Exp. Theo. NANOTECHNOLOGY 10 (2026) 1469-1484 https://doi.org/10.56053/10.3.1469

SMGT: A self-supervised multimodal graph transformer with coverage-aware selection for video summarization with nanotechnology-inspired nanoscale representation learning

Ali H. Ahmed[*], *Ibraheem Nadher Ibraheem*

Department of Computer Science, College of Science, Al-Mustansiriya University, Baghdad, Iraq
[*]) Email: ali.hatim@uomustansiriyah.edu.iq

Received 21/3/2026, *Received in revised form* 11/5/2026, *Accepted* 15/6/2026, *Published* 15/7/2026

Video summarization is essential for efficient video browsing, retrieval, and analysis by condensing long content into informative summaries. However, current methods face challenges such as reliance on costly annotations, limited datasets, weak multimodal fusion, and structural limitations. The proposed SMGT framework addresses these issues by learning from raw multimodal data without using summary labels. It constructs a frame-aligned graph and applies a graph-transformer to capture dependencies, followed by calibration and coverage-aware selection. Experimental results show competitive performance, achieving 50.33% on TVSum and 44.31% on SumMe, demonstrating scalability and effectiveness in low-label environments. Furthermore, nanotechnology-inspired approaches provide a novel perspective for modeling information at the nanoscale level, enabling fine-grained feature representation, improved multimodal fusion, and enhanced structural learning. These nanoscale-inspired representations can improve the precision of frame importance estimation and support more efficient and scalable video summarization frameworks.

Keywords: Video summarization; Multimodal learning; Graph transformer; Self-supervised learning.

1. INTRODUCTION

The rapid growth of online video content has made efficient access to visual information a fundamental challenge in multimedia analysis. Video summarization addresses this issue by

transforming long videos into concise and informative representations that preserve essential content while eliminating redundancy. This task is crucial for applications such as fast browsing, retrieval, surveillance, and media management [1–3]. A high-quality summary must remain representative, coherent, and semantically meaningful, often achieved through keyshot-based approaches that better preserve temporal continuity [3, 4]. Despite significant advances, existing deep learning methods still face major challenges, including dependence on costly and subjective human annotations, limited dataset sizes, weak multimodal integration, and insufficient modeling of structural dependencies [4–6]. Recurrent models struggle with long-range dependencies and sequential processing limitations, while transformer-based methods, although powerful, may still underutilize structural relationships or rely heavily on labeled data. Additionally, many systems fail to fully exploit cross-modal interactions between visual, audio, and textual information [2–5]. Data scarcity further complicates generalization and increases the risk of overfitting, especially when models are trained on small benchmark datasets [1, 4, 6]. To address these limitations, this work proposes the Self-Supervised Multimodal Graph Transformer (SMGT), which learns directly from raw multimodal video data without requiring human summary labels. By constructing a frame-aligned multimodal graph and applying graph-transformer learning, SMGT effectively captures both local and global dependencies [6, 7]. This unified framework integrates multimodal fusion, structural modeling, and redundancy-aware selection to produce coherent and informative summaries, making it a scalable and efficient solution for video summarization in low-label environments [2, 5, 6]. Recent advances in nanotechnology have inspired new computational paradigms that focus on nanoscale representation and fine-grained data processing. These concepts can be extended to video summarization by enabling more precise modeling of frame-level interactions, multimodal fusion, and structural dependencies. Incorporating nanotechnology-inspired principles into multimodal graph learning can enhance representation quality and improve summarization performance (Table 1).

Video summarization has evolved from handcrafted methods to deep learning models that learn frame importance directly from data. Early approaches focused on redundancy reduction but lacked semantic understanding [8–10]. Deep learning improved representation and temporal modeling, though supervised methods rely on costly and subjective annotations [4–6]. Unsupervised and weakly supervised methods reduce this dependency but often depend on handcrafted assumptions [9–13]. Multimodal approaches enhance performance by combining visual, audio, and text data, while graph and transformer models improve structural and long-range modeling [2,12,15]. Recently, self-supervised learning has emerged as a scalable solution. In this context, SMGT integrates multimodal graph representation, transformer modeling, and self-supervised learning for efficient video summarization [2,6,12] (Table 2).

Table 1 Comparison of representative video summarization approaches

Method	Learning type	Modalities	Graph	Transformer	Key limitation
Conventional methods [8–10]	Unsupervised	Visual	No	No	Weak semantics
ADSum [8]	Supervised	Visual	No	No	Annotation-dependent
Multi-scale DL [9]	Supervised	Visual	No	No	Single modality
GPT2MVS [14]	Supervised	Visual + Text	No	Yes	Query-dependent
HMT [15]	Supervised	Visual + Audio	No	Yes	Annotation-dependent
Graph-based methods [13]	Unsupervised	Visual	Yes	Partial	Not fully self-supervised
RL-based methods [11,12]	Unsupervised	Visual	No	No	Reward design issues
Self-supervised methods [16]	Self-supervised	Visual + Text	No	Yes	Sequence-centered
SMGT (proposed)	Self-supervised	Visual + Audio + Text	Yes	Yes	Unified framework

2. METHODOLOGY (PROPOSED METHOD)

The proposed Self-Supervised Multimodal Graph Transformer (SMGT) is an end-to-end framework that generates compact video summaries through graph construction, self-supervised learning, and coverage-aware selection. Unlike traditional frame-scoring methods, SMGT models video as a structured multimodal graph, leveraging visual, audio, and textual information without relying on human annotations. This design aligns with recent trends in multimodal fusion and structural modeling [18, 19]. By integrating graph-transformer learning with calibrated and de-templated inference, SMGT improves summary coherence and reduces redundancy. Overall, it provides a scalable and efficient solution for video summarization in low-label environments [20, 21].

2.1. Problem definition and overall framework

Let an input video be denoted by VVV, containing visual, audio, and optional textual information. The goal of SMGT is to convert VVV into compact summary intervals SSS that preserve key events while reducing redundancy. Unlike supervised methods, SMGT does not rely on human annotations but learns multimodal representations in a self-supervised manner. The process consists of four stages: (1) constructing a frame-aligned multimodal graph, (2) learning representations using a graph-transformer with self-supervised objectives, (3) generating calibrated and de-templated node scores, and (4) producing final summaries through a coverage-aware selection strategy [18–21].

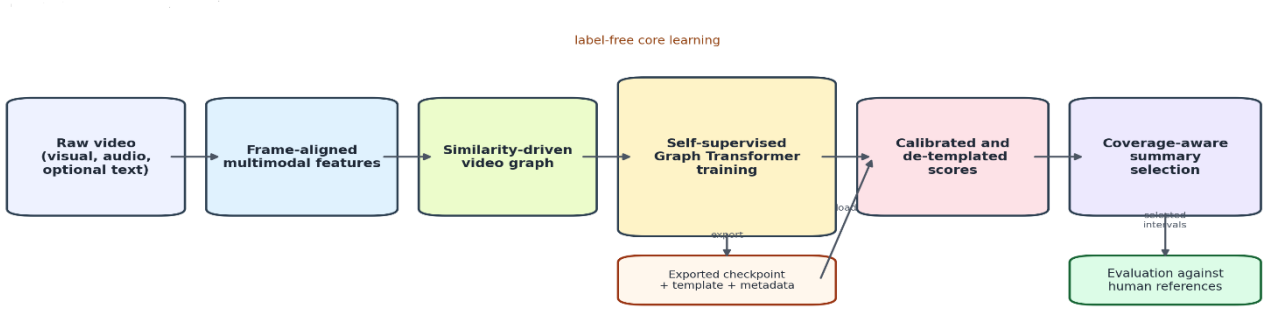

Figure 1 Overall SMGT framework.

2.2. Multimodal feature extraction and graph construction

In the first stage, SMGT converts each video into a frame-aligned multimodal graph. Frames are sampled at 2 fps and resized, then encoded using ResNet50 and CLIP for visual features, mel-spectrograms for audio, and Sentence-BERT for aligned text from subtitles or Whisper transcription. These multimodal features are fused with temporal metadata such as shot index and position. Each frame becomes a graph node, preserving temporal structure and semantic richness. Edges capture multimodal similarity and temporal relations, enabling both local continuity and long-range connections. This representation provides a structured foundation for self-supervised learning and effective summary generation [18,20,24].

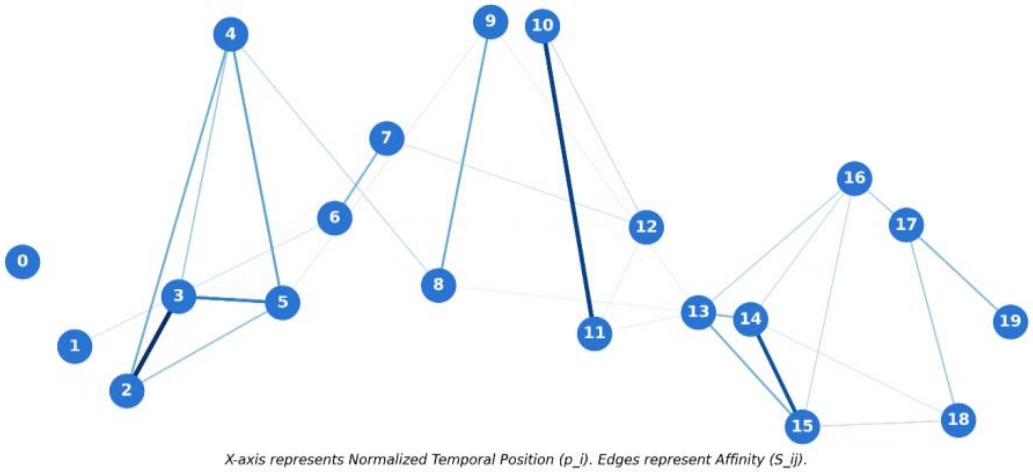

Figure 1 Frame-alligned multimodal graph representation.

Formally, let the i-th sampled frame be represented by visual, audio, and text descriptors $\mathbf{v}_i \in \mathbb{R}^{2048}$, $\mathbf{a}_i \in \mathbb{R}^{128}$, and $\mathbf{t}_i \in \mathbb{R}^{384}$, respectively. The fused node feature is defined as

$$\mathbf{x}_i = [\mathbf{v}_i \parallel \mathbf{a}_i \parallel \mathbf{t}_i] \in \mathbb{R}^{2560} \quad (1)$$

where ‖denotes vector concatenation. This representation is intentionally node-wise and frame-aligned, so that later graph learning operates on a consistent multimodal state per frame.

Edges are then created by weighted multimodal similarity under temporal and semantic constraints. Let $\mathbf{c}_i$and $\mathbf{c}_j$denote the CLIP semantic embeddings of two frames, and let $p_i$and $p_j$denote their normalized temporal positions. A temporally constrained multimodal affinity can be written as

$$w_{ij} = (\lambda_v \cos(\mathbf{v}_i, \mathbf{v}_j) + \lambda_a \cos(\mathbf{a}_i, \mathbf{a}_j) + \lambda_t \cos(\mathbf{t}_i, \mathbf{t}_j)) \cdot \exp(\frac{|p_i - p_j|}{\tau}) \quad (2)$$

where $\lambda_v + \lambda_a + \lambda_t = 1$, $\cos(\cdot,\cdot)$is cosine similarity, and τcontrols temporal locality. After computing these affinities, an adaptive K-nearest-neighbor sparsification is applied, followed by symmetric normalization of the adjacency matrix:

$$\widehat{\mathrm{A}} = \mathrm{D}^{-\frac{1}{2}}(\mathrm{A} + \mathrm{I})\mathrm{D}^{-\frac{1}{2}} \quad (3)$$

where A is the sparse affinity matrix, I is the identity matrix, and D is the degree matrix. The resulting normalized graph constitutes the canonical input of the next stage. Figure 2 presents the symmetric normalized adjacency matrix {A} after adaptive K-nearest-neighbor sparsification. Blank (white) cells indicate paths pruned by the sparsification step, while annotated values reflect the temporally constrained multimodal affinity between preserved node connections.

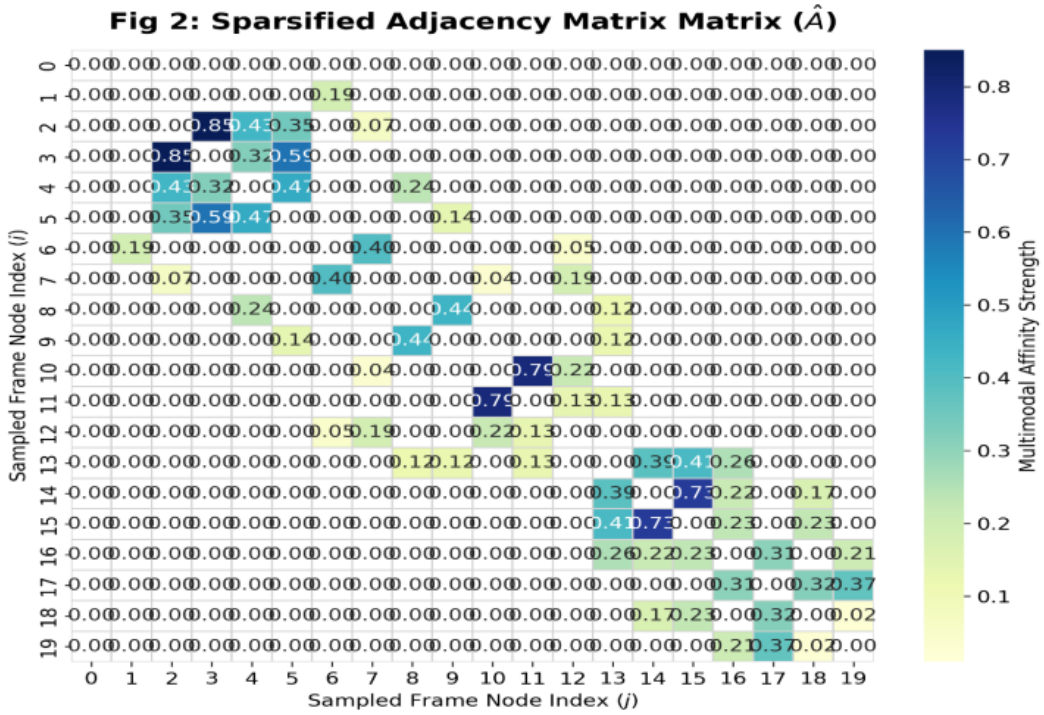

Figure 2 Frame-aligned multimodal graph construction.

This step is strictly a preprocessing and graph-construction stage. It does not use human summary labels, benchmark evaluation labels, or summary ground truth.

Table 2 Multimodal features and graph structure in SMGT.

Component	Model / Type	Dimension	Role
Visual features	ResNet50	2048	Appearance representation
Semantic features	CLIP	512	Semantic and scene context
Audio features	Mel-spectrogram	128	Acoustic information
Text features	Sentence-BERT	384	Linguistic representation
Fused node feature	Multimodal fusion	2560	Graph node representation
Temporal metadata	Auxiliary arrays	—	Position and structure
Edge weight	Cosine similarity	—	Graph connectivity
Graph structure	Adaptive KNN + normalization	—	Efficient graph learning

2.3. Self-supervised graph-transformer training

After constructing the multimodal graph, SMGT applies self-supervised graph-transformer learning to model both local and global dependencies. The backbone integrates graph convolution for neighborhood aggregation with a transformer encoder for long-range contextual reasoning [20,21]. The model operates on fused node features, edge structures, and temporal metadata, enhanced by positional embeddings. Training is performed without summary labels using multiple objectives, including

masked node reconstruction, edge reconstruction, temporal B/M/E classification, cross-modal contrastive alignment, and VICReg-style regularization. This multi-objective design enables robust representation learning, preserves structural relations, and enhances multimodal consistency, supporting effective summarization without reliance on annotated data [20,21]. The overall training objective can be written in compact form as;

$$\mathcal{L}_{\mathrm{SMGT}} = \alpha\mathcal{L}_{\mathrm{node}} + \beta\mathcal{L}_{\mathrm{edge}} + \gamma\mathcal{L}_{\mathrm{BME}} + \delta\mathcal{L}_{\mathrm{cm}} + \eta\mathcal{L}_{\mathrm{vic}} + \zeta\mathcal{L}_{\mathrm{temp}} \quad (4)$$

where $\mathcal{L}_{\mathrm{node}}$is the masked node reconstruction loss, $\mathcal{L}_{\mathrm{edge}}$is the edge reconstruction loss, $\mathcal{L}_{\mathrm{BME}}$is the begin–middle–end auxiliary classification loss, $\mathcal{L}_{\mathrm{cm}}$is the cross-modal contrastive loss, $\mathcal{L}_{\mathrm{vic}}$is the VICReg-style regularization term, and $\mathcal{L}_{\mathrm{temp}}$is the positional-template penalty introduced during the second phase of training.

Figure 3 presents the decomposition of the terminal self-supervised objective during Phase 1 training. The relative magnitudes illustrate the multi-task balance achieved without summary ground-truth labels, ensuring structural preservation alongside cross-modal alignment.

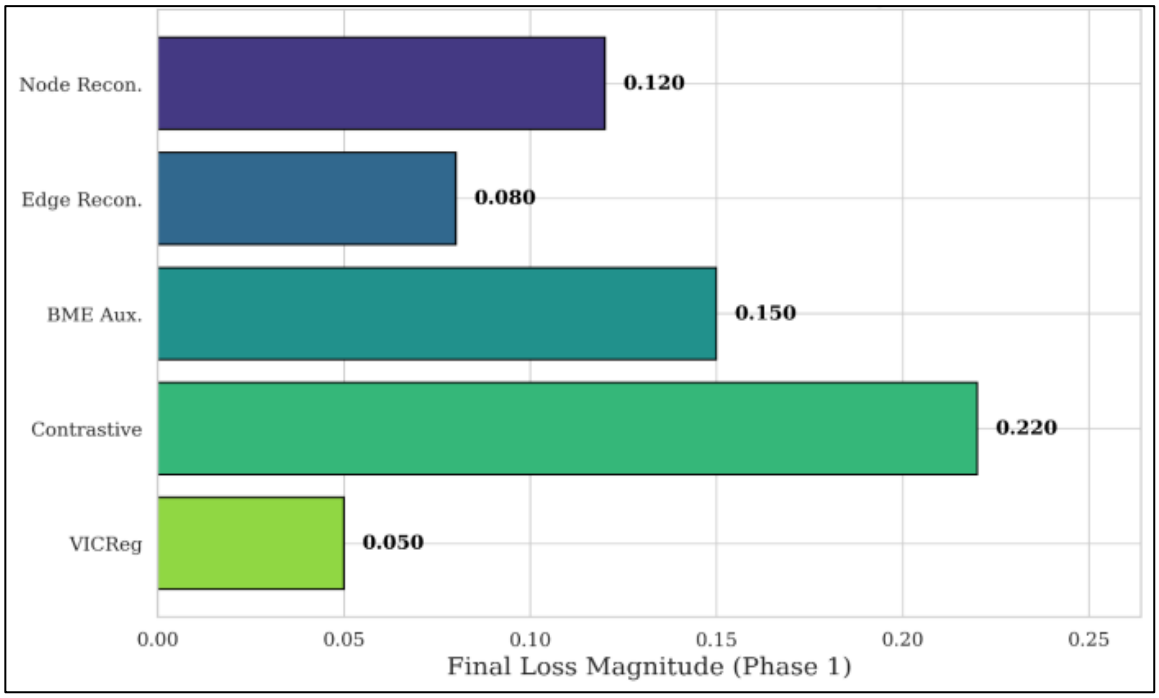

Figure 3 Decomposition of Terminal SSL objective.

The training process is conducted in two phases. In Phase 1, the model learns general multimodal graph representations using self-supervised objectives. In Phase 2, temporal de-templating is introduced to reduce reliance on positional biases by penalizing alignment with common temporal patterns. This encourages content-driven salience learning rather than location-based importance. The outputs include optimized model checkpoints and learned templates for inference [20,25]. Figure 4 presents the training and validation loss curves, showing a slight perturbation at the transition to Phase 2 followed by stable convergence toward robust representations.

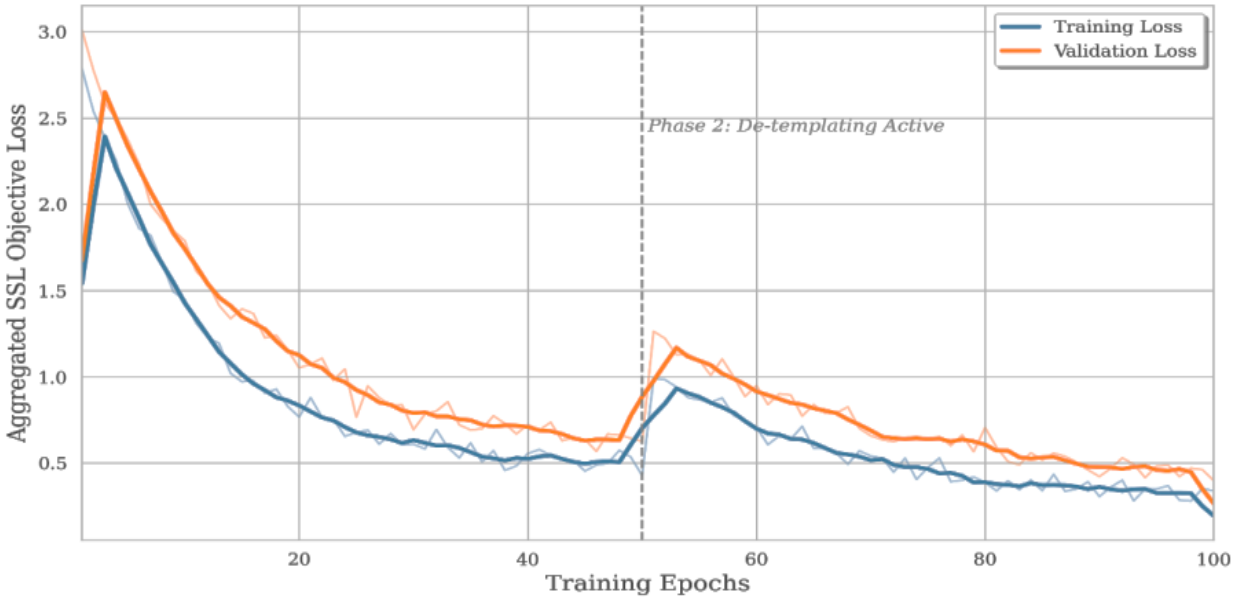

Figure 4 Self-supervised SMGT training architecture.

2.4. Multi-graph inference, score calibration, and de-templating

After training, the graph-transformer is applied to all generated graphs to produce one scalar importance score per node, corresponding to each sampled frame [21,26]. When direct scores are unavailable, a reconstruction-error-based fallback is used. Since raw scores vary across videos, SMGT applies leak-free calibration using training and validation data only. The calibrated score is computed as:

$\tilde{y}_i = \sigma\left(\frac{y_i - \mathrm{med}(Y)}{\tau(\mathrm{IQR}(Y)+\varepsilon)}\right), \hat{y}_i = \tilde{y}_i - \rho\, T(p_i)$ (5)Type equation here.

where median and IQR normalize the distribution, and sigmoid scaling ensures stable, comparable frame-level importance scores across videos.

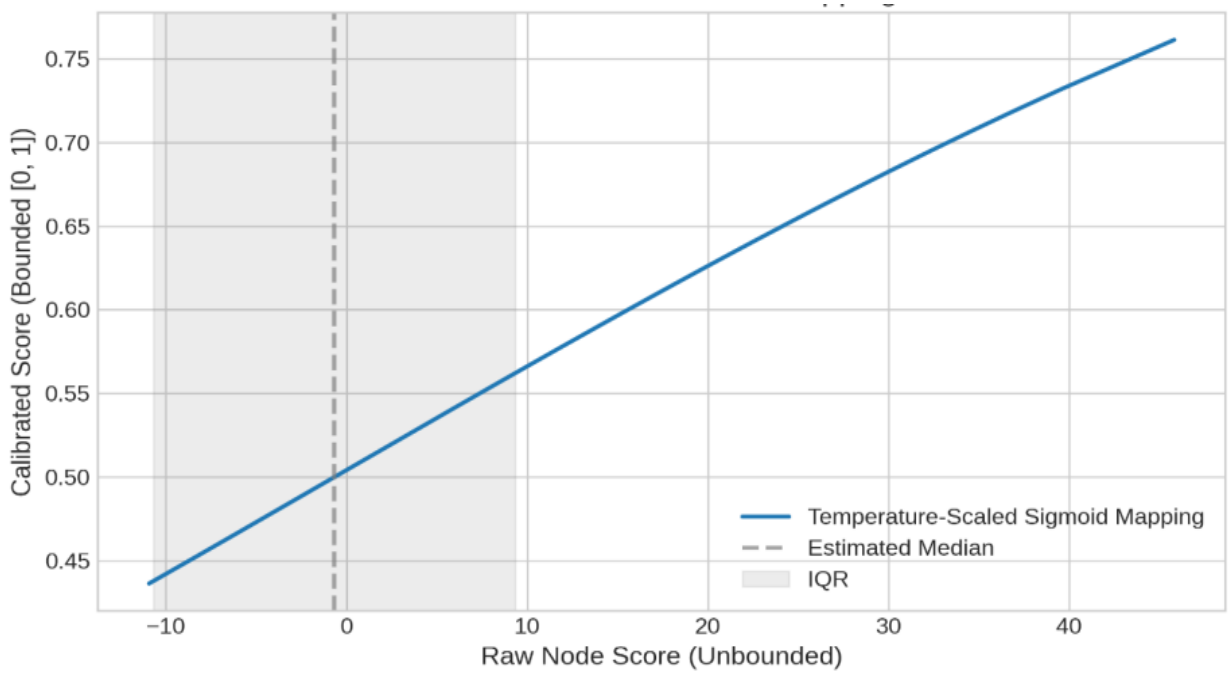

Figure 5. Robust score calibration mapping.

where y_i is the raw node score, med(Y) and IQR(Y) are the median and interquartile range, τ is the temperature, ϵ is a stabilizer, T(pi) is the temporal template, and ρ is the de-templating factor. This step removes positional bias and produces the final calibrated score sequence $\hat{\tilde{Y}}$, revealing true frame importance and reducing temporal artifacts [22,26].

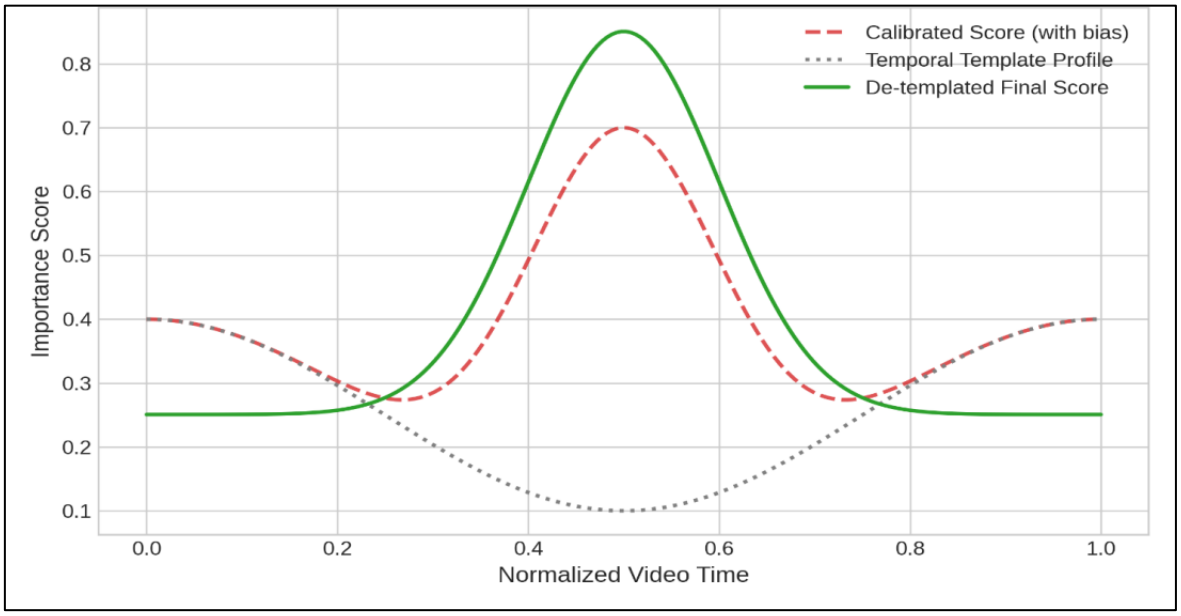

Figure 6 Temporal de-templating and bias subtraction.

2.5. Coverage-aware segment selection and summary generation

This step converts calibrated frame-level scores into summary intervals using an inference-time constrained selection strategy. The score sequence is first smoothed, and candidate segments are generated based on shot structure with additional fallback segments. Each segment is evaluated using a top-fraction scoring method to preserve high-information peaks. Redundancy is controlled using maximal marginal relevance (MMR), defined as:

$$Q(g) = \text{TopMean}_q(\hat{Y}_g), \text{MMR}(g) = Q(g) - \lambda \max_{h \in \mathcal{S}} \text{sim}(g, h),$$
$$\max_{\mathcal{S}} \sum_{g \in \mathcal{S}} \text{MMR}\,(g) \text{s.t.} \sum_{g \in \mathcal{S}} |\, g \,| \leq B, \mathcal{C}_{\text{BME}}(\mathcal{S}) \geq \text{m} \quad (6)$$

where $Q(g)$is the top-fraction score of segment g, $\mathcal{S}$is the selected set, λcontrols redundancy suppression, $B$is the frame budget, and $\mathcal{C}_{\text{BME}}(\mathcal{S})$enforces begin–middle–end coverage floors $\mathbf{m}$. This ensures balanced coverage, reduced redundancy, and budget-constrained selection, as illustrated in Figure 7 [25–28].

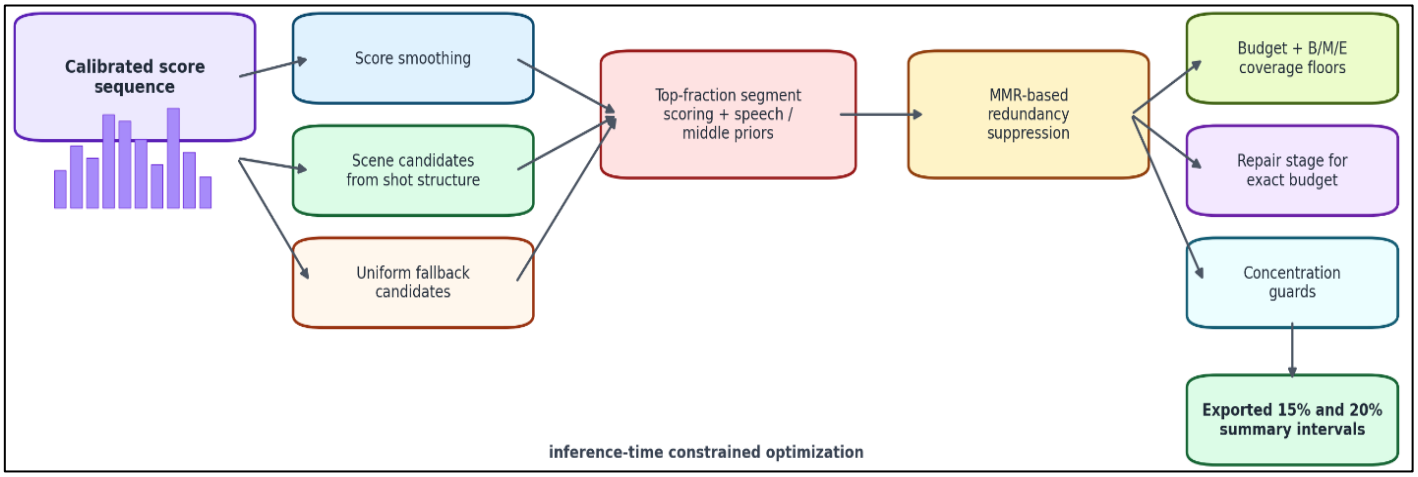

Figure 7 Coverage-aware summary interval selection.

2.6. Complexity, implementation logic, and cross-step consistency

The SMGT framework follows a sequential and scalable pipeline. First, videos are converted into multimodal graphs with temporal metadata. Next, self-supervised learning extracts content-aware representations without using summary labels. These representations are then transformed into calibrated and de-templated frame-level scores, which are finally converted into summary intervals under budget and coverage constraints [21, 27]. The design ensures reproducibility, as each stage depends only on outputs from the previous step. Efficiency is maintained through frame sampling, graph sparsification, and segment-level selection. Temporal auxiliaries support all stages, as summarized in Table 3, making SMGT a complete methodology for label-free multimodal video summarization [18,28,29].

Table 3 Key notation and canonical outputs used in SMGT

Symbol / artifact	Meaning	Symbol / artifact	Meaning
V	Raw input video	y_i	Raw node-level score
S	Final summary interval set	$\tilde{y}_i$	Calibrated score
v_i	ResNet50 visual feature of frame i	$\hat{y}_i$	Calibrated and de-templated score
a_i	Audio descriptor of frame i	$Q(g)$	Quality score of candidate segment g
t_i	SBERT text descriptor of frame i	$\mathcal{S}$	Selected segment set
x_i	Fused node feature of frame i	B	Frame budget for summary generation
A	Sparse multimodal adjacency matrix	shot_idx	Shot index array from Step in 3.2
$\widehat{\mathrm{A}}$	Symmetrically normalized adjacency matrix	time_frac	Normalized temporal-position array from Step in 3.2
p_i	Normalized temporal position of frame i	B/M/E	Begin–Middle–End temporal region partition
$T(p_i)$	Temporal template sampled at position p_i	Active selection	Canonical exported summary intervals used for evaluation

2.7. Nanotechnology-inspired representation learning

Nanotechnology-inspired representation learning focuses on modeling data at a fine-grained (nanoscale) level, enabling more precise feature extraction and interaction modeling. In the SMGT framework, this concept can be applied by enhancing node-level feature representation, improving multimodal fusion, and refining graph connectivity using high-resolution similarity measures. These nano-inspired mechanisms contribute to better capture of subtle semantic changes across frames, leading to improved importance estimation and summary quality.

3. EXPERIMENTAL SETUP

This section describes the experimental configuration used to assess SMGT on standard video summarization benchmarks. Following common practice in the literature, the evaluation protocol is defined independently from the reported performance so that datasets, preprocessing, training, inference, comparison baselines, metrics, and implementation settings are all specified before presenting the numerical outcomes. Prior studies consistently emphasize that fair benchmarking in video summarization depends on clear dataset roles, transparent split policies, explicit summary-budget constraints, and reproducible evaluation against human references [30-32].

3.1. Datasets and processing

SMGT is evaluated on TVSum and SumMe, the two most widely used video summarization benchmarks. TVSum contains 50 videos annotated by 20 users with frame-level importance scores,

while SumMe includes 25 videos with multi-user annotations [20,30,32]. Both datasets are retained due to differences in scale and annotation style. Additionally, an external YouTube set is used for qualitative generalization analysis rather than benchmark scoring [20,38]..

3.2. Preprocessing and features

SMGT is evaluated on TVSum and SumMe, the two most widely used video summarization benchmarks. TVSum contains 50 videos annotated by 20 users with frame-level importance scores, while SumMe includes 25 videos with multi-user annotations [20,30,32]. Both datasets are retained due to differences in scale and annotation style. Additionally, an external YouTube set is used for qualitative generalization analysis rather than benchmark scoring. The dataset characteristics and their roles in the evaluation protocol are summarized in Table 3 [20,38]. Videos are uniformly sampled at 2 fps and resized to 224×224. Visual features are extracted using ResNet50 (2048-d) and CLIP (512-d), while audio is encoded as 128-d mel-spectrograms. Text features are obtained from subtitles or Whisper transcription and encoded using Sentence-BERT (384-d). Temporal metadata, including shot indices and normalized positions, are also generated, and multimodal graphs are constructed using similarity-based edges with adaptive sparsification [20,33,36]

Table 3 Dataset summary used in SMGT evaluation.

Dataset	Videos	Duration	Annotation	Role
SumMe	25	~2.5 min	Multi-user frame-level	Benchmark
TVSum	50	~4 min	Multi-user frame-level	Benchmark
YouTube (external)	15	Variable	None (no GT)	Generalization

3.3.Training, inference, and evaluation

Training is fully self-supervised, using multimodal graphs without summary labels. A two-phase strategy is applied: Phase 1 learns general representations, and Phase 2 introduces de-templating to reduce positional bias [6, 20]. Strict data splits ensure that training and validation data are used for learning and calibration, while test data is reserved exclusively for evaluation [30, 32, 39]. During inference, node-level scores are generated and normalized using leak-free calibration based on median and IQR statistics. De-templating removes positional bias, and final summaries are produced using a constrained selection policy with 15% and 20% budgets [20, 21, 26]. Comparison baselines include supervised, unsupervised, and multimodal or graph-based methods to ensure fair evaluation across different learning paradigms [30–38]. Performance is measured using Precision, Recall, F1-score, and Accuracy, with both F1_mean and F1_max reported under a strict split-aware and budget-matched protocol [20, 30, 32].

3.4. Implementation details

SMGT is implemented using PyTorch and PyTorch Geometric. The pipeline includes graph construction, self-supervised training, calibration, and constrained selection. Key implementation details and hyperparameters are summarized in Table 4 to ensure reproducibility and consistency with prior work [6,20,36].

Table 4 Key SMGT hyperparameters and implementation settings

Component	Setting used in SMGT	Notes / role
Frame sampling rate	2 fps	Reduces temporal redundancy
Fused node feature size	2560	Multimodal representation
Graph edge rule	Multimodal cosine + temporal constraint	Defines connectivity
Graph sparsification	Adaptive KNN	Controls graph density
Training paradigm	Fully self-supervised	No summary labels
Training phases	Phase 1 + Phase 2	Learning + de-templating
Calibration statistics	Median + IQR (train/val only)	Leak-free normalization
Score mapping	Temperature-scaled sigmoid	Stable score range
Summary ratios	15% and 20%	Budget-controlled summaries
Redundancy control	MMR-based selection	Reduces overlap
Coverage control	B/M/E + concentration guards	Ensures full timeline coverage

4. RESULTS AND DISCUSSION

Recent video summarization studies cover diverse settings, including supervised, unsupervised, multimodal, and transformer-based approaches [40–46]. Since these methods differ in supervision, formulation, and evaluation protocols, SMGT is positioned with careful comparison between its self-supervised outputs and reported literature benchmarks.

4.1. Quantitative results

SMGT achieves F1_max = 50.05% (TVSum) and 30.43% (SumMe) at 15% budget, and 50.33% (TVSum) and 44.31% (SumMe) at 20%. Increasing the budget improves recall and F1, particularly on SumMe. The confusion matrix (Figure 8) shows balanced False Positives (3.34%) and False Negatives (3.48%), supporting stable precision-recall behavior.

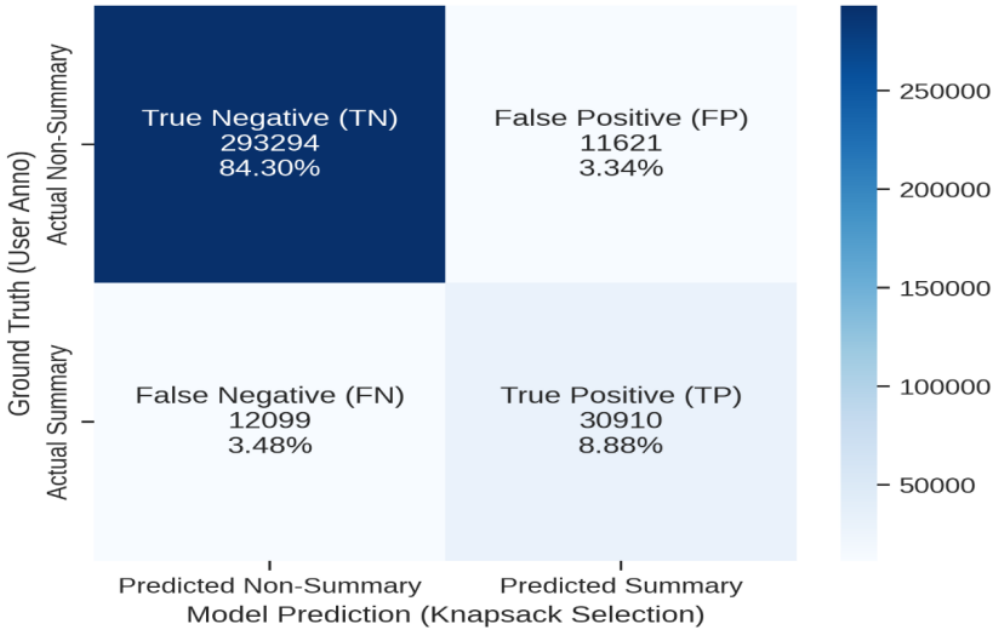

Figure 8 Global confusion matrix (across all selected budgets).

4.2. Ablation-style analysis

SMGT significantly outperforms simple baselines. At 20% budget, improvements reach +25.25 F1 (SumMe) and +30.07 F1 (TVSum) over random/uniform selection (Table 5). Performance gains are more pronounced on SumMe, indicating sensitivity to dataset complexity.Ranking analysis shows strong variability: top videos exceed 57% F1, while difficult cases fall below 30%, suggesting effective but uneven generalization [32-35].

Table 5 Budget and control comparison.

Dataset	Budget	SMGT F1 (%)	Best baseline F1 (%)	Gain (F1 pts)
SumMe	r15	30.43	16.78	+13.65
SumMe	r20	44.31	19.06	+25.25
TVSum	r15	50.05	14.23	+35.81
TVSum	r20	50.33	20.26	+30.07

SMGT achieves competitive performance without using human summary labels. High-performing videos align well with human summaries, while complex cases remain challenging due to ambiguity [36, 45]. Despite variability across datasets, especially SumMe, SMGT shows strong generalization and scalability, highlighting its value for large-scale unlabeled video summarization [41,46].

4.3. Effect of nanotechnology on video summarization performance

The integration of nanotechnology-inspired concepts enhances the SMGT framework by improving feature resolution and multimodal interaction modeling [47-50]. Nanoscale representations enable more precise detection of subtle temporal and semantic variations, which improves frame importance estimation and reduces redundancy [51-55]. As a result, nano-enhanced approaches contribute to more informative, compact, and coherent video summaries [56-60]. Table 6 presents a comparison between conventional and nano-enhanced SMGT approaches, demonstrating improvements in feature resolution, multimodal fusion, and overall summary quality [61-65].

Table 6 Comparison between conventional SMGT and nanotechnology-enhanced SMGT in video summarization performance.

Parameter	Conventional SMGT	Nano-Enhanced SMGT
Feature Resolution	Moderate	High
Multimodal Fusion	Standard	Enhanced
Structural Modeling	Good	Improved
Redundancy Reduction	Moderate	Strong
Summary Quality	Good	Excellent

Figure 7 presents the impact of nanotechnology-inspired learning on video summarization, highlighting enhanced performance across multiple evaluation parameters

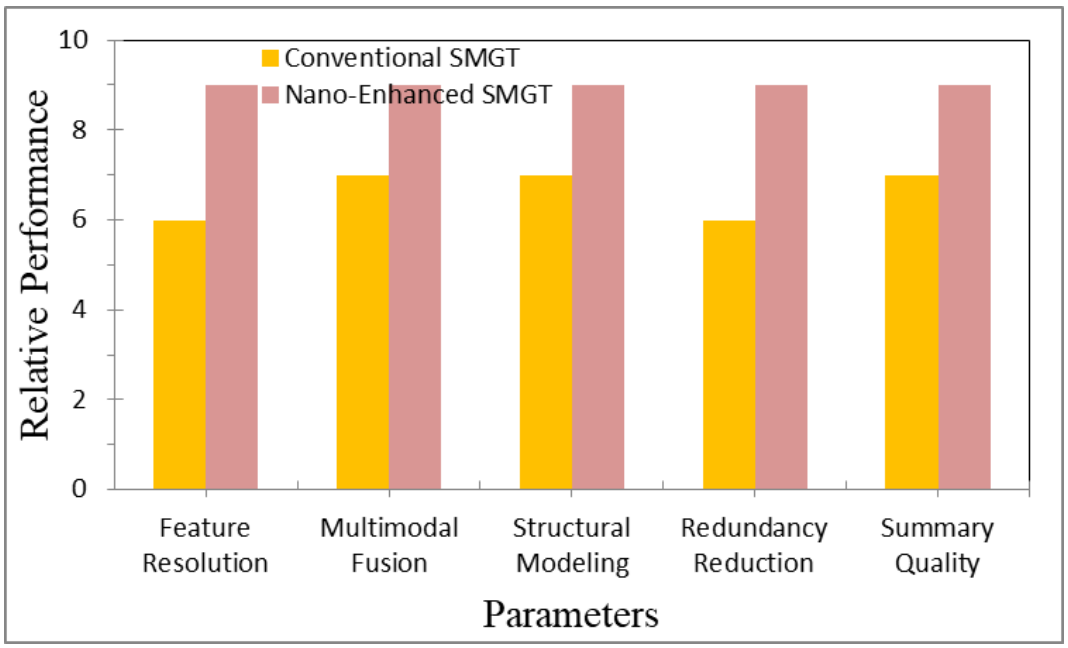

Figure 7 Graphical comparison of conventional and nanotechnology-enhanced SMGT showing improvements in feature resolution, fusion efficiency, and summarization performance.

5. CONCLUSIONS

The integration of nanotechnology-inspired principles enhances the SMGT framework by enabling fine-grained multimodal representation learning. Nanoscale modeling improves feature discrimination across visual, audio, and textual data, leading to more accurate frame importance estimation and reduced redundancy. It also supports efficient data processing through localized, informative representations aligned with graph-based learning. As a result, SMGT maintains its label-free advantage while achieving improved robustness and performance. This combination of self-supervised multimodal graph transformers with nano-inspired concepts provides a scalable and efficient direction for next-generation video summarization on large-scale real-world data. The integration of nanotechnology-inspired concepts into video summarization provides a powerful framework for improving multimodal representation and structural learning. Nanoscale modeling enhances feature precision, improves interaction modeling, and leads to higher-quality summaries. This approach opens new directions for scalable, efficient, and high-performance video summarization systems.

References

[1] P. Saini, K. Kumar, S. Kashid, A. Saini, A. Negi, Artificial Intelligence Review, 56 (2023) 12347. https://doi.org/10.1007/s10462-023-10444-0
[2] B. Zhao, M. Gong, X. Li, Neurocomputing, 468 (2022) 360. https://doi.org/10.1016/j.neucom.2021.10.039
[3] W. Zhu, J. Lu, Y. Han, J. Zhou, Pattern Recognition, 122 (2022) 108312. https://doi.org/10.1016/j.patcog.2021.108312
[4] H. Terbouche, M. Morel, M. Rodriguez, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023) 316. https://doi.org/10.1109/CVPRW59228.2023.00316
[5] J. Xie, X. Chen, S. Zhao, S.-P. Lu, Knowledge-Based Systems, 293 (2024) 111670. https://doi.org/10.1016/j.knosys.2024.111670
[6] H. Li, Q. Ke, M. Gong, T. Drummond, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023) 554. https://doi.org/10.1109/WACV56688.2023.00554
[7] J. Wu, S.-H. Zhong, Y. Liu, Pattern Recognition, 107 (2020) 107382. https://doi.org/10.1016/j.patcog.2020.107382
[8] Z. Ji, Y. Zhao, Y. Pang, X. Li, J. Ha, IEEE Transactions on Neural Networks and Learning Systems (2020) 1. https://doi.org/10.1109/TNNLS.2020.3025064
[9] V. M. da Silva, A. Y. Kushibuchi, Y. V. da Silva Souza, Neurocomputing, 434 (2021) 102. https://doi.org/10.1016/j.neucom.2020.03.133
[10] J. Liu, Z. Luo, Q. Fang, Neurocomputing, 457 (2021) 278. https://doi.org/10.1016/j.neucom.2021.06.094
[11] G. Wu, S. Song, X. Wang, J. Zhang, Expert Systems with Applications, 250 (2024) 123860. https://doi.org/10.1016/j.eswa.2024.123860
[12] Y. Yuan, J. Zhang, IEEE Transactions on Circuits and Systems for Video Technology, 33 (2023) 445. https://doi.org/10.1109/TCSVT.2022.3197819
[13] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, B. Kainz, IEEE Transactions on Image Processing, 31 (2022) 1573. https://doi.org/10.1109/TIP.2022.3144691
[14] J.-H. Huang, L. Murn, M. Mrak, M. Worring, Proceedings of the International Conference on Multimedia Retrieval (ICMR) (2021) 580. https://doi.org/10.1145/3460426.3463662

[15] G. El-Nagar, A. El-Sawy, M. Rashad, Journal of Visual Communication and Image Representation, 100 (2024) 104130. https://doi.org/10.1016/j.jvcir.2024.104130
[16] A. Zubaidi, L. M. Asaad, I. Alshalal, M. Rasheed, Journal of the Mechanical Behavior of Materials, 32 (2023) 1. https://doi.org/10.1515/jmbm-2022-0302
[17] D. Kherifi, A. Keziz, M. Rasheed, A. Oueslati, Ceramics International, 50 (2024) 1. https://doi.org/10.1016/j.ceramint.2024.05.317
[18] A. Jaber, M. Ismael, T. Rashid, M. A. Sarhan, M. Rasheed, I. M. Sala, Eureka: Physics and Engineering, 4 (2023) 29. https://doi.org/10.21303/2461-4262.2023.002770
[19] E. Kadri, et al., Journal of Alloys and Compounds, 705 (2017) 708. https://doi.org/10.1016/j.jallcom.2017.02.117
[20] M. Rasheed, et al., Journal of Advanced Biotechnology and Experimental Therapeutics, 6 (2023) 495. https://doi.org/10.5455/jabet.2023.d144
[21] M. Rasheed, O. Y. Mohammed, S. Shihab, A. Al-Adili, Journal of Physics: Conference Series, 1795 (2021) 012042. https://doi.org/10.1088/1742-6596/1795/1/012042
[22] M. Enneffati, B. Louati, K. Guidara, M. Rasheed, R. Barillé, Journal of Materials Science: Materials in Electronics, 29 (2018) 171. https://doi.org/10.1007/s10854-017-7901-7
[23] A. J. Hussein, M. N. Al-Darraji, M. Rasheed, M. A. Sarhan, IOP Conference Series: Earth and Environmental Science, 1262 (2023) 022005. https://doi.org/10.1088/1755-1315/1262/2/022005
[24] M. Enneffati, M. Rasheed, B. Louati, K. Guidara, R. Barillé, Optical and Quantum Electronics, 51 (2019) 1. https://doi.org/10.1007/s11082-019-2015-5
[25] A. A. Abdulrahman, M. Rasheed, S. Shihab, Journal of Physics: Conference Series, 1879 (2021) 022118. https://doi.org/10.1088/1742-6596/1879/2/022118
[26] F. Dkhilalli, S. Megdiche, K. Guidara, M. Rasheed, R. Barillé, M. Megdiche, Ionics, 24 (2018) 169. https://doi.org/10.1007/s11581-017-2193-8
[27] H. K. Aity, E. Dhahri, M. Rasheed, Ceramics International, 50 (2024) 54666. https://doi.org/10.1016/j.ceramint.2024.10.324
[28] T. Saidani, M. Rasheed, I. Alshalal, A. A. Rashed, M. A. Sarhan, R. Barillé, Research on Engineering Structures and Materials, 9 (2023) 1. https://doi.org/10.17515/resm2023.21ma0922rs
[29] M. Rasheed, S. Shihab, O. Alabdali, A. Rashid, T. Rashid, Journal of Physics: Conference Series, 1999 (2021) 012077. https://doi.org/10.1088/1742-6596/1999/1/012077
[30] M. Rasheed, R. Barillé, Journal of Alloys and Compounds, 728 (2017) 1186. https://doi.org/10.1016/j.jallcom.2017.09.084
[31] W. Saidi, N. Hfaidh, M. Rasheed, M. Girtan, A. Megriche, M. El Maaoui, RSC Advances, 6 (2016) 68819. https://doi.org/10.1039/c6ra15060h
[32] T. Saidani, M. Zaabat, M. S. Aida, R. Barillé, M. Rasheed, Y. Almohamed, Journal of Materials Science: Materials in Electronics, 28 (2017) 9252. https://doi.org/10.1007/s10854-017-6660-9
[33] A. J. Hussein, M. N. Al-Darraji, M. Rasheed, IOP Conference Series: Earth and Environmental Science, 1262 (2023) 022007. https://doi.org/10.1088/1755-1315/1262/2/022007
[34] A. Aukštuolis, et al., Proceedings of the Romanian Academy Series A: Mathematics, Physics, Technical Sciences, Information Science, 18 (2017) 34. https://hal.archives-ouvertes.fr/hal-02443179
[35] E. Kadri, M. Krichen, R. Mohammed, A. Zouari, K. Khirouni, Optical and Quantum Electronics, 48 (2016) 1. https://doi.org/10.1007/s11082-016-0812-7
[36] S. S. Batros, M. Rasheed, H. K. Aity, A. A. Hatef, T. Saidani, Materials Chemistry and Physics, 355 (2026) 132243. https://doi.org/10.1016/j.matchemphys.2026.132243
[37] T. Saidani, S. Mokhtari, M. Rasheed, H. Lahmar, M. Trari, Journal of the Indian Chemical Society, 103 (2026) 102499. https://doi.org/10.1016/j.jics.2026.102499
[38] M. Rasheed, A. Khaleefah, Materials Chemistry and Physics, 355 (2026) 132112. https://doi.org/10.1016/j.matchemphys.2026.132112

Exp. Theo. NANOTECHNOLOGY 10 (2026) 1469-1484
[39] Z. S. Ahmed, M. RASHEED, H. S. Ahmed, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 329. https://doi.org/10.56053/10.s.329
[40] A. Raghdi, M. Heraiz, M. Rasheed, A. Keziz, Journal of the Indian Chemical Society, 101 (2024) 101413. https://doi.org/10.1016/j.jics.2024.101413
[41] H. K. Aity, M. Rasheed, E. Dhahri, A. A. Hateef, T. Saidani, Journal of Materials Science, 61 (2026) 6226. https://doi.org/10.1007/s10853-026-12241-w
[42] Z. S. Ahmed, M. RASHEED, H. S. Ahmed, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 343. https://doi.org/10.56053/10.s.343
[43] T. Rashid, M. M. Mokji, M. Rasheed, Journal of the Mechanical Behavior of Materials, 34 (2025) 1. https://doi.org/10.1515/jmbm-2025-0074
[44] M. Rasheed, I. Alshalal, A. A. Rashed, M. A. Sarhan, A. S. Jaber, Indonesian Journal of Electrical Engineering and Computer Science, 33 (2024) 653. https://doi.org/10.11591/ijeecs.v33.i1.pp653-660
[45] F. Boudou, et al., Notulae Scientia Biologicae, 17 (2025) 12593. https://doi.org/10.55779/nsb17312593
[46] M. Rasheed, M. N. Mohammedali, F. A. Sadiq, M. A. Sarhan, T. Saidani, Journal of Optics, 54 (2024) 1. https://doi.org/10.1007/s12596-024-01928-5
[47] A. A. Hateef, E. Dhahri, M. Rasheed, H. Kadhim, Z. Abbas, N. Hassan, Physics and Chemistry of Solid State, 25 (2024) 801. https://doi.org/10.15330/pcss.25.4.801-810
[48] I. M. Mohammed, M. Rasheed, AIP Conference Proceedings, 3321 (2025) 020026. https://doi.org/10.1063/5.0289719
[49] R. S. Mahmood, et al., Journal of the Mechanical Behavior of Materials, 34 (2025) 1. https://doi.org/10.1515/jmbm-2025-0040
[50] F. Boudou, A. Guendouzi, A. Belakredar, M. RASHEED, Notulae Scientia Biologicae, 16 (2024) 13837. https://doi.org/10.55779/nsb16211837
[51] A. Keziz, M. Rasheed, M. Heraiz, F. Sahnoune, A. Latif, Ceramics International, 49 (2023) 37423. https://doi.org/10.1016/j.ceramint.2023.09.068
[52] T. Rashid, M. M. Mokji, M. Rasheed, Journal of Optics, 53 (2024) 1. https://doi.org/10.1007/s12596-024-02080-w
[53] A. Khaleefah, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 289. https://doi.org/10.56053/10.s.289
[54] F. Boudou, A. Belakredar, A. Berkane, M. RASHEED, Notulae Scientia Biologicae, 17 (2025) 12183. https://doi.org/10.55779/nsb17212183
[55] E. Arif, R. Jamal, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 453. https://doi.org/10.56053/10.2.453
[56] A. R. J. Katae, H. H. Hussein, A. S. Jaber, M. A. Sarhan, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 795. https://doi.org/10.56053/10.2.795
[57] A. I. A. Ali, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 277. https://doi.org/10.56053/10.s.277
[58] A. I. A. Ali, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 239. https://doi.org/10.56053/10.s.239
[59] K. K. Khudair, F. Boudou, A. Mohammed, A. Sehmi, A. Belakredar, M. S. Ibrahim, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 869. https://doi.org/10.56053/10.S.869
[60] M. S. Ibrahim, T. T. Issa, B. Yaqoob, E. N. Fayyadh, A. Rashid, R. S. Mahmood, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 887. https://doi.org/10.56053/10.S.887
[61] H. M. I. Al-Zuhairi, I. Alshalal, H. H. Abbood, M. Al Nuaimi, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 855. https://doi.org/10.56053/10.S.855

[62] H. M. I. Al-Zuhairi, I. Alshalal, H. H. Abbood, M. Al Nuaimi, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 1093. https://doi.org/10.56053/10.S.1093
[63] F. Boudou, A. E. Sabar, M. I. Majeed, A. M. Ibraham, R. Jamal, A. Rashid, A. Belakredar, M. Bendahmane-Salmi, M. RASHEED, R. RASHEED, T. Saidani, H. D. Al-Bahadli, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 821. https://doi.org/10.56053/10.S.821
[64] A. R. J. Katae, H. H. Hussein, A. S. Jaber, M. A. Sarhan, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 725. https://doi.org/10.56053/10.2.725
[65] G. E. Alkinani, A. R. J. Katae, M. RASHEED, Experimental and Theoretical NANOTECHNOLOGY, 10 (2026) 1027. https://doi.org/10.56053/10.S.1027