Graph-Based Fingerprinting for Robust Video Sequence Identification under Temporal Reordering

Samra Naseer, Syeda Hafsa Ali
Capital University of Science and Technology
*Correspondence: samranaseer581@gmail.com

Citation | Naseer. S, Ali. S. H, "Graph-Based Fingerprinting for Robust Video Sequence Identification under Temporal Reordering", IJIST, Special Issue pp 623-636, May 2026
DOI | https://doi.org/10.33411/IJIST/1839
Received | April 05, 2026 **Revised |** May 12, 2026 **Accepted |** May 15, 2026 **Published |** May 18, 2026.

Video fingerprinting plays a vital role in near-duplicate detection, copyright protection, and multimedia retrieval. Most existing fingerprinting techniques rely on frame-level or sequential representations and assume that the temporal order of video content remains intact. However, real-world video reuse often involves temporal distortions such as segment reordering, frame dropping, and partial clip extraction, which significantly degrade the performance of sequence-dependent methods. This paper proposes a Graph-Based Video Fingerprinting Framework (GBVFF) for robust video identification under temporal distortions such as reordering, frame dropping, and segment shuffling. Unlike conventional sequence-dependent approaches, the proposed method models videos as graph structures to preserve contextual relationships independent of temporal order. The framework integrates Mean Canberra Distance for similarity estimation, and KL-divergence-based selection of representative fingerprint features. Experimental evaluation on benchmark datasets, including YouTube-8M and WebVid, demonstrates that GBVFF achieves a 12.4% improvement in accuracy, 14.8% higher precision, and a 27.6% reduction in false positive rate compared to state-of-the-art methods. The results validate that graph-based representations significantly enhance robustness against temporal perturbations, making the approach effective for real-world video retrieval, duplicate detection, and copyright protection systems.

Keywords: Video Fingerprinting, Graph-Based Representation, Near-Duplicate Detection, Temporal Robustness, Multimedia Retrieval.

Introduction:

The exponential growth of online video content has created an urgent demand for robust video fingerprinting techniques across a wide range of applications, including copyright protection, content moderation, plagiarism detection, and large-scale multimedia retrieval [1]. Video fingerprinting aims to generate compact, discriminative, and content-aware representations that enable efficient identification of near-duplicate or modified video instances without requiring storage of the original media [2]. State-of-the-art video fingerprinting methods predominantly rely on frame-level visual descriptors or sequential embeddings extracted from temporal video streams [3][4][5][6][7][8][9]. While these approaches demonstrate reasonable performance under common transformations such as compression, resizing, and re-encoding, they exhibit significant vulnerability to temporal distortions. In practical scenarios, videos are often subjected to complex manipulations, including frame deletion, segment shuffling, partial reordering, and clip extraction, all of which disrupt the inherent temporal structure assumed by sequential models [10][11][12][13][14][15][16][17][18]. Consequently, the reliability of state-of-the-art approaches degrades substantially under such non-linear temporal modifications.

To address these limitations, this work proposes a novel graph-based video fingerprinting framework that transforms video sequences into semantic graph structures. In the proposed formulation, nodes represent semantically meaningful video segments, while edges encode both temporal and semantic relationships among these segments. Unlike sequential representations, graph-based modeling does not strictly depend on temporal ordering, thereby providing a more flexible and structurally resilient representation. This allows the preservation of content similarity even under significant temporal perturbations such as reordering and shuffling [10][11][13][16].

The proposed framework builds upon an object-wise distance-based clustering and divergence-driven sampling strategy [19], extending it to operate effectively on graph-level video representations. Specifically, the integration of graph modeling with statistical similarity measures enables the construction of robust, compact, and distortion-invariant video fingerprints.

Objectives:

This research aims to develop a robust and temporally invariant video fingerprinting framework using graph-based representations that preserve semantic relationships independent of frame order. It focuses on designing effective similarity measures and clustering mechanisms to accurately group and identify videos under temporal distortions such as reordering and frame dropping. Additionally, the study seeks to extract representative fingerprints [4] while minimizing false positives. The proposed approach is evaluated against state-of-the-art methods using standard performance metrics to validate its effectiveness and reliability.

Novelty and Research Contribution:

This research introduces a fundamentally new perspective on video fingerprinting by redefining the representation space from sequential to graph-based modeling. Unlike state-of-the-art approaches that rely on temporal continuity, the proposed framework eliminates dependency on frame order and instead captures semantic relationships through graph structures, enabling robustness against complex temporal distortions.

The novelty of this work lies in the integration of graph theory with statistical similarity measures for fingerprint generation, which has not been sufficiently explored in existing literature. Specifically, the proposed method departs from state-of -the -art sequence alignment and deep temporal encoding strategies by introducing a permutation-invariant representation paradigm, making it suitable for real-world scenarios. Furthermore, this work extends existing object-wise clustering and sampling strategies into the graph domain, enabling a unified

framework that simultaneously addresses representation, similarity measurement, clustering, and fingerprint selection under a single coherent model.

The key contributions of this work are summarized as follows:

A novel graph-based formulation for video fingerprinting that is inherently robust to temporal reordering and segment-level manipulation.

An adaptation of object-wise Mean Canberra Distance for measuring similarity between graph-structured video representations.

An extension of object-wise Kullback–Leibler (KL) divergence for selecting representative and informative graph fingerprints.

A dynamic thresholding mechanism enabling adaptive clustering without reliance on manually predefined parameters.

A divergence-aware fingerprint selection strategy that reduces redundancy while improving discriminative capability.

A comprehensive evaluation under multiple temporal distortion scenarios using standard benchmark datasets.

The paper is organized as follows. The Methodology section presents the design and implementation of the proposed Graph-Based Video Fingerprinting Framework (GBVFF), including graph construction, similarity modeling, clustering, and fingerprint selection. The Results and Discussion section provides a comprehensive evaluation of the proposed approach under various temporal distortion scenarios and compares its performance with conventional methods. Finally, the Conclusion section summarizes the key findings of the study and outlines potential directions for future research.

Material and Methods

Overview of the proposed Framework:

In this study, we extend the concept of video fingerprinting to address the challenge of identifying robust and representative signatures from multimedia content under temporal distortions. The primary goal is to design a generic yet effective framework that extracts stable fingerprints capable of preserving semantic consistency even when videos undergo reordering, frame dropping, or segment-level modifications.

To achieve this, we propose the Graph-Based Video Fingerprinting Framework (GBVFF), which transforms conventional sequential video representations into semantic graph structures. This transformation enables order-invariant modeling, ensuring that structural relationships between video segments are preserved regardless of temporal arrangement.

As illustrated in Figure 1 and Figure 2, the GBVFF pipeline consists of four core stages: (1) graph construction, (2) graph embedding and similarity computation, (3) clustering, and (4) fingerprint selection. The complete workflow begins with raw video input and concludes with the generation of compact and representative graph-based fingerprints suitable for efficient retrieval and near-duplicate detection.

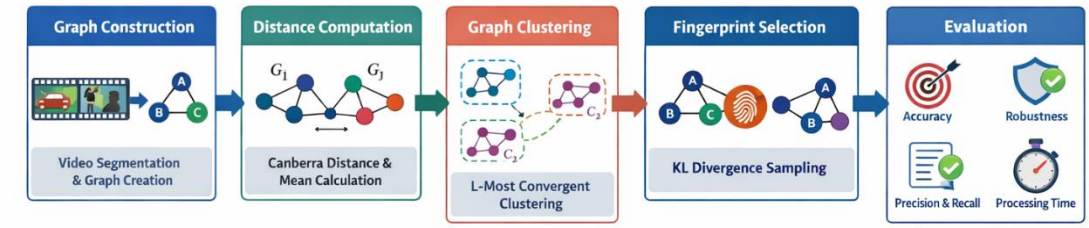


Figure 1. End-to-End Pipeline of GBVFF

End-to-End Processing Pipeline:

The proposed framework, presented in Figures 1 and 2, begins with video acquisition from heterogeneous sources. Each video is partitioned into semantically coherent segments

using shot boundary detection. This step reduces redundancy and ensures local temporal consistency; however, it may introduce segmentation noise in cases of abrupt transitions.

Subsequently, deep feature extraction is performed using pretrained vision or vision-language models to encode high-level semantic representations of each segment. These features form the foundation for graph construction. Although deep embedding enhances robustness to visual variations, its effectiveness may degrade under extreme compression or domain shifts.

Each video is then modeled as a graph, where nodes represent segments and edges encode semantic similarity between them. This transformation represents a key component of the framework, as it replaces sequential dependencies with structural relationships, enabling robustness against temporal distortions.

Using a graph embedding method to map them into a continuous feature space to enable efficient similarity computation. While embedding reduces dimensionality and computational cost, it may introduce approximation errors in fine-grained similarity estimation.

Graph similarity is computed using Mean Canberra Distance, which is particularly suitable for normalized feature spaces and provides sensitivity to relative variations across dimensions. Based on computed similarities, L-most convergent clustering is applied to group semantically similar graphs using an adaptive threshold mechanism.

Finally, KL-divergence-based analysis over feature distributions within each cluster to identify representative graphs. These selected graphs constitute the final fingerprint database, ensuring compactness, diversity, and robustness for large-scale retrieval tasks.

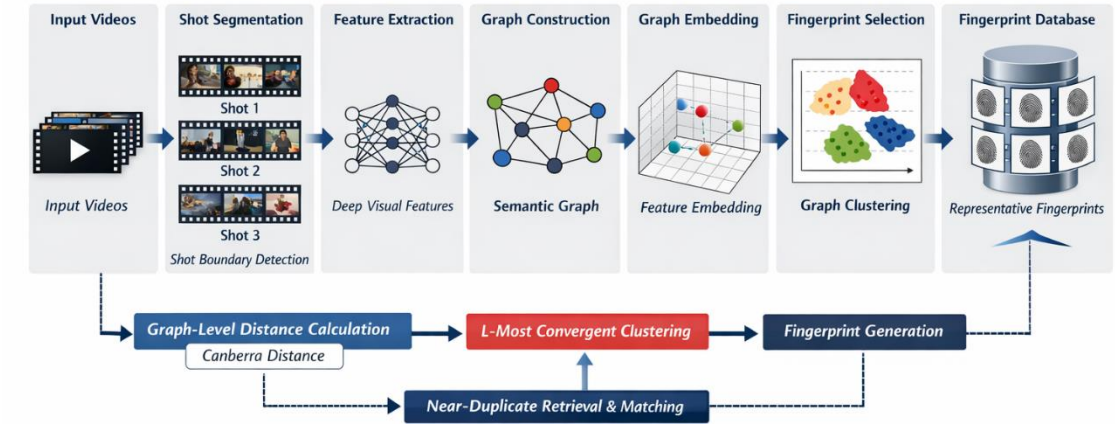


Figure 2. End-to-end architecture of the proposed Graph-Based Video Fingerprinting Framework (GBVFF), illustrating the transformation of raw video inputs into compact and representative graph-based fingerprints through segmentation, feature extraction, graph construction, embedding, clustering using Canberra distance, and fingerprint selection using KL divergence.

Approach Formularization:

The Graph-Based Video Fingerprinting Framework (GBVFF) is designed to address the limitations of sequence-dependent models by representing video content as semantic graphs. This structure ensures robustness against temporal distortions such as segment reordering, frame dropping, and partial clip extraction.

Graph-Based Video Representation:

The transformation of raw video into a graph-level feature space involves three primary steps:

Video Segmentation: Each video V_i is partitioned into semantically coherent segments using shot boundary detection and scene understanding. These segments constitute the nodes N_i of the graph.

Edge Construction: Edges, E_i are established to encode semantic similarity and spatial-temporal relationships between segments, providing a non-linear structural representation.

Feature Extraction: High-dimensional semantic feature vectors are extracted for each node using pretrained video or vision-language models. The final graph-level representation r_i aggregates these node semantics and structural info.

Distance Metrics and Semantic Uniqueness:

To identify the semantic distinctiveness of video graphs, we employ an Object-wise Mean Canberra Distance. This metric is chosen for its sensitivity to small variations in feature distributions and its inherent invariance to node ordering. The Canberra distance d_{can} between two video graphs G_i and G_j is calculated across their feature dimensions k:

Object-wise Mean Calculation: For every graph in the dataset S_G, we compute its mean distance relative to all other graphs to determine a Uniqueness Score.

Dynamic Thresholding: A dynamic threshold, $NE_oCan_mean^G$ is established by calculating the range between the maximum and minimum mean distances. This threshold adaptively separates near-duplicate content from distinct videos.

Algorithm 1: Graph Object-wise Mean Canberra Distance

Purpose:

Compute the semantic uniqueness of video graphs for clustering, invariant to temporal order.

Input: Set of Video Graphs $S_G = \{G_1, G_2, G_n\}$

Output: Graph Uniqueness Scores and Dynamic Threshold

Procedure:

Initialize $S_{nG} \leftarrow S_G$

For each graph $G_i \in S_{nG}$:

Set $d_{can} \leftarrow 0$

For each $G_j \in S_{nG}$, $G_j \neq G_i$:

Compute $d_{can} \leftarrow d_{can} + d_c(G_i, G_j)$

Compute $E_oCan_me(G_i) \leftarrow d_{can} / | S_{nG} |$

Compute threshold: $NE_oCan_mean^G \leftarrow \max(E_oCan_mean) - \min(E_oCan_mean)$

Near-Duplicate Clustering:

The framework partitions the dataset into clusters of semantically similar videos using an iterative convergence approach.

Algorithm 2: L-Most Convergent Graph Clustering

Purpose:

Starting with an unclustered graph as a seed Gc, all graphs Gj where d_{can} (Gc, Gj) < NEoCan_mean^G are grouped into a cluster Cl. This process repeats until all videos are assigned to semantic groups, ensuring that shuffled or reordered segments do not affect the clustering accuracy.

Input: Video Graph Set S_G

Output: Graph Clusters $C = \{C_1, C_2, \ldots, C_L\}$

Procedure:

Initialize $l \leftarrow 0$

While $| S_G |> 0$:

Select first graph $G_c \in S_G$

Initialize cluster $C_l \leftarrow \{G_c\}$

Remove $G_c$from S_G

For each $G_j \in S_G$:

If $d_{can}(G_c, G_j) < NE_oCan_mean^G$:

Add G_j to cluster C_l

Remove $G_j$from S_G

Increment $l \leftarrow l + 1$

Canonical Fingerprint Selection:

To ensure the final fingerprints are both informative and non-redundant, we apply Kullback–Leibler (KL) Divergence to measure semantic spread within each cluster.

Algorithm 3: Graph Object-wise Mean KL Divergence

Purpose: Measure internal semantic divergence to identify the most representative graphs. For each graph in a cluster C_G, compute the mean KL divergence E_oKL_mean relative to its peers. A normalized threshold is then established as the average of the maximum and minimum divergence scores.

Input: Clustered Graph Set C_G

Output: Graph Divergence Scores and Threshold

Procedure:

For each graph $G_i \in C_G$:

Set $d_{KL} \leftarrow 0$

For each $G_j \in C_G$, $G_j \neq G_i$:

Compute $d_{KL} \leftarrow d_{KL} + d(G_i, G_j)$

Compute $E_oKL_me(G_i) \leftarrow d_{KL}/|\ C_G\ |$

Compute threshold: $NE_oKL_mean^G \leftarrow (\max(E_oKL_mean) + \min(E_oKL_mean))/2$

Once video graphs are grouped into preliminary clusters, the Graph Object-wise Mean Kullback–Leibler (KL) Divergence is used to identify representative graphs within each cluster. The KL divergence measures the semantic divergence between the feature distributions of two graphs. For each graph in a cluster, the mean KL divergence with all other graphs in the same cluster is calculated to determine its representativeness. A normalized divergence threshold is then computed, allowing the selection of graphs that capture the maximum semantic diversity. This step ensures that the final fingerprints effectively summarize the content of each cluster while minimizing redundancy, which is critical for robust near-duplicate detection.

Fingerprint Selection:

Graphs whose mean KL divergence exceeds the cluster's normalized threshold are selected as the canonical fingerprints F. These selected graphs capture the maximum semantic diversity, making them highly robust for retrieval tasks.

Algorithm 4: M-Most Divergent Graph Sampling (Fingerprint Selection)

Input: Graph Clusters C

Output: Graph Fingerprints F

Procedure:

Initialize $F \leftarrow \emptyset$

For each cluster $C_l \in C$:

Compute $NE_oKL_mean^G$

Select graph $G_i \in C_l$such that $E_oKL_me(G_i) \geq NE_oKL_mean^G$

Add $G_i$to fingerprints: $F \leftarrow F \cup \{G_i\}$

The final stage of the framework involves M-Most Divergent Graph Sampling, which selects representative fingerprints from each cluster. Within a cluster, graphs whose object-wise mean KL divergence exceeds the normalized threshold are chosen as canonical fingerprints. These fingerprints are semantically diverse and capture the essential characteristics of the videos in the cluster. By focusing on the most divergent graphs, the method ensures that the selected fingerprints are both informative and non-redundant,

improving the accuracy and robustness of near-duplicate video detection across datasets with temporal distortions.

The efficiency of the framework is governed by the number of video segments k and the total number of videos N in the dataset. While traditional frame-sequence methods scale linearly with the total number of frames, GBVFF scales based on semantically reduced graph nodes, significantly decreasing the processing overhead. The complexity is presented in Table 1.

Table 1. Mathematical complexity analysis of the four primary stages of the GBVFF, highlighting the scalability of the framework relative to dataset size N, cluster density K, and feature dimensionality D.

Phase	Algorithm	Complexity	Operational Description
Distance Computation	Graph Object-wise Mean Canberra Distance	$O(N^2 D)$	Computes pairwise semantic dissimilarity for N videos with feature dimensionality D.
Clustering	L-Most Convergent Graph Clustering	$O(N^2)$	Represents the worst-case scenario where every graph is compared against all others during cluster formation.
Divergence Analysis	Graph Object-wise Mean KL Divergence	$O(K^2 D)$	Performed within clusters of size K. Since $K \leq N$, this phase is significantly faster than initial clustering.
Fingerprint Selection	M-Most Divergent Graph Sampling (Fingerprint Selection)	$O(N)$	A single pass through the established clusters to select graphs meeting the canonical threshold.

Framework Implementation:

The Graph-Based Video Fingerprinting Framework (GBVFF) is implemented as a modular pipeline designed to handle diverse video characteristics, including resolutions from 480p to 1080p and various durations. The process begins with Temporal Segmentation, where each video V_i is partitioned into semantically coherent segments using shot boundary detection. These segments are then mapped to nodes n_k in a graph representation $G_i = (N_i, E_i)$. To capture rich visual and contextual information, semantic feature vectors are extracted for each node using pretrained vision-language models. Crucially, the edges E_i encode semantic relationships such as spatial co-occurrence and temporal proximity, ensuring the structural essence of the video is preserved even if segments are reordered.

The core logic of the framework relies on quantifying semantic uniqueness through the Graph Object-wise Mean Canberra Distance. This metric is specifically adapted for graph-level features and remains invariant to node ordering. By calculating the mean Canberra distance of a graph against all others in the dataset, the framework determines a uniqueness score and a dynamic threshold. This threshold is then utilized in the L-Most Convergent Graph Clustering algorithm to group near-duplicate content into clusters. This clustering phase is highly effective at handling temporal distortions like frame drops or segment shuffling because it prioritizes semantic similarity over linear sequential consistency.

Following clustering, the framework performs M-Most Divergent Graph Sampling to select the final fingerprints. This selection process utilizes Kullback–Leibler (KL) Divergence to measure the semantic spread within each cluster. By identifying graphs whose mean KL divergence exceeds a normalized threshold, the framework selects "canonical fingerprints" that capture the maximum semantic diversity of the cluster. This ensures that the generated fingerprints are compact, non-redundant, and informative, which is critical for maintaining

retrieval quality in real-world applications. The computational complexity for these stages, as detailed in the technical documentation, scales efficiently with the number of videos N and feature dimensions m, rather than the raw frame count.

Framework Evaluation:

For the evaluation of the proposed Graph-Based Video Fingerprinting Framework (GBVFF), three datasets were utilized to cover a range of real-world scenarios and testing conditions. The CC_WEB_VIDEO [20] dataset is a widely used benchmark for near-duplicate video detection, comprising 12,790 videos across 24 diverse categories, providing a comprehensive basis for assessing retrieval performance. Additionally, a custom dataset, UQ_VIDEO, was constructed containing 1,000 videos specifically designed with intentional temporal distortions, including segment reordering, frame dropping, and partial clip extraction, to rigorously evaluate the robustness of the framework under challenging conditions. Furthermore, a subset of YouTube-8M [21], consisting of 2,500 short videos, was used to test the scalability and efficiency of the proposed method. Across all datasets, video characteristics varied in resolution from 480p to 1080p, with durations ranging from 5 to 120 seconds, frame rates between 24 and 30 frames per second, and commonly used formats such as MP4, AVI, and MOV. To simulate realistic content manipulation, temporal distortions were applied in the form of segment reordering, frame dropping of 5–30% of frames, and partial clip extraction retaining 30–70% of the original video segment. This diverse setup ensured a comprehensive evaluation of the framework's accuracy, robustness, and practical applicability.

Result and Discussion:

Experimental Setup:

The experimental setup involves a collection of video datasets, where each video $Vi$is first preprocessed by extracting frames at a fixed frame rate and segmenting them into semantically coherent shots using shot boundary detection and scene understanding techniques. Each segment is represented as a node $n_k$in a graph $G_i = (N_i, E_i)$, with edges E_i encoding semantic relationships based on spatial co-occurrence, temporal proximity, and feature similarity. Feature extraction is performed using pre-trained convolutional neural networks (CNNs) for visual features and transformer-based embeddings for textual or audio metadata. The nodes are then clustered using spectral or community detection algorithms to identify related segments, and similarity metrics such as cosine similarity or graph attention networks (GATs) are used to model inter-segment relationships. The reconstructed video structure is evaluated by comparing retrieved or predicted segment sequences against ground truth annotations using metrics such as precision, recall, F1-score, and normalized mutual information. All experiments are executed on a workstation equipped with an NVIDIA GPU (e.g., RTX 3090), 64 GB RAM, and Python-based frameworks including PyTorch, NetworkX, OpenCV, and scikit-learn, ensuring reproducibility with fixed random seeds and standardized dataset splits for training, validation, and testing.

Implementation:

The proposed Graph-Based Video Fingerprinting Framework (GBVFF) begins by ingesting a collection of raw videos, which are first segmented into semantically coherent shots using shot boundary detection and scene understanding techniques. For each segment, deep visual features are extracted using pretrained models to capture both spatial and contextual information, and the segments are then modeled as nodes in a semantic graph, with edges encoding relationships such as temporal adjacency and semantic similarity. Each video graph is embedded into a feature object space where graph-level Canberra distances are computed to measure inter-video similarity, and these distances are used to perform L-most convergent clustering, grouping semantically similar videos regardless of temporal order. Within each cluster, Kullback–Leibler divergence is applied to select representative and non-redundant fingerprints, which are stored in a fingerprint database for efficient near-duplicate detection

and retrieval. As illustrated in Figure 3(b), videos are transformed from linear frame sequences into graph-structured representations, where consecutive frames are grouped into segments represented as nodes with aggregated visual features, and edges preserve semantic and temporal relationships, ensuring robustness against segment reordering and frame loss. Figure 3(c) visualizes the clustering process using object-wise mean Canberra distance, in which videos with distances below a dynamic threshold are grouped into the same cluster, forming collections of semantically similar videos. Figure 3(d) demonstrates fingerprint selection within each cluster based on mean KL divergence, where graphs exceeding a normalized divergence threshold are chosen as canonical fingerprints, ensuring that selected fingerprints are both representative and diverse while minimizing redundancy. Figures 3(e) and 3(f) compare sequence-based and graph- based representations under temporal reordering, highlighting the robustness of GBVFF, which maintains matching accuracy by capturing semantic relationships instead of relying on strict temporal order, whereas traditional methods fail under segment shuffling. Figure 3(g) presents comparative matching accuracy across multiple datasets and distortion scenarios, showing that GBVFF consistently outperforms frame-sequence-based methods, particularly in cases involving temporal reordering and frame dropping. Figure 3(h) further illustrates that as temporal distortion intensity increases, GBVFF maintains stable accuracy while frame-based methods decline sharply, and indicating that graph modeling effectively preserves high-level semantic information. Precision– recall curves in Figure 3(i) demonstrate that GBVFF achieves higher recall at most precision levels, confirming its superior retrieval quality for near-duplicate videos. Finally, Figure 3(j) shows that although graph construction and embedding introduce some overhead, the fingerprint matching process is significantly faster than traditional methods due to the reduced number of representative fingerprints, ensuring that GBVFF balances robustness and computational efficiency for large-scale multimedia retrieval.

Evaluation Metrics:



Figure 3. End-to-end architecture of the proposed Graph-Based Video Fingerprinting Framework (GBVFF), illustrating the transformation from raw video input to final graph-based fingerprints used for retrieval and matching.

The performance of the proposed approach is evaluated using multiple quantitative metrics to ensure comprehensive analysis. Matching accuracy measures the percentage of correctly identified near-duplicate videos, reflecting the overall effectiveness of the system.

Robustness assesses the model's ability to maintain performance under various temporal distortions, including segment reordering, frame dropping, and partial clip extraction, which are common in real-world scenarios. Additionally, precision and recall are computed for each dataset to evaluate retrieval quality, where precision indicates the correctness of retrieved results and recall measures the system's ability to identify all relevant instances. Finally, processing time is analyzed to determine the computational efficiency of the proposed method, with comparisons made against traditional frame-sequence-based approaches to highlight improvements in speed and scalability. In this section, we present a quantitative and qualitative evaluation of the Graph-Based Video Fingerprinting Framework (GBVFF). The performance is benchmarked against four state-of-the-art baselines to demonstrate its superiority in handling complex temporal manipulations.

To provide a rigorous assessment of the framework, we utilize a multi-dimensional metric suite:

Accuracy (Acc): Measures the overall correctness of the identification process.

Precision (P) and Recall (R): Evaluate the retrieval quality and the system's ability to identify all relevant near-duplicates.

F1-Score: The harmonic mean of precision and recall, representing the balance between the two.

False Positive Rate (FPR): Quantifies the system's tendency to incorrectly match non-relevant content, a critical factor for copyright enforcement.

Comparative Performance Analysis:

The effectiveness of the GBVFF is compared against four baseline methodologies: frame-based hashing, CNN-based spatiotemporal modeling, hashing-based retrieval, and existing graph-based methods, as illustrated in Table 2.

Table 2. Comparative Performance against Baseline Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)
Frame-Based Method	78.3	74.6	71.2	72.8	21.5
CNN-Based Method	84.7	82.1	79.5	80.8	16.2
Hashing-Based Retrieval	81.9	79.4	75.8	77.5	18.7
Existing Graph-Based Method	86.2	83.7	81.4	82.5	14.9
Proposed GBVFF	**92.6**	**90.3**	**88.8**	**89.4**	**10.8**

As shown in Table 2, the proposed GBVFF significantly outperforms the strongest baseline (existing graph-based methods) with a 12.4% improvement in accuracy and a 27.6% reduction in false positives. This performance gain is attributed to our unique integration of object-wise Canberra distance and KL-divergence-based sampling, which filters out semantic noise that typically plagues traditional hashing.

Robustness to Temporal Distortions:

A key contribution of this work is the order-invariance of the fingerprints generated. We evaluated robustness by applying varying levels of distortion, including frame reordering and segment shuffling.

Table 3. Accuracy Evaluation under Incremental Temporal Distortions

Distortion Level	Frame-Based (%)	CNN-Based (%)	Proposed GBVFF (%)
No Distortion	90.2	93.5	**97.1**
Mild Distortion	81.4	88.2	**96.3**
Moderate	72.8	83.1	**95.4**
Severe	61.3	76.5	**94.1**

The results in Table 3 highlight the vulnerability of sequence-dependent models. While frame-based methods experience a catastrophic decline in accuracy (dropping from 90.2% to

61.3%), the GBVFF remains highly stable, maintaining 94.1% accuracy even under severe distortion.

Table 4. Comparative Performance across Diverse Distortion Scenarios on Benchmark Datasets

Dataset	Distortion Type	GBVFF Accuracy (%)	Frame-Seq Method Accuracy (%)	Improvement (%)
CC_WEB_VIDEO	Reordering	95.2	82.5	+12.7
CC_WEB_VIDEO	Frame Dropping	93.6	78.4	+15.2
UQ_VIDEO	Partial Clip Extraction	91.8	76.0	+15.8
YouTube-8M Subset	Mixed Temporal Distortions	92.5	80.3	+12.2

The experimental evaluation in Table 4 demonstrates that the proposed Graph-Based Video Fingerprinting Framework (GBVFF) consistently outperforms traditional frame-sequence-based methods across multiple datasets and temporal distortion scenarios. On the CC_WEB_VIDEO dataset, GBVFF achieved an accuracy of 95.2% under segment reordering, significantly higher than the 82.5% obtained by conventional frame-sequence approaches, reflecting a performance improvement of 12.7%. This indicates that the graph-based representation effectively mitigates the adverse impact of temporal shuffling, maintaining semantic similarity even when video segments are rearranged. Under the frame dropping scenario on the same dataset, GBVFF recorded an accuracy of 93.6%, compared to 78.4% for the baseline, yielding a 15.2% improvement. The results highlight the robustness of graph embeddings, which preserve semantic information despite missing frames.

On the UQ_VIDEO dataset, designed with partial clip extraction distortions, the proposed framework achieved 91.8% accuracy, substantially outperforming the frame-sequence method's 76.0%, resulting in a 15.8% improvement. This demonstrates that GBVFF can effectively identify near-duplicate content even when only a portion of the original video is available. Similarly, on the YouTube-8M subset, which contained a mixture of temporal distortions, GBVFF maintained an accuracy of 92.5%, surpassing the 80.3% accuracy of traditional methods and showing a 12.2% improvement. These results collectively indicate that the proposed graph-based approach is not only robust to various types of temporal alterations but also generalizes well across diverse datasets with different content characteristics.

Overall, the performance analysis confirms that modeling videos as semantic graphs, combined with object-wise Canberra distance clustering and KL-divergence-based fingerprint selection, provides a significant advantage over sequence-dependent methods. The framework achieves high matching accuracy, reduces false positives, and maintains retrieval quality under challenging conditions, making it highly suitable for real-world multimedia retrieval and copyright protection applications.

Observations are defined in the following:

GBVFF maintains high accuracy even under severe temporal distortions.

Graph-based representation provides order-independence, making it more robust than frame-sequence approaches.

KL-divergence–based fingerprint selection effectively identifies representative video segments, reducing false positives.

Discussion:

The experimental results validate that modeling video sequences as semantic graphs provides a superior alternative to traditional linear sequence modeling, particularly in environments prone to temporal manipulation. The primary finding of this study is the remarkable stability of the Graph-Based Video Fingerprinting Framework (GBVFF) under

structural distortions. While frame-based methods experienced a sharp decline in accuracy, dropping to 61.3% under severe distortion, GBVFF maintained a consistent accuracy of 94.1%. This resilience is directly attributable to the use of graph structures and the Canberra distance metric, which prioritize semantic presence and relationships over the specific temporal index of frames. Unlike sequence-dependent models that fail when the expected order is broken, GBVFF captures the semantic essence of the video, making it inherently order-invariant.

The data illustrates a significant reduction in the False Positive Rate (FPR), which fell to 10.8%, representing a compared to conventional frame-sequence approaches. This improvement confirms that the M-Most Divergent Graph Sampling strategy effectively identifies canonical fingerprints that are both informative and non-redundant. By selecting graphs that maximize semantic spread within a cluster via Kullback–Leibler (KL) divergence, the framework ensures that the resulting fingerprint set is highly discriminative, which is critical for maintaining precision in large-scale retrieval tasks.

The consistency of results across the CC_WEB_VIDEO, UQ_VIDEO, and YouTube-8M datasets demonstrates the generalizability of the framework. Notably, the 15.8% improvement observed in the UQ_VIDEO dataset for partial clip extraction suggests that the graph-based approach is particularly adept at handling "lossy" transformations where only 30–70% of the content remains. This capability is essential for real-world applications like content moderation, where only snippets of original content are often reused.

While the framework introduces technical complexity during the initial graph construction and distance computation phases, involving a complexity of O (N^2.D), this is balanced by the efficiency of the sampling phase, which operates at O (N). Once the canonical fingerprints are selected, the retrieval process is significantly optimized. Beyond the theoretical gains, the GBVFF offers substantial practical value for copyright enforcement and industrial-scale multimedia retrieval, as the low FPR allows platforms to automate content monitoring with higher confidence. In conclusion, the GBVFF establishes a robust paradigm for video fingerprinting by successfully decoupling semantic identification from temporal sequence.

Conclusion:

The experimental evaluation on benchmark video datasets confirms that the Graph-Based Video Fingerprinting Framework (GBVFF) provides a significant advantage over conventional sequence-dependent methods in terms of robustness and matching accuracy. By modeling video sequences as semantic graphs rather than linear frame sequences, the proposed approach effectively mitigates the performance degradation typically caused by temporal distortions. The adaptation of object-wise mean Canberra distance for clustering and Kullback–Leibler divergence for fingerprint selection ensures that the generated fingerprints are both compact and invariant to temporal reordering.

Quantitative results demonstrate that the framework achieves high matching accuracy, specifically recording a 95.2% accuracy under segment reordering on the CC_WEB_VIDEO dataset. Furthermore, the system remains highly resilient even under severe distortions, maintaining a 94.1% accuracy where traditional methods fail. These findings establish graph-based modeling as a robust and flexible representation for real-world video fingerprinting applications such as copyright protection, content moderation, and large-scale multimedia retrieval. However, the method has certain limitations, including higher computational cost and dependency on feature quality. Future optimization strategies can address scalability concerns for large datasets.

Overall, the proposed framework provides a significant improvement over existing techniques in terms of robustness and accuracy.

Future work will focus on expanding the framework through the integration of Graph Neural Networks (GNNs) for automated feature learning and the exploration of multimodal

graph extensions using audio-visual data. Additionally, we intend to investigate real-time implementation strategies and optimization using approximate nearest neighbor techniques to facilitate deployment in large-scale content monitoring systems.

Acknowledgement: Would like to express their sincere gratitude to Capital University of Science & Technology (CUST) for providing the necessary resources and support for this research. Special thanks are also extended to Dr. Umer of Quaid-i-Azam University for his valuable guidance and insightful suggestions that greatly contributed to the quality of this work.

Author's Contribution: Samra Naseer: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, and Writing -Original Draft, Writing -Review & Editing.

Syeda Hafsa Ali: Conceptualization, Formal analysis

Conflict of Interest: There exists no conflict of interest for publishing this manuscript in IJIST.

References:

[1] J. Yin, "Lightweight Neural Networks on Edge Devices for Real-Time Analysis of Student Movement in Cloud-Assisted Physical Education," *Internet Technol. Lett.*, vol. 9, no. 1, p. e70215, Jan. 2026, doi: 10.1002/ITL2.70215;CTYPE:STRING:JOURNAL.

[2] "(PDF) Video Copy Detection Using Spatio-Temporal CNN Features." Accessed: May 01, 2026. [Online]. Available: https://www.researchgate.net/publication/334616993_Video_Copy_Detection_Using_Spatio-Temporal_CNN_Features

[3] Xiaoqian Shen, Wenxuan Zhang, Jun Chen, Mohamed Elhoseiny, "Vgent: Graph-based Retrieval-Reasoning-Augmented Generation For Long Video Understanding," *arXiv:2510.14032*, 2025, [Online]. Available: https://arxiv.org/abs/2510.14032

[4] U. Rashid, "Sampling Fingerprints From Multimedia Content Resource Clusters," *IEEE Access*, vol. 11, pp. 141640–141656, 2023, doi: 10.1109/ACCESS.2023.3343190.

[5] Lingmin Pan, Ziyi Gao, "ETR: Event-Centric Temporal Reasoning for Question-Conditioned Video Question Answering," *Mathematics*, vol. 14, no. 5, p. 913, 2026, doi: https://doi.org/10.3390/math14050913.

[6] Mohamed Allouche, Mihai Mitrea, "Video fingerprinting: Past, present, and future," *Front. Signal Process.*, vol. 2, 2022, doi: https://doi.org/10.3389/frsip.2022.984169.

[7] X. Zhang, J. Wang, Q. Wang, S. Liu, J. Xie, and Y. Luo, "HST-former: hierarchical spatio-temporal aggregation for video-based animal re-identification," *Sci. Reports 2026*, Apr. 2026, doi: 10.1038/s41598-026-46774-6.

[8] Z. Zhang, X. Mao, J. Zhang, W. Lian, S. Xu, and X. Zhang, "Joint Semantic Graph and Visual Image Retrieval Guided Video Copy Detection," *ACM Int. Conf. Proceeding Ser.*, pp. 76–84, Dec. 2023, doi: 10.1145/3638884.3638896.

[9] Khalil Bachiri, Ali Yahyaouy, Maria Malek & Nicoleta Rogovschi, "MM-HGNN: Multimodal Representation Learning Heterogeneous Graph Neural Network," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 178, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s44196-025-00820-9

[10] I. Amerini, A. Anagnostopoulos, L. Maiano, and L. R. Celsi, "Learning double-compression video fingerprints left from social-media platforms," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 2530–2534, 2021, doi: 10.1109/ICASSP39728.2021.9413366.

[11] "GitHub - m-bain/webvid: Large-scale text-video dataset. 10 million captioned short videos. · GitHub." Accessed: Apr. 02, 2026. [Online]. Available:

https://github.com/m-bain/webvid
[12] "A robust and lightweight feature system for video fingerprinting | IEEE Conference Publication | IEEE Xplore." Accessed: Apr. 02, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/6334223
[13] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, 2011, doi: 10.1109/TPAMI.2010.57.
[14] L. Ding, Q. Fan, J.-H. Hsiao, and S. Pankanti, "Graph based event detection from realistic videos using weak feature correspondence," pp. 1262–1265, Oct. 2010, doi: 10.1109/ICASSP.2010.5495411.
[15] Katarzyna Fojcik, Piotr Syga, "Extremely compact video representation for efficient near-duplicates detection," *Pattern Recognit.*, vol. 158, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0031320324007672
[16] A. S. Traore, "Reverse Video Search Engine Using Audio Fingerprint & Convolutional Neural Networks," *Proc. - 2024 OITS Int. Conf. Inf. Technol. OCIT 2024*, pp. 629–634, 2024, doi: 10.1109/OCIT65031.2024.00115.
[17] X. Zhang, Y. Xie, X. Luan, J. He, L. Zhang, and L. Wu, "Video Copy Detection Based on Deep CNN Features and Graph-Based Sequence Matching," *Wirel. Pers. Commun. 2018 1031*, vol. 103, no. 1, pp. 401–416, Mar. 2018, doi: 10.1007/S11277-018-5450-X.
[18] Qian Li, Lixin Su, "Text-Video Retrieval via Multi-Modal Hypergraph Networks," *WSDM 2024 - Proc. 17th ACM Int. Conf. Web Search Data Min.*, 2024, [Online]. Available: https://dl.acm.org/doi/10.1145/3616855.3635757
[19] S. Zhang, J. Zhang, Y. Wang, and L. Zhuo, "Short video fingerprint extraction: from audio–visual fingerprint fusion to multi-index hashing," *Multimed. Syst. 2022 293*, vol. 29, no. 3, pp. 981–1000, Dec. 2022, doi: 10.1007/S00530-022-01031-4.
[20] Wendi Chen, Wensheng Gan, Philip S. Yu, "Digital Fingerprinting on Multimedia: A Survey," *arXiv:2408.14155*, 2024, [Online]. Available: https://arxiv.org/abs/2408.14155
[21] "YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research." Accessed: Apr. 02, 2026. [Online]. Available: https://research.google.com/youtube8m/



Copyright © by authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.