribution of Ratings of Judges

The boxplot [Figure 6] visualisation represents the distribution of each judge's ratings across all the evaluated videos. It shows the median (the line within the box), the interquartile range (the box itself), and the range excluding outliers (the "whiskers" extending from the box).

#### Inferences from the ratings of Judges:

- The average (mean) rating across all videos ranges from 3.54 to 3.98, with Judge 4 having the highest average rating.
- The standard deviation, which indicates the variability of the ratings, is relatively similar among the judges, ranging from approximately 0.82 to 0.98.
- The minimum ratings given by any judge to any video range from 2.2 to 2.5, with Judge 2 giving the lowest rating of 2.2.
- The maximum ratings are at or slightly above 5, with Judge 5 giving a rating of 5.2, which is actually above the scale's maximum of 5.
- The interquartiles range (IQR), represented by the difference between the 75th percentile and the 25th percentile, shows the spread of the middle 50% of the ratings. These also vary slightly but are within a comparable range.
- The range of ratings (max-min) for each judge shows how wide the ratings spread is, with all judges having a range of 2.5 or higher, indicating there is some diversity in how they rated the videos.

#### 5.2. Agreement between the Judges

When dealing with continuous data, one way to assess the agreement between judges is to use similarity or correlation metrics. One common approach is to use Pearson correlation for pairwise comparisons between judges. Pearson correlation measures the linear correlation between two sets of data and provides a value between -1 and 1, where 1 means total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. Another option could be to use Spearman's rank correlation, which assesses how well the relationship between two variables (x,y) can be described using a monotonic function, if the data doesn't meet the assumptions of Pearson's correlation. Hence, for this research work, we have computed [Figure 8] both and drawn inferences for making the evaluation process reliable.

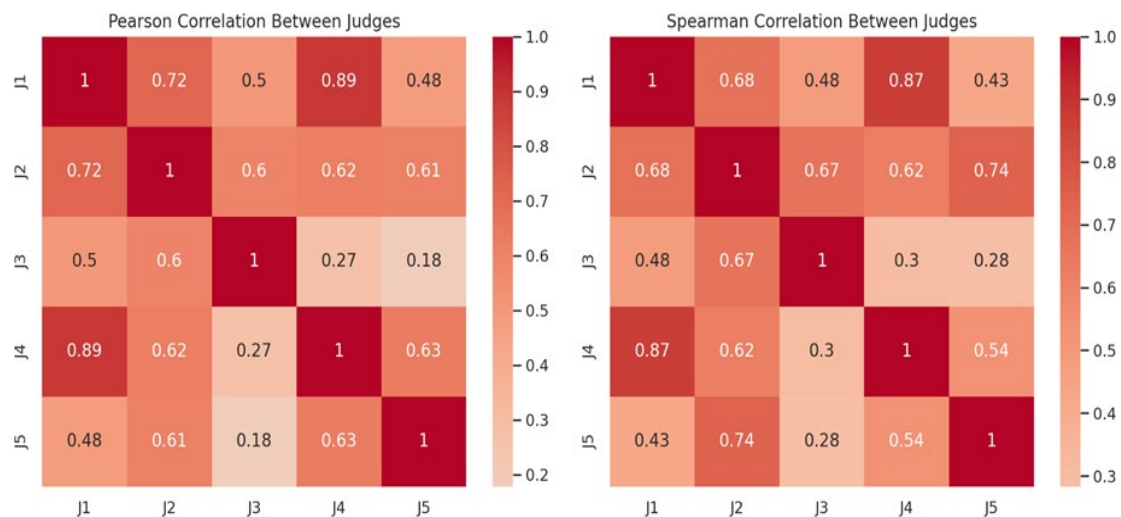


Figure 7: Agreement Correlation between Judges

**Inferences: Pearson Correlation**

- Judge 1 and Judge 4 show the highest Pearson correlation (0.888), indicating a strong positive linear relationship between their ratings. Suggesting that they independently agree with each other on the view of the quality of the videos they assessed.
- Judge 2 and Judge 5 also have a relatively high correlation (0.613), suggesting a moderate positive linear relationship. Implying they moderately agree with each other's assessment results.
- The lowest Pearson correlation is between Judge 3 and Judge 5 (0.178), which indicates a weak relationship. Implying that they think differently about the videos in terms of the standard 10 questions given to all the judges.

**Inferences: Spearman Correlation**

The heatmaps [Figure 7] visually represent these relationships. In both heatmaps, warmer colours (towards red) indicate stronger positive correlations, while cooler colours (towards blue) indicate weaker correlations.

- Judge 1 and Judge 4 again show the highest Spearman correlation (0.872), suggesting a strong monotonic relationship. This indicates that judges 1 and 4 have similar thoughts and agreement on the quality of the videos they are assessing.
- Judge 2 and Judge 5 also show a high Spearman correlation (0.740), indicating a strong monotonic relationship between them. In other words, both of them have similar views on the quality of the videos.
- The lowest Spearman correlation is between Judge 3 and Judge 5 (0.283), indicating a weak monotonic relationship. This implies that these judges think differently on perceived quality of the videos.

In nutshell, Judges 1 and 4 tend to rate videos more similarly to each other than to other judges, according to both correlation measures. Judges 3 and 5 tend to have the least agreement in their ratings. These results can help understand which judges have similar rating patterns and which do not, providing a form of agreement assessment.

**5.3. Evaluation of Quality of Videos**

The assessment of the video was done based on the following parameters suggested to judges.

Table 5: Questionnaire with Judges

For Judges : Ten Multiple Choice Questions for Content Analysis of Love Letter Videos : These questions are designed to assess how well the automatically generated videos capture the emotional essence and key themes of the original love letters, while also considering visual coherence and aesthetics.

Instructions: Please rate each video on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree) based on the following statements:

S.no	Questions
1	The overall mood of the video accurately reflects the emotions expressed in the love letter.
2	The visual elements like colours, and scene changes complement the tone and sentiment of the writing.
3	The scenes and elements chosen enhance the understanding and interpretation of the love letter.
4	The pace and flow of the video feel appropriate and match the emotional rhythm of the love letter.
5	The use of text overlays (if any) is well-integrated and adds value to the visual storytelling.
6	The video avoids clichés or generic imagery that detract from the uniqueness of the love letter.
7	The overall aesthetic of the video is pleasing and enhances the emotional impact of the love letter.
8	The video feels personal and conveys a sense of intimacy, mirroring the nature of a love letter.
9	The video successfully evokes an emotional response in the viewer that aligns with the intended feelings of the love letter.
10	The starting and ending of the video are appropriate.

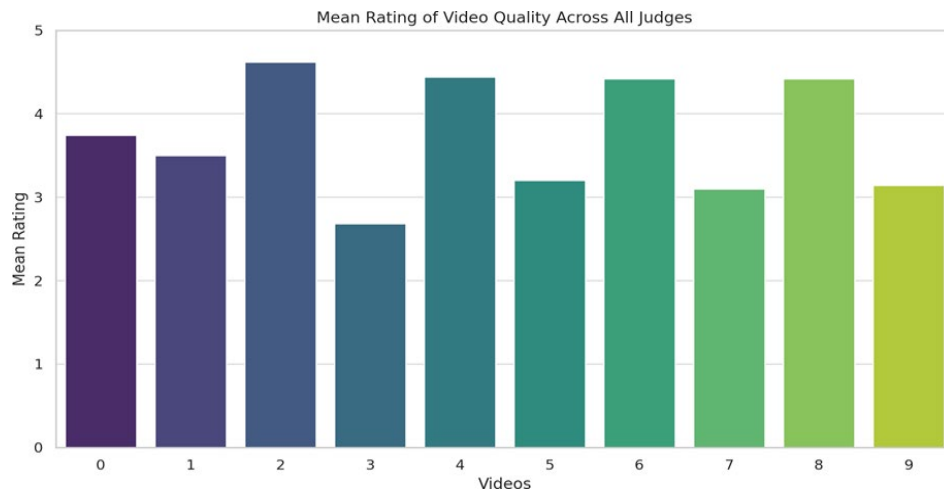


Figure 8. Average rating given by Judges

Based on the calculated mean ratings, we can draw several conclusions about the video quality as perceived by the judges:

- Videos [3](#), [5](#), [7](#), and [9](#) received the highest mean ratings, all above 4.4 on average, indicating that the judges were generally perceiving the videos as high quality. Videos 4 and 6 have the lowest mean ratings, at approximately 2.68 and 3.20 respectively, suggesting the judges were perceiving the videos as lower quality.
- Videos [1](#), [2](#), [8](#), and [10](#) have mean ratings that are close to the middle of the scale, indicating a more moderate or mixed perception of the videos quality.

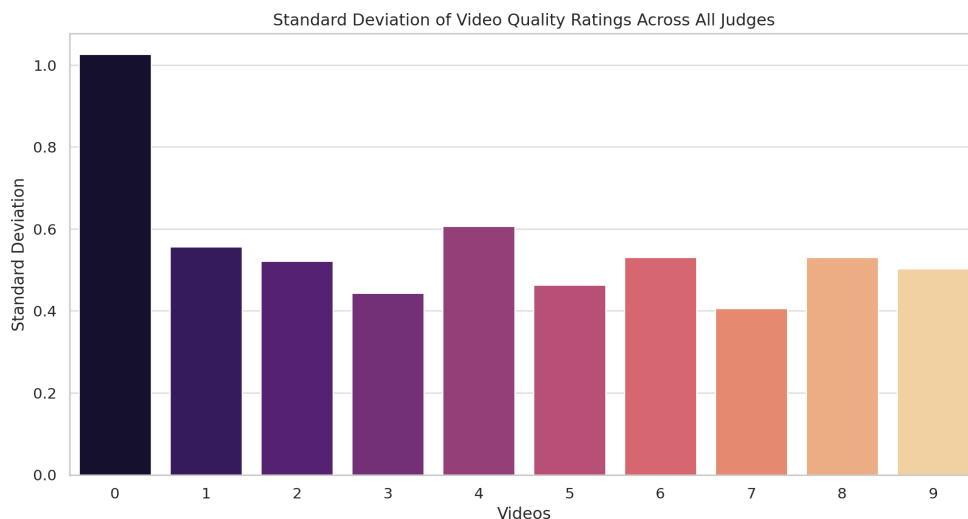


Figure 9. Standard Deviation of rating

The barplot [X] visualises the mean rating for each video, showing at a glance which videos were rated higher on average and might be considered of better quality.

The barplot [Y] visualises the standard deviation of the ratings for each video, which gives us an idea of the agreement between the judges' perceptions:

- Videos [4](#), [6](#), [8](#), and [10](#) have relatively low standard deviations, suggesting that the judges generally agreed about the quality of these videos.
- Video [1](#) has the highest standard deviation, indicating that there were less agreement among the judges

about its quality.

The Pearson and Spearman correlation coefficients indicated varying degrees of agreement between pairs of judges, with some showing strong positive correlations in their ratings (notably Judges 1 and 4), while others showed weaker relationships. This variability in correlation suggests that while there is some level of consensus and the same it can be observed that the judges also have individual criteria or thresholds for quality assessment. The descriptive statistics and visualisations of video quality provided further insights. The mean ratings showed which videos were generally perceived as higher quality, and the standard deviations gave an indication of how consistent those perceptions were among the judges. Videos with higher mean ratings and lower standard deviations were considered higher quality with greater consensus among judges. The final outcome is that the Videos 3, 5, 9, 7, and 1 are considered the top 5 in terms of quality based on the judges' ratings. The overall agreement between judges, as measured by correlation and standard deviation, was mixed, with some videos showing strong consensus and others more variability.

#### **5.4. Comparative Analysis and Validation**

In terms of methodological framework comparison it can clearly be observed that the paper has distinct approaches toward automatic video generation. Our paper introduces a novel, context-driven video generation framework that processes unstructured textual inputs (e.g., scripts, narratives). The framework consists of multiple stages, starting with the segmentation of the input script into meaningful units. It then applies sentiment, pragmatic, and semantic analysis to derive the entities, emotions, actions, and locations mentioned in the text. These extracted elements guide the dynamic querying and selection of media components, which are then assembled into a video composition. Technically, our work is designed to understand the context and emotions embedded in free-form text, allowing for a comprehensive reflection of the narrative in the generated video. It emphasizes a holistic approach by integrating different aspects of Natural Language Processing (NLP) to infer the mood, tone, and temporal aspects of the script [38]. Further it should be noticed that the proposed system generates dynamic queries to search and retrieve appropriate video clips and media components from a universal media set, ensuring that the composition is contextually relevant. In terms of evaluation of the work, the quality of the generated videos is assessed through human evaluation, focusing on narrative coherence and emotional accuracy, highlighting the ability of the framework to translate complex textual content into video form.

However, in the case of [55], the baseline paper presents an automatic video generation framework designed specifically for structured data storytelling. It uses a fact-driven approach, where structured data facts serve as the input to a clip selection algorithm. A predefined library of animated visualizations (clips) is mapped to various data facts. The core of the framework is an algorithm that selects the optimal clip for each data fact, arranges these clips into a coherent sequence, and optimizes the duration of the final video. Its prime focus is drawing on established visual storytelling principles. Its clip selection and arrangement algorithm aim to convey data insights effectively and linearly. In terms of content selection, it relies on a pre-built clip library where each data fact (e.g., value, proportion, and trend) is mapped to specific visual representations. The content selection process involves choosing the most suitable clip from this library for each data fact, guided by factors such as transition cost and visual consistency. Evaluation in this paper is done through user studies "subjective", assessing their comprehensibility, engagement, and quality compared to human-made videos. Table 6 gives a comparative view between these two research approaches.

Table 6.Comparative View of Different Approaches

Aspect	Proposed Method	Baseline Paper[53]	Validation Insight
Framework Approach	Context-driven video generation using NLP (Sentiment, Pragmatic, and Semantic Analysis).	Fact-driven video generation using a predefined clip library and selection algorithm.	Our Paper handles diverse narrative text with context-awareness; the baseline focuses on structured data storytelling.
Input Data Type	Unstructured text (narratives, scripts).	Structured data facts(value, proportion, trend).	Our research work has more flexibility to interpret varied narrative inputs, while the baseline is restricted to well-defined, structured data.
Content Selection	Dynamic media querying based on context and emotions extracted from text.	Clip selection from a pre-built library, guided by data facts and optimization criteria (e.g., transition cost).No audio or background audio for' narration.	Dynamic content selection allows personalized output in Our work, whereas the baseline ensures efficiency and consistency through a fixed library.
Narrative Understanding	Deep contextual and emotional analysis to capture mood, tone, and relationships for comprehensive narrative understanding.	Focused on visualizing data facts; narrative coherence is achieved through arrangement algorithms and transitions between visual clips.	Our research work enables deeper emotional and contextual storytelling; the baseline provides clarity and consistency in presenting factual data stories.



Video Composition Process	Video scenes are constructed dynamically from segmented script elements and aligned based on narrative flow and sentiment.	Algorithm selects and arranges optimal clips to create coherent data videos with minimized transition cost and maintained visual consistency.	Our research work offers more adaptable scene composition, allowing flexibility; baseline provides efficient sequencing for structured data with predefined templates.
Optimization & Output Quality	Evaluated by human judges for narrative coherence, fidelity, and emotional alignment using inter-rater agreement (intraclass correlation coefficient).	Evaluated through user studies focusing on comprehensibility, engagement, and comparison with manually composed videos.	For our work the evaluation emphasizes narrative fidelity and coherence; baseline focuses on viewer engagement and alignment with human-created content
Flexibility & Scalability	Adaptable to a wide range of narratives due to context-aware processing and dynamic querying.	Efficient for generating fact-based stories, limited by fixed visual templates, less adaptable to nuanced narratives.	Our research work can handle a broader range of content but may have computational overhead; the baseline is highly efficient but lacks content diversity.
Algorithm Complexity	High computational complexity due to multi-layered NLP analyses (sentiment, pragmatic, semantic) and dynamic query generation.	Moderate complexity involving clip selection optimization and transition cost minimization using a predefined set of visual templates.	Your approach is more computationally intensive due to complex analyses, while the baseline is streamlined for efficiency.

Content Adaptability	High adaptability, with the ability to interpret various narrative structures, moods, and emotions in free-form text.	Limited adaptability, tailored specifically for structured data visualization using predefined clips.	Your method is versatile for different storytelling contexts, while the baseline is tailored for structured data, restricting adaptability.
System Architecture	Designed with a dynamic query engine and NLP modules for sentiment, pragmatic, and semantic processing, allowing flexible media retrieval.	Built on a fact-driven clip library with an algorithm for optimal clip selection and arrangement. Uses JavaScript (D3.js) for video clip animation.	Your architecture is more modular and complex, enabling context-driven querying, while the baseline's architecture is simpler, focusing on clip arrangement from a fixed set.
Evaluation and Validation Methods	Uses human judges to assess video quality based on narrative coherence, emotional fidelity, and alignment with input script. Employs inter-rater agreement (intraclass correlation coefficient) to ensure objectivity.	Employs two user studies to evaluate videos for comprehensibility, engagement, and comparison with manually composed videos. Relies on qualitative user feedback to refine the output.	Your validation approach focuses on the narrative and emotional accuracy, while the baseline's method emphasizes user comprehension and engagement, highlighting different success metrics.

Limitations	Requires significant computational resources for complex NLP processing; variability in video quality due to dynamic content selection.	Restricted to predefined clip library and structured data, limiting depth in narrative and emotional representation.	Our research work 's adaptability comes with computational demands; baseline's efficiency is countered by lack of content personalization and context sensitivity.
-------------	---	--	--

From the recent works on diffusion models, it is amply clear that these models excel at synthesizing novel visuals from abstract prompts, our framework prioritizes contextual fidelity and efficient assembly of real-world media to mirror structured narratives. These approaches address fundamentally different problems: diffusion models "create" content, whereas our system "curates" it. Comparing them is akin to contrasting a documentary editor (proposed method) with a painter (diffusion models)—both valuable but serving distinct purposes.

Table 7. Explicit Comparison Between Proposed Framework and Diffusion Models

Aspect	Proposed Framework	Diffusion Models (e.g., Sora, Stable Video Diffusion)[64]	Distinction Rationale
Core Methodology	Template-based assembly: Dynamically queries and sequences pre-existing media clips using NLP-driven contextual analysis.	Generative synthesis: Creates novel visual content from noise via iterative denoising.	Assembly vs. generation: The former repurposes existing assets; the latter synthesizes new content.
Input Flexibility	Requires structured textual input (narratives) with explicit entities, emotions, and context.	Accepts free-form text prompts (e.g., "a couple dancing under cherry blossoms at dusk").	Context-driven assembly relies on structured narrative analysis; diffusion models prioritize open-ended creativity.
Media Source	Dependent on a predefined media database (e.g., Pixabay clips, images, sounds).	Generates entirely new visuals/textures not present in training data.	Proposed method is media-library-bound; diffusion models hallucinate novel pixels.

Output Type	Composite videos stitched from retrieved clips.	Coherent, synthesized videos with smooth transitions and novel scenes.	Assembly preserves real-world media consistency; diffusion enables fictional/unseen visuals.
Computational Demand	Low inference cost (querying and sequencing pre-rendered clips).	High computational cost (iterative denoising, GPU-heavy training).	Efficiency vs. creativity: Proposed method suits resource-constrained applications.
Contextual Fidelity	High alignment with input script (pragmatic/semantic analysis ensures narrative coherence).	Variable fidelity: May diverge from prompts due to stochastic generation.	Structured analysis ensures fidelity; diffusion models prioritize aesthetics over script adherence.
Creativity	Limited to media library diversity; no novel scene generation.	Unbounded creativity: Generates imaginative, non-existent scenes/objects.	Trade-off: Assembly ensures realism; diffusion enables artistic freedom.
Use Cases	Factual storytelling, educational content, love letter visualization.	Artistic/entertainment videos, abstract concept visualization, and prototyping.	Contextual accuracy vs. open-ended creativity.

The approach adopted here can maintain strict alignment with input scripts, leveraging pragmatic and semantic analysis to ensure narrative coherence, while diffusion models trade controllability for unbounded creativity. Future work could explore hybrid systems, combining template-based assembly for contextual accuracy with diffusion-generated fillers for missing media components. Since an apple can be compared with apples only, we have not done a comparative study regarding our approach and diffusion models. However, Table 7 gives an explicit contrasting view of these two different approaches. The next section finally gives the inferences and conclusions drawn from this study.

## 6. Conclusions & Future Directions

The research developed an Automated Video Composition (AVC) system to visualize love letters using media elements like gifs, video snippets, and images obtained through APIs and manual annotation, and 300 preprocessed love letters for a real-world use case.

Text preprocessing removes special characters, stop words, and grammar checks for clarity. Refined text is segmented into tables for named entity recognition, enabling dynamic queries to retrieve media elements and narrative context for structured computational scenes.

The solution uses algorithms to create logical scene assembly, utilizing media datasets, text analytics, query generation, and structural computation techniques to maximize video cohesion, and integrates FFmpeg library for programmatic video construction.

Five judges evaluated the AVG output quality, showing promise for improving analytics, video composition,

and alignment. Future work should focus on incorporating contextual voice-overs into the AVC framework to advance computational reproduction of coherent video narratives.

The integration of pragmatic analysis into video composition is a technological advancement, ensuring that videos not only reflect the literal content of the text but also encapsulate its implied meanings and emotional undertones. This novel integration enhances the coherence, narrative flow, and emotional impact of the generated videos, offering a more sophisticated and human-like video creation process.

### References

- [1] Juraska, J., Bowden, K., & Walker, M. (2019). ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*. Retrieved December 16, 2023, from <http://dx.doi.org/10.18653/v1/w19-8623>
- [2] Mazaheri, A., & Shah, M. (2022). Video Generation from Text Employing Latent Path Construction for Temporal Modeling. In *2022 26th International Conference on Pattern Recognition (ICPR)*. Retrieved December 16, 2023, from <http://dx.doi.org/10.1109/icpr56361.2022.9956706>
- [3] Hu, Y., Luo, C., & Chen, Z. (2022). Make It Move: Controllable Image-to-Video Generation with Text Descriptions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Retrieved December 17, 2023, from <http://dx.doi.org/10.1109/cvpr52688.2022.01768>
- [4] Li, X., Yang, J., & Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). *Neurocomputing*, 452, 675–689. <https://doi.org/10.1016/J.NEUCOM.2020.07.139>
- [5] Mortha, M. D. D., Maddala, S., & Raju, V. (2021). Data Preprocessing for Learning, Analyzing and Detecting Scene Text Video based on Rotational Gradient. In *International Conference on Data Science, E-learning and Information Systems 2021*. Retrieved December 17, 2023, from <http://dx.doi.org/10.1145/3460620.3460621>
- [6] Lienhart, R., & Effelsberg, W. (2000). Automatic text segmentation and text recognition for video indexing. *Multimedia Systems*, 8(1), 69–81. <https://doi.org/10.1007/s005300050006>
- [7] Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1), 541–547. <https://doi.org/10.1609/aaai.v27i1.8679>
- [8] Marqués, F., Pardàs, M., & Salembier, P. (1996). Coding-Oriented Segmentation of Video sequences. In *Video Coding*. Boston, MA: Springer US. Retrieved December 17, 2023, from [http://dx.doi.org/10.1007/978-1-4613-1337-3\\_3](http://dx.doi.org/10.1007/978-1-4613-1337-3_3)
- [9] Churchill, R., & Singh, L. (2021). textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data. In *Proceedings of the 10th International Conference on Data Science, Technology and Applications*. Retrieved December 21, 2023, from <http://dx.doi.org/10.5220/00105590006000070>
- [10] Li, Y., et al. (2023). Automatic Context Pattern Generation for Entity Set Expansion. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12458–12469. <https://doi.org/10.1109/tkde.2023.3275211>
- [11] Rahman, N., & Borah, B. (2017). Context Sensitive Query Correction Method for Query-Based Text Summarization. In *Computational Science and Its Applications – ICCSA 2017*. Cham: Springer International Publishing, 17–30. Retrieved December 21, 2023, from [http://dx.doi.org/10.1007/978-3-319-62407-5\\_2](http://dx.doi.org/10.1007/978-3-319-62407-5_2)
- [12] Pu, T. (2023). Video Scene Graph Generation with Spatial-Temporal Knowledge. In *Proceedings of the 31st ACM International Conference on Multimedia*. Retrieved December 22, 2023, from <http://dx.doi.org/10.1145/3581783.3613433>
- [13] Shi, D., Sun, F., Xu, X., Lan, X., Gotz, D., & Cao, N. (2021). AutoClips: An Automatic Approach to Video Generation from Data Facts. *Computer Graphics Forum*, 40(3), 495–505. <https://doi.org/10.1111/cgf.14324>
- [14] S, R. (2023). Text-to-Image Generation using Generative AI. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 07(08). <https://doi.org/10.55041/ijssrem25320>

- [15] Raja, S., S, M., & J, P. (2023). Text to Video Generation using Deep Learning. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM). Retrieved December 22, 2023, from <http://dx.doi.org/10.1109/iconstem56934.2023.10142725>
- [16] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet*, 15(8), 260. <https://doi.org/10.3390/fi15080260>
- [17] Kulkarni, A., Shivananda, A., Kulkarni, A., & Gudivada, D. (2023). Diffusion Model and Generative AI for Images. In *Applied Generative AI for Beginners*. Berkeley, CA: Apress, 155–177. Retrieved December 22, 2023, from [http://dx.doi.org/10.1007/978-1-4842-9994-4\\_8](http://dx.doi.org/10.1007/978-1-4842-9994-4_8)
- [18] Duan, J., Zhao, H., Zhou, Q., Qiu, M., & Liu, M. (2020). A Study of Pre-trained Language Models in Natural Language Processing. In 2020 IEEE International Conference on Smart Cloud (SmartCloud). Retrieved December 22, 2023, from <http://dx.doi.org/10.1109/smartcloud49737.2020.00030>
- [19] Sanjeeva, P., Reddy, V. B. N., Goud, J. I., Prasad, A. G., & Pathani, A. (2023). TEXT2AV – Automated Text to Audio and Video Conversion. *E3S Web of Conferences*, 430, 01027. <https://doi.org/10.1051/e3sconf/202343001027>
- [20] Khalilian, M., Ehsaei, M., & Fard, S. T. (2019). Recommendation algorithms for unstructured big data such as text, audio, image and video. In *Big Data Recommender Systems - Volume 1: Algorithms, Architectures, Big Data, Security and Trust*. Institution of Engineering and Technology, 133–168. Retrieved December 22, 2023, from [http://dx.doi.org/10.1049/pbpc035f\\_ch7](http://dx.doi.org/10.1049/pbpc035f_ch7)
- [21] Özköse, Y. E., Gökçe, Z., & Duygulu, P. (2023). Alignment of Image-Text and Video-Text Datasets. In 2023 31st Signal Processing and Communications Applications Conference (SIU). Retrieved December 22, 2023, from <http://dx.doi.org/10.1109/siu59756.2023.10224043>
- [22] Galitsky, B. (2022). Improving open domain content generation by text mining and alignment. In *Artificial Intelligence for Healthcare Applications and Management*. Elsevier, 489–521. Retrieved December 22, 2023, from <http://dx.doi.org/10.1016/b978-0-12-824521-7.00011-9>
- [23] Kuo, C. (2023). *The Handbook of NLP with Gensim: Leverage topic modelling to uncover hidden patterns, themes, and valuable insights within textual data*. Packt Publishing Ltd.
- [24] Kulkarni, A., Shivananda, A., & Kulkarni, A. (2022). *Natural Language Processing Projects: Build Next-Generation NLP Applications Using AI Techniques*.
- [25] Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1), 33–64.
- [26] Ren, X., et al. (2017). CoType. In *Proceedings of the 26th International Conference on World Wide Web*. Retrieved December 27, 2023, from <http://dx.doi.org/10.1145/3038912.3052708>
- [27] Wang, R., Hou, F., Cahan, S., Chen, L., Jia, X., & Ji, W. (2022). Fine-Grained Entity Typing with a Type Taxonomy: a Systematic Review. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/tkde.2022.3148980>
- [28] Nair, I., et al. (2023). A Neural CRF-based Hierarchical Approach for Linear Text Segmentation. In *Findings of the Association for Computational Linguistics: EACL 2023*. Retrieved February 07, 2024, from <http://dx.doi.org/10.18653/v1/2023.findings-eacl.65>
- [29] Chien, J.-T., & Chueh, C.-H. (2012). Topic-Based Hierarchical Segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 55–66. <https://doi.org/10.1109/tasl.2011.2143405>
- [30] Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29, 21–43.
- [31] Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2022). Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*, 13(1), 342.

- [32] Brito, G., Mombach, T., & Valente, M. T. (2019, February). Migrating to GraphQL: A practical assessment. In 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER) (pp. 140-150). IEEE.
- [33] Hua, X.-S., Lu, L., & Zhang, H.-J. (2004). Automatic music video generation based on temporal pattern analysis. In Proceedings of the 12th annual ACM international conference on Multimedia. <http://dx.doi.org/10.1145/1027527.1027641>
- [34] West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of personality*, 59(3), 609-662.
- [35] Ren, W., Singh, S., Singh, M., & Zhu, Y. S. (2009). State-of-the-art on spatio-temporal information-based video retrieval. *Pattern recognition*, 42(2), 267-282.
- [36] Khurana, K., & Deshpande, U. (2021). Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: a comprehensive survey. *IEEE Access*, 9, 43799-43823.
- [37] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., & Trancoso, I. (2011). Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8), 1163-1177.
- [38] Pu, T., Chen, T., Wu, H., Lu, Y., & Lin, L. (2023). Spatial-temporal knowledge-embedded transformer for video scene graph generation. *IEEE Transactions on Image Processing*.
- [39] Callemein, T., Roussel, T., Diba, A., De Feyter, F., Boes, W., Van Eycken, L., ... & Goedeme, T. (2021). Show me where the action is! Automatic capturing and timeline generation for reality TV. *Multimedia Tools and Applications*, 80, 383-408.
- [40] Lindley, C. A. (2001). A Video Annotation Methodology for Interactive Video Sequence Generation. In *Digital Content Creation*. London: Springer London, 163-183. [http://dx.doi.org/10.1007/978-1-4471-0293-9\\_13](http://dx.doi.org/10.1007/978-1-4471-0293-9_13)
- [41] Chen, H., He, K., Liang, P., & Li, R. (2010). Text-based requirements preprocessing using natural language processing techniques. In 2010 International Conference On Computer Design and Applications. <http://dx.doi.org/10.1109/icdda.2010.5540935>
- [42] Huge Collection of Famous Celebrity Romantic Love Letters (theromantic.com) 15 Famous Love Letters - Love Notes Written By Celebrities and Historical Figures. (countryliving.com)
- [43] Mashtalir, S., & Mashtalir, V. (2020). Spatio-temporal video segmentation. *Advances in Spatio-Temporal Segmentation of Visual Data*, 161-210.
- [44] Lim, C. G., Jeong, Y. S., & Choi, H. J. (2019). Survey of temporal information extraction. *Journal of Information Processing Systems*, 15(4), 931-956. Top of Form
- [45] Kumar, R. R., Kumar, K., Jain, A. K., Sharma, V., Sharma, N., & Jain, N. (2023, October). QVD-Querying Video Databases for Event Related Frames Using Text Keywords. In 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 1060-1064). IEEE.
- [46] Kaushal, R. K., & Panda, S. N. (2016). Effective Teaching Methods and Proposed Web Libraries for Designing Animated Course Content: A Review. *International Journal of Advanced Computer Science and Applications*, 7(2).
- [47] [https:// Love Letter a Day \(quora.com\)](https://LoveLetteraDay.quora.com)
- [48] Kaur, G., Kaur, A., Khurana, M., & Damaševičius, R. (2024). Sentiment polarity analysis of love letters: Evaluation of TextBlob, Vader, Flair, and Hugging Face transformer. *Computer Science and Information Systems*, (00), 40-40.
- [49] Kaur, A., & Khurana, M. (2024, March). Multimodal Sentiments: Unraveling Text and Emoji Dynamics Through Deep Learning. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-6). IEEE.
- [50] Pratibha, Kaur, A., Khurana, M., & Damaševičius, R. (2024). Multimodal hinglish tweet dataset for deep pragmatic analysis. *Data*, 9(2), 38.



- [51] Kaur, G., Kaur, A., & Khurana, M. (2023, October). Exploring the Role of Mathematical Modelling in Automatic Scene Generation amidst Rapid Technological Advances. In 2023 4th International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 391-397). IEEE.
- [52] Kaur, G., Kaur, A., & Khurana, M. (2024, March). A Survey of Computational Techniques for Automated Video Creation and their Evaluation. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-6). IEEE.
- [53] gagan, gaganpreet (2024), "Script Dataset for Computational Scene Generation", Mendeley Data, V3, doi: 10.17632/rd5bjbnm35.3
- [54] [5 million+ Stunning Free Images to Use Anywhere - Pixabay](#)
- [55] Shi, D., Sun, F., Xu, X., Lan, X., Gotz, D., & Cao, N. (2021, June). Autoclips: An automatic approach to video generation from data facts. In Computer Graphics Forum (Vol. 40, No. 3, pp. 495-505). A2, etc.
- [56] Li, Z., Li, Z., Zhang, J., Feng, Y., & Zhou, J. (2021). Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2476-2483.
- [57] Hong, X., Sayeed, A., Mehra, K., Demberg, V., & Schiele, B. (2023). Visual writing prompts: Character-grounded story generation with curated image sequences. Transactions of the Association for Computational Linguistics, 11, 565-581.
- [58] Solanki, S. R., & Khublani, D. K. (2024). From Script to Screen: Unveiling Text-to-Video Generation. In Generative Artificial Intelligence: Exploring the Power and Potential of Generative AI (pp. 81-112). Berkeley, CA: Apress.
- [59] Kamatala, S., Jonnalagadda, A. K., & Naayini, P. (2025). Transformers Beyond NLP: Expanding Horizons in Machine Learning. *Yes it was accepted by IRE Journals*, 8(7).
- [60] Priyadarsini, N. I., Yenumula, S., & Reddy, K. V. (2025). MCQ Generation using NLP Techniques. In *ITM Web of Conferences* (Vol. 74, p. 01016). EDP Sciences.
- [61] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). Imagen Video: High Definition Video Generation with Diffusion Models. ArXiv. <https://arxiv.org/abs/2210.02303>
- [62] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S. and Kreis, K., 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 22563-22575).
- [63] Wang, X., Chan, K. C., Yu, K., Dong, C., & Change Loy, C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 0-0).
- [64] Sun, R., Zhang, Y., Shah, T., Sun, J., Zhang, S., Li, W., ... & Wei, B. From Sora What We Can See: A Survey of Text-to-Video Generation.