

REINFORCED ATTENTION FOR VIDEO SUMMARISATION

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

SIMRAN RAY

(23/AFI/07)

Under the supervision of

Prof. ANIL SINGH PARIHAR



Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042, India

JUNE, 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Simran Ray**, Roll No – **23/AFI/07** student of M.Tech - AFI, hereby declare that the project Dissertation titled “**Reinforced Attention for Video Summarisation**” which is submitted by us to the Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Simran Ray

Date: 27.05.2025

(23/AFI/07)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Thesis titled “**Reinforced Attention for Video Summarisation**” which is submitted by **Simran Ray**, Roll No – **23/AFI/07**, **Computer Science and Engineering**, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Prof. Anil Singh Parihar

Date: 27.05.2025

SUPERVISOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express heartfelt gratitude to **Prof. Anil Singh Parihar** for his unwavering guidance, mentorship, and encouragement throughout the course of this project. His expertise and insights were invaluable in shaping my understanding of the problem and in navigating the challenges we encountered. Prof. Parihar's support extended far beyond technical advice — he helped us appreciate the industrial relevance of this work, guided us with a clear vision, and inspired us to aim higher. His constant motivation, constructive feedback, and availability to clarify doubts at every stage were instrumental in the successful completion of this project.

I would also like to extend our sincere thanks to **Dr. Kavinder Singh** for his valuable inputs and encouragement during this project. His insightful suggestions helped us refine our approach, and his expertise provided a broader perspective on the challenges of video summarization. I am grateful for his time, patience, and willingness to share his knowledge, which greatly enriched the learning experience.

Place: Delhi

Simran Ray

Date: 27.05.2025

(23/AFI/07)

Abstract

Video summarization is a critical task for enabling efficient browsing, retrieval, and storage of large-scale video content by generating concise yet informative summaries. In this paper, we propose the Global and Local Attention-based Video Summarization Network (GLASN), a novel framework that combines global and local attention mechanisms with positional encoding to model both long-range dependencies and local temporal dynamics within video sequences. By leveraging the combination attention framework, GLASN selectively focuses on semantically important frames while maintaining the global context necessary for coherent summaries. We formulate video summarization as a sequential decision-making problem and adopt a reinforcement learning (RL) framework, optimizing GLASN with reward functions that promote both diversity and representativeness—key factors for high-quality summaries. Importantly, our approach is fully unsupervised, eliminating the need for labor-intensive, human-annotated labels, which is crucial for scalability in real-world applications where annotating large volumes of data is infeasible. Extensive experiments on benchmark datasets demonstrate that GLASN effectively captures the essence of video content and outperforms or competes with state-of-the-art methods, showcasing the benefits of attention-based architectures and unsupervised RL training for video summarization.

Contents

| | |
|--|-----------|
| Candidate's Declaration | i |
| Certificate | ii |
| Acknowledgement | iii |
| Abstract | iv |
| Content | v |
| List of Tables | vi |
| List of Figures | vii |
| List of Symbols, Abbreviations | viii |
| 1 INTRODUCTION | 1 |
| 1.1 Video Summarisation | 1 |
| 1.2 Reinforcement Learning for Video Summarization | 2 |
| 1.3 Attention Mechanism for Video Summarization | 3 |
| 2 LITERATURE REVIEW | 5 |
| 3 METHODOLOGY | 9 |
| 3.1 Input | 9 |
| 3.2 Attention Mechanism | 10 |
| 3.3 Reward function for Reinforcement Learning | 15 |
| 3.4 Training | 16 |
| 3.5 Regularisation | 18 |
| 3.6 Optimisation | 19 |
| 3.7 Summary Generation | 19 |
| 4 EXPERIMENTS AND RESULTS | 21 |
| 4.1 Datasets | 21 |
| 4.2 Evaluation Metrics | 21 |
| 4.3 Implementation Details | 22 |
| 4.4 Performance Comparisons | 23 |
| 4.5 Ablation Study | 23 |
| 5 CONCLUSION AND FUTURE SCOPE | 26 |
| A Plagiarism Verification | 29 |

List of Tables

| | | |
|-----|--|----|
| I | Performance comparison (F-score %) on SumMe and TVSum datasets . . . | 22 |
| II | Ablation study showing the variation in performance (F-score %) of our model on SumMe and TVSum with different numbers of video segments and data fusion strategies. | 24 |
| III | Ablation study on the performance (F-score %) of our model on SumMe and TVSum using different numbers of attention heads in the global and local attention mechanisms. | 24 |
| IV | Component-wise ablation study of GLASN model. | 24 |

List of Figures

| | | |
|---|---|----|
| 1 | Multi-head attention mechanism. | 11 |
| 2 | Regressor Network Architecture. It processes attention-encoded features and outputs frame-level importance scores, indicating the probability of each frame being included in the final video summary. | 12 |
| 3 | Training Global and Local Attention-based Summarisation Network (GLASN) via Reinforcement Learning (RL). GLASN takes frame-level feature representation of a video, passed through a CNN model (GoogLeNet) and generates a set of actions for each frame determining which ones of them are included in the video summary. The agent of our RL learns with the help of our Reward Function. | 13 |
| 1 | Plot of rewards while training video-1 | 25 |
| 2 | Ground Truth and importance scores (probabilities) of video-4 predicted by GLASN | 25 |

List of Symbols

| | |
|------------------|---|
| A | Attention |
| Q | Query, input to attention head |
| K | Key, input to attention head |
| V | Vector, input to attention head |
| S | Softmax function |
| d | is scale factor, to scale the dot product. |
| F | Frames. |
| PE | Positional Encoding |
| MA | Multi-head Attention |
| p_t | Probability of each frames to be included in summary |
| a_t | action, to decide whether a frame should be selected or not |
| V | Video Summary |
| R_{div} | Diversity Reward |
| R_{rep} | Representativeness Reward |
| g_t | contextualized feature vector |
| θ | policy parameters |
| J | average reward over the action sequence |
| π_θ | policy network |
| R(S) | Reward Function |
| ∇_θ | gradient of $J\theta$ |
| $L_{percentage}$ | Regularisation term controlling percentage of frames selected |
| $L_{penalty}$ | Regularisation term penalising heavy weight of the network |
| β_1 | hyperparameters balancing $L_{percentage}$ |
| β_2 | hyperparameters balancing $L_{penalty}$ |

Chapter 1

INTRODUCTION

1.1 Video Summarisation

With the explosion of video data across social media, surveillance systems, and entertainment platforms, efficient video content management has become a pressing challenge. Every minute, thousands of hours of video are uploaded online, creating a deluge of content that is neither scalable nor practical for manual browsing or processing. In this context, video summarization (VS) has emerged as a crucial task — aiming to generate short, representative summaries that capture the essential content of videos without losing semantic meaning.

Video summarization refers to the task of creating a shorter version of a video that preserves the most important and informative content. Summaries can be extractive (selecting original frames or shots) or abstractive (generating new content), with extractive summaries being more common due to their ease of implementation and lower risk of semantic distortion.

A well-generated video summary offers enormous practical benefits — from facilitating quick content retrieval and indexing to enabling surveillance monitoring, news clipping, sports highlights generation, and assisting accessibility in education or healthcare [1]. However, summarizing videos is inherently challenging due to the complex, redundant, and dynamic nature of video data, which is further compounded by its temporal dimension and the subjective nature of "importance" in different contexts.

Early approaches relied on hand-crafted features like color histograms, motion vectors, and shot boundaries to identify significant frames or segments [1]. While simple, these methods struggled with generalization and semantic understanding. With the advent of machine learning, researchers began exploring supervised models trained on human-annotated datasets to predict frame importance scores.

However, supervised methods such as Gygli et al.'s approach — which proposed using user preferences for summary generation — were limited by the subjective nature of annotations. The inherent diversity in human perception made it challenging to define a "ground truth" summary. Moreover, collecting large-scale labeled datasets for video summarization proved costly and time-consuming.

Traditionally, video summarization methods have relied on supervised or unsupervised

learning techniques. Supervised approaches leverage human-labeled data to learn frame-level importance, but they suffer from significant annotation overhead and subjectivity [1]. Unsupervised methods reduce reliance on labels but struggle to capture the rich semantics and diversity required for human-aligned summaries [2]. Moreover, most classical models fail to model long-range temporal dependencies and semantic relationships across video segments, which are critical for understanding narrative flow and context.

The introduction of deep learning models brought a paradigm shift. Models like adversarial LSTM networks [3] enabled unsupervised summarization by optimizing adversarial objectives to generate summaries indistinguishable from human-created ones. Similarly, Panda and Roy-Chowdhury [4] proposed collaborative summarization techniques that leverage topic-related videos to improve summary quality. Hsaio et al. [5] introduced HSA-RNN, which dynamically adapts hierarchical temporal structures for more efficient video summarization.

A significant milestone in the field came with attention-based methods, which have shown exceptional ability to model spatio-temporal dynamics and semantic relationships. Gao et al. [6] proposed a method that combines global and local attention mechanisms with positional encoding to enhance the temporal structure and semantic fidelity of video summaries. Their model not only captures frame-wise relevance but also encodes the relative position of frames, thereby improving coherence and reducing abrupt transitions between selected segments. This dual-attention mechanism allows for a more context-aware and structurally consistent summary — critical for maintaining narrative flow in long or complex videos.

Recent techniques such as the Spatiotemporal Vision Transformer (STVT) [7] leverage attention mechanisms to further enhance summarization quality by modeling inter-frame and intra-frame relationships. Together, the use of these attention-based frameworks have significantly improved the semantic and structural quality of video summaries, making them more consistent with human expectations.

1.2 Reinforcement Learning for Video Summarization

Reinforcement learning (RL) seems like a good alternative to the usual supervised learning, mainly because it lets models figure out how to summarize through trial and error using a reward signal. In video summarization, this means the model can try out lots of different summaries and keep tweaking its strategy to do better based on a reward that measures stuff like how diverse or representative the summary is, and how short it can be too.

Zhou et al. [1] introduced a deep reinforcement learning framework for unsupervised video summarization that maximized a diversity-representativeness (DR) reward. This reward function encourages the selected frames to be both representative of the video’s content and diverse from each other. The model uses a RL agent to select frames based on video features and optimises the summary generation using policy gradient methods.

This helps us not need for expensive manual annotations.

Other reinforcement-based approaches have tried out different ways of designing the reward function. Like, Mahasseni et al. [3] came up with an adversarial RL setup — basically, one part of the model (the generator) picks key frames, and another part (the discriminator) checks if the summary looks real or not by comparing it with the actual one. This kind of training helps make the summaries look more natural and flow better. RL also gives the model some flexibility — it can learn how to handle videos of different lengths, types of content, and even scenes that are more complex.

Moreover, reinforcement learning uses exploration, which allows the model to consider non-greedy selections and long-term rewards. This is particularly useful in the context of video summarization, as local frame importance might not always lead to a globally pre-size summary. By optimizing the selection of frame sequences, rather than just focusing on individual frame scores, RL encourages the model to think more in terms of semantic flow and overall information content.

The flexibility of reinforcement learning also makes it possible to integrate additional modules, such as attention mechanisms and temporal encoders. These components further enhance the model’s ability to focus on important segments and maintain the continuity of the storyline. Overall, RL-based approaches have continued to advance the field of automatic video summarization, offering both theoretical soundness and practical effectiveness.

1.3 Attention Mechanism for Video Summarization

The rise of attention mechanisms has significantly transformed the field of computer vision, enabling models to dynamically focus on the most informative parts of an image or video sequence. At its core, attention allows a model to weigh different input regions differently, emulating how humans naturally prioritize certain visual cues over others. This becomes crucial in tasks involving large amounts of spatio-temporal data, such as video summarization, where not all frames or regions contribute equally to the underlying semantics.

The seminal work on the Vision Transformer (ViT) by Dosovitskiy et al. [8] demonstrated that pure attention-based architectures could rival and even surpass convolutional neural networks (CNNs) on large-scale image classification tasks. Unlike CNNs that rely on localized receptive fields, ViT processes images by dividing them into patches and feeding them as sequences into transformer layers. These layers compute pairwise attention between all patches, enabling the model to capture long-range dependencies and holistic scene understanding. This global receptive field of attention mechanisms provides a significant advantage, especially when capturing relationships between distant but semantically related regions within an image.

Extending this concept to video data, the Spatiotemporal Vision Transformer (STVT) proposed by Hsu et al. [7] introduces attention mechanisms capable of handling both spatial and temporal dimensions. STVT integrates inter-frame (temporal) and intra-

frame (spatial) attention, allowing the model to effectively reason about how objects and scenes evolve over time while simultaneously focusing on salient spatial regions within each frame. This dual attention structure is particularly well-suited for video summarization tasks, where selecting temporally spread-out but contextually relevant segments is key to generating informative summaries.

Gao et al. [6] took a complementary approach by combining global and local attention in a more structured manner and incorporating positional encoding directly into the attention process. Their framework captures global temporal patterns while still attending to local variations within scenes. Notably, the positional encoding allows the model to maintain awareness of the chronological order of frames, thus avoiding disjointed or contextually inconsistent summaries. This is particularly important in long videos where temporal coherence is crucial for narrative fidelity. Gao et al.’s hybrid attention model effectively enhances both representativeness and diversity by ensuring the selection of semantically rich yet temporally distributed frames.

In the context of video summarization, attention mechanisms help the model selectively focus on frames or shots that contribute the most to the story flow or the overall meaning of the video. Traditional approaches often rely on handcrafted features or simple uniform sampling, which can miss subtle but important segments. Attention-based models, on the other hand, learn to prioritize frames based on their context and relevance, which makes the generated summaries more aligned with how humans perceive and remember important events in a video.

Moreover, attention mechanisms offer interpretability, as they provide insights into which frames the model considered most relevant. This transparency furthers in understanding the decision-making process of the summarization model. Fajtl et al. [9] emphasized this aspect in their work, demonstrating how attention visualizations can provide actionable explanations for summary decisions.

GLASN integrates lightweight attention mechanisms in a reinforcement learning environment. This design allows the model to both focus selectively on semantically rich content and optimize summary quality based on diversity and representativeness rewards. Specifically, from PGL-sum by Gao et al. [6] we ensure that both local semantic structure and global temporal structure are taken into consideration for in the summary generation process.

Chapter 2

LITERATURE REVIEW

Deep learning has profoundly transformed the field of video summarization by enabling data-driven learning paradigms, in contrast to earlier rule-based or heuristic-driven methods. At its core, video summarization aims to extract the most informative and representative content from videos, and deep models excel in learning such abstract representations from large-scale data.

One of the earliest and most influential approaches in deep video summarization uses Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to capture temporal dependencies across video frames. LSTM-based models treat video summarization as a sequential prediction problem, allowing the model to learn when to include a frame in the summary based on past information. Zhang et al. [2] proposed a bidirectional LSTM (BiLSTM) model, which processes video sequences both forward and backward to better understand temporal dependencies. This design helps the model maintain contextual awareness across the entire video.

To handle the hierarchical structure of videos, Hsiao et al. [5] introduced the Hierarchical Structure-Adaptive RNN (HSA-RNN), which uses a two-level LSTM architecture. The first level identifies shots, and the second level determines the importance of those shots, effectively modeling the multi-scale structure of videos and reducing redundancy at both frame and segment levels.

Building on this, Mahasseni et al. [3] introduced the Deep Summarization Network (DSN), an unsupervised adversarial model that uses an LSTM summarizer and a discriminator. The summarizer generates keyframe selections, while the discriminator tries to distinguish between generated and ground-truth summaries. The model is trained with a diversity-representativeness reward, ensuring the summaries are both varied and informative.

Similarly, Yoon et al. [10] propose an unsupervised RL model with temporal consistency and interpolation: they combine a transformer-CNN backbone with new consistency rewards, achieving state-of-the-art results on SumMe and TVSum. These methods eschew hand-crafted rewards in favor of learned reconstruction or consistency signals. For example, Wang et al. (2022) incorporate an auxiliary summarization loss to capture long-term dependencies, showing that this loss “significantly improve[s] the performance” of LSTM-based summarizers on SumMe/TVSum. Afzal and Tahir [11] similarly enhanced deep RL by combining ResNet-152 features with a GRU-based policy, reporting improved F1 on SumMe over the baseline DR-DSN. In summary, these unsupervised RL models demon-

strate that learned rewards—whether via reconstruction, diversity/representativeness, or feature consistency—can substantially boost summary quality on standard benchmarks. [12].

Despite their strengths, LSTM-based models have limitations, particularly in modeling long-range dependencies and handling complex relationships over extended sequences. Attention mechanisms address these shortcomings by allowing models to selectively focus on key parts of a video without being constrained by sequence length.

Hsu et al. [7] proposed the Spatiotemporal Vision Transformer (STVT), which leverages transformer-based attention to capture both spatial and temporal patterns. Unlike RNNs, transformers rely entirely on self-attention mechanisms and are particularly effective in modeling long-range dependencies. Inspired by the success of the Vision Transformer (ViT) in image classification [8], many subsequent video summarization models have adopted transformer-based architectures. This STVT model achieves state-of-the-art performance on SumMe/TVSum by jointly modeling both temporal dependencies and spatial context.

Transformer and attention mechanisms have been adapted for video summarization with impressive results. For example, Hsu [7] introduce a Frame Index Vision Transformer (FIVT) that treats frame segments as “words” and adds explicit index and class embeddings. This purely transformer-based, supervised model outperforms previous RNN/CNN methods on SumMe and TVSum

Attention-based methods have also branched into handling multiple annotations or modalities. Terbouche propose MAAM [13], a probabilistic attention model that aggregates multiple human summaries via an EM framework. MAAM embeds frames with a Vision Transformer and uses an attention network to predict importance scores; the learned “average-annotation” attention significantly outperforms single-annotation models on SumMe and TVSum.

Other recent works leverage multimodal inputs. Zhu [14] introduce a topic-aware summarization task: they build the TopicSum dataset (with video, audio, and textual annotations) and design a multimodal transformer that fuses vision, audio, and text to generate multiple topic-specific summaries.

Gao et al. [6] further refined the transformer architecture for video summarization by proposing a model that integrates global and local attention mechanisms with learnable positional encoding. Their model captures coarse global structure as well as fine-grained temporal cues, enabling a balanced summary that preserves narrative flow and detail. Notably, their use of learnable positional encoding enhances the model’s ability to reason about frame order and importance, which is critical in video summarization.

Traditional supervised learning approaches require large amounts of annotated data, which is expensive and time-consuming to collect. Reinforcement Learning (RL) has emerged as a compelling alternative, particularly for unsupervised or weakly supervised video summarization.

Zhou et al. [1] introduced a reinforcement learning-based framework that formulates video summarization as a Markov Decision Process (MDP). In this setup, an agent se-

lects keyframes to maximize a diversity-representativeness reward. This approach circumvents the need for human-labeled data while still learning effective summarization policies.

Similarly, Mahasseni et al. [3] utilized adversarial reinforcement learning to refine summary quality, combining the benefits of generative modeling with reward optimization. Their summarizer-discriminator framework helps the model align with human-like summarization without relying on explicit supervision.

Beyond deep learning, early video summarization relied on heuristic methods. Gygli et al. [15] developed one of the first formalized frameworks that scores frames based on handcrafted features such as motion, aesthetics, and uniqueness. Though simplistic by modern standards, these methods established foundational principles around representativeness and diversity.

Hybrid approaches that combine deep learning with traditional methods have also shown promise. Panda and Roy-Chowdhury [4] introduced a collaborative summarization framework that merges deep visual features with high-level semantic metadata (e.g., tags, text). This multi-source approach improves topic relevance and cross-video summarization.

Potapov et al. [16] contributed a category-specific video summarization approach that utilizes domain knowledge to tailor summaries based on the semantic content of the video. This method leverages classifiers trained for different video categories, enabling the generation of summaries that are not only visually concise but also contextually relevant. Their approach was among the first to demonstrate the benefits of aligning summarization strategies with content types, thus enhancing user relevance and interpretability. Building upon the idea of contextual awareness, Zhu et al. [17] proposed DSNet, a detect-to-summarize framework that incorporates object detection features from pre-trained detectors such as Faster R-CNN. DSNet effectively identifies salient objects and events across frames, and integrates them into a summarization pipeline, ensuring that important visual cues are preserved while maintaining compactness and diversity. This approach bridges high-level object understanding with frame-level importance prediction, enhancing the semantic fidelity of generated summaries.

Fajtl et al. [9] introduced a dynamic attention-based encoder-decoder model, where frame-level importance scores are computed using temporal attention over the entire sequence. This allows the model to capture global temporal relationships and dynamically adjust attention based on evolving context within the video, leading to more informative and coherent summaries. Their model demonstrated that learning attention weights directly from data outperforms traditional fixed-window or heuristic approaches in temporal summarization. Lastly, Metelli et al. [15] addressed the adaptability of summarization systems by exploring configurable environments in reinforcement learning. Their work suggests frameworks where summarization agents can be customized or fine-tuned based on user preferences, domains, or task constraints. This direction points toward more interactive and personalized summarization systems, where the summarization process itself can adaptively learn from feedback or shifting objectives.

In conclusion, deep learning has propelled video summarization by introducing models that can learn both low-level visual features and high-level semantic patterns. LSTM-

based methods initially demonstrated success in modeling sequential dependencies. The emergence of attention mechanisms and transformer-based models significantly improved the ability to capture global context and long-range dependencies. Reinforcement learning, especially in unsupervised settings, further expands the scalability and adaptability of these models. Going forward, hybrid models that integrate transformers, RL, and multimodal learning—incorporating audio, text, and metadata—are likely to shape the future of video summarization.

Chapter 3

METHODOLOGY

3.1 Input

Video summarization relies on capturing and processing the most salient visual information from video frames. Instead of feeding raw frames directly to the model, we use deep feature extraction to transform each frame into a compact yet meaningful representation. Convolutional Neural Networks (CNNs) have been widely adopted for image and video analysis due to their ability to learn hierarchical feature representations. These networks operate by applying a series of convolutional filters to an image, progressively learning low-level features such as edges and textures, and high-level features such as object categories, spatial arrangements, and scene context. CNN-based feature extraction has been instrumental in tasks like image classification, object detection, and action recognition in videos [mettl]. For this purpose, we leverage GoogLeNet, a deep convolutional neural network (CNN) pretrained on the ImageNet dataset [18]. We take the penultimate layer of the network as the feature input for our model.

Among various CNN architectures, GoogLeNet (also known as Inception v1) stands out due to its efficiency in extracting meaningful features while maintaining a reduced computational footprint compared to deeper networks like ResNet. GoogLeNet, introduced by Szegedy et al. [18], incorporates the Inception module, which enables multi-scale feature extraction within a single layer. The Inception module processes input at multiple scales simultaneously, allowing the model to capture both fine-grained details and high-level scene representations. This is especially beneficial for video summarization, where diverse visual patterns need to be analyzed. Unlike VGG-16 or ResNet-101, GoogLeNet achieves high accuracy with fewer parameters, making it suitable for large-scale video datasets. The final layer of GoogLeNet provides an abstract, semantically rich feature representation of each frame, preserving key spatial relationships while discarding unnecessary details. For our summarization framework, GoogLeNet strikes a balance between accuracy, efficiency, and generalization, making it well-suited for extracting meaningful frame-level features.

Feature extraction using CNNs like GoogLeNet plays a crucial role in video summarization by Reducing Dimensionality. Instead of raw pixel data (millions per frame), a compact feature representation (e.g., 1024-dimensional vector) preserves relevant content while making processing efficient. This also helps in enhancing temporal understanding, since High-level visual embeddings enable the model to focus on meaningful transitions, avoiding redundant frames. The extracted feature vectors serve as input for attention

mechanisms, ensuring that the most important frames are identified effectively. It has significantly reduces the time complexity for our model since we do not have to deal with huge amounts of raw pixel data anymore.

3.2 Attention Mechanism

The attention mechanism has revolutionized sequential data processing by enabling models to focus on the most relevant parts of the input while maintaining a global context. In video summarization, this capability is essential for identifying and emphasizing key frames or segments that capture both spatial and temporal dynamics. Unlike traditional recurrent models like LSTMs, which process frames sequentially and struggle with long-range dependencies, attention-based architectures—particularly Transformers—can model relationships among all frames directly. Self-attention computes pairwise affinities between frames, assigning importance scores that highlight salient moments and facilitate the creation of concise, informative summaries.

At the core of our model we leverage stacked self-attention [19] and feed-forward layers to learn inter-frame relationships effectively. Each video frame is embedded and processed through this architecture, allowing the model to capture both local nuances and global context simultaneously. This approach not only enhances computational efficiency but also ensures the resulting summaries are cohesive and meaningful. By focusing on visually and semantically important frames, the attention mechanism helps distill long videos into engaging summaries that retain the essence of the original content, filtering out redundancy and emphasizing critical events and transitions.

Our architecture utilises a multi-head attention mechanism and a fully connected regressor network [9] . These layers use multi-head self attention to identify both long-term connections as well as short-term relationships between frames, helping the model to pay “attention” on the most important parts of our input. The feedforward neural network (FFN) further transform the attention-weighted inputs, along with dropout regularization and ReLU activation for improved generalization and non-linear feature learning. Our setup, is configured with a total of 8 attention heads and a hidden dimension of 512, making it capable of capturing diverse relational patterns among the frames.

Self-attention computes a representation of a sequence by relating different positions within it. In the context of video summarization, this mechanism allows the model to evaluate the contextual importance of each frame by comparing it to every other frame, thereby capturing long-range temporal dependencies effectively. The self-attention projects this input into three learned matrices - Queries Q, Keys K and Vectors V. The scaled dot-product attention is defined as [19] :

$$A(Q, K, V) = S \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (1)$$

Where Q , K and V are the “query, key, and value matrices” respectively, obtained by combining the input feature with learnable weight matrices vectors, d is scale factor of K , used to scale the dot product. Intuitively, each frame’s query vector compares (dot-products) against the key vectors of all frames; after softmax normalization, these attention weights multiply the value vectors to produce a context-aware output for each frame. S is the softmax function which normalizes the attention weights so that they sum to 1, allowing the model to assign importance to each frame according to its relevance. This mechanism allows every frame in a video sequence to attend to every other frame, capturing global dependencies irrespective of their temporal distance.

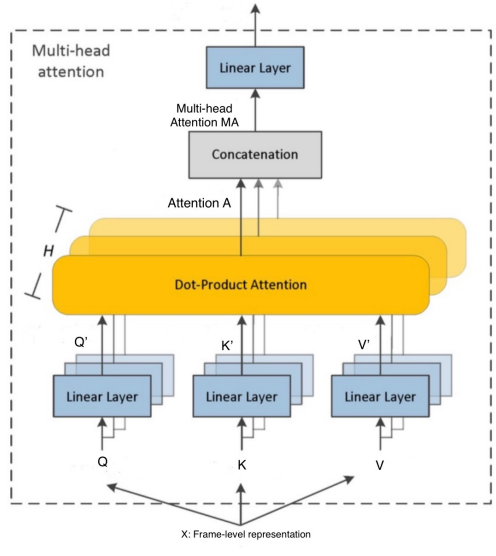


Figure 1: Multi-head attention mechanism.

While dot-product attention captures relationships between frames, it lacks inherent positional awareness. To address this, positional encoding is incorporated. The classical approach, as introduced in "Attention is All You Need" [19], uses sine and cosine functions of varying frequencies:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (3)$$

where pos denotes the position in the sequence, and i denotes the dimension index. This results in an $F \times D$ matrix, where F is the number of frames (or tokens) and D is the feature dimension.

For our task of video summarisation, Instead of applying positional encoding over the feature dimensions, we compute a $F \times F$ positional encoding matrix, explicitly modeling

positional relationships between frames in the sequence [9] .

This matrix $PE_{F \times F}$, is added directly to the dot-product term of Query and Key before the softmax operation. It is given as

$$A(Q, K, V) = S \left(\frac{QK^\top}{\sqrt{d}} + PE_{F \times F} \right) V \quad (4)$$

This influences how each frame attends to every other frame, explicitly making the attention scores aware of positional relationships, hence more intuitive for video, as it models inter-frame temporal dependencies directly at the attention score level. This modification enhances the model’s ability to capture inter-frame temporal dependencies by influencing the attention scores based on relative frame positions, making it more suitable for tasks like video summarization.

To further refine the attention mechanism, [6] introduces a combination of global and local attention. For computing local attention, we need to perform an additional step of data segmentation, which splits the frame feature vectors into P non-overlapping segments. Each of these segments is then individually forwarded to a local multi-head attention block that focuses on the corresponding locality of the video within each segment. This step also helps us have low computational complexity since it allows it to apply a dimensionality reduction.

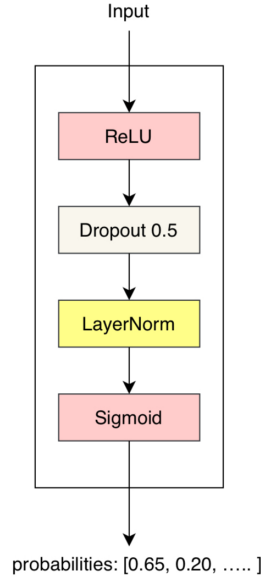


Figure 2: Regressor Network Architecture. It processes attention-encoded features and outputs frame-level importance scores, indicating the probability of each frame being included in the final video summary.

The global attention captures dependencies across the entire sequence, while local attention focuses on frame relationships within segmented intervals. The outputs of global and local attention are fused using strategies such as addition, multiplication, averaging, or

maximum selection. For the purpose of efficient back propagation, a residual connection is added to this result. In the model, we use multi-head attention: multiple such attentions run in parallel with different learned linear projections, here multiple such attentions are computed in parallel to learn different subspace relationships between video frames [7].

For multi-head attention, the model computes multiple sets of Q, K, V and combines the results as follows:

$$\text{MA}(Q, K, V) = \text{Concat}(h_1, h_2, \dots, h_h)W^O \quad (5)$$

Here each head $h_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ and W_i^Q, W_i^K and W_i^V are learnable projection matrices. The heads are concatenated and linearly transformed to yield the final output. This allows the model to capture diverse patterns of similarity (e.g. attending to motion vs. color) simultaneously.

In video summarization, modeling temporal dependencies is critical since the importance of a given frame or scene often depends on events that occurred earlier or later in the video. Traditional models like LSTMs and GRUs tend to struggle with such long-range dependencies due to issues like gradient vanishing and the inherently sequential nature of their computation [2].

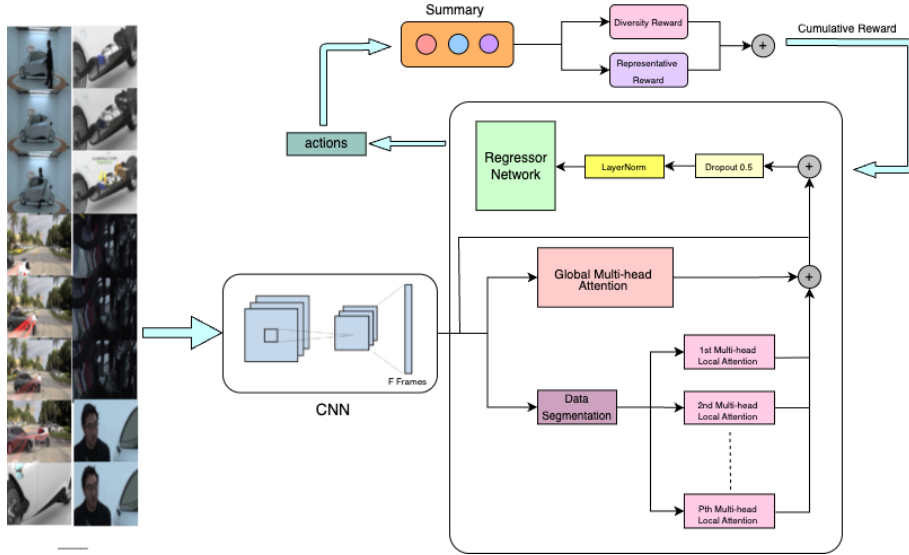


Figure 3: Training Global and Local Attention-based Summarisation Network (GLASN) via Reinforcement Learning (RL). GLASN takes frame-level feature representation of a video, passed through a CNN model (GoogLeNet) and generates a set of actions for each frame determining which ones of them are included in the video summary. The agent of our RL learns with the help of our Reward Function.

Following the attention modules, the model uses a regressor network [9] to assign a continuous importance score to each frame which acts as the probability for the particular frame to be included in the summary. This regressor is composed of a feedforward layer

with a sigmoid activation function, enabling output values in the range $[0, 1]$, suitable for interpreting as probabilities in the subsequent Bernoulli sampling step. The simplicity and differentiability of this regressor structure make it especially useful for training with backpropagation in an unsupervised setting. Its lightweight design and compatibility with reinforcement learning or sampling-based optimization methods make it ideal for integrating into transformer-style or attention-based summarization models.

The final stage of GLASN includes a probabilistic decision-making mechanism that controls which video frames we select for the summary and which ones we exclude. Instead of using a fixed, deterministic policy, the model uses a probabilistic one, which gives it more flexibility to try out different frame selections during training. This kind of exploration is really helpful in reinforcement learning (RL) because it lets the model evaluate and improve different summarization strategies based on the reward feedback. Using a probabilistic policy encourages the model to create summaries that are not only short but also diverse and representative of the original video. This approach is inspired by earlier work in unsupervised video summarization with deep RL, where diversity and representativeness were included explicitly in the reward function [1]. It is given as:

$$p_t = \text{sigmoid}(Wh_t) \quad (6)$$

To decide whether a frame should be selected or not, we employ Bernoulli sampling, where the action $a_t \in \{0, 1\}$ is drawn from a Bernoulli distribution parameterized by the frame’s selection probability p_t , given as:

$$a_t \sim \text{Bernoulli}(p_t) \quad (7)$$

This stochastic decision process offers several advantages in an RL-based video summarization setting:

- **Exploration–Exploitation Trade-off:** By sampling actions according to p_t , the agent naturally balances exploration of new summary compositions with exploitation of high-confidence selections. Unlike deterministic thresholding, which can become trapped in suboptimal patterns, Bernoulli sampling allows the model to occasionally select low-probability frames, potentially discovering richer or more representative summaries during training.
- **Policy Gradient Compatibility:** Bernoulli sampling fits seamlessly into policy gradient algorithms. Gradients of the expected reward with respect to model parameters can thus be estimated directly via the score-function (log-probability) trick, enabling end-to-end optimization without requiring a differentiable approximation of the discrete decision.
- **Discrete Action Semantics:** Video summarization fundamentally involves a binary choice—whether to include a frame or not. Bernoulli sampling provides a principled probabilistic model for such binary actions, preserving the discrete nature of the task while still allowing smooth gradient-based learning.

To summarise, stochastic sampling process introduces exploration into the learning process, which is critical when applying reinforcement learning to video summarization. Unlike deterministic selection (e.g., greedy or threshold-based methods), the Bernoulli-based action sampling enables the model to explore different frame combinations, promoting diverse and representative summaries over time [1].

If $a_t = 1$ the frame is included in the summary; otherwise, it is skipped. The composed video summary is given as:

$$Summary = \{V_{y_i} | a_{y_i} = 1, i = 1, 2, 3, \dots\} \quad (8)$$

Several prior works, have adopted similar probabilistic policies in the context of reinforcement learning for video summarization. Like Yao et al. [20] proposed using stochastic frame selection combined with adversarial training to encourage realism in the generated summaries. Sampling allows the agent to explore new frame combinations, avoiding local optima and leads to diverse summary candidates across training episodes, improving generalization.

3.3 Reward function for Reinforcement Learning

The strength of RL in video summarization lies in its ability to optimize non-differentiable evaluation metrics such as diversity and representativeness, which are crucial for high-quality summaries but are not easily captured in conventional loss functions. Zhou et al. [1] introduced a pioneering unsupervised framework that utilizes a diversity-representativeness reward to guide the agent in generating balanced summaries without any labeled ground truth. Similarly, Mahasseni et al. [3] proposed an adversarial reinforcement learning model where a summarizer network is trained alongside a discriminator to distinguish generated summaries from real ones, pushing the model to produce more human-like outputs. These strategies enable the agent to internalize the trade-off between retaining salient content and avoiding redundancy, resulting in summaries that are both concise and semantically rich. Crucially, the design of the reward function plays a pivotal role; it must encode domain-specific goals and human preferences effectively, as poorly constructed rewards can misguide the learning process and produce irrelevant summaries. By enabling sequential reasoning and optimizing task-specific criteria, RL offers a flexible and scalable approach for generating personalized and high-quality video summaries.

In our approach, we utilize the “Diversity-Representativeness Reward Function”, as used in DSN [1], to guide the training of our video summarization model. This reward function is such that it assess and enhance the effectiveness of the generated summaries returned with our model. A good video summary should not only capture the main content of the input video but also minimize redundancy by incorporating diverse and representative elements. By incorporating this reward function, our model learns to strike a balance between these two crucial aspects, leading to more meaningful and informative summaries.

The diversity reward motivates the model to choose frames that are varied from each other. This is done by assessing the difference between the chosen frames within our vector space. The rationale is simple: a summary that repeatedly highlights similar content

adds little value. By maximizing diversity, the model ensures that a wide range of unique events or scenes from the original video is captured, offering a thorough summary. The R_{div} is determined by calculating the average dissimilarity between each pair of chosen frames:

$$R_{\text{diversity}} = \frac{1}{|\gamma|(|\gamma| - 1)} \sum_{t_1 \in \gamma} \sum_{\substack{t_2 \in \gamma \\ t_1 \neq t_2}} \text{distance}(x_{t_1}, x_{t_2}) \quad (9)$$

Where distance computes the compliment of cosine similarity or the distance between two frames x_{t_1} and x_{t_2} .

The representativeness reward, focuses on ensuring that the selected frames effectively represent the entire video. This is modeled as a clustering problem, specifically the “k-medoids” problem [21]. Our objective here is to select all the frames which becomes the medoids that minimize the mean-square error between all video frames and the closest selected frames. In other words, the chosen frames serve as cluster centroids or medoids, representing the most significant features of a input. This ensures that the summary is not only diverse but also representative of the original video’s main themes and events. It is given as:

$$R_{\text{rep}} = \exp \left(\frac{1}{T} \sum_{t=1}^T \min \|\mathbf{x}_{t_1} - \mathbf{x}_{t_2}\|_2 \right) \quad (10)$$

We train our model by taking the cumulative reward:

$$R(S) = R_{div} + R_{rep} \quad (11)$$

During the training process, both the diversity reward R_{div} and R_{rep} are of comparable magnitude. This ensures either of the reward is not dominating the gradient calculations. In cases where no frames are selected, meaning the action sample consists solely of zeros, the network receives a reward of zero, effectively discouraging this behavior [1].

3.4 Training

We employ Reinforcement Learning (RL) with the REINFORCE algorithm for training our video summarization model, where the model learns a policy to select frames based on rewards derived from the Diversity-Representativeness Reward Function. The core of our approach is based on attention heads, which produce contextualized feature vectors m_t for each frame at time t . These vectors are used as the input to the policy function, which in turn predicts actions (i.e., the frames to be included in the summary).

The training process in RL revolves around maximizing the cumulative reward over a sequence of actions. For video summarization, the reward function is often designed to

evaluate the generated summary in terms of representativeness and diversity. The agent explores various frame selections to find the optimal combination that balances these two factors. This trial-and-error approach allows the model to learn strategies for summary generation that align closely with human preferences, even in the absence of explicit supervision. By framing video summarization as a sequence of decisions, RL allows our network to evaluate the effect of any frame selected to be included in video summary. This ability is important for generating concise and relevant summaries which preserve the core message of the original content without redundancy.

The REINFORCE algorithm is a foundational approach in reinforcement learning, often used for training models with policy-based methods. It is a Monte Carlo approach where the agent learns by sampling trajectories—sequences of actions and their associated rewards—and uses this data to refine its policy. The main concept is to modify the parameters of the policy such that probability of selecting frames that yield in higher rewards is increased [22]. Doing this ensures only frames that are most relevant is taken for summary. This is achieved by weighting the gradient of the log-probability of an action by the cumulative reward associated with that action. In simpler terms, the algorithm encourages actions that yield better outcomes while discouraging less favorable ones. Despite being straightforward, REINFORCE is powerful for scenarios like video summarization, where the reward signal is often sparse, and the model must consider the long-term impact of its decisions on the overall summary.

Use of attention in our network enables the agent, in context of RL to focus on the most relevant frames from a video, producing a detailed set of feature vectors for each frame. These feature vectors are essential for assessing the significance of frames in the video summary. Contextualized feature vectors are obtained from the outputs of the transformer module. For each time step t , contextualized feature vector g_t is given as the aggregate of outputs from each attention head:

$$\mathbf{g}_t = \text{Attention}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T) \quad (12)$$

The policy gradient method is employed to train our network. The idea is we maximize the expected reward, which is defined as the average reward over the action sequence:

$$J(\theta) = E_{\pi_{\theta}(a_{1:T})}[\text{Reward}(S)] \quad (13)$$

Here $\pi_{\theta}(a_{1:T})$ is the probability distribution over the sequence of actions $a_{1:T}$ predicted by the policy network π_{θ} and $R(S)$ is the reward computed using the Diversity-Representativeness Reward function [1].

For optimising the model, we compute the gradient of the $J\theta$ for the policy parameters θ . This gradient provides information on how to adjust the policy to maximize the expected reward. For the objective function gradient is computed as:

$$\nabla_{\theta} J(\theta) = E_{\pi_{\theta}(a_{1:T})} \left[R(S) \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(a_t | \mathbf{g}_t) \right] \quad (14)$$

Since calculating the expectation over complex action sequences is computationally expensive, we approximate the gradient using Monte Carlo sampling. We simulate N episodes of action sequences and then take the average gradient across these episodes. The gradient estimate becomes:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^T R_n \nabla_{\theta} \log \pi_{\theta}(a_t | \mathbf{g}_t) \quad (15)$$

The gradient estimate above can exhibit high variance, which can lead to unstable updates and slow convergence. To reduce variance, we subtract a baseline b from the reward, which helps to stabilize the gradient updates. This technique is known as baseline subtraction, and it has been shown to improve the performance of policy gradient methods by reducing the variance of the gradients. The modified final gradient update becomes:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^T (R_n - b) \nabla_{\theta} \log \pi_{\theta}(a_t | \mathbf{g}_t) \quad (16)$$

The agent’s objective is to optimize the reward function, which motivates the model to choose frames that would have both diversity as well as representativeness w.r.t the input video, resulting in a high-quality video summary.

3.5 Regularisation

Since the goal of a RL model is to maximise the reward, it may have a tendency to select more number of frames which eventually leads to higher reward score. To prevent this a regularization term inspired by Mahasseni, Lam, and Todorovic 2017 [3] . This term penalizes the probability distributions produced by the Attention based Video Summarization Network (AVSN). This ensures that the selected frame percentage remains within a reasonable range. It is given as:

$$L_{\text{percentage}} = \left\| \frac{1}{T} \sum_{t=1}^T p_t - \epsilon \right\|^2 \quad (17)$$

Where ϵ is a hyperparameter that controls the intended proportion of frames to be included in the summary. By minimizing this term, we effectively constrain the number of frames that can be selected, ensuring that the summaries produced are concise and not overly long.

Additionally, to avoid overfitting and enhance the our network’s ability to generalization capacity, we apply an L_2 regularization term on the parameters θ f the network. This penalizes excessively large values of the weights, helping the model avoid memorization and ensuring it learns more robust features. The L_2 regularization term is given as:

$$L_{\text{penalty}} = \sum_{i,j} \theta_{i,j}^2 \quad (18)$$

Together, these regularization techniques help balance the frame selection process and keep the model from overfitting, leading to better performance on unseen data.

3.6 Optimisation

The policy parameters θ are optimised using stochastic gradient approach. This is central to training the agent to make effective frame-selection decisions. The objective is to maximize the expected cumulative reward over a sequence of actions. Since direct optimization is infeasible due to the non-differentiable nature of the reward, policy gradient methods—such as the REINFORCE algorithm are employed to estimate the gradient of $J(\theta)$ with respect to theta. This estimated gradient is then used to update the policy parameters through stochastic gradient ascent. To reduce the variance of gradient estimates and improve training stability, techniques like baseline subtraction or entropy regularization are commonly incorporated. It is given as:

$$\theta = \theta - \alpha \Delta_{\theta}(-J + \beta_1 L_{\text{percentage}} + \beta_2 L_{\text{penalty}}), \quad (19)$$

Where α , β_1 and β_2 are hyper-parameters balancing the regularisation.

3.7 Summary Generation

To generate summaries, we start by using the trained AVSN to predict frame-level importance scores for a given test video. These scores represent the likelihood of each frame being selected for the summary. Shot-level scores are calculated ntaking the mean scores of all frames within a given shot. This step ensures that the importance is evaluated at a higher, more meaningful level, aligning with the structure of the video.

For temporal segmentation, we rely on Kernel Temporal Segmentation (KTS) [16] a technique designed to detect significant transitions in video content. KTS works by grouping visually similar frames into clusters, effectively segmenting the video into coherent shots. This method captures natural boundaries in the content, providing well-defined units for subsequent shot-level scoring and selection. By applying KTS, we ensure the segmentation process aligns with the video’s inherent narrative flow [12].

To create a concise summary, we select shots that maximize the total importance score while adhering to a constraint: the summary’s total duration cannot exceed 15% of the input video length. This selection of frames can be considered like a 0/1 Knapsack puzzle,

a classic optimization challenge where items (in this case, shots) must be chosen to maximize value (importance score) without exceeding a capacity limit (summary duration). Since it is an “NP hard” problem, hence it is computationally expensive to compute all the solutions for large instances. To address this, we use a dynamic programming approach to achieve a near-optimal solution efficiently. This ensures the summary is both informative and adheres to the length constraints.

Chapter 4

EXPERIMENTS AND RESULTS

4.1 Datasets

To assess the effectiveness of our network and train it using two widely recognized benchmark datasets - SumMe and TVSum. These datasets are widely recognized in video summarization research due to their diversity and high-quality annotations. SumMe comprises a dataset containing 25 human-generated videos that span a variety of themes, like sports, travel, and holidays. Each video is relatively short, ranging from 1 to 6 minutes in length, and is accompanied by multiple human-annotated summaries, with 15 to 18 individuals providing ground truth annotations for each video. This diversity of input allows SumMe to serve as a robust testing ground for evaluating summarization techniques against human-curated benchmarks.

Similarly, TVSum offers a larger dataset with 50 videos covering a broad range of topics, including news broadcasts, documentaries, and more. The videos in TVSum are longer, lasting upto 10 minutes with a minimum of 2 minutes, and each is annotated by 20 individuals who assign frame-level importance scores. These annotations reflect collective preferences, offering a reliable measure of what viewers consider significant in the content. Together, SumMe and TVSum provide complementary evaluation platforms, one emphasizing user-generated diversity and the other collective viewer insights allowing us to thoroughly test the generalizability and performance of our summarization approach.

4.2 Evaluation Metrics

To ensure an impartial analysis with respect to existing approaches, we adopt the widely recognized evaluation protocol introduced in [2]. This method employs the F-score as the main metric to examine the alignment between the summary produced by the model and the human-labeled ground truth summary. By balancing precision and recall, the F-score effectively measures how well our video summaries capture the most significant parts of the video. It is given as:

$$P = \frac{\text{overlapped}(A,B)}{A}, R = \frac{\text{overlapped}(A,B)}{B} \quad (20)$$

$$Fscore = \frac{2PR}{P + R} \times 100\% \quad (21)$$

Where A is the ground truth summary and B is the generated summary. There we can see that this metric describes the overlapped duration of the generated summary and its ground truth.

4.3 Implementation Details

To enhance computational efficiency, we uniformly sample one frame per second from all training and testing videos. Frame-level features are extracted using the Inception-V3 network. Frame-level deep features are extracted using the pool5 layer output from a pre-trained GoogleNet model [18] on the ImageNet dataset. The resulting feature vectors of dimensionality $D = 1024$. The number of video segments P on which local attention is applied is set to 4. The number of heads H for global attention equals 8. To segment visually consistent sequences, we employ the Kernel Temporal Segmentation (KTS) algorithm [16], which groups consecutive similar frames into distinct shots. We cap the number of generated shots at 50 for each video. For videos with fewer than 50 frames, segmentation is skipped, and the extracted frame-level features are directly used as shot-level representations. To optimize our model parameters, we take a learning rate of 0.001 along with a weight decay of 10^{-5} .

The experiments were conducted on a Mac system running macOS 15.5 Sequoia, equipped with an Apple M2 Pro chip (12-core CPU, 19-core GPU) and 16GB of unified memory. The implementation utilized PyTorch version 1.13.0, leveraging the Metal Performance Shaders (MPS) backend for GPU acceleration.

Table I: Performance comparison (F-score %) on SumMe and TVSum datasets

| Model | SumMe | TVSum |
|---------------------|-------------|-------------|
| DSN | 41.4 | 57.6 |
| Bi-LSTM | 37.6 | 54.2 |
| DPP-LSTM | 38.6 | 57.1 |
| GAN (DPP) | 39.1 | 51.7 |
| Hierarchical RL | 43.6 | 58.4 |
| CosNet | 47.8 | 59.7 |
| TR-Sum | 54.5 | 62.3 |
| GLASN (Ours) | <u>50.6</u> | 66.2 |

4.4 Performance Comparisons

To evaluate the effectiveness of our proposed model, we conducted a detailed performance comparison against several prominent reinforcement learning (RL)-based video summarization methods. These include DSN [1], DPP-LSTM and Bi-LSTM [2], SUM-GAN(dpp) [3], and more recent RL-based frameworks such as the hierarchical model by Chen et al. [23], CoSNet [24], and TR-SUM [12]. The comparison was carried out using the widely adopted SumMe and TVSum datasets, with the F-score metric serving as the primary evaluation criterion due to its balanced consideration of precision and recall. Our model demonstrates competitive performance, closely aligning with or outperforming several of these existing methods. The details are given in Table 1.

4.5 Ablation Study

To gain a deeper understanding of the architectural choices in our proposed AVSN model, we conducted a series of ablation experiments focused on key design components. These experiments aim to identify the optimal configuration for video segmentation, feature fusion, the number of attention heads used in both local and global attention mechanisms and the use of core components of our model [6].

Our first experiment investigates how the number of video segments—used to control the granularity of local attention, affects summarization performance. In parallel, we explore the fusion strategies for integrating the outputs of global and local attention modules. Table II summarizes the performance across various configurations. Results demonstrate that dividing videos into four segments and using addition-based fusion yields the most favorable results on both TVSum and SumMe datasets. This combination likely allows the model to preserve richer localized details while maintaining a coherent global context.

Subsequently, we analyze the impact of varying the number of attention heads in both attention modules. We evaluate local attention with 2, 4, and 8 heads, and observe that using 4 attention heads consistently produces robust performance across datasets. For global attention, usage of 8 heads gives the best result whereas for local attention the optimal number of heads is 4. This setting offers a strong balance, delivering consistently competitive results across both datasets. The performance with different number of heads is shown in Table III.

To further examine the contribution of individual architectural components within AVSN, we designed three model variants by selectively removing key mechanisms: one without the global attention module, relying solely on local attention to capture temporal dependencies. Another without the local attention module, depending only on global attention for sequence modeling and a third variant excluding the positional encoding used in attention score computation. These variants were evaluated under the same experimental conditions using five randomly generated data splits. The results, summarized in Table IV, reveal that both global and local attention mechanisms are essential for strong performance, with the absence of global attention having a particularly detrimental effect on the SumMe dataset. The removal of local attention also caused noticeable performance drops, demonstrating that both global and localized temporal structures are necessary for comprehensive video understanding. Furthermore, excluding positional encoding reduced

performance, especially on SumMe, confirming its role in enhancing the model’s ability to capture the temporal structure of videos. Together, these findings highlight the synergistic importance of each component in maintaining the effectiveness of the full AVSN architecture.

Table II: Ablation study showing the variation in performance (F-score %) of our model on SumMe and TVSum with different numbers of video segments and data fusion strategies.

| | SumMe | | | TVSum | | |
|----------------------------------|----------|-------------|----------|----------|-------------|----------|
| Segments Fusion | 2 | 4 | 8 | 2 | 4 | 8 |
| Addition | 44.3 | 50.6 | 47.8 | 64.4 | 66.2 | 65.5 |
| Average pooling | 45.5 | 49.1 | 48.7 | 66.1 | 65.9 | 59.9 |
| Max pooling | 45.1 | 48.9 | 47.8 | 64.3 | 62.6 | 61.5 |
| Multiplication | 39.3 | 45.1 | 48.6 | 49.9 | 49.5 | 42.7 |

Table III: Ablation study on the performance (F-score %) of our model on SumMe and TVSum using different numbers of attention heads in the global and local attention mechanisms.

| | SumMe | | | TVSum | | |
|-------------------------------|----------|-------------|----------|----------|-------------|----------|
| Local Global | 2 | 4 | 8 | 2 | 4 | 8 |
| 2 | 45.3 | 45.4 | 47.8 | 64.4 | 66.1 | 65.5 |
| 4 | 45.5 | 49.1 | 48.7 | 66.1 | 65.9 | 59.9 |
| 8 | 45.1 | 49.9 | 47.8 | 64.3 | 66.1 | 61.5 |
| 16 | 39.3 | 48.2 | 40.6 | 63.9 | 64.5 | 62.8 |

Table IV: Component-wise ablation study of GLASN model.

| Model Variant | SumMe F-score (%) | TVSum F-score (%) |
|-------------------------------|-------------------|-------------------|
| GLASN w/o Global Attention | 39.2 | 59.1 |
| GLASN w/o Local Attention | 42.0 | 62.4 |
| GLASN w/o Positional Encoding | 47.3 | 65.7 |
| GLASN (full model) | 50.6 | 66.2 |

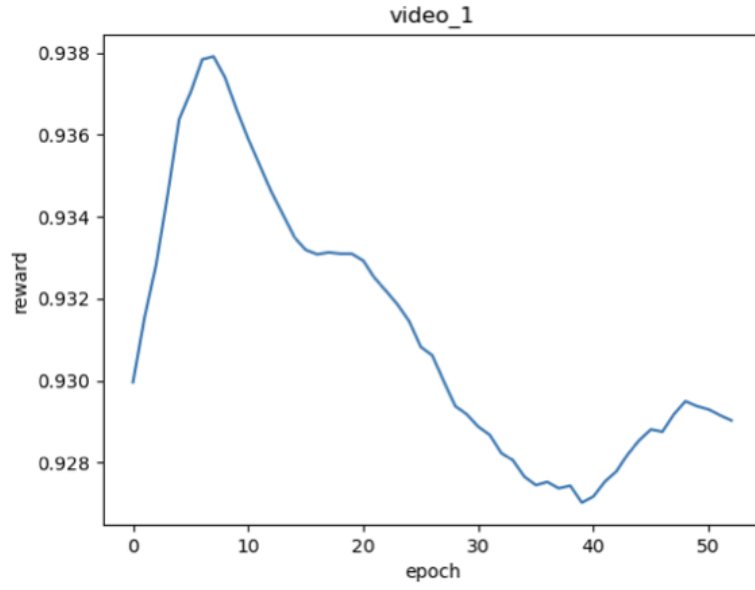


Figure 1: Plot of rewards while training video-1

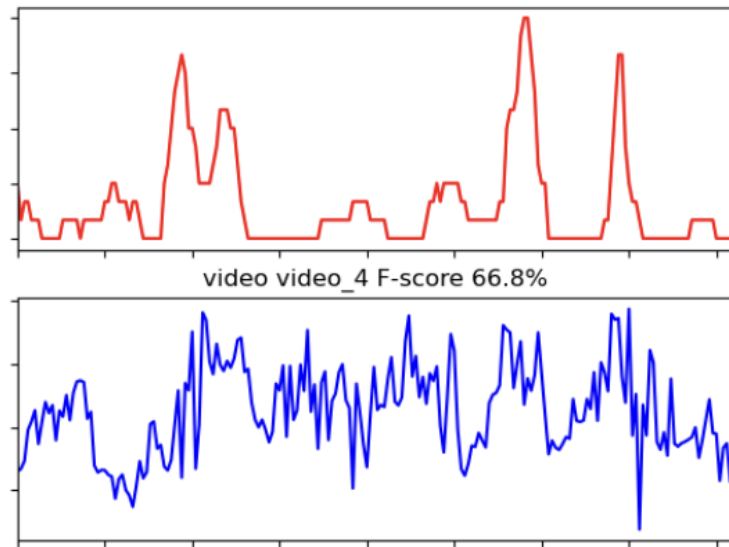


Figure 2: Ground Truth and importance scores (probabilities) of video-4 predicted by GLASN

Chapter 5

CONCLUSION AND FUTURE SCOPE

The domain of video summarization has experienced significant advancement, driven by the urgent need for efficient techniques to condense and interpret vast amounts of video data. Although supervised methods have demonstrated strong performance, their dependence on large-scale labeled datasets presents a major limitation, as such annotations are costly and often impractical to obtain. This challenge highlights the critical importance of unsupervised learning approaches, which offer greater scalability and adaptability by operating without extensive human supervision.

Attention mechanisms, especially those derived from transformer architectures, have shown remarkable effectiveness in modeling temporal dependencies within sequential data. In particular, combining global and local attention enables models to capture both the overarching context of the entire video and the fine-grained details within smaller temporal segments. By selectively focusing on the most informative parts at multiple scales, these hybrid attention mechanisms provide a richer and more nuanced understanding of video content.

In this project, we have explored the application of combined attention-based techniques for unsupervised video summarization, aiming to enhance summary quality and representativeness. Global attention captures long-range dependencies by attending to the entire video sequence, while local attention focuses on short-range, fine-grained interactions within smaller segments. This multi-scale approach effectively models relationships at different granularities, allowing the summarization system to better preserve both the global storyline and important local details. Our findings indicate that integrating global and local attention significantly improves summarization performance, producing outputs that are more coherent, concise, and representative compared to traditional unsupervised methods.

The results obtained from our evaluation on benchmark datasets like TVSum and SumMe demonstrate that our model can generate summaries that are competitive with, and in some cases exceed, the performance of existing state-of-the-art unsupervised approaches. Importantly, the architecture’s design does not rely on ground-truth labels, making it particularly suitable for real-world applications where manual annotation is unfeasible. The ablation studies further confirm the contribution of each model component, validating the importance of combining local and global attention and optimizing the number of attention heads and video segmentation granularity.

Future work can explore several promising directions to build upon the current model. One such direction involves the integration of multimodal data—including audio signals, speech transcripts, and textual metadata—which can provide complementary contextual information to guide the summarization process. The inclusion of such modalities has the potential to further enhance the semantic richness and relevance of generated summaries, particularly in videos containing dialogues, background music, or on-screen text.

Another fruitful area for expansion is the use of reinforcement learning to optimize summary selection policies over time. Although our current approach is entirely unsupervised and static, reinforcement learning allows models to learn from user feedback or predefined reward functions, offering dynamic adaptability and personalization of summaries. Combining the strengths of attention-based architectures with reinforcement-based optimization can result in more intelligent, context-aware summarization systems that align better with human preferences.

Additionally, incorporating user-specific preferences and feedback mechanisms into the model can personalize summaries based on individual viewing habits, domain interests, or attention patterns. This opens up opportunities for adaptive video summarization systems that not only summarize but also curate content according to the intended audience. Such systems could prove invaluable in educational platforms, streaming services, or professional video review applications.

On the technical side, further research can focus on model compression and optimization methods to make sure the model runs in real-time, especially when deployed in resource-limited devices like mobile phones, drones, or embedded systems. Efficient transformer versions or sparse attention techniques could be explored to keep good performance while cutting down on the computational cost.

From a social impact perspective, effective video summarization can bring about transformative changes across multiple domains. In education, for instance, automatic summarization of long lecture videos can help students quickly revise key topics and enhance learning efficiency. For journalists and media analysts, condensed versions of lengthy broadcasts can facilitate rapid content review and curation. In law enforcement and surveillance, automated video summaries can significantly reduce the burden of manual footage review, helping with quicker incident response and decision-making.

Moreover, video summarization holds the potential to promote digital accessibility and inclusion. For individuals with limited attention spans, cognitive impairments, or time constraints, summarization tools can make video content more digestible and engaging. By tailoring summaries to emphasize essential content, such technologies ensure broader reach and inclusivity in digital communication. In humanitarian settings, rapid summarization of aerial footage captured during natural disasters or conflicts can assist relief workers in prioritizing areas for intervention, thereby saving lives and resources.

In conclusion, the future of video summarization lies at the intersection of interpretability, adaptability, and scalability. The promising results achieved through attention-based, unsupervised architectures reinforce the viability of such models for practical deployment. By continuing to enhance model efficiency, integrate multimodal inputs, and incorporate

user feedback, the field can move closer to realizing intelligent, context-aware summarization systems that benefit diverse user communities across the globe. The ongoing evolution of this technology holds immense promise for reshaping how we interact with, analyze, and derive value from the ever-expanding universe of video data.

Appendix A

Plagiarism Verification

Title of the Thesis “Reinforced Attention for Video Summarisation”

Total Pages 29 Name of the Scholar “Simran Ray”

Supervisor(s)
(1) “Prof. Anil Singh Parihar”

Department of “Computer Science and Engineering”

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: “Turnitin” Similarity Index: 8% Total Word Count: 9450

Date: May 2025

Candidate’s Signature
(Simran Ray)

Signature of Supervisor
Prof. Anil Singh Parihar
Computer Science & Engineering
Delhi Technological University


Bibliography

- [1] K. Zhou, Y. Qiao, and T. Xiang, “Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9222–9231.
- [2] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, “Video summarization with long short-term memory,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 211–226.
- [3] B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised video summarization with adversarial lstm networks,” in *CVPR*, 2017.
- [4] R. Panda and A. K. Roy-Chowdhury, “Collaborative summarization of topic-related videos,” in *CVPR*, 2017.
- [5] Y. Hsiao, Y. Wang, S. Zhao, and B. Ghanem, “Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7405–7414.
- [6] J. Gao, Y. Ma, P. Zhang, K. Chen, W. Wang, and X. Wang, “Combining global and local attention with positional encoding for video summarization,” in *IEEE ISM*, 2021.
- [7] T.-C. Hsu, Y. Liao, and C.-R. Huang, “Video summarization with spatiotemporal vision transformer,” *IEEE Transactions on Image Processing*, vol. 32, pp. 2082–2097, 2023.
- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [9] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, “Summarizing videos with attention,” in *Asian Conf. on Comp. Vision 2018 Workshops*. Springer International Publishing, 2018, pp. 39–54.
- [10] U. N. Yoon, M. D. Hong, and G.-S. Jo, “Unsupervised video summarization based on deep reinforcement learning with interpolation,” *Sensors*, vol. 23, no. 7, p. 3384, 2023.
- [11] M. S. Afzal and M. A. Tahir, “Reinforcement learning based video summarization with combination of resnet and gated recurrent unit,” in *Proc. 16th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2021.

- [12] M. Abbasi, H. Hadizadeh, and P. Saeedi, “Unsupervised video summarization via reinforcement learning and a trained evaluator,” *CoRR*, vol. abs/2407.04258, 2024.
- [13] H. Terbouche, M. Morel, M. Rodriguez, and A. Othmani, “Multi-annotation attention model for video summarization,” in *Proc. CVPR Workshops (LSHVU Track)*, 2023.
- [14] Y. Zhu, W. Zhao, R. Hua, Y. Yin, X. Bi, and X. Wu, “Topic-aware video summarization using multimodal transformer,” *Pattern Recognition*, vol. 140, p. 109578, 2023.
- [15] A. M. Metelli, “Configurable environments in reinforcement learning: An overview,” *Springer*, 2022.
- [16] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, “Category-specific video summarization,” in *European Conference on Computer Vision (ECCV)*, 2014, pp. 540–555.
- [17] W. Zhu, J. Lu, J. Li, and J. Zhou, “Dsnet: A flexible detect-to-summarize network for video summarization,” *IEEE Transactions on Image Processing*, vol. 30, pp. 948–962, 2021.
- [18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *CVPR*, 2015, pp. 1–9.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [20] T. Yao, T. Mei, and Y. Rui, “Highlight detection with pairwise deep ranking for first-person video summarization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 982–990.
- [21] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating summaries from user videos,” in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 505–520.
- [22] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
- [23] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, “Weakly supervised video summarization by hierarchical reinforcement learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 4002–4011.
- [24] T. Liu, “Compare and select: Video summarization with multi-agent reinforcement learning,” 2020.

Simran Ray

REINFORCED ATTENTION FOR VIDEO SUMMARISATION

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98553206

Submission Date

May 30, 2025, 3:52 PM GMT+5:30

Download Date

May 30, 2025, 3:55 PM GMT+5:30

File Name

plagCheckV2.2.pdf

File Size

859.8 KB

29 Pages

9,450 Words

55,172 Characters





8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Small Matches (less than 8 words)

Match Groups

-  **77 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 4%  Publications
- 4%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 77 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%** Internet sources
- 4%** Publications
- 4%** Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

| | | |
|---|------------------------|-----|
| 1 | Internet | |
| arxiv.org | | 1% |
| 2 | Internet | |
| qmro.qmul.ac.uk | | 1% |
| 3 | Publication | |
| Zutong Li, Lei Yang. "Weakly Supervised Deep Reinforcement Learning for Video ... | | <1% |
| 4 | Submitted works | |
| University of Thessaly on 2025-01-25 | | <1% |
| 5 | Internet | |
| ojs.aaai.org | | <1% |
| 6 | Submitted works | |
| University of Warwick on 2025-05-04 | | <1% |
| 7 | Publication | |
| "ECAI 2020", IOS Press, 2020 | | <1% |
| 8 | Internet | |
| idt.iti.gr | | <1% |
| 9 | Submitted works | |
| Beirut Arab University on 2025-02-26 | | <1% |
| 10 | Publication | |
| Karan Makhija, Thi-Nga Ho, Eng-Siong Chng. "Transfer Learning for Punctuation ... | | <1% |

| | | | |
|----|-----------------|---|-----|
| 11 | Internet | www.frontiersin.org | <1% |
| 12 | Submitted works | Macao Polytechnic Institute on 2025-03-15 | <1% |
| 13 | Internet | labstic.univ-guelma.dz | <1% |
| 14 | Submitted works | University of Sheffield on 2023-09-13 | <1% |
| 15 | Internet | fastercapital.com | <1% |
| 16 | Publication | Jing Zhang, Guangli Wu, Shanshan Song. "Video Summarization Generation Base... | <1% |
| 17 | Internet | mdpi-res.com | <1% |
| 18 | Internet | beckie-khmer.com | <1% |
| 19 | Publication | Rui Li, Fan Zhang, Tong Li, Ning Zhang, Tingting Zhang. "DMGAN: Dynamic Multi-... | <1% |
| 20 | Submitted works | University College London on 2023-09-11 | <1% |
| 21 | Submitted works | University of Oxford on 2024-05-14 | <1% |
| 22 | Internet | drum.lib.umd.edu | <1% |
| 23 | Submitted works | IUBH - Internationale Hochschule Bad Honnef-Bonn on 2024-03-30 | <1% |
| 24 | Publication | Lecture Notes in Computer Science, 2015. | <1% |

| | | | |
|----|-----------------|---|-----|
| 25 | Publication | Lei Cao, Qikai Zhang, Chunjiang Fan, Yongnian Cao. "Not Another Dual Attention ..." | <1% |
| 26 | Submitted works | Nanyang Technological University on 2024-04-14 | <1% |
| 27 | Submitted works | Otto-von-Guericke-Universität Magdeburg on 2023-11-07 | <1% |
| 28 | Submitted works | University of Leeds on 2024-08-11 | <1% |
| 29 | Publication | Gunuganti, Jeshmitha. "Unsupervised Video Summarization Using Adversarial Gr..." | <1% |
| 30 | Publication | Khushboo Khurana, Umesh Deshpande. "Two stream multi-layer convolutional n..." | <1% |
| 31 | Submitted works | Leiden University on 2023-08-30 | <1% |
| 32 | Submitted works | Letterkenny Institute of Technology on 2021-09-09 | <1% |
| 33 | Publication | Qiuxia Lai, Wenguan Wang, Hanqiu Sun, Jianbing Shen. "Video Saliency Prediction..." | <1% |
| 34 | Publication | Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ..." | <1% |
| 35 | Submitted works | University of Surrey on 2024-08-28 | <1% |
| 36 | Submitted works | University of Westminster on 2025-04-16 | <1% |
| 37 | Internet | serp.ai | <1% |

Simran Ray

REINFORCED ATTENTION FOR VIDEO SUMMARISATION



Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98553206

Submission Date

May 30, 2025, 3:52 PM GMT+5:30

Download Date

May 30, 2025, 3:55 PM GMT+5:30

File Name

plagCheckV2.2.pdf

File Size

859.8 KB

29 Pages

9,450 Words

55,172 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

